*M*odel selecti*o*n and appr*o*ximati*o*n in high-dimensi*o*nal mixtures *o*f experts m*o*dels : from theory to practice

Trung Tin Nguyen

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité MATHEMATIQUES

Préparée au sein de l'Université de Caen Normandie

Model Selection and Approximation in High-dimensional Mixtures of Experts Models: From Theory to Practice

Présentée et soutenue par
TRUNG TIN NGUYEN

Thèse soutenue le 14/12/2021 devant le jury composé de		
M. SYLVAIN ARLOT	Professeur des universités, Université Paris-Saclay	Rapporteur du jury
MME JUDITH ROUSSEAU	Professeur des universités, Université Paris-Dauphine	Rapporteur du jury
MME GAËLLE CHAGNY	Chargé de recherche, Université Rouen Normandie	Membre du jury
M. HIEN DUY NGUYEN	Maître de conférences, University of Queensland	Membre du jury
M. CHRISTOPHE BIERNACKI	Professeur des universités, Université de Lille	Président du jury
M. FAICEL CHAMROUKHI	Professeur des universités, Université Caen Normandie	Directeur de thèse

Thèse dirigée par FAICEL CHAMROUKHI, Laboratoire de Mathématiques 'Nicolas Oresme' (Caen)

UNIVERSITÉ CAEN NORMANDIE

Abstract

Mixtures of experts (MoE) models are a ubiquitous tool for the analysis of heterogeneous data across many fields including statistics, bioinformatics, pattern recognition, economics, and medicine, among many others. They provide conditional constructions for regression in which the mixture weights, along with the component densities, are explained by the predictors, allowing for flexibility in the modeling of data arising from complex data generating processes. In this thesis, we study the approximation capabilities and model estimation and selection properties, of a wide variety of mixture distributions, with a particular focus on a rich family of MoE models in a high-dimensional setting, including MoE models with Gaussian experts and softmax or Gaussian gating functions, which are the most popular choices and are powerful tools for modeling complex non-linear relationships between responses and predictors that arise from different subpopulations. We consider both the theoretical statistical and methodological aspects, and the numerical tools, related to the conception of these models, as well as to their data-driven estimation and model selection.

More precisely, in this thesis, we first review the universal approximation properties of classical mixture distributions in order to prepare the theoretical framework and to clarify some unclear and vague statements in the literature, before considering them in the context of MoE models. In particular, we prove that, to an arbitrary degree of accuracy, location-scale mixtures of a continuous probability density function (PDF) can approximate any continuous PDF, uniformly, on a compact set; and location-scale mixtures of an essentially-bounded PDF can approximate any PDF in Lebesgue spaces. Then, after improving upon approximation results in the context of unconditional mixture distributions, we study the universal approximation capabilities of MoE models in a variety of contexts, including conditional density approximation and approximate Bayesian computation (ABC). Given input and output variables are both compactly supported, we provide denseness results in Lebesgue spaces for conditional PDFs. Moreover, we prove that the quasi-posterior distribution resulting from ABC with surrogate posteriors built from finite Gaussian mixtures using an inverse regression approach, converges to the true one, under standard conditions. Finally, we establish non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds on the penalties hold true, in high-dimensional regression scenarios for a variety of MoE regression models, including Gaussian-gated and softmax-gated Gaussian MoE, based on an inverse regression strategy or a Lasso penalization, respectively. In particular, our oracle inequalities show that the performance in Jensen–Kullback–Leibler type loss of our penalized maximum likelihood estimators are roughly comparable to that of oracle models if we take large enough the constants in front of the penalties, whose forms are only known up to multiplicative constants and proportional to the dimensions of models. Such theoretical justifications of the penalty shapes motivate us to make use of the slope heuristic criterion to select several hyperparameters, including the number of mixture components, the amount of sparsity (the coefficients and ranks sparsity levels), the degree of polynomial mean functions, and the potential hidden block-diagonal structures of the covariance matrices of the multivariate predictor or response variable. To support our theoretical results and the statistical study of non-asymptotic model selection in a variety of MoE models, we perform numerical studies by considering simulated and real data, which highlight the performance of our finite-sample oracle inequality results.

Keywords: Mixture of experts; mixture models; universal approximation; penalized maximum likelihood; feature selection; non-asymptotic model selection; high-dimensional statistics; Lasso; regularization; inverse regression; Wasserstein distance; EM algorithm; MM algorithm; proximal-Newton; coordinate ascent; clustering; classification; prediction; approximate Bayesian computation.

Acknowledgement

The completion of this thesis could not have been achieved without the support and help of my advisors, co-authors, Ph.D. committee, colleagues, family, and friends whom I wish to acknowledge wholeheartedly.

First of all, I would like to thank my Ph.D. advisor, Professor Faïcel Chamroukhi, for having constantly supported and guided me throughout my thesis. I gratefully acknowledge all of the time you have spent discussing to all of my ideas, and all of the effort that you have put into fixing my bad grammar. Most of all, I appreciate your friendship, confidence, patience, motivation and for having guided me during these three unforgettable years. Faïcel has been supportive and has given me the freedom to pursue various projects. He has also provided insightful discussions about the research and helpful career advice. I had the immense pleasure of working with you both professionally as well as personally. Secondly, I wish to express my special gratitude to my co-advisor, Doctor (Senior Lecturer) Hien Duy Nguyen, for his support and his availability. I am also very grateful to Hien for his scientific advice, his knowledge and many insightful discussions and suggestions. I am thankful for all of the times you let me rant and the way you respectfully acknowledge my crazy opinions. I greatly appreciate all of your support and your encouragement for me to do better, even when I am resistant to the idea. Hien has had a massive impact on my research, my thinking, and my professional development. His contributions extend far beyond this thesis. I look forward to working with both of you in the future.

Next, I am grateful to my co-authors, Professor Geoffrey McLachlan, Senior Researcher (Directeur de Recherche) Florence Forbes and Doctor (Chargé de Recherche) Julyan Arbel, for their help as well as contribution on my publications and especially for giving me the chance to collaborate on their research. Geoffrey and Florence were and remain one of my best role models for scientists, mentors, and teachers. Special thanks must be given to Doctor Hien Duy Nguyen and Senior Researcher Florence Forbes for providing me with financial support for a Visiting PhD Fellowship at the Centre Inria Grenoble-Rhône-Alpes and giving me the opportunity to participate in the LANDER (Latent Analysis, Adversarial Networks, and DimEnsionality Reduction) research project.

In particular, I would like to express very special thanks to the two rapporteurs of my thesis: Professor Sylvain Arlot and Professor Judith Rousseau, for the time spent reading my manuscript and for the efforts in providing very enriching comments, which have helped me produce a much-improved manuscript. I wish also to express my gratitude to the other members of the jury for my defense: Professor Christophe Biernacki and Doctor Gaëlle Chagny, for agreeing to be part of my committee.

I wish also to express my thanks to the members of my follow-up committee ("*comité de suivi individuel*" or "CSI"), Professor Mustapha Lebbah and Professor Francesco Amoroso, for several helpful comments concerning on the progression of my Ph.D. and the working environment during the preparation of this thesis.

Furthermore, a very special thanks to Professor Eric Ricard for his kindness and support as a director of LMMO (*Laboratoire de Mathématiques Nicolas Oresme*) which belongs to *Université de Caen Normandie* and *Centre National de la Recherche Scientifique*, in particular, for interesting discussions with him and for his suggestions in functional analysis. Then I also would like to express my gratitude to the secretaries, Marie, Anita, Sonia, Axelle and Marie, who have accompanied and always patiently answered my administrative problems during the last three years. Furthermore, thank you to everyone at the *Université de Caen Normandie*, for providing me a valuable *contrat doctoral* granted by MESRI (*Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation*) and such a wonderful environment in which to do my research.

Moreover, I wish to express my special thanks to Professor Le Thi Hoai An for providing me the

valuable master internship at Laboratoire d'Informatique Théorique et Appliquée, Université de Lorraine and Professor Dang Duc Trong for guiding me the bachelor thesis, which play important steps on my academic career ladder. Next, I am very grateful to Professor Long Nguyen, Professor Vincent Rivoirard, Professor Duong Minh Duc, Professor Marc Peigné, Professor Mounir Haddou, Doctor (INRIA Research scientist) Jing-Rebecca Li, Professor François James, Professor Laurence Halpern, Doctor (Maître de conférences) Juliette Ryan, Professor Jean-Yves Dauxois, Professor Nabendu Pal, Doctor (Maître de conférences) Xavier Gendre, Doctor (CNRS Research engineer) Laurent Risser, Doctor (Lecturer), Associate Professor Ngo Hoang Long, Professor Hien Tran, Professor Stefan Ankirchner, Professor Andrew Ng, for their important academic lectures during my undergraduate and graduate programs, which has motivated me to pursue the doctoral program in Statistics and Data Science, particularly in mixture models and high-dimensional statistics. Then, I would like to thank Assistant Professor Nhat Ho and Directeur de Recherche Tuan Ngo Dac, for thoughtful and insightful discussions about my future career research projects. I wish also to express my thanks to Doctor Van Hà Hoang and Doctor Vinh Thanh Ho for their help and useful information when I started to do master internship and to look for my doctoral scholarship.

Next, I would like to thank *Statify team*, in particular, Directeur de Recherche Stéphane Girard, Assistant Professor Jean-Baptiste Durand, Senior researcher Sophie Achard, Doctor Pierre Wolinski, Doctor Pascal Alain Dkengne Sielenou, Doctor Benoit Kugler, Ph.D. Student Alexandre Constantin, Ph.D. Student Mariia Vladimirova, Ph.D. Student Meryem Bousebata, Ph.D. Student Daria Bystrova, Ph.D Student Theo Moins, Ph.D Student Minh-Tri Le; *LMNO team*, in particular, Ph.D. Student Nhat Thien Pham, Doctor Huy Hung Le, Doctor Bao Tuyen Huynh, Doctor Marius Bartcus, Research Engineer Florian Lecoq, Doctor José G. Gómez García, Doctor Van Thanh Nguyen, Doctor André Sesboüé, Ph.D. Student Nacer Sellila, Doctor Etienne Menard, Doctor Angelot Behajaina, Doctor Julien Poirier, Doctor Arnaud Plessis, Doctor Guillaume Gandolfi, Doctor Frank Taipe Huisa, Doctor Vlerë Mehmeti, Doctor (Maître de conférences) Léonard Cadilhac, Doctor Mostafa Kadiri, Doctor Rubén Muñoz-Bertrand, Ph.D. Student Stavroula Makri, Ph.D. Student Emmanuel Graff, Ph.D. Student Tiphaine Beaumont, Ph.D. Student Etienne Emmelin, Doctor Ndeye Coumba Sarr, Ph.D. Student Dorian Berger; *many other friends, classmates, alumni friends from French-Vietnamese Master 2 program in Applied Mathematics, Vietnam National University-Ho Chi Minh Univeristy of Science, Vietnam, Hung Vuong High School for The Gifted, Nguyen Quoc Phu Secondary School,* in particular, Ph.D. Student Phat Van Thai, Ph.D. Student Minh Quan Hoang Vu, Ph.D. Student Thuy Trang Ngoc Dinh, Doctor Thu Nguyen, Doctor Binh Nguyen, Ph.D. Student Thai-Son Tu, Nhu Do, Ph.D. Student Minh Toan Nguyen, Minh Nguyen, Doctor Thao Thi Minh Le, Ph.D. Student Arum Lee, Doctor Manh Khang Dao, Doctor Tien Dat Nguyen, Doctor Tan-Binh Phan, Doctor Duc Thach Son Vu, Ph.D. Student Van Thanh Nguyen, Dinh Quoc Thai, Ph.D. Student Minh Hoang Le, Doctor (Maitre de conférences) Le Hoai Minh, Ph.D. Student Luu Hoang Phuc Hau, Ph.D. Student Thanh Huan Vo, Ph.D. Student Ly-Duyen Tran, Ph.D. Student Huy Huynh, Ph.D. Student Cong Bang Trang, Ph.D. Student Van Hoi Nguyen, Tu Tam, Tam Vuong, Nhu Tran, Thi Nguyen, Hong Thai, T1ers 2010-2013, and Hung Vuong High school dormitory 2010-2013; for your support, your advice and all the good times we spent together. It has been a pleasure knowing you during such years and getting to know you more in the future.

Finally, a very special thanks to my parents Van Gung Nguyen and Thi Ra Tran, my elder sisters Thi Cap Nguyen and Thi Cam Nhan Nguyen, my inspirational secondary and high school mathematics teachers, Nguyen Thi Le Hang, Tran Van Tri, Vu Mai Trang, my lovely girlfriend, Ph.D. Student Dung Ngoc Nguyen, thank you for being such strong anchors in my life. Dung, thank you very much for your companionship through these trying years. I am very fortunate to have such strong love and to be so supported. Thank you for always being there for me. There is not enough space for me to fully convey my full appreciation for all of you.

Caen, December 14, 2021
TrungTin Nguyen

Contents

List of Figures

List of abbreviations

Below, the list of abbreviations that has been used throughout this thesis, can be found.

ABC	Approximate Bayesian Computation
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BLLiM	Block-diagonal covariance for Gaussian Locally-Linear Mapping
BLoME	Block-diagonal Localized Mixture of Experts
CAPUSHE	CAlibrating Penalty Using Slope HEuristics
EM	Expectation-Maximization (algorithm)
GLLiM	Gaussian Locally Linear Mapping
GLoME	Gaussian-gated Localized Mixture of Experts
i.i.d.	Independent and Identically Distributed
HGLLiM	Hierarchical Gaussian Locally Linear Mapping
LASSO	Least Absolute Shrinkage and Selection Operator
LinBoSGaBloME	Linear-combination-of-Bounded-functions Softmax-Gated Block-diagonal Mixture of Experts
LinBoSGaME	Linear-combination-of-Bounded-functions Softmax-Gated Mixture of Experts
MMD	Maximum Mean Discrepancy
MW_2	Mixture-Wasserstein distance
MLE	Maximum Likelihood Estimate/Estimation/Estimator
MM	Minorization–Maximization
MoE	Mixture of Experts
MS	MisSpecified
NLL	Negative Log-Likelihood
NVS	Normed Vector Space
PDF	Probability Density Function
PMLE	Penalized Maximum Mikelihood Estimators
PSGaBloME	Polynomial Softmax-Gated Block-diagonal Mixture of Experts
WS	Well-Specified

Chapter 1

Introduction

Contents

1.1 Mixture of experts modelling framework

Mixture of experts (MoE) models, originally introduced as neural networks (NNs) in Jacobs et al. (1991) and Jordan & Jacobs (1994), where they were used to model complex and heterogeneous data generating processes (DGPs). The main idea of MoE is a divide-and-conquer principle that proposes dividing a complex problem into a set of simpler subproblems and then one or more specialized problem-solving tools, or experts, are assigned to each of the subproblems. A schematic diagram of an MoE model as a NN is provided in Figure 1.1.

MoE are flexible models that generalize the classical finite mixture models as well as finite mixtures of regression models (McLachlan & Peel, 2000, Section 5.13), and are widely used in statistics and machine learning, thanks to their flexibility and the abundance of applicable statistical estimation and model selection tools. Their flexibility comes from allowing the mixture weights (or the gating functions, gating networks) to depend on the explanatory variables, along with the component densities (or experts). This permits the modeling of data arising from more complex data generating processes than the classical finite mixtures and finite mixtures of regression models, whose mixing parameters are independent of the covariates. Due to their flexibility, MoE can be used in many statistical problems, namely to cluster or classify data, to estimate conditional densities, to conduct regression analysis and to analyze regression outcomes. Detailed reviews on practical and theoretical aspects of MoE models can be found in Yuksel et al. (2012), Masoudnia & Ebrahimpour (2014) and Nguyen & Chamroukhi (2018).

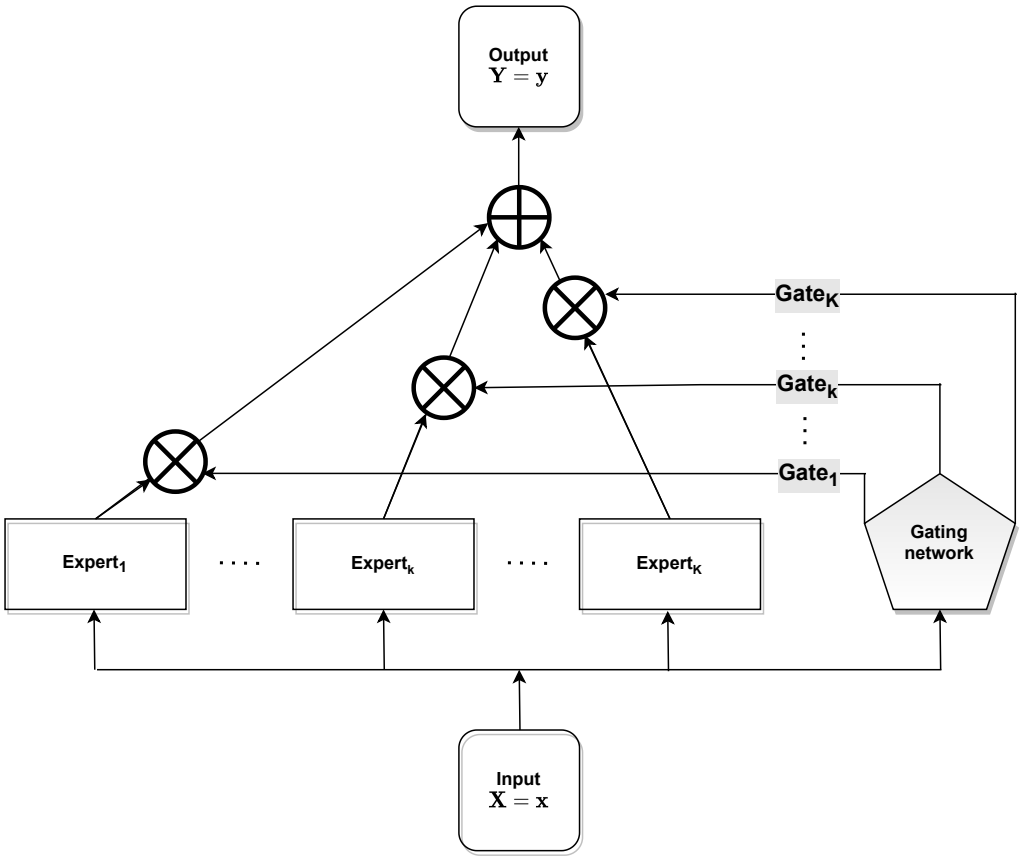


Figure 1.1: Schematic diagram of the NN architecture of a K-component MoE model.

Statistically, we consider a regression framework and aim at capturing the potential nonlinear relationship between a multivariate response $\mathbf{Y}$ and a vector of covariates $\mathbf{X}$. Here and subsequently, $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := \left(\mathbf{X}_i, \mathbf{Y}_i\right)_{i\in[n]}$, $[n] = \{1, \ldots, n\}, n \in \mathbb{N}^\star$, denotes a random sample, and $\mathbf{x}$ and $\mathbf{y}$ stands for the observed values of the random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. We assume that the response variable $\mathbf{Y}$ depends on the explanatory variable $\mathbf{X}$ through a regression-type model. The explanatory variable goes by different names, such as covariate, predictor, independent variable, feature, or sometimes just variable. The response variable is often called the output or dependent variable. Throughout this thesis, we will use all of these terms interchangeably.

Firstly, we focus on a natural starting point for setting up a more probabilistic MoE model, often called "direct application" of mixture modelling; see Titterington et al., 1985 for more details. Then, we will specify an alternative perspective on mixture modelling in Sections 1.1.1 and 1.1.2 that provide an "analytic" view, complementary to this "synthetic" or "generative" view. Regarding the former perspective, suppose that the population from which we are sampling is heterogeneous, *i.e.,* given an input $\mathbf{X} = \mathbf{x}$, there are multiple groups (that can be interpreted as clusters), indexed by $k \in [K]$, present in the population in proportions $g_k\left(\mathbf{x}; \boldsymbol{\omega}\right), k \in [K]$, where $\boldsymbol{\omega}$ is some parame-

ter vector parametrising the input-dependent proportions function $g_k(\mathbf{x};\cdot)$, with $g_k(\mathbf{x};\boldsymbol{\omega}) > 0$, and $\sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\omega}) = 1$. In this way, there is a latent unobserved random variable representation of the mixture model (usually called *latent* or *allocation variables*), involving the latent cluster membership of each observation, denoted by $Z \in [K], K \in \mathbb{N}^\star$, where $Z = k$ if observation $\mathbf{y}$ given the input $\mathbf{x}$ belongs to cluster k for $k \in [K]$. Next, the conditional relationship between Z and the input $\mathbf{X}$ can be characterized by

$$p(Z = k|\mathbf{X} = \mathbf{x}) = g_k(\mathbf{x};\boldsymbol{\omega}). \tag{1.1.1}$$

Then, we can characterize the relationship between the output $\mathbf{y}$ and the input $\mathbf{X}$ by

$$p(\mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k) = \phi_k(\mathbf{y};\boldsymbol{\theta}_k(\mathbf{x})), \tag{1.1.2}$$

where $\boldsymbol{\theta}_k(\mathbf{x})$ is some parameter vector and, given an input $\mathbf{x}$, $\phi_k(\cdot;\boldsymbol{\theta}_k(\mathbf{x}))$ is a probability density function (PDF). We can imagine that given an input $\mathbf{X} = \mathbf{x}$, the observed output $\mathbf{y}$, drawn from the population, is generated in two step: firstly, the group Z is drawn from a multinomial distribution with a single trial and probabilities equal to $\mathbf{g}(\mathbf{x};\boldsymbol{\omega}) = (g_k(\mathbf{x};\boldsymbol{\omega}))_{k\in[K]}$; and secondly, given $\mathbf{X} = \mathbf{x}, Z = k$, the output $\mathbf{y}$ is drawn from $\phi_k(\mathbf{y};\boldsymbol{\theta}_k(\mathbf{x}))$.

Note that this two-stage sampling gives exactly the same models in analytic views, see (1.1.4) and (1.1.15) for more details, for the conditional distribution of $\mathbf{y}|\mathbf{x}$. Indeed, via characterizations (1.1.1) and (1.1.2), and by using the law of total probability, we can characterize the marginal relationship between the response and the input, unconditional on Z, via the expression

$$\begin{aligned} p(\mathbf{y}|\mathbf{X} = \mathbf{x}) &= \sum_{k=1}^{K} p(\mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k)\, p(Z = k|\mathbf{X} = \mathbf{x}) \\ &= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\omega})\, \phi_k(\mathbf{y};\boldsymbol{\theta}_k(\mathbf{x})) =: s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}), \end{aligned} \tag{1.1.3}$$

where $\boldsymbol{\psi} = (\boldsymbol{\omega}, \boldsymbol{\theta})$ is the vector of all parameter elements that are required in characterizing (1.1.3), see Figure 1.2 for more details.

For a better comparison between a standard MoE regression model (in which all model parameters are functions of covariates) and the special cases, where some of the model parameters do not depend on covariates, the four models in the MoE framework are presented in Figure 1.3; see also Fruhwirth-Schnatter et al. (2019, Chapter 12) for more detail. In the context of regression, MoE models with Gaussian experts and softmax or Gaussian gating functions are the most popular choices and are powerful tools for modeling more complex non-linear relationships between responses and predictors that arise from different subpopulations. This is largely studied because of their universal approximation properties, see Chapter 2 for more details, which have been extensively studied for not only finite mixture models (Genovese & Wasserman, 2000, Rakhlin et al., 2005, Nguyen, 2013, Ho et al., 2016a,b, Nguyen et al., 2020d,b) but also conditional densities of MoE models (Jiang & Tanner, 1999a, Norets et al., 2010, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2019, 2021a).

More precisely, Chapter 2 (see also Section 1.4) provides a detailed exposition of these universal approximation properties. In particular, we prove that, to an arbitrary degree of accuracy, location-scale mixtures of a continuous PDF can approximate any continuous PDF, uniformly, on a compact set; and for any finite $p \geq 1$, location-scale mixtures of an essentially-bounded PDF can approximate any PDF, in the L_p norm. Moreover, given input and output variables that are both compactly supported, we demonstrate the richness of the class of MoE models by proving denseness results in Lebesgue spaces for conditional PDFs.

In this thesis, we wish to firstly investigate MoE models with Gaussian gating functions for clustering and regression, first introduced by Xu et al. (1995), which extended the original MoE models of Jacobs et al. (1991). Based on the works of Nguyen et al. (2021c,b), we refer to these models as *Gaussian-gated localized MoE* (GLoME) models and *block-diagonal covariance for localized mixture of experts* (BLoME) models, to be developed in Chapter 3. It is worth pointing out that the BLoME models generalize GLoME models by utilizing a parsimonious covariance structure, via block-diagonal structures for covariance matrices in the Gaussian experts. It is also interesting to point out

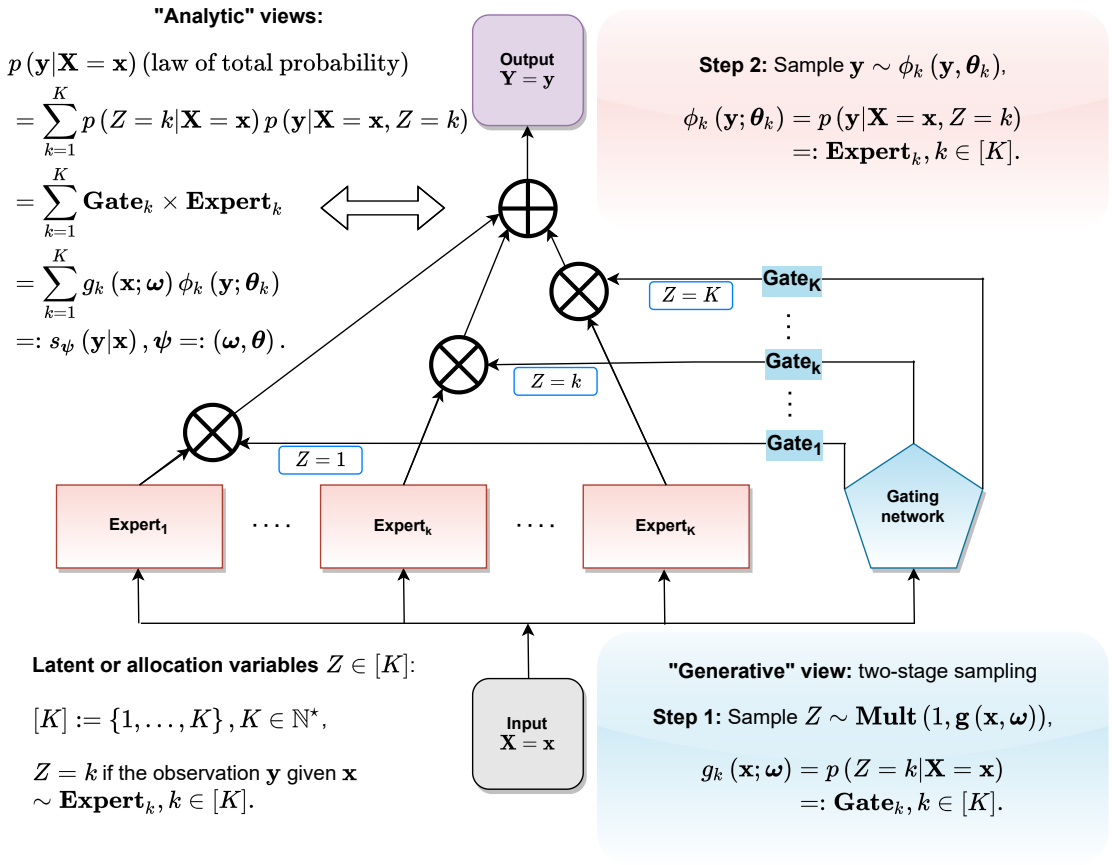
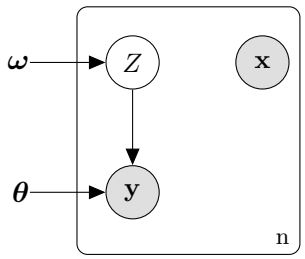
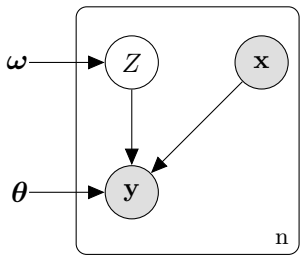


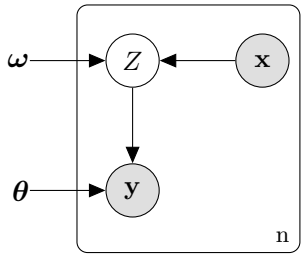
Figure 1.2: Analytic and generative views of a K-component MoE model.



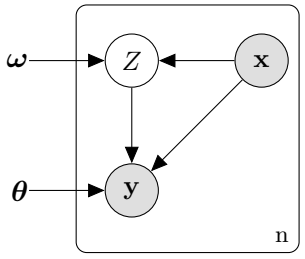
(a) Mixture model



(b) MoE regression model



(c) Simple MoE regression model



(d) Standard MoE regression model

Figure 1.3: Based on the graphical model representation, namely the presence or absence of edges between the covariates $\mathbf{x}$ and the latent variable Z and response variable $\mathbf{y}$, there are four special cases of MoE regression models. More precisely, in Figure 1.3a, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|Z)p(Z)$; in Figure 1.3b, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, Z)p(Z)$; in Figure 1.3c, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|Z)p(Z|\mathbf{x})$; and in Figure 1.3d, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, Z)p(Z|\mathbf{x})$. This figure is inspired from Fruhwirth-Schnatter et al. (2019, Chapter 12, Fig. 12.2).

that supervised *Gaussian locally-linear mapping* (GLLiM) and *block-diagonal covariance for Gaussian locally-linear mapping* (BLLiM) models in Deleforge et al. (2015c) and Devijver et al. (2017) are affine

instances of GLoME and BLoME models, respectively, where linear combination of bounded functions are considered instead of affine for mean functions of Gaussian experts.

Next, Chapter 4 is devoted to the study of MoE models with softmax gating functions based on the idea of the original MoE models of Jacobs et al. (1991). In Chapter 4, we obtain what will be referred to as *linear-combination-of-bounded-functions softmax-gated block-diagonal mixture of experts* (LinBoSGaBloME) regression models. In particular, we simply refer to affine instances of LinBoSGaBloME models as *softmax-gated mixture of experts* (SGaME) regression models. One of the main drawbacks of LinBoSGaBloME models is the difficulty of applying an EM algorithm, which requires an internal iterative numerical optimization procedure (*e.g.,* MM algorithm, iteratively-reweighted least squares, proximal Newton-type procedure, Newton-Raphson algorithm) to update the softmax parameters. GLoME and BLoME models overcome this problem by using the Gaussian gating network that enables us to link GLoME with finite mixtures of Gaussian models. Given its mixture model foundation, the maximization with respect to the parameters of the gating network can be solved analytically within the EM algorithm framework, which decreases the computational complexity of the estimation routine. Furthermore, we then can also make use of well established theoretical results for finite mixture models.

In spite of the fact that the MoE nomenclature has its origins in the machine learning literature (Jacobs et al., 1991), SGaME models have been broadly applied to numerous areas of science, technology and business, for the tasks of classification, clustering, and regression: switching regression models (Quandt, 1972), concomitant variable latent class models (Dayton & Macready, 1988), latent class regression models (DeSarbo & Cron, 1988), mixed models (Wang et al., 1996), functional data analysis and signal processing (Chamroukhi et al., 2009, Samé et al., 2011, Chamroukhi et al., 2013a), finite smooth mixtures (Li et al., 2011), image classification and semantic segmentation tasks (Wang et al., 2020), modeling neural connectivity Bock & Fine (2014), segmentation of spectral images (Cohen & Le Pennec, 2014), climatic change modeling (Nguyen & McLachlan, 2014), phone activity recognition (Lee & Cho, 2014), heterogeneity modeling in neural connectivity data (Eavani et al., 2016), reinforcement learning (He et al., 2016), the tasks of language modeling and machine translation (Shazeer et al., 2017), multi-modal deep generative models on different sets of modalities including a challenging image-language dataset (Shi et al., 2019), anomaly detection (Yu et al., 2021), just to name a few.

The important point to note here is that both GLoME and BLoME models have been also thoroughly studied in the statistics and machine learning literatures and their forms appear in many different guises, including localized MoE (Ramamurti & Ghosh, 1996, 1998, Moerland, 1999, Bouchard, 2003), normalized Gaussian networks (Sato & Ishii, 2000), MoE modeling of priors in Bayesian nonparametric regression (Norets & Pelenis, 2014, Norets & Pati, 2017), cluster-weighted modeling (Ingrassia et al., 2012), GLLiM in inverse regression (Deleforge et al., 2015c), BLLiM model (Devijver et al., 2017), deep mixture of linear inverse regressions (Lathuilière et al., 2017), hierarchical Gaussian locally linear mapping structured mixture (HGLLiM) model (Tu et al., 2019), multiple-output Gaussian gated mixture of linear experts (Nguyen et al., 2019), and approximate Bayesian computation with surrogate posteriors using GLLiM (Forbes et al., 2021).

From now on, we are interested in estimating the law of the random variable $\mathbf{Y}$ conditionally on $\mathbf{X}$. The following assumptions will be needed throughout the chapter. We assume that the covariates $\mathbf{X}$ are independent but not necessarily identically distributed. The assumptions on the responses $\mathbf{Y}$ are stronger: conditional on $\mathbf{X}_{[n]}$, the $\mathbf{Y}_i, i \in [n]$, are independent, and each $\mathbf{Y}$ follows a law with true (but unknown) PDF $s_0(\cdot|\mathbf{X} = \mathbf{x})$, which is approximated via MoE models.

1.1.1 GLoME and BLoME models

Motivated by an inverse regression framework where the role of predictor and response variables should be exchanged such that $\mathbf{Y} = (\mathbf{Y}_j)_{j\in[L]}, [L] = \{1, \ldots, L\}$, becomes the input and $\mathbf{X} = (\mathbf{X}_j)_{j\in[D]}$ plays the role of a multivariate output, we consider the following GLoME model, defined by (1.1.4) (see also in Nguyen et al., 2021c). This construction goes back to the work of Li (1991), Deleforge et al. (2015c),

and Perthame et al. (2018). In this way, we define its corresponding conditional PDF as follows:

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y};\boldsymbol{\omega}\right)\phi_D\left(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}),\boldsymbol{\Sigma}_k\right), \tag{1.1.4}$$

$$g_k\left(\mathbf{y};\boldsymbol{\omega}\right) = \frac{\pi_k\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K}\pi_j\phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}. \tag{1.1.5}$$

Here, $g_k(\cdot;\boldsymbol{\omega})$ and $\phi_D\left(\cdot;\boldsymbol{\upsilon}_{k,d}(\cdot),\boldsymbol{\Sigma}_k\right)$, $k\in[K]$, $K\in\mathbb{N}^\star$, $d\in\mathbb{N}^\star$, are called Gaussian gating functions (networks) and Gaussian experts, respectively. Furthermore, we decompose the parameters of the model as follows: $\boldsymbol{\psi}_{K,d}=(\boldsymbol{\omega},\boldsymbol{\upsilon}_d,\boldsymbol{\Sigma})\in\boldsymbol{\Omega}_K\times\boldsymbol{\Upsilon}_{K,d}\times\mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega}=(\boldsymbol{\pi},\boldsymbol{c},\boldsymbol{\Gamma})\in(\boldsymbol{\Pi}_{K-1}\times\mathbf{C}_K\times\boldsymbol{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi}=(\pi_k)_{k\in[K]}$, $\boldsymbol{c}=(\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d=(\boldsymbol{\upsilon}_{k,d})_{k\in[K]}\in\boldsymbol{\Upsilon}_{K,d}$, and $\boldsymbol{\Sigma}=(\boldsymbol{\Sigma}_k)_{k\in[K]}\in\mathbf{V}_K$. Note that $\boldsymbol{\Pi}_{K-1}=\left\{(\pi_k)_{k\in[K]}\in(\mathbb{R}^+)^K,\sum_{k=1}^{K}\pi_k=1\right\}$ is a $K-1$ dimensional probability simplex, $\mathbf{C}_K$ is a set of K-tuples of mean vectors of size $L\times 1$, $\boldsymbol{V}'_K$ is a set of K-tuples of elements in $\mathcal{S}_L^{++}$, where $\mathcal{S}_L^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ is a set of K-tuples of mean functions from $\mathbb{R}^L$ to $\mathbb{R}^D$ depending on a degree d (*e.g.,* a degree of polynomials), and $\mathbf{V}_K$ is a set containing K-tuples from $\mathcal{S}_D^{++}$.

Recall that GLLiM and BLLiM models are affine instances of GLoME and BLoME models and are especially useful for high-dimensional regression data since there exist link functions between the inverse and forward conditional density, see Figure 1.4 for comprehensive classification and nomenclature of standard MoE regression models with Gaussian gating networks. Note that the principle of inverse regression is only useful when the functions $\boldsymbol{\upsilon}_{k,d}(\mathbf{y})$ are linear, because there is then no explicit way to express the law of $\mathbf{Y}|\mathbf{X}$ from that of $\mathbf{X}|\mathbf{Y}$ for higher degree of polynomials. However, to have more consistent notations with the previous affine results of GLLiM, BLLiM models from Deleforge et al. (2015c), Devijver et al. (2017), we decide to use the inverse regression frameworks instead of the forward one.

Next, we describe a characterization of GLLiM and BLLiM models. A GLLiM model, as originally introduced in Deleforge et al. (2015c), is used to capture the nonlinear relationship between the response and the set of covariates in high-dimensional regression data, typically in the case when $D\gg L$, by the K locally affine mappings:

$$\mathbf{Y} = \sum_{k=1}^{K}\mathbb{I}\left(Z=k\right)\left(\mathbf{A}_k^*\mathbf{X}+\mathbf{b}_k^*+\mathbf{E}_k^*\right). \tag{1.1.6}$$

Here, $\mathbb{I}$ is an indicator function and Z is a latent variable capturing a cluster relationship, such that $Z=k$ if $\mathbf{Y}$ originates from cluster $k\in[K]$. Cluster specific affine transformations are defined by matrices $\mathbf{A}_k^*\in\mathbb{R}^{L\times D}$ and vectors $\mathbf{b}_k^*\in\mathbb{R}^L$. Furthermore, $\mathbf{E}_k^*$ are error terms capturing both the reconstruction error due to the local affine approximations and the observation noise in $\mathbb{R}^L$.

Following the common assumption that $\mathbf{E}_k^*$ is a zero-mean Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_k^*\in\mathbb{R}^{L\times L}$, it holds that

$$p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x},Z=k;\boldsymbol{\psi}_K^*\right)=\phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right), \tag{1.1.7}$$

where we denote by $\boldsymbol{\psi}_K^*$ the vector of model parameters and ϕ_L is the PDF of a Gaussian distribution of dimension L. In order to enforce the affine transformations to be local, $\mathbf{X}$ is defined as a mixture of K Gaussian components as follows:

$$p\left(\mathbf{X}=\mathbf{x}|Z=k;\boldsymbol{\psi}_K^*\right)=\phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right), p\left(Z=k;\boldsymbol{\psi}_k^*\right)=\boldsymbol{\pi}_k^*, \tag{1.1.8}$$

where $\mathbf{c}_k^*\in\mathbb{R}^D, \boldsymbol{\Gamma}_k^*\in\mathbb{R}^{D\times D}$, $\boldsymbol{\pi}^*=(\boldsymbol{\pi}_k^*)_{k\in[K]}\in\boldsymbol{\Pi}_{K-1}^*$, and $\boldsymbol{\Pi}_{K-1}^*$ is the $K-1$ dimensional probability simplex. Then, according to formulas for conditional multivariate Gaussian variables and the following

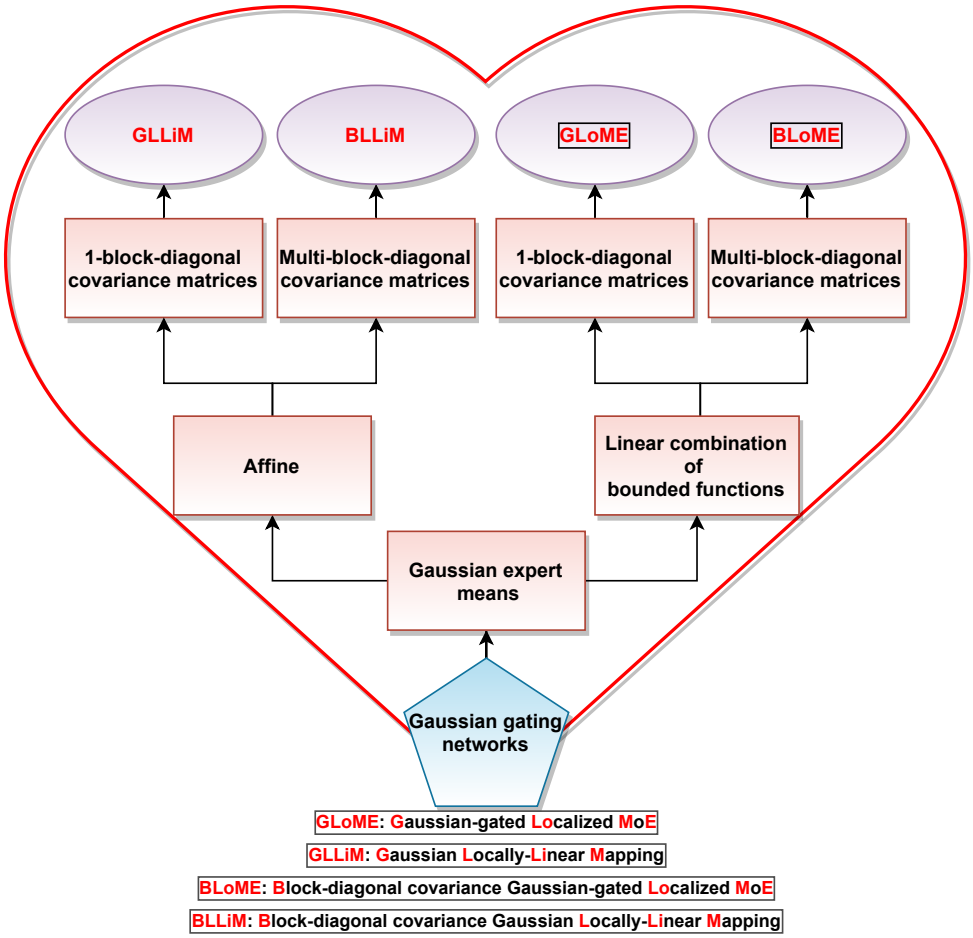


Figure 1.4: A comprehensive classification and nomenclature of standard MoE regression models with Gaussian gating networks.

hierarchical decomposition

$$\begin{aligned} p\left(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x};\boldsymbol{\psi}_K^*\right) &= \sum_{k=1}^{K} p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x},Z=k;\boldsymbol{\psi}_K^*\right) p\left(\mathbf{X}=x|Z=k;\boldsymbol{\psi}_K^*\right) p\left(Z=k;\boldsymbol{\psi}_K^*\right), \\ &= \sum_{k=1}^{K} \boldsymbol{\pi}_k^* \phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right) \phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right), \end{aligned}$$

we obtain the following *forward conditional density* (Deleforge et al., 2015c):

$$p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x};\boldsymbol{\psi}_K^*\right) = \sum_{k=1}^{K} \frac{\boldsymbol{\pi}_k^* \phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right)}{\sum_{j=1}^{K} \boldsymbol{\pi}_j^* \phi_D\left(\mathbf{x};\mathbf{c}_j^*,\boldsymbol{\Gamma}_j^*\right)} \phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right), \tag{1.1.9}$$

where $\boldsymbol{\psi}_K^* = (\boldsymbol{\pi}^*, \boldsymbol{\theta}_K^*) \in \Pi_{K-1} \times \boldsymbol{\Theta}_K^* =: \boldsymbol{\Psi}_K^*$. Here, $\boldsymbol{\theta}_K^* = \left(\mathbf{c}_k^*, \boldsymbol{\Gamma}_k^*, \mathbf{A}_k^*, \mathbf{b}_k^*, \boldsymbol{\Sigma}_k^*\right)_{k\in[K]}$ and

$$\boldsymbol{\Theta}_K^* = \left(\mathbb{R}^D \times \mathcal{S}_D^{++}(\mathbb{R}) \times \mathbb{R}^{L\times D} \times \mathbb{R}^L \times \mathcal{S}_L^{++}(\mathbb{R})\right)^K.$$

Without assuming anything further on the structure of the parameters, the dimension of the model (denoted by $\dim(\cdot)$), is defined as the total number of parameters that have to be estimated, as follows:

$$\dim\left(\boldsymbol{\Psi}_K^*\right) = K\left(1 + D(L+1) + \frac{D(D+1)}{2} + \frac{L(L+1)}{2} + L\right) - 1.$$

It is worth mentioning that $\dim(\boldsymbol{\Psi}_K)$ can be very large compared to the sample size (see, e.g., Deleforge et al., 2015c, Devijver et al., 2017, Perthame et al., 2018 for more details in their real data sets) whenever D is large and $D \gg L$. Furthermore, it is more realistic to make assumptions on the residual covariance matrices $\boldsymbol{\Sigma}_k^*$ of error vectors $\mathbf{E}_k^*$ rather than on $\boldsymbol{\Gamma}_k^*$ (cf. Deleforge et al., 2015c, Section 3). This justifies the use of the inverse regression trick from Deleforge et al. (2015c), which leads a drastic reduction in the number of parameters to be estimated.

More specifically, in (1.1.9), the roles of input and response variables should be exchanged such that $\mathbf{Y}$ becomes the covariates and $\mathbf{X}$ plays the role of the multivariate response. Therefore, its corresponding *inverse conditional density* is defined as a Gaussian locally-linear mapping (GLLiM) model, based on the previous hierarchical Gaussian mixture model, as follows:

$$p(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y}, Z=k;\boldsymbol{\psi}_K) = \phi_D(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k), \tag{1.1.10}$$

$$p(\mathbf{Y}=\mathbf{y}|Z=k;\boldsymbol{\psi}_K) = \phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k), p(Z=k;\boldsymbol{\psi}_k) = \pi_k, \tag{1.1.11}$$

$$p(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}_K) = \sum_{k=1}^{K} \frac{\pi_k \phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K}\pi_j\phi_L(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j)} \phi_D(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k), \tag{1.1.12}$$

where $\boldsymbol{\Sigma}_k$ is a $D\times D$ covariance structure (usually diagonal, chosen to reduce the number of parameters) automatically learnt from data, and $\boldsymbol{\psi}_K$ is the set of parameters, denoted by $\boldsymbol{\psi}_K = (\boldsymbol{\pi},\boldsymbol{\theta}_K) \in \boldsymbol{\Pi}_{K-1}\times \boldsymbol{\Theta}_K =: \boldsymbol{\Psi}_K$. An intriguing feature of the GLLiM model is described in Lemma 1.1.1, which is proved in Section 3.2.5.1.

Lemma 1.1.1. *The parameter $\boldsymbol{\psi}_K^*$ in the forward conditional PDF, defined in (1.1.9), can then be deduced from $\boldsymbol{\psi}_K$ in (1.1.12) via the following one-to-one correspondence:*

$$\boldsymbol{\theta}_K = \begin{pmatrix} \mathbf{c}_k \\ \boldsymbol{\Gamma}_k \\ \mathbf{A}_k \\ \mathbf{b}_k \\ \boldsymbol{\Sigma}_k \end{pmatrix}_{k\in[K]} \mapsto \begin{pmatrix} \mathbf{c}_k^* \\ \boldsymbol{\Gamma}_k^* \\ \mathbf{A}_k^* \\ \mathbf{b}_k^* \\ \boldsymbol{\Sigma}_k^* \end{pmatrix}_{k\in[K]} = \begin{pmatrix} \mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k \\ \boldsymbol{\Sigma}_k+\mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top \\ \boldsymbol{\Sigma}_k^*\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} \\ \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1}\mathbf{c}_k - \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k) \\ \left(\boldsymbol{\Gamma}_k^{-1}+\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1} \end{pmatrix}_{k\in[K]} \in \boldsymbol{\Theta}_K^*, \tag{1.1.13}$$

with the note that $\boldsymbol{\pi}^* \equiv \boldsymbol{\pi}$.

We wish to provide some simulated examples of GLoME regression models on 1-dimensional data sets, that is, with $L = D = 1$. We construct simulated data sets following two scenarios: a *well-specified* (WS) case in which the true forward conditional density s_0^*, which can be estimated via an affine Gaussian mean instance of GLoME, namely GLLiM model, based on inverse conditional PDF $s_{\boldsymbol{\psi}_{K,d}}$ using inverse regression strategy, belongs to the class of proposed models:

$$\begin{aligned} s_0^*(y|x) = &\frac{\phi(x;0.2,0.1)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi(y;-5x+2,0.09) \\ &+ \frac{\phi(x;0.8,0.15)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi(y;0.1x,0.09), \end{aligned}$$

and a *misspecified* (MS) case, whereupon such an assumption is not true:

$$\begin{aligned} s_0^*(y|x) = &\frac{\phi(x;0.2,0.1)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi\left(y;x^2-6x+1,0.09\right) \\ &+ \frac{\phi(x;0.8,0.15)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi\left(y;-0.4x^2,0.09\right). \end{aligned}$$

Here we assume that the true number of mixture components $K_0 = 2$.

Figures 1.5a and 1.5e show some typical realizations of 2000 data points arising from the WS and MS scenarios. Note that by using GLLiM, the penalized maximum likelihood estimator of GLLiM, as introduced in Section 1.2.3 and Chapter 3, performs well in the WS setting (Figures 1.5b to 1.5d). In the MS case, we expect our procedure, namely the GLLiM-EM algorithm introduced in Section 3.2.3,

using slope heuristic, see Section 1.2.4 for more details, to automatically balance the model bias and its variance (Figures 1.5f to 1.5h), which leads to the choice of a complex model, with 4 mixture components. This observation will be elaborated upon in the subsequent descriptions and experiments, see Sections 1.2.3, 1.2.4 and 3.2.3, respectively.

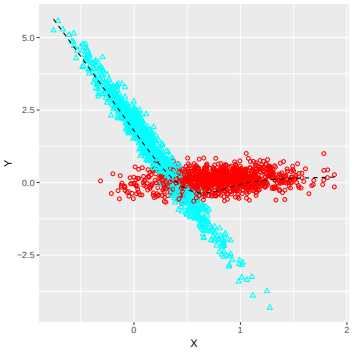
In the BLoME model, we wish to make use of block-diagonal structures by replacing $\mathbf{\Sigma}_k$ and $\mathbf{V}_K$ by $\mathbf{\Sigma}_k(\mathbf{B}_k)$ and $\mathbf{V}_K(\mathbf{B})$, defined in (1.1.14), respectively (see, *e.g.,* Devijver et al., 2017, Devijver & Gallopin, 2018, Nguyen et al., 2021b). This block-diagonal structures for covariance matrices are not only used for a trade-off between complexity and sparsity but also motivated by some real applications, where we want to perform prediction on data sets with heterogeneous observations and hidden graph-structured interactions between covariates; for instance, for gene expression data sets in which conditionally on the phenotypic response, genes interact with few other genes only, *i.e.,* there are small modules of correlated genes (see Devijver et al., 2017, Devijver & Gallopin, 2018 for more details). To be more precise, for $k \in [K]$, we decompose $\mathbf{\Sigma}_k(\mathbf{B}_k)$ into G_k blocks, $G_k \in \mathbb{N}^\star$, and we denote by $d_k^{[g]}$ the set of variables into the gth group, for $g \in [G_k]$, and by $\text{card}\left(d_k^{[g]}\right)$ the number of variables in the corresponding set. Then, we define $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$ to be a block structure for the cluster k, and $\mathbf{B} = (\mathbf{B}_k)_{k\in[K]}$ to be the covariate indexes into each group for each cluster. In this way, to construct the block-diagonal covariance matrices, up to a permutation, we make the following definition: $\mathbf{V}_K(\mathbf{B}) = (\mathbf{V}_k(\mathbf{B}_k))_{k\in[K]}$, for every $k \in [K]$,

$$\mathbf{V}_k(\mathbf{B}_k) = \left\{ \mathbf{\Sigma}_k(\mathbf{B}_k) \in \mathcal{S}_D^{++} \;\middle|\; \begin{array}{l} \mathbf{\Sigma}_k(\mathbf{B}_k) = \mathbf{P}_k \begin{pmatrix} \mathbf{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \mathbf{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\text{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}, \tag{1.1.14}$$

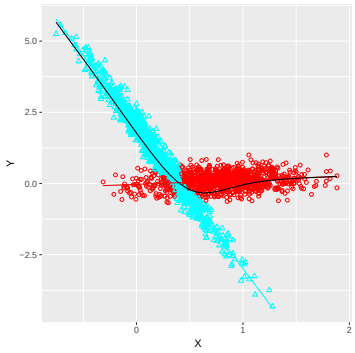
where $\mathbf{P}_k$ corresponds to the permutation leading to a block-diagonal matrix in cluster k. It is worth pointing out that outside the blocks, all coefficients of the matrix are zeros and we also authorize reordering of the blocks: *e.g.,* $\{(1,3);(2,4)\}$ is identical to $\{(2,4);(1,3)\}$, and the permutation inside blocks: *e.g.,* the partition of 4 variables into blocks $\{(1,3);(2,4)\}$ is the same as the partition $\{(3,1);(4,2)\}$.

It is interesting to point out that GLLiM and BLLiM models in Deleforge et al. (2015c), Devijver et al. (2017) are affine instances of GLoME and BLoME models, respectively, where linear combination of bounded functions (*e.g.,* polynomials) are considered instead of affine mean functions for the Gaussian experts. The BLLiM framework aims to model a sample of high-dimensional regression data issued from a heterogeneous population with hidden graph-structured interaction between covariates. In particular, the BLLiM model is considered as a good candidate for performing model-based clustering and for predicting the response in situations affected by the "curse of dimensionality" phenomenon, where the number of parameters could be larger than the sample size. Indeed, to deal with high-dimensional regression problems, the BLLiM model is based on an inverse regression strategy, which inverts the role of the high-dimensional predictor and the multivariate response. Therefore, the number of parameters to estimate is drastically reduced. More precisely, BLLiM utilizes GLLiM, described in Deleforge et al. (2015a,c), in conjunction with a block-diagonal structure hypothesis on the residual covariance matrices to make a trade-off between complexity and sparsity.

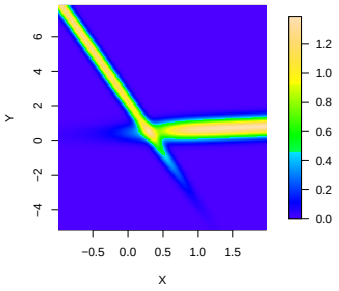
This prediction model is fully parametric and highly interpretable. For instance, it might be useful for the analysis of transcriptomic data in molecular biology to classify observations or predict phenotypic states, as for example disease versus non disease or tumor versus normal (Golub et al., 1999, Nguyen & Rocke, 2002, Lê Cao et al., 2008). Indeed, if the predictor variables are gene expression data measured by microarrays or by the RNA-seq technologies and the response is a phenotypic variable, situations affected by the BLLiM not only provides clusters of individuals based on the relation between gene expression data and the phenotype, but also implies a gene regulatory network specific to each cluster of individuals (see Devijver et al., 2017 for more details).



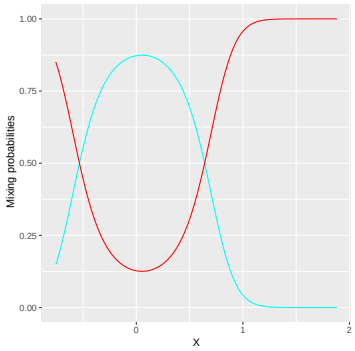
(a) Typical realization of an WS example



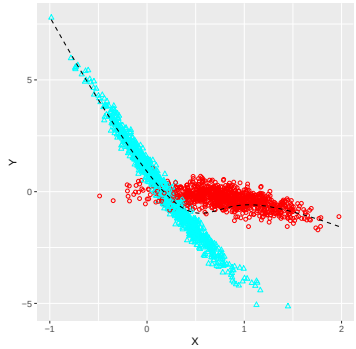
(b) Clustering by GLoME



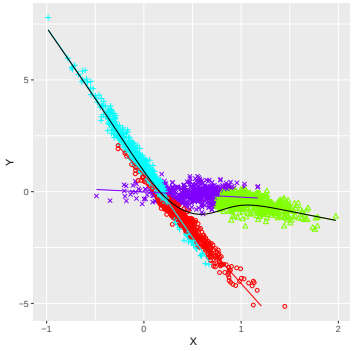
(c) 2D view of the resulting conditional density with the 2 regression components



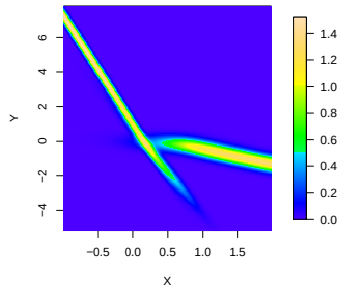
(d) Gating network probabilities



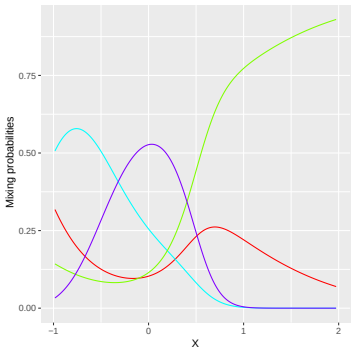
(e) Typical realization of an MS example



(f) Clustering by GLoME



(g) 2D view of the resulting conditional density with the 4 regression components



(h) Gating network probabilities

Figure 1.5: Clustering deduced from the estimated conditional density of GLoME by a maximum a posteriori probability (MAP) principle with 2000 data points of the examples from the WS and MS scenarios. The dash and solid black curves present the true and estimated mean functions.

1.1.2 SGaME and LinBoSGaBloME models

We will consider the statistical frameworks in which we model a sample of high-dimensional regression data issued from a heterogeneous population via a suitable MoE model with softmax gating functions. We emphasize that the dimension of the input $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ and/or the output $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^q$ variable are/is typically much higher than the sample size n. In this thesis, based on the original MoE models from Jacobs et al. (1991), we aim to establish a MoE model with softmax gating functions as generic as possible such that it can be used to handle with high-dimensional regression datasets and to study oracle inequalities. To do that, we first define $s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x})$ to be a conditional PDF of MoE model as follows:

$$s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} g_{\mathbf{w},k}(\mathbf{x})\, \phi_q\left(\mathbf{y}; \boldsymbol{\upsilon}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), \text{ where,} \tag{1.1.15}$$

$$g_{\mathbf{w},k}(\mathbf{x}) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\mathbf{x})\right)}, \mathbf{w}(\mathbf{x}) = \left(w_k(\mathbf{x})\right)_{k\in[K]}. \tag{1.1.16}$$

Here, $g_{\mathbf{w},k}(\cdot)$ and $\phi_q\left(\cdot; \boldsymbol{\upsilon}_k(\cdot), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), k \in [K]$, are called softmax gating functions (or gating networks) and Gaussian experts, respectively. Note that for every $\mathbf{x} \in \mathcal{X}$, $\left(g_{\mathbf{w},k}(\mathbf{x})\right)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$. Furthermore, we decompose the set of model parameters as follows: $\boldsymbol{\psi}_K = (\mathbf{w}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \mathbf{W}_K \times \boldsymbol{\Upsilon}_K \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_K$, $\mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_K$, $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}_k)_{k\in[K]} \in \boldsymbol{\Upsilon}_K$, and $\boldsymbol{\Sigma}(\mathbf{B}) = \left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)\right)_{k\in[K]} \in \boldsymbol{\Sigma}_k(\mathbf{B}_k)$. It is worth noting that $\mathbf{W}_K$ and $\boldsymbol{\Upsilon}_K$ are sets of K-tuples of functions defined in logistic schemes (weights) and mean functions from $\mathbb{R}^p$ to $\mathbb{R}^+$ and $\mathbb{R}^p$ to $\mathbb{R}^q$, respectively; and $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$ is a set containing K-tuples of $\mathcal{S}_q^{++}$ with the block-diagonal structures defined in (1.1.14), where $\mathcal{S}_q^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^q$. Since we need to bound the model complexity using the dimension of model, we have to restrict our attention to LinBoSGaBloME models, where $\mathbf{W}_K$ and $\boldsymbol{\Upsilon}_K$ are defined as the linear combination of a finite set of bounded functions whose coefficients belong to a compact set. When the dimension of both inputs and outputs are not too large, we do not need to select relevant variables. Then, we can work on the previous LinBoSGaBloME models with general structures for means, weights and multi-block-diagonal covariance matrices. In some situation, we do not need to take into account the trade-off between complexity and sparsity for covariance matrices, in LinBoSGaBloME models, we can consider 1-block-diagonal covariance matrices, which is well studied in Montuelle et al. (2014) and will be referred to be as *linear-combination-of-bounded-functions softmax-gated mixture of experts* (LinBoSGaME) regression models. However, to deal with high-dimensional data and to simplify the interpretation of sparsity, in LinBoSGaBloME model, we propose to utilize polynomials for the weights of the softmax gating functions and the Gaussian expert means, which will be referred to as *polynomial softmax-gated block-diagonal mixture of experts* (PSGaBloME) regression models. In particular, we simply refer to affine instances of LinBoSGaBloME models as *softmax-gated mixture of experts* (SGaME) regression models. Compared to the general PSGaBloME model, SGaME model is defined based on 1-block-diagonal covariance matrices. This means that we do not impose the potential hidden graph-structured interactions between covariate variables. Furthermore, instead of using a linear combination of a finite set of bounded functions whose coefficients belong to a compact set, SGaME models utilize the linear functions for both the weights of softmax gating networks and the means of Gaussian experts. The readers are referred to Figure 1.6 for comprehensive classification and nomenclature of standard MoE regression models with softmax gating networks.

1.2 Model selection in mixtures of experts regression models

It is worth pointing out that several hyperparameters must be estimated to construct BLoME and PSGaBloME regression models, including the number of mixtures components (or clusters), the block structure of large covariance matrices specific of each cluster (the size and the number of blocks), the degree of polynomials appearing in gating networks and Gaussian mean experts, the relevant variables and rank sparse models in PSGaBloME. Data driven choices of hyperparameters of learning algorithms belong to the model selection class of problems, which has attracted much attention in statistics and

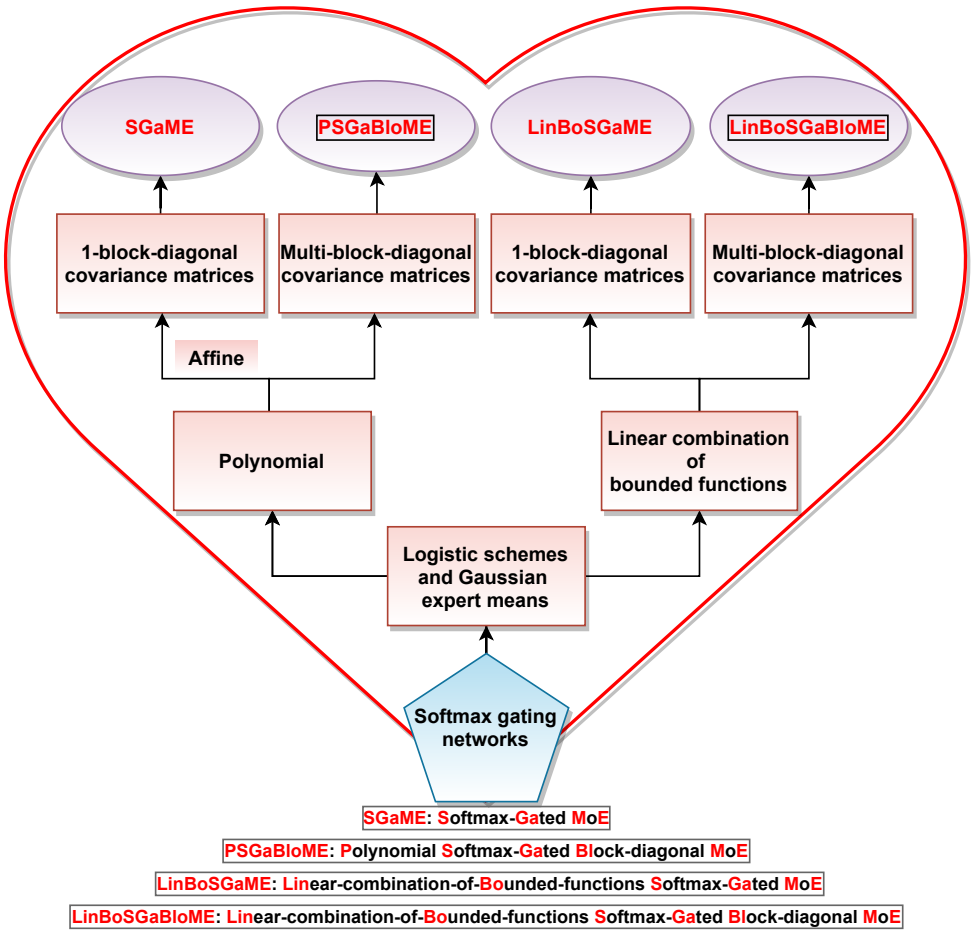


Figure 1.6: A comprehensive classification and nomenclature of standard MoE regression models with softmax gating networks.

machine learning over the last 50 years (Akaike, 1974, Mallows, 1973, Anderson & Burnham, 2002, Massart, 2007). This is a particular instance of the estimator (or model) selection problem: given a family of estimators, how do we choose, using data, one among them whose risk is as small as possible? Note that penalization is one of the main strategies proposed for model selection. It suggests to choose the estimator minimizing the sum of its empirical risk and some penalty terms corresponding to how well the model fits the data, while avoiding overfitting.

In this thesis, we are interested in controlling and accounting for model complexity when selecting the best number of mixture components of a model. In general, model selection is often performed using the Akaike information criterion (AIC; Akaike, 1974) or the Bayesian information criterion (BIC; Schwarz et al., 1978). An important limitation of these criteria, however, is that they are only valid asymptotically. This implies that there are no finite sample guarantees when using AIC or BIC, for choosing between different levels of complexity. Their use in small sample settings is thus ad hoc. To overcome such difficulties, Birgé & Massart (2007) proposed a novel approach, called slope heuristics, supported by a non-asymptotic oracle inequality. This method leads to an optimal data-driven choice of multiplicative constants for penalties. Recent reviews and practical issues regarding the slope heuristic can be found in Baudry et al. (2012), Arlot (2019), and the references given therein.

It should be stressed that a general model selection result, originally established by Massart (2007, Theorem 7.11), guarantees a penalized criterion leads to a good model selection and the penalty being only known up to multiplicative constants and proportional to the dimensions of models. In particular, such multiplicative constants can be calibrated by the slope heuristic approach in a finite sample setting. Then, in the spirit of the concentration inequality-based methods developed in Massart (2007), Massart & Meynet (2011), and Cohen & Le Pennec (2011), a number of finite-sample oracle results have been established for the least absolute shrinkage and selection operator (LASSO)

(Tibshirani, 1996) and general penalized maximum likelihood estimators (PMLE). These results include the works for high dimensional Gaussian graphical models (Devijver & Gallopin, 2018), Gaussian mixture model selection (Maugis & Michel, 2011b,a), finite mixture regression models (Meynet, 2013, Devijver, 2015a,b, 2017b,a), SGaME models without considering high-dimensional setting (Montuelle et al., 2014).

No attempt has been made in the literature to develop a finite-sample oracle inequality for MoE regression models framework for high-dimensional data. In this thesis, to the best of our knowledge, we are the first to provide finite-sample oracle inequalities for several high-dimensional MoE regression models, including the GLoME model (Nguyen et al., 2021c, Section 3.2), BLoME model (Nguyen et al., 2021b, Section 3.3), SGaME model using LASSO (Nguyen et al., 2020c, Section 4.2), and PSGaBloME model (Section 4.3). In particular, our proof strategy makes use of recent novel approaches comprising a model selection theorem for the maximum likelihood estimator (MLE) among a random subcollection (Devijver, 2015b), a non-asymptotic model selection result for detecting a good block-diagonal structure in high-dimensional graphical models (Devijver & Gallopin, 2018) and a reparameterization trick to bound the metric entropy of the Gaussian gating parameter space in GLoME models (Nguyen et al., 2021c), see also Section 3.2 for more details. Note that for the Gaussian gating parameters, the technique for handling the logistic weights in the SGaME models of Montuelle et al. (2014) is not directly applicable to the GLoME or BLoME framework, due to the quadratic form of the canonical link. Therefore, we propose a *reparameterization trick*[1] to bound the metric entropy of the Gaussian gating parameters space; see Equation (3.2.25) and Section 3.2.5.2 for more details. Furthermore, in Nguyen et al. (2021c, Theorem 3.2.3), see also Section 3.2, we extend one of corollaries (in which the authors used linear combination of bounded functions for the functions in softmax gating networks) from Montuelle et al. (2014, Theorem 1) to the quadratic form of the canonical link from gating networks, see more details in Equation (3.2.25) and Lemma 3.2.10.

Among of the main contributions of this thesis are the important theoretical results: finite-sample oracle inequalities that provide non-asymptotic bounds on the risks, and lower bounds on the penalty functions that ensure non-asymptotic theoretical controls on the estimators under the Jensen–Kullback–Leibler loss. These oracle inequalities also provide some theoretical justifications of the penalty shapes when using the slope heuristic for GLLiM, GLoME, BLLiM, BLoME, SGaME, and PSGaBloME models. We emphasize that although the finite-sample oracle inequalities compare performances of our estimators with the best model in the collection, they also allow us to well approximate a rich class of conditional densities if we take enough degree of polynomials of Gaussian expert means (belongs to $\mathcal{D}_{\boldsymbol{\Upsilon}}$) and/or enough clusters (among the set $\mathcal{K}$) in the context of mixture of Gaussian experts (Jiang & Tanner, 1999a, Mendes & Jiang, 2012, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2021a). This leads to the upper bounds on the risks being small, for $\mathcal{D}_{\boldsymbol{\Upsilon}}$ and $\mathcal{K}$ well-chosen.

Especially, aside from important theoretical issues regarding the tightness of the bounds, the way to integrate a priori information and the minimax analysis of our proposed PMLE, we hope that our finite-sample oracle inequalities and corresponding interesting numerical experiments help to partially answer the two following important questions raised in the area of MoE regression models: (1) What number of mixture components K should be chosen, given the sample size n, and (2) Whether it is better to use a few complex experts or combine many simple experts, given the total number of parameters. Note that, such problems are considered in the work of Mendes & Jiang (2012, Proposition 1), where the authors provided some qualitative insights and only suggested a practical method for choosing K and d involving a complexity penalty or cross-validation. Furthermore, their model is only for a non-regularized maximum-likelihood estimation, and thus is not suitable in the high-dimensional setting.

In this thesis, we will consider the parameter selection problem as a model selection problem, by constructing a collection of models, with more or less clusters, complex or simple experts controlling via the orders of polynomial of weights and Gaussian experts, high or low rank sparse models, and

[1]Note that we only use this nomenclature to perform a change of variables of the Gaussian gating parameters space of GLoME models via the logistic weights of SGaME models. This reparameterization trick does not stand for the well-known one of Variational Autoencoders (VAEs) in the deep learning literature (see Kingma & Welling, 2013, for more details).

more or less active coefficients. Then, it will remain to choose a model among this collection. In general, let us denote by $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ the collection of models that we consider, indexed by $\mathcal{M}$. It is worth pointing out that, in contrary to what one might think, having a collection of models that is too large can be detrimental, for example by selecting inconsistent estimators (Bahadur, 1958) or suboptimal estimators (Birgé & Massart, 1993). This is the called model selection paradigm.

Before discussing the finite-sample oracle inequalities for model selection via penalization in MoE regression models, we review some standard facts regarding estimation by contrast minimization.

1.2.1 Minimum contrast estimation

The minimum contrast estimation method is based on the existence of a contrast function, denoted by γ, fulfilling the fundamental property that the unknown conditional PDF satisfies

$$s_0 = \arg\min_{t\in\mathcal{S}} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right].$$

In this way we obtain what will be referred to as the associated loss function, denoted l, that is defined via

$$l\left(s_0, t\right) = \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right], \quad \forall t \in \mathcal{S}.$$

Let us define some empirical contrast γ_n (based on the observation $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := (\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]}$) such that

$$\forall t \in \mathcal{S}, \quad \gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma\left(t, \mathbf{X}_i, \mathbf{Y}_i\right).$$

For the model $\mathbf{m}$, a *minimum contrast estimator* $\widehat{s}_{\mathbf{m}}$ of s_0 is a minimizer of the empirical contrast γ_n over $S_{\mathbf{m}}$, *i.e.*, $\widehat{s}_{\mathbf{m}} = \arg\min_{t\in S_{\mathbf{m}}} \gamma_n(t)$. The idea is that, under reasonable conditions, $\gamma_n(t)$ converges to $\mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right]$, and that there is some hope to get a sensible estimator of s_0, at least if s_0 belongs (or is close enough) to model $S_{\mathbf{m}}$. To measure the quality of such an estimator, we make use of the following *risk* $\mathcal{R}\left(\widehat{s}_{\mathbf{m}}\right) = \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$.

For example, in the density estimation framework, the popular maximum likelihood estimator is a minimum contrast estimator. Indeed, we assume that the sample $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]}$ has the density s_0 w.r.t. a measure μ and consider another density t w.r.t. the same measure. Then, the negative log-likelihood $-\ln\left[t\left(\mathbf{y}|\mathbf{x}\right)\right]$ is the maximum likelihood contrast, and the corresponding loss function is the Kullback–Leibler divergence defined by $\mathrm{KL}\left(s_0, t\right) = \int s_0 \ln\left(\frac{s_0}{t}\right) d\mu$. For a fuller treatment regarding other examples of contrast for regression, classification and Gaussian white noise, we refer the reader to Massart (2007).

1.2.2 The model choice paradigm

The purpose is to select the "best" estimator among the collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$. Let $S_{\widehat{\mathbf{m}}}$ be the model selected by a given model selection procedure. We will denote by $\widehat{s}_{\widehat{\mathbf{m}}}$ the selected estimator and emphasize that both $\widehat{s}_{\mathbf{m}}$ (for any $\mathbf{m}$) and $\widehat{\mathbf{m}}$ are built from the same sample $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right)$. This procedure has been well studied from both an asymptotic and a non-asymptotic point of view.

Ideally, for a given n and a given dataset, one would like to consider $\mathbf{m}^*$ minimizing the risk $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$, with respect to $\mathbf{m} \in \mathcal{M}$. In other words,

$$\mathbf{m}^* \in \arg\min_{\mathbf{m}\in\mathcal{M}} l\left(s_0, \widehat{s}_{\mathbf{m}}\right). \tag{1.2.1}$$

The minimum contrast estimator $\widehat{s}_{\mathbf{m}^*}$ on the corresponding model $S_{\mathbf{m}^*}$ is called an *oracle*. This terminology has previously been introduced by Donoho & Johnstone (1994). Unfortunately, since the loss $l\left(s_0, \widehat{s}_{\mathbf{m}}\right)$ depends on the unknown sample distribution s_0, thus so does $\mathbf{m}^*$ and the oracle $\widehat{s}_{\mathbf{m}^*}$ should not be an estimator of s_0. However, this oracle can serve as a benchmark for building any data driven selection procedure among the collection of estimators $\left(\widehat{s}_{\mathbf{m}}\right)_{m\in\mathcal{M}}$. It is now natural to consider

data-driven criteria to select an estimator which tends to mimic an oracle. In other words, we would like the risk of the selected estimator $\widehat{s}_{\widehat{\mathbf{m}}}$, *i.e.*, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[l(s_0,\widehat{s}_{\widehat{\mathbf{m}}})\right]$, to be as close as possible to the risk of an oracle, *i.e.*, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[l(s_0,\widehat{s}_{\mathbf{m}^*})\right]$.

It is worth pointing out that the non-asymptotic approach (see, *e.g.*, Massart, 2007, Wainwright, 2019 for the complete bibliography) differs from the usual asymptotic point of view in the sense that the number as well as the dimensions of the models in $\mathcal{M}$ may depend on n. We wish to construct a model selection procedure such that the selected model $S_{\widehat{\mathbf{m}}}$ is *optimal*. For instance, it fulfills the following oracle inequality

$$l\left(s_0,\widehat{s}_{\widehat{\mathbf{m}}}\right) \leq C_1 l\left(s_0,\widehat{s}_{\mathbf{m}^*}\right) + \frac{C_2}{n} \tag{1.2.2}$$

with C_1 as close to 1 as possible and C_2/n a remainder term. The oracle inequality is said to be exact if $C_1 = 1$. We expect that this inequality holds either in expected value or with high probability. In particular, when such results are too difficult to be achieved, it suffices to obtain a weaker form:

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[l\left(s_0,\widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}} \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[l\left(s_0,\widehat{s}_{\mathbf{m}^*}\right)\right] + \frac{C_2}{n}. \tag{1.2.3}$$

1.2.3 Model selection via penalization

Now, let us describe how to select a model via minimizing a penalized criterion, to reach a bias/variance compromise. Indeed, we can decompose the loss into a bias and a variance parts as follows:

$$\begin{aligned} l\left(s_0,\widehat{s}_{\mathbf{m}}\right) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}},\mathbf{X},\mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0,\mathbf{X},\mathbf{Y})\right] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}},\mathbf{X},\mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0,\mathbf{X},\mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}},\mathbf{X},\mathbf{Y})\right] + \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}},\mathbf{X},\mathbf{Y})\right] \\ &= \underbrace{l\left(s_0,s_{\mathbf{m}}\right)}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}},\mathbf{X},\mathbf{Y}) - \gamma(s_{\mathbf{m}},\mathbf{X},\mathbf{Y})\right]}_{\text{variance}}, \end{aligned}$$

where $s_{\mathbf{m}} = \arg\min_{t\in S_{\mathbf{m}}} \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(t,\mathbf{X},\mathbf{Y})\right]$ is one of the best approximations of s_0 in $S_{\mathbf{m}}$. It is worth pointing out that in order to minimize the bias, we need a complex model, which fits very closely to the data; and to minimize the variance, we should not consider too complex models, in order to avoid overfitting of the data.

The main methods to account for such model selection procedures are cross-validation and hold-out (see, *e.g.*, Arlot & Celisse, 2010, Maillard, 2020 for the complete bibliography), or penalized criteria. It is emphasized that the main difficulty in carrying out the cross-validation and hold-out is time complexity, particularly in high-dimensional setting. Therefore, the choice of penalization criteria seems to the best adapted to our high-dimensional MoE regression models.

Let us describe the method in more details. The *model selection via penalization* procedure consists in considering some proper *penalty function* pen: $\mathcal{M} \to \mathbb{R}_+$ and taking $\widehat{\mathbf{m}}$ that minimizes the *penalized criterion*, defined as $\gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m})$ over $\mathcal{M}$. This means that we select

$$\widehat{\mathbf{m}} = \arg\min_{\mathbf{m}\in\mathcal{M}} \left\{\gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m})\right\}. \tag{1.2.4}$$

In other words, in the context of the maximum likelihood estimator for the regression case, for a given choice of pen($\mathbf{m}$), the *selected model* $S_{\widehat{\mathbf{m}}}$ is chosen as the one whose index is an η'-almost minimizer of the sum of the negative log-likelihood (NLL) and this penalty:

$$\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \text{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m}\in\mathcal{M}} \left(\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \text{pen}(\mathbf{m})\right) + \eta'. \tag{1.2.5}$$

Here, $\widehat{s}_{\mathbf{m}}$ is defined as the η-minimizer of the NLL:

$$\sum_{i=1}^{n} -\ln\left(s_{\widehat{s}_{\mathbf{m}}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \sum_{i=1}^{n} -\ln\left(s_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \eta, \tag{1.2.6}$$

where the error term η is necessary when the infimum may not be unique or even not be reached. Note that $\widehat{s}_{\widehat{\mathbf{m}}}$ is then called the η'-penalized likelihood estimator and depends on both the error terms η and η'. From hereon in, the terms *"best data-driven model or estimate"* and *"selected model or estimator"* are both used to indicate that it satisfies (1.2.5).

We emphasize that choosing the penalty is tricky but obviously necessary. The construction of such functions in the context of maximum likelihood estimator goes back to the work of Akaike and Schwarz, see respectively Akaike (1974) and Schwarz et al. (1978). They proposed the now classic AIC and BIC criteria, the two most widely known and pervasively used tools in statistical model selection, where the penalty is respectively establish as follows:

$$\begin{aligned}\text{pen}_{\text{AIC}}(\mathbf{m}) &= D_{\mathbf{m}},\\ \text{pen}_{\text{BIC}}(\mathbf{m}) &= \frac{\ln(n) D_{\mathbf{m}}}{2},\end{aligned}$$

where $D_{\mathbf{m}}$ is the dimension of the model $\mathbf{m}$, and n is the size of the sample considered. These well-known penalized criteria have been widely studied (see, *e.g.,* Anderson & Burnham, 2002) and based on asymptotic approximations. Therefore, such criteria may be wrong in a non-asymptotic context. More precisely, AIC and BIC are based on Wilks' theorem and a Bayesian approach, see, *e.g.,* Cavanaugh & Neath (2019) and Neath & Cavanaugh (2012), respectively, for recent reviews on the conceptual and theoretical foundations. At the same time, Mallows (1973), and later Craven & Wahba (1978) proposed other famous penalized criteria: Mallows's C_p and generalized cross-validation (GCV), respectively, in the context of linear regression. Mathematically, Mallows obtained

$$\text{pen}_{\text{Mallows}}(\mathbf{m}) = \frac{2 D_{\mathbf{m}} \sigma^2}{n},$$

where σ^2 is noise level of the true regression model which is unknown (if it does exist) and σ^2 is thus difficult to estimate. Similarly, the solution proposed by the GCV method is based on cross-validation to choose the unknown tunning parameter (which best value is actually σ^2). Thus, again, we have to estimate an unknown parameter.

1.2.4 Slope heuristics

Motivated by some recent works on concentration inequalities, Birgé & Massart (2001) introduced the slope heuristic, which is a non-asymptotic methodology to select a model from a collection of models. This slope heuristics allows us to choose an optimal penalties from data which are known up to a multiplicative constant κ. Let us describe the ideas of this heuristic. In this framework, the penalty shape is then adapted as $\text{pen}_{\text{shape}}(\cdot)$ and an unknown constant κ_{opt} exists such that

$$\text{pen}_{\text{opt}} : \mathbf{m} \in \mathcal{M} \mapsto \kappa_{\text{opt}} \text{pen}_{\text{shape}}(\mathbf{m})$$

is an optimal penalty. In order to select the oracle model via using (1.2.1) and (1.2.4), we are looking for a penalty close to the following penalty function:

$$\mathcal{M} \ni \mathbf{m} \mapsto \text{pen}(\mathbf{m}) = l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}).$$

However, since s_0 is unknown in practice, we will try to approach this quantity by decomposing it into:

$$\begin{aligned} l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right] - \gamma_n(\widehat{s}_{\mathbf{m}}) \\ &= \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right]}_{\nu_{\mathbf{m}}} + \underbrace{\left[\gamma_n(s_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}})\right]}_{\widehat{\nu}_{\mathbf{m}}} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right]}_{(1)} - \underbrace{\left[\gamma_n(s_{\mathbf{m}}) - \gamma_n(s_0)\right]}_{(2)} - \gamma_n(s_0). \end{aligned} \tag{1.2.7}$$

Here, $\nu_{\mathbf{m}}$ is an "estimation error" term, $\widehat{\nu}_{\mathbf{m}}$ is an empirical "estimation error" term. We will denote by $\Delta_n(s_{\mathbf{m}}) = (1) - (2)$, which corresponds to the difference between the "bias" term and its empirical

version. Note that $\gamma_n(s_0)$ does not depend on $\mathbf{m}$, the main idea is to estimate the following *ideal penalty*, defined as $\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}})$, from the data in order to build an optimal penalty function. Then, (1.2.7) implies that

$$l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}) = \text{pen}^*(\mathbf{m}) - \gamma_n(s_0), \quad \forall \mathbf{m} \in \mathcal{M}. \tag{1.2.8}$$

Next, we wish to prove the following oracle inequality

$$l(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) + [\text{pen}(\widehat{\mathbf{m}}) - \text{pen}^*(\widehat{\mathbf{m}})] \leq \inf_{\mathbf{m} \in \mathcal{M}} \{l(s_0, \widehat{s}_{\mathbf{m}}) + [\text{pen}(\mathbf{m}) - \text{pen}^*(\mathbf{m})]\}. \tag{1.2.9}$$

Indeed, by definition of the ideal penalty by (1.2.8), and the fact that $\widehat{m} \in \mathcal{M}$, it holds that for all $\mathbf{m} \in \mathcal{M}$,

$$\begin{aligned} l(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) + [\text{pen}(\widehat{\mathbf{m}}) - \text{pen}^*(\widehat{\mathbf{m}})] &= \gamma_n(\widehat{s}_{\widehat{\mathbf{m}}}) + \text{pen}(\widehat{\mathbf{m}}) - \gamma_n(s_0) \\ &\leq \gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m}) - \gamma_n(s_0) \text{ (using (1.2.4))} \\ &= l(s_0, \widehat{s}_{\mathbf{m}}) + [\text{pen}(\mathbf{m}) - \text{pen}^*(\mathbf{m})] \text{ (using (1.2.8))}. \end{aligned}$$

The important point to note here is the form of (1.2.9) motivates us to look for a penalty close to the ideal penalty to obtain an oracle inequality. According to the expression of the ideal penalty $\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}})$, both $\nu_{\mathbf{m}}$ and $\Delta_n(s_{\mathbf{m}})$ depend on the unknown conditional PDF s_0. Therefore, it is natural to try to relate the penalty function to the empirical estimation error term $\widehat{\nu}_{\mathbf{m}}$.

To accomplish this task, from a theoretical point of view, Birgé & Massart, in Birgé & Massart (2001, 2007) proposed and proved for the first time the slope heuristics method in the context of Gaussian homoscedastic least squares regression with fixed design. They prove that there exists a *minimal penalty*, $\text{pen}_{\min}(\mathbf{m}) = \widehat{\nu}_{\mathbf{m}}$, namely such that the dimension and the risk of models selected with lighter penalties become very large, whereas higher penalties should select models with "reasonable" complexity. Furthermore, they show that considering a penalty equal to *twice this minimal penalty* allows to select a model close to the oracle model in terms of risk.

More precisely, given the chosen penalty as $\text{pen}(\mathbf{m}) = \kappa\widehat{\nu}_{\mathbf{m}}$, the penalized criterion can be written as

$$\text{crit}(\mathbf{m}) = (1-\kappa)\gamma_n(\widehat{s}_{\mathbf{m}}) + \kappa\gamma_n(s_{\mathbf{m}}).$$

Therefore, three cases occur:

- if $\kappa = 1$ then $\text{crit}(\mathbf{m}) = \gamma_n(s_{\mathbf{m}})$, which concentrates around its expectation $\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})] = \underbrace{l(s_0, s_{\mathbf{m}})}_{\text{bias}} + \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\gamma(s_0, \mathbf{X}, \mathbf{Y})]$ for large n: this procedure selects a model minimizing the bias and does not take into account the variance, which leads to such criterion has "a significant probability"[2] of selecting a too complex model;
- if $\kappa < 1$ then when the complexity goes up, the criterion always goes down because of the two terms in $\text{crit}(\mathbf{m})$ being dropping: the selected models is always one of the most complex ones;
- if $\kappa > 1$ then the criterion rises with the complexity of the most complex models due to the fact that of ruling out the corresponding bias terms (these models almost have the same bias): the dimension of the selected models will be more reasonable.

The first point of the slope heuristics is $\widehat{\nu}_{\mathbf{m}} \approx \nu_{\mathbf{m}}$ since $\widehat{\nu}_{\mathbf{m}}$ is the empirical counterpart of $\nu_{\mathbf{m}}$. In particular, it is expected that we can control the fluctuation of $\Delta_n(s_{\mathbf{m}})$ around its zero expectations through concentration results. Therefore, we can approximate the ideal penalty as twice the minimal penalty due to the fact that

$$\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}}) \approx 2\widehat{\nu}_{\mathbf{m}}.$$

[2]Note that when $\kappa = 1$, the probability to select a too complex model is in general positive but strictly below 1. Actually, this is general with slope heuristics / minimal penalty algorithms: with a data-driven criterion (hence a bit random), when the penalty is exactly at the minimal level, the selected model complexity is highly random (we are just at the critical state of the phase transition) Arlot (2019).

Thus, in practice, the main remaining issue is to determine the minimal penalty $\widehat{\nu}_{\mathbf{m}}$. To this end, on the data set $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right)$; either we look for the greatest jump of complexity of the selected model as a function of the multiplicative constant κ in the penalty or we look at the asymptotic slope of a linear regression between the penalty shape $\text{pen}_{\text{shape}}(\cdot)$ and the contrast value $\gamma_n(s_{\mathbf{m}})$ for the most complex models, which will be referred to as *dimension jump* or *data-driven slope estimation*, respectively. In this way, we obtain the minimal penalty, and we multiply it by two to obtain the optimal penalty. For a deeper discussion of the principle of slope heuristics, we refer the reader to Baudry et al. (2012), Arlot (2019) and the references given therein. Figures 1.7a and 1.7b illustrate these ideas.

We emphasize that from a practical point of view, we use the so-called CAPUSHE (CAlibrating Penalty Using Slope HEuristics) package in R (Arlot et al., 2016, Baudry et al., 2012) to implement the dimension jump and the data-driven slope estimation approaches. In practice, using slope heuristic is effective when an optimal penalty $\text{pen}_{\text{opt}}(\cdot) = \kappa_{\text{opt}}\text{pen}_{\text{shape}}(\cdot)$ is known up to a multiplicative factor. It is worth pointing out that the keystone of the slope heuristics is that $\frac{\kappa_{\text{opt}}}{2}\text{pen}_{\text{shape}}(\mathbf{m})$ is a good estimate of $\widehat{s}_{\mathbf{m}}$ and provides a minimal penalty. Generally speaking, the $\text{pen}_{\text{shape}}(\cdot)$ can be chosen as the complexity measure, when its definition is not obvious a priori. This complexity measure is typically the model dimension $D_{\mathbf{m}}$ or the number of free parameters needed to be estimated.

From a theoretical point of view, in this thesis, we contribute several non-asymptotic oracle inequalities, Theorems 1.2.2 and 1.2.3, which provide non-asymptotic bounds on the risks, and lower bounds on the penalty functions that ensure non-asymptotic theoretical controls on the estimators under the Jensen–Kullback–Leibler loss. These oracle inequalities also provide some theoretical justifications of the penalty shapes when using the slope heuristic for the corresponding MoE regression models. More precisely, penalized likelihood criteria are proposed in Chapters 3 and 4 to select best data-driven MoE regression models among a specific model collection. These criteria depend on unknown constants which can be calibrated in practical situations by slope heuristic. In particular, in order to work with conditional PDF in several MoE regression models, we wish to make use of a model selection theorem for MLE among a random subcollection (cf. Devijver, 2015b, Theorem 5.1 and Devijver & Gallopin, 2018, Theorem 7.3), which is an extension of a whole collection of conditional densities from Cohen & Le Pennec (2011, Theorem 2), and of Massart (2007, Theorem 7.11), working only for density estimation.

1.2.5 Asymptotic analysis of a parametric model

By using (1.2.2), (1.2.3) and (1.2.9), we prove that it is natural to try to relate the non-asymptotic model selection procedure, in particular the slope heuristic, to the oracle inequalities, which are described in Sections 1.2.6 to 1.2.8. Section 1.2.5 was intended as an attempt to explain in more details why we should pay more attention to the non-asymptotic upper bound via providing the drawbacks of asymptotic analysis of a parametric model.

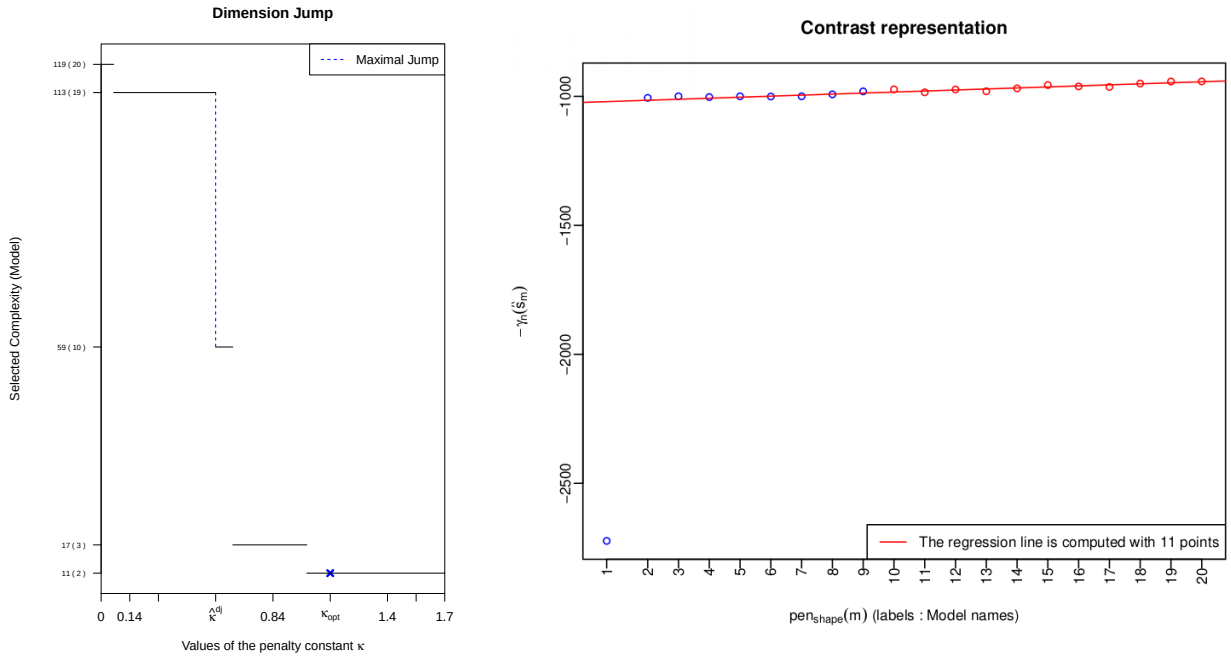
We should now specify our *goodness* criteria. In the maximum likelihood approach, the Kullback–Leibler divergence is the most natural loss function, which is defined for two densities s and t by

$$\text{KL}(s,t) = \begin{cases} \int_{\mathbb{R}^D} \ln\left(\frac{s(\mathbf{y})}{t(\mathbf{y})}\right) s(\mathbf{y})d\mathbf{y} & \text{if } sd\mathbf{y} \text{ is absolutely continuous w.r.t. } td\mathbf{y}, \\ +\infty & \text{otherwise.} \end{cases}$$

However, to take into account the structure of inverse conditional densities and the random covariates $\mathbf{Y}_{[n]}$, we consider the *tensorized Kullback–Leibler divergence* $\text{KL}^{\otimes \text{n}}$, defined as:

$$\text{KL}^{\otimes \text{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\text{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right], \tag{1.2.10}$$

if sdy is absolutely continuous w.r.t. tdy, and $+\infty$ otherwise. Note that if the predictors are fixed, this divergence is the classical fixed design type divergence in which there is no expectation. We refer to our result as a *weak oracle inequality*, because its statement is based on a smaller divergence, when



(a) Dimension jump

(b) Data-driven slope estimation

Figure 1.7: Illustration of the slope heuristic with 2000 data points of the examples from the WS scenario. In Figure 1.7a we estimate κ via using $\widehat{\kappa}^{\mathrm{dj}}$ the largest jump of complexity. We then select a model that minimizes the penalized log-likelihood by $\kappa_{\mathrm{opt}} = 2\widehat{\kappa}^{\mathrm{dj}}$. In Figure 1.7b, we estimate κ by looking for the asymptotic slope of a linear regression between the penalty shape $\mathrm{pen}_{\mathrm{shape}}(\cdot)$ and the contrast value $\gamma_n(s_{\mathbf{m}})$ for the most complex models.

compared to $\mathrm{KL}^{\otimes \mathrm{n}}$, namely the *tensorized Jensen-Kullback–Leibler divergence*:

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\,\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right),(1-\rho)\,s\left(\cdot|\mathbf{Y}_i\right)+\rho t\left(\cdot|\mathbf{Y}_i\right)\right)\right], \tag{1.2.11}$$

with $\rho \in (0,1)$. We note that $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ was first used in Cohen & Le Pennec (2011). However, a version of this divergence appears explicitly with $\rho = \frac{1}{2}$ in Massart (2007), and it is also found implicitly in Birgé et al. (1998). This loss is always bounded by $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ but behaves like $\mathrm{KL}^{\otimes \mathrm{n}}$, when t is close to s. The main tools in the proof of such a weak oracle inequality are deviation inequalities for sums of random variables and their suprema. These tools require a boundedness assumption on the controlled functions which is not satisfied by $-\ln\frac{s_{\mathbf{m}}}{s_0}$, and thus also not satisfied by $\mathrm{KL}^{\otimes \mathrm{n}}$. Therefore, we consider instead the use of $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$. In particular, in general, it holds that $C_\rho\,\mathrm{d}^{2\otimes \mathrm{n}} \leq \mathrm{JKL}_{\rho}^{\otimes \mathrm{n}} \leq \mathrm{KL}^{\otimes \mathrm{n}}$, where $C_\rho = \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho},1\right)\left(\ln\left(1+\frac{\rho}{1-\rho}\right)-\rho\right)$ (see Cohen & Le Pennec 2011, Prop. 1) and $\mathrm{d}^{2\otimes \mathrm{n}}$ is a tensorized extension of the squared Hellinger distance $\mathrm{d}^{2\otimes \mathrm{n}}$, defined by

$$\mathrm{d}^{2\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}^2\left(s\left(\cdot|\mathbf{Y}_i\right),t\left(\cdot|\mathbf{Y}_i\right)\right)\right].$$

Moreover, if we assume that, for any $\mathbf{m} \in \mathcal{M}$ and any $s_{\mathbf{m}} \in S_{\mathbf{m}}$, $s_0 d\lambda \ll s_{\mathbf{m}} d\lambda$, then (see Montuelle et al., 2014, Cohen & Le Pennec, 2011)

$$\frac{C_\rho}{2+\ln\left\|s_0/s_{\mathbf{m}}\right\|_\infty}\,\mathrm{KL}^{\otimes \mathrm{n}}(s_0,s_{\mathbf{m}}) \leq \mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0,s_{\mathbf{m}}). \tag{1.2.12}$$

We will consider a parametric model of inverse conditional PDFs to which the true inverse conditional PDF s_0 does not necessary belongs as follows:

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{\mathbf{m}}}(\mathbf{x}|\mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}} \subset \mathbb{R}^{\dim(S_{\mathbf{m}})}\right\}.$$

This construction of *misspecified model*, *i.e.*, $s_0 \notin S_{\mathbf{m}}$, goes back to the work of White (1982) for density estimation. For a treatment of a more general case for conditional PDFs, we refer the reader to Cohen & Le Pennec (2011). Section 1.2.5 contains a brief summary of such classical results without proofs via Theorem 1.2.1.

Theorem 1.2.1 (White, 1982, Cohen & Le Pennec, 2011)**.** *Suppose that the model $S_{\mathbf{m}}$ is identifiable (for MoE regression models, see,* e.g., *Jiang & Tanner, 1999c, Hennig, 2000) and there are the existences of the* $\dim(S_{\mathbf{m}}) \times \dim(S_{\mathbf{m}})$ *matrices* $\mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}})$ *and* $\mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}})$ *defined by:*

$$[\mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}})]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{-\partial^2 \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},k}\partial\psi_{\mathbf{m},l}}(\mathbf{x}|\mathbf{Y}_i)\, s_0(\mathbf{x}|\mathbf{Y}_i)\, d\lambda\right],$$
$$[\mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}})]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},k}}(\mathbf{x}|\mathbf{Y}_i)\frac{\partial \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},l}}(\mathbf{x}|\mathbf{Y}_i)\, s_0(\mathbf{x}|\mathbf{Y}_i)\, d\lambda\right].$$

We define $\boldsymbol{\psi}_{\mathbf{m}}^*$ *to be the elements of* $\arg\min_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}})$*. Then, under some strong regularity assumptions on* $\boldsymbol{\psi}_{\mathbf{m}} \mapsto s_{\boldsymbol{\psi}_{\mathbf{m}}}$, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}[\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})]$ *is asymptotically equivalent to*

$$\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}^*}\right) + \frac{1}{2n} tr\left(\mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}}^*)\,\mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}}^*)^{-1}\right).$$

In particular, when $s_0 \in S_{\mathbf{m}}$*, it holds that* $s_0 = s_{\boldsymbol{\psi}_{\mathbf{m}}^*}, \mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}}^*) = \mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}}^*)$*. Therefore, the previous asymptotic equivalent of* $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}[\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})]$ *becomes the classical parametric one, namely,*

$$\underbrace{\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}^*}\right)}_{=0} + \frac{1}{2n}\dim(S_{\mathbf{m}}).$$

Theorem 1.2.1 depends heavily on the asymptotic normality of $\sqrt{n}\left(\widehat{\boldsymbol{\psi}}_{\mathbf{m}} - \boldsymbol{\psi}_{\mathbf{m}}^*\right)$. One may ask whether this is still true if this normality does not hold. Several works are devoted to the study of the non-asymptotic normality: extension in non parametric case or non-identifiable model, often called Wilk's phenomenon (see Wilks, 1938 for more details); generalization of the corresponding Chi-Square goodness-of-fit test (Fan et al., 2001); finite sample deviation of the corresponding empirical quantity in a bounded loss setting (Boucheron & Massart, 2011). Motivated by the work of Cohen & Le Pennec (2011, 2013), with as few assumptions on the collection of conditional PDFs $S_{\mathbf{m}}$ as possible, we are initially interested in finding a non-asymptotic upper bound of type

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})\right] \leq \left(\inf_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}}) + \frac{1}{2n}\dim(S_{\mathbf{m}})\right) + C_2\frac{1}{n}.$$

However, in reality, we obtained the following weaker upper bound (oracle inequalities)

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})\right] \leq C_1\left(\inf_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}}) + \frac{\kappa}{n}\mathfrak{D}_{\mathbf{m}}\right) + C_2\frac{1}{n}.$$

Indeed, by technical problems, we have to replace the left hand $\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})$ by a smaller divergence $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}})$ and the constant $C_1 = 1+\epsilon, \epsilon > 0$, can not be equal 1. Furthermore, κ is a constant that depends on ϵ and the model complexity term $\mathfrak{D}_{\mathbf{m}}$ replaces the dimension term $\dim(S_{\mathbf{m}})$. However, this result allows us to have the right bias/variance trade-off flavor and recover usual minimax properties of specific estimators.

Here and subsequently, in order to establish our oracle inequalities, we need to assume that the input space is a bounded set and make explicit some classical boundedness conditions on the parameter space.

1.2.6 Weak oracle inequality for GLoME models

In GLoME regression models, we choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, where $d_{\max} \in \mathbb{N}^\star$ and $K_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. We wish to estimate the unknown inverse conditional density s_0 by conditional densities belonging to the following collection of inverse models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \{(K,d) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\right\}. \tag{1.2.13}$$

Here, $\widetilde{\boldsymbol{\Omega}}_K$ are bounded Gaussian gating parameter vectors, $\boldsymbol{\Upsilon}_{K,d}$ is defined as a linear combination of a finite set of bounded functions whose coefficients belong to a compact set, and $\mathbf{V}_K$ are bounded positive definite covariance matrices, see (1.2.14), (1.2.16) (or more general (1.2.15)), and (1.2.17), respectively, for more details.

More precisely, assume that there exist deterministic positive constants $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}}, A_{\boldsymbol{\Gamma}}$, $\widetilde{\boldsymbol{\Omega}}_K$ is defined by

$$\widetilde{\boldsymbol{\Omega}}_K = \{\boldsymbol{\omega} \in \boldsymbol{\Omega}_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}_k) \leq M(\boldsymbol{\Gamma}_k) \leq A_{\boldsymbol{\Gamma}}, a_{\boldsymbol{\pi}} \leq \pi_k\}, \tag{1.2.14}$$

where $m(\mathbf{A})$ and $M(\mathbf{A})$ stand for, respectively, the modulus of the smallest and largest eigenvalues of any matrix $\mathbf{A}$. Following the same structure for the means of Gaussian experts from Montuelle et al. (2014), the set $\boldsymbol{\Upsilon}_{K,d}$ will be chosen as a tensor product of compact sets of moderate dimension (*e.g.,* a set of polynomials of degree smaller than d, whose coefficients are smaller in absolute values than $T_{\boldsymbol{\Upsilon}}$). More specifically, $\boldsymbol{\Upsilon}_{K,d} = \otimes_{k\in[K]}\boldsymbol{\Upsilon}_{k,d} =: \boldsymbol{\Upsilon}_{k,d}^K$, where $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_{b,d}$, $\forall k \in [K]$, and

$$\boldsymbol{\Upsilon}_{b,d} = \left\{\mathbf{y} \mapsto \left(\sum_{i=1}^{d} \boldsymbol{\alpha}_i^{(j)} \varphi_{\boldsymbol{\Upsilon},i}(\mathbf{y})\right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \tag{1.2.15}$$

Here, $d \in \mathbb{N}^\star, T_{\boldsymbol{\Upsilon}} \in \mathbb{R}^+$, and $(\varphi_{\boldsymbol{\Upsilon},i})_{i\in[d]}$ is a collection of bounded functions on $\mathcal{Y}$. In particular, we focus on the bounded $\mathcal{Y}$ case and assume that $\mathcal{Y} = [0,1]^L$, without loss of generality. In this case, $\varphi_{\boldsymbol{\Upsilon},i}$ can be chosen as monomials with maximum (non-negative) degree d: $\mathbf{y}^{\mathbf{r}} = \prod_{l=1}^{L} \mathbf{y}_l^{\mathbf{r}_l}$. Recall that a multi-index $\mathbf{r} = (\mathbf{r}_l)_{l\in[L]}, \mathbf{r}_l \in \mathbb{N}^\star \cup \{0\}, \forall l \in [L]$, is an L-tuple of nonnegative integers. We define $|\mathbf{r}| = \sum_{l=1}^{L} \mathbf{r}_l$ and the number $|\mathbf{r}|$ is called the order or degree of $\mathbf{y}^{\mathbf{r}}$. Then, $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$, where

$$\boldsymbol{\Upsilon}_{p,d} = \left\{\mathbf{y} \mapsto \left(\sum_{|\mathbf{r}|=0}^{d} \boldsymbol{\alpha}_{\mathbf{r}}^{(j)} \mathbf{y}^{\mathbf{r}}\right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \tag{1.2.16}$$

Note that any covariance matrix $\boldsymbol{\Sigma}_k$ can be decomposed into the form $B_k\mathbf{P}_k\mathbf{A}_k\mathbf{P}_k^\top$, such that $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$ is a positive scalar corresponding to the volume, $\mathbf{P}_k$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_k$ and $\mathbf{A}_k$ the diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$; $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$, $\mathcal{A}(\lambda_-, \lambda_+)$ is a set of diagonal matrices $\mathbf{A}_k$, such that $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$; and $SO(D)$ is the special orthogonal group of dimension D. In this way we obtain what is known as the classical covariance matrix sets described by Celeux & Govaert (1995) for Gaussian parsimonious clustering models, defined by

$$\mathbf{V}_K = \left\{\left(B_k\mathbf{P}_k\mathbf{A}_k\mathbf{P}_k^\top\right)_{k\in[K]} : \forall k \in [K], B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}(\lambda_-, \lambda_+)\right\}. \tag{1.2.17}$$

The following Theorem 1.2.2 provides a lower bound on the penalty function, pen($\mathbf{m}$), which guarantees that the PMLE for GLoME models selects a model that performs almost as well as the best model. Note that, in Section 1.2.10.2, we briefly prove Theorem 1.2.2, which is then restated as Theorem 3.2.3. The readers are referred to Section 3.2.4 for a comprehensive proof, see also Nguyen et al. (2021c).

Theorem 1.2.2 (Oracle inequality for GLoME models)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, arising from an unknown conditional density* s_0*. Given a collection of GLoME models,* $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$*, defined by* (1.2.13)*, there is a constant* C *such that for any* $\rho \in (0,1)$*, for any* $\mathbf{m} \in \mathcal{M}$*,* $z_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$ *and any* $C_1 > 1$*, there is a constant* κ_0 *depending only on* ρ *and* C_1*, such that if for every index* $\mathbf{m} \in \mathcal{M}$*,*

$$\text{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n) \dim\left(S_{\mathbf{m}}\right) + z_{\mathbf{m}}\right] \text{ with } \kappa > \kappa_0,$$

then the η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined in* (1.2.5)*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\text{JKL}_{\rho}^{\otimes \text{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}} \left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \text{KL}^{\otimes \text{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{\text{pen}(\mathbf{m})}{n} \right) + \frac{\kappa_0 C_1 \Xi}{n} + \frac{\eta + \eta'}{n}. \tag{1.2.18}$$

1.2.7 Weak oracle inequality for BLoME models

In the framework of BLoME models, we choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, where $d_{\max} \in \mathbb{N}^\star$ and $K_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. Moreover, $\mathbf{B}$ is selected among a list of candidate structures $(\mathcal{B}_k)_{k\in[K]} \equiv (\mathcal{B})_{k\in[K]}$, where $\mathcal{B}$ denotes the set of all possible partitions of the covariables indexed by $[D]$, for each cluster of individuals. We wish to estimate the unknown conditional density s_0 by conditional densities belonging to the following collection of models: $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \left\{(K, d, \mathbf{B}) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}, \mathbf{B} \in (\mathcal{B})_{k\in[K]}\right\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} = \left(\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}\left(\mathbf{B}\right)\right) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\left(\mathbf{B}\right)\right\}, \tag{1.2.19}$$

where $\widetilde{\boldsymbol{\Omega}}_K$, $\boldsymbol{\Upsilon}_{K,d}$, and $\mathbf{V}_K\left(\mathbf{B}\right)$ are defined in (1.2.14), (1.2.16) (or more general (1.2.15)), and (1.1.14), respectively.

For the block-diagonal covariances of Gaussian experts, we assume that there exist some positive constants λ_m and λ_M such that, for every $k \in [K]$,

$$0 < \lambda_m \leq m\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq M\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq \lambda_M. \tag{1.2.20}$$

Note that this is a quite general assumption and is also used in the block-diagonal covariance selection for Gaussian graphical models of Devijver & Gallopin (2018).

In theory, we would like to consider the whole collection of models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$. However, the cardinality of $\mathcal{B}$ is large; its size is a Bell number. Even for a moderate number of variables D, it is not possible to explore the set $\mathcal{B}$, exhaustively. We restrict our attention to a random subcollection $\mathcal{B}^R$ of moderate size. For example, we can consider the BLLiM procedure from Devijver et al. (2017, Section 2.2).

Note that the constructed collection of models with block-diagonal structures for each cluster of individuals is designed, for example, by the BLLiM procedure from Devijver et al. (2017), where each collection of partition is sorted by sparsity level. Nevertheless, our finite-sample oracle inequality, Theorem 1.2.3, still holds for any random subcollection of $\mathcal{M}$, which is constructed by some suitable tools in the framework of BLoME regression models. Note that Theorem 1.2.3 is restated as Theorem 3.3.2 and is fully proved in Section 3.3.3, see also Nguyen et al. (2021b). We highlight our main contribution and briefly establish its proof in Section 1.2.11.

Theorem 1.2.3 (Oracle inequality for BLoME models)**.** *Let* $(\mathbf{x}_{[n]}, \mathbf{y}_{[n]})$ *be the observations coming from an unknown conditional density* s_0*. For each* $\mathbf{m} = (K, d, \mathbf{B}) \in (\mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} \times \mathcal{B}) \equiv \mathcal{M}$*, let* $S_{\mathbf{m}}$ *be define by* (1.2.19)*. Assume that there exists* $\tau > 0$ *and* $\epsilon_{KL} > 0$ *such that, for all* $\mathbf{m} \in \mathcal{M}$*, one can find* $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$*, such that*

$$\text{KL}^{\otimes \text{n}}\left(s_0, \bar{s}_{\mathbf{m}}\right) \leq \inf_{t\in S_{\mathbf{m}}} \text{KL}^{\otimes \text{n}}\left(s_0, t\right) + \frac{\epsilon_{KL}}{n}, \text{ and } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0.$$

Next, we construct some random subcollection $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *of* $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *by letting* $\widetilde{\mathcal{M}} \equiv \left(\mathcal{K}\times\mathcal{D}_{\boldsymbol{\Upsilon}}\times\mathcal{B}^R\right) \subset \mathcal{M}$ *such that* $\mathcal{B}^R$ *is a random subcollection* $\mathcal{B}$*, of moderate size. Consider the collection* $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *of* η*-log likelihood minimizers satisfying* (1.2.6) *for all* $\mathbf{m}\in\widetilde{\mathcal{M}}$*. Then, there is a constant* C *such that for any* $\rho\in(0,1)$*, and any* $C_1>1$*, there are two constants* κ_0 *and* C_2 *depending only on* ρ *and* C_1 *such that, for every index,* $\mathbf{m}\in\mathcal{M}$*,* $\xi_{\mathbf{m}}\in\mathbb{R}^+$*,* $\Xi=\sum_{\mathbf{m}\in\mathcal{M}}e^{-\xi_{\mathbf{m}}}<\infty$ *and*

$$\operatorname{pen}(\mathbf{m})\geq\kappa\left[(C+\ln n)\dim(S_{\mathbf{m}})+(1\vee\tau)\xi_{\mathbf{m}}\right],$$

with $\kappa>\kappa_0$*, the* η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined as in* (1.2.5) *on the subset* $\widetilde{\mathcal{M}}\subset\mathcal{M}$*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes\mathrm{n}}\left(s_0,\widehat{s}_{\widehat{\mathbf{m}}}\right)\right]\leq C_1\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}}\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,t\right)+2\frac{\operatorname{pen}(\mathbf{m})}{n}\right)\right]+C_2(1\vee\tau)\frac{\Xi^2}{n}+\frac{\eta'+\eta}{n}.$$

1.2.8 Weak oracle inequality for PSGaBloME models

1.2.8.1 Linear combination of bounded functions for the weights and the means

We follow the idea from Montuelle et al. (2014) to restrict our attention on a finite set of bounded functions whose coefficients belong to a compact set. It is worth mentioning that such quite general setting includes the polynomial basis when the predictors are bounded, the suitable renormalized wavelet dictionaries as well as the Fourier basis on an interval. More precisely, we first define the following two collection of bounded functions for the weights and means: $\mathcal{X}\ni\mathbf{x}\mapsto\left(\theta_{\mathbf{W},d}\left(\mathbf{x}\right)\right)_{d\in[d_{\mathbf{W}}]}\in[-1,1]^{d_{\mathbf{W}}}$ and $\mathcal{X}\ni\mathbf{x}\mapsto\left(\theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right)\right)_{d\in[d_{\boldsymbol{\Upsilon}}]}\in[-1,1]^{d_{\boldsymbol{\Upsilon}}}$, where $d_{\mathbf{W}}\in\mathbb{N}^\star$ and $d_{\boldsymbol{\Upsilon}}\in\mathbb{N}^\star$ indicate its degrees, respectively. Then, by making use of these collections, we are able to define the corresponding desired bounded spaces via tensorial constructions as follows:

$$\begin{aligned}
\mathbf{W}_{K,d_{\mathbf{W}}}&=\{0\}\otimes\mathbf{W}^{K-1},\mathbf{W}=\left\{\mathcal{X}\ni\mathbf{x}\mapsto\sum_{d=1}^{d_{\mathbf{W}}}\omega_d\theta_{\mathbf{W},d}\left(\mathbf{x}\right)\in\mathbb{R}:\max_{d\in[d_{\mathbf{W}}]}|\omega_d|\leq T_{\mathbf{W}}\right\},\\
\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}}&=\boldsymbol{\Upsilon}^K,\boldsymbol{\Upsilon}=\left\{\mathcal{X}\ni\mathbf{x}\mapsto\left(\sum_{d=1}^{d_{\boldsymbol{\Upsilon}}}\beta_d^{(z)}\theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right)\right)_{z\in[q]}:\max_{d\in[d_{\boldsymbol{\Upsilon}}],z\in[q]}\left|\beta_d^{(z)}\right|\leq T_{\boldsymbol{\Upsilon}}\right\}.
\end{aligned}\tag{1.2.21}$$

When p and q are not too large, we do not need to select relevant variables and/or use rank sparse models. We do not need to select relevant variables. Then, we can work on the previous LinBoSGaBloME models with general structures for means, weights and multi-block-diagonal covariance matrices or with LinBoSGaME models as in Montuelle et al. (2014). However, in PSGaBloME models, to handle with high-dimensional data and to simplify the interpretation of sparsity, we propose to utilize polynomials for weights and polynomial regression models for the softmax gating functions and the means of Gaussian experts as follows:

$$\begin{aligned}
\mathbf{W}_{K,d_{\mathbf{W}}}&=\{0\}\otimes\mathbf{W}^{K-1},\mathbf{W}=\left\{\mathcal{X}\ni\mathbf{x}\mapsto\sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}}\omega_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}}\in\mathbb{R}:\max_{\boldsymbol{\alpha}\in\mathcal{A}}|\omega_{\boldsymbol{\alpha}}|\leq T_{\mathbf{W}}\right\},\\
\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}}&=\left\{\mathcal{X}\ni\mathbf{x}\mapsto\left(\boldsymbol{\beta}_{k0}+\sum_{d=1}^{d_{\boldsymbol{\Upsilon}}}\boldsymbol{\beta}_{kd}\mathbf{x}^d\right)_{k\in[K]}:\max\left\{\left\|\boldsymbol{\beta}_{kd}\right\|_\infty:k\in[K],d\in(\{0\}\cup[d_{\boldsymbol{\Upsilon}}])\right\}\leq T_{\boldsymbol{\Upsilon}}\right\}.
\end{aligned}\tag{1.2.22}$$

Here, note that the multi-index $\boldsymbol{\alpha}=(\alpha_t)_{t\in[p]},\alpha_t\in\mathbb{N}^\star\bigcup\{0\}=:\mathbb{N},\forall t\in[p]$, is an p-tuple of nonnegative integers that satisfies $\mathbf{x}^{\boldsymbol{\alpha}}=\prod_{j=1}^p x_j^{\alpha_j}$ and $|\boldsymbol{\alpha}|=\sum_{t=1}^p\alpha_t$. Then, for all $l\in[d_{\mathbf{W}}]$, we define $\mathcal{A}=\bigcup_{l=0}^{d_{\mathbf{W}}}\mathcal{A}_l$, $\mathcal{A}_l=\left\{\boldsymbol{\alpha}=(\alpha_t)_{t\in[p]}\in\mathbb{N}^p,|\boldsymbol{\alpha}|=l\right\}$. The number $|\boldsymbol{\alpha}|$ is called the order or degree of monomials $\mathbf{x}^{\boldsymbol{\alpha}}$. By using the well-known stars and bars methods, *e.g.,* Feller (1957, Chapter 2), the cardinality of the set $\mathcal{A}$, denoted by $\operatorname{card}(\mathcal{A})$, equals $\binom{d_{\mathbf{W}}+p}{p}$. Note that, for all $d\in[d_{\boldsymbol{\Upsilon}}]$, we

define $\mathbf{x}^d$ as $\left(x_j^d\right)_{j\in[p]}$ for the means, which are often used for polynomial regression models. Moreover, given any matrix $\mathbf{A} \in \mathbb{R}^{q\times p}$, the following notations are used for matrix norms: the *max norm* $|||\mathbf{A}|||_\infty = \max_{i\in[q],j\in[p]} \left|[\mathbf{A}]_{i,j}\right|$, the 2-*norm* $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2=1} \left|\mathbf{x}^\top \mathbf{A}\mathbf{x}\right| = \sup_{\lambda\in \mathrm{vp}(\mathbf{A})} |\lambda|$, where $\mathrm{vp}(\mathbf{A})$ denotes the spectrum of $\mathbf{A}$, and the *Frobenius norm* $\|\mathbf{A}\|_F^2 = \sum_{i=1}^q \sum_{j=1}^p \left|[\mathbf{A}]_{i,j}\right|^2$. Then, it holds that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$, $\|\mathbf{A}\|_2 \leq \sqrt{qp}|||\mathbf{A}|||_\infty$, and for any $\mathbf{x} \in \mathbb{R}^p$, $\|\mathbf{x}\|_2 \leq \sqrt{p}\,\|\mathbf{x}\|_\infty$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2\,\|\mathbf{x}\|_2$, *e.g.,* Golub & Van Loan (2013, Chapter 2).

1.2.8.2 Variable selection via selecting relevant variables

The Lasso estimator, originally established by Tibshirani (1996), is a classical choice for variable selection and has been extended to deal with multiple multivariate regression models for column sparsity using the Group-Lasso estimator (Yuan & Lin, 2006). Note that the Group-Lasso penalty can be used to select a subset of variables for one choice of regularization parameter in the Lasso-Rank procedure, as done, *e.g.,* Devijver (2015b, 2017a,b) or to get a ranking of the variables, as done, e.g., in Bach (2008).

Recall that, for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $\boldsymbol{\beta}_{kd}$ is the matrix of d-th term of regression coefficients, $\boldsymbol{\Sigma}_k$ ($\mathbf{B}_k$) is the covariance matrix in the mixture component k, and the g_k is the mixture proportion k with the $\boldsymbol{\alpha}$-th order term of its monomials is $\omega_{k\boldsymbol{\alpha}}$. Furthermore, given a regressor $\mathbf{x}$, for all $k \in [K]$, for all $d \in [d_{\boldsymbol{\Upsilon}}]$ and for all $z \in [q]$, $\left[\boldsymbol{\beta}_{kd}\mathbf{x}^d\right]_z = \sum_{j=1}^p [\boldsymbol{\beta}_{kd}]_{z,j}\, x_j^d$ is the z-th component of the d-th terms of means for the mixture components k. In particular, for all $l \in [d_{\mathbf{W}}]$, $j \in [p]$, we define $\boldsymbol{\omega}_k^{[j,l]} = \left\{\omega_{k\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} = (\alpha_t)_{t\in[p]} \in \mathcal{A}_l, \boldsymbol{\alpha}_j > 0\right\}$.

We have to deal with high-dimensional data where we estimate many coefficients while given a small number of target variables. Therefore, we need to focus on selecting relevant variables via the notion of irrelevant indices in Definition 1.2.4.

Definition 1.2.4 (Relevant variables in PSGaBloME models)**.** A couple $(\mathbf{Y}_z, \mathbf{X}_j)$ and its corresponding indices $(z, j) \in [q] \times [p]$ are said to be *irrelevant* if, for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $l \in [d_{\mathbf{W}}]$, $[\boldsymbol{\beta}_{kd}]_{z,j} = 0, \boldsymbol{\omega}_k^{[j,l]} = \mathbf{0}$. This means that the variable $\mathbf{X}_j$ does not explain the variable $\mathbf{Y}_z$ for the regression models. A couple and its corresponding indices are relevant if they are not irrelevant. A model is said to be sparse if there are few of relevant variables. We denote by $\mathbf{J}$ the set of indices (z, j) of relevant couples $(\mathbf{Y}_z, \mathbf{X}_j)$. Then, we define the set of relevant variables (columns) as $\mathbf{J}_{\boldsymbol{\omega}} = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$. We denote by $\mathbf{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ and $\mathbf{b}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ the matrix and vector with vectors $\mathbf{0}$ on the columns indexed by the set $\mathbf{J}_{\boldsymbol{\omega}}^C$ and values 0 on the set $\mathbf{J}_{\boldsymbol{\omega}}^C$, respectively. Here, $\mathbf{J}_{\boldsymbol{\omega}}^C$ is the complement of the set $\mathbf{J}_{\boldsymbol{\omega}}$.

Remark that $\mathbf{J} \subset \mathcal{P}([q] \times [p])$ and $\mathbf{J}_{\boldsymbol{\omega}} \subset \mathcal{P}([q])$, where $\mathcal{P}([q] \times [p])$ contains all subsets of $[q] \times [p]$. In our context, we focus on the Group-Lasso estimator to detect relevant variables, where the groups correspond to the columns. Therefore, if for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, a matrix $\boldsymbol{\beta}_{kd}$ has $\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})$ relevant columns, there are $q\,\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})$ coefficients to be estimated instead of qp per clusters and coefficient matrices. This leads to the number of parameters to be estimated is then drastically reduced when $\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}}) \ll p$. Furthermore, such column sparsity may enhance the interpretation since the responses are described by only few relevant columns. To construct the regularization for coefficients of polynomial functions, we can consider the sparse Group-Lasso estimator from Simon et al. (2013) and Hastie et al. (2015, Chapter 4).

1.2.8.3 Rank sparse models

This approach is based on rank sparse models, introduced by Anderson et al. (1998). More precisely, if regression matrices have low-rank or at least can be well approximated by low-rank matrices, then its corresponding regression models are called rank sparse. In the PSGaBloME model, for every $k \in [K]$, $d \in [L]$, the matrix $\boldsymbol{\beta}_{kd}$ is fully determined by $R_{kd}(p + q - R_{kd})$ coefficients if it has rank R_{kd}. This advantage will be very useful because the total parameters to estimate may be smaller than the sample size nq. It is worth noting that such low-rank estimation generalizes the classical principal

component analysis for reducing the dimension of multivariate data and appears in many applications: *e.g.,* Friston et al. (2003, 2019, analysis of fMRI image data), Anderson et al. (1998, analysis of EEG data decoding).

By combining the previous rank and column sparsity, we consider the matrices of regression coefficients $\boldsymbol{\beta}_{kd}$ of rank R_{kd} and a vector of ranks $\mathbf{R} = (R_{kd})_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}]}$ belongs to $[\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{d_{\boldsymbol{\Upsilon}}K}$, where in general, $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$.

We describe in more detail the collection of PSGaBloME models with relevant variables and rank sparse models in the sequel.

1.2.8.4 Collection of models

To simplify the notations, L and D stand for $\binom{d_{\mathbf{W}}+\text{card}(\mathbf{J}_{\boldsymbol{\omega}})}{\text{card}(\mathbf{J}_{\boldsymbol{\omega}})}$ and $d_{\boldsymbol{\Upsilon}}$, which are related to the dimensions of $\mathbf{W}_{K,d_{\mathbf{W}}}$ and $\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}}$, respectively. Combining all the previous structures defined in Sections 1.2.8.1 to 1.2.8.3 , given $\mathbf{m} = (K, L, D, \mathbf{B}, \mathbf{J}, \mathbf{R}) \in \mathbb{N}^\star \times \mathbb{N}^\star \times \mathbb{N}^\star \times (\mathcal{B}_k)_{k\in[K]} \times \mathcal{P}([q]\times[p]) \times [\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{DK}$, some real positive constants $A_{\mathbf{u},\mathbf{v}} > 0$, $A_\sigma > 0$, we obtain the following model:

$$
\begin{aligned}
S_{\mathbf{m}} &= \left\{ (\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}}(\mathbf{y}|\mathbf{x}) =: s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\psi}_{(K,L_{\mathbf{W}},D_{\boldsymbol{\Upsilon}},\mathbf{B},\mathbf{J},\mathbf{R})} \in \boldsymbol{\Psi}_{(K,L_{\mathbf{W}},D_{\boldsymbol{\Upsilon}},\mathbf{B},\mathbf{J},\mathbf{R})} \right\}, \\
\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \left((\omega_{k\boldsymbol{\alpha}})_{k\in[K],\boldsymbol{\alpha}\in\mathcal{A}}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}}\right)_{d\in[D]} \right)_{k\in[K]}, (\boldsymbol{\Sigma}_k(\mathbf{B}_k))_{k\in[K]} \right) \\
&\in \left(\mathbb{R}^L\right)^{K-1} \times \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}, \\
\boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \Bigg\{ \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}}\right)_{d\in[D]}^{[\mathbf{J}_{\boldsymbol{\omega}}]} \right)_{k\in[K]} \in \left(\mathbb{R}^{q\times 1} \times \left(\mathbb{R}^{q\times p}\right)^D \right)^K : \forall k \in [K], \forall d \in [D], \\
&\qquad \boldsymbol{\beta}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} [\sigma_{kd}]_r [\mathbf{u}_{kd}]_{\bullet,r} \left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}, \text{rank}\left(\boldsymbol{\beta}_{kd}^{R_{kd}}\right) = R_{kd}, \forall r \in [R_{kd}], [\sigma_{kd}]_r < A_\sigma, \\
&\qquad \max_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}],r\in[R_{kd}]} \left\{ \|\boldsymbol{\beta}_{k0}\|_\infty, \left\|[\mathbf{u}_{kd}]_{\bullet,r}\right\|_\infty, \left\|\left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}\right\|_\infty \right\} \leq A_{\mathbf{u},\mathbf{v}} \Bigg\}. \qquad (1.2.23)
\end{aligned}
$$

In the above, for $k \in [K], d \in [D]$, $[\sigma_{kd}]_r, r \in [R_{kd}]$, denote the singular values of $\boldsymbol{\beta}_{kd}^{R_{kd}}$, with corresponding orthogonal unit vectors $\left([\mathbf{u}_{kd}]_{\bullet,r}\right)_{r\in[R_{kd}]}$ and $\left(\left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}\right)_{r\in[R_{kd}]}$ (Strang, 2019, I.8). The dimension of $S_{\mathbf{m}}$ is

$$
\dim(S_{\mathbf{m}}) = (K-1)L + qK + \sum_{k=1}^{K}\sum_{d=1}^{D} R_{kd}(\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) + q - R_{kd}) + \sum_{k=1}^{K}\sum_{g=1}^{G_k} \frac{\text{card}\left(d_k^{[g]}\right)\left(\text{card}\left(d_k^{[g]}\right)+1\right)}{2}.
$$

Remark that the collection of models in (1.2.23) is generally large and therefore not tractable in practice. This motivates us to restrict the numbers of components K, the orders of mononomial weights L and polynomial means D among finite sets $\mathcal{K} = [K_{\max}]$, $\mathcal{L} = [L_{\max}]$ and $\mathcal{D} = [D_{\max}]$, respectively, where $K_{\max} \in \mathbb{N}^\star$, $L_{\max} \in \mathbb{N}^\star$ and $D_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. Furthermore, we focus on a (potentially random) subcollection $\mathcal{J}$ of $\mathcal{P}([q]\times[p])$, the controlled size being required in high-dimension case. Moreover, the number of possible vectors of ranks considered is reduced by working on a subset (potentially random) $\mathcal{R}_{(K,\mathbf{J},D)}$ of $[\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{DK}$.

In particular, recall that $\mathbf{B}$ is selected among a list of candidate structures $(\mathcal{B}_k)_{k\in[K]} \equiv (\mathcal{B})_{k\in[K]}$, where $\mathcal{B}$ denotes the set of all possible partitions of the covariables indexed by $[p]$ for each cluster of individuals. It is worth mentioning that the size of $\mathcal{B}$ (Bell number) is very large even for a moderate number of variables p. This prevents us to consider an exhaustive exploration of the set $\mathcal{B}$. Motivated by the recent novel work from Devijver & Gallopin (2018), for each cluster $k \in [K]$, we restrict our attention to the sub-collection $\mathcal{B}_{k,\boldsymbol{\Lambda}} = (\mathcal{B}_{k,\lambda})_{\lambda\in\boldsymbol{\Lambda}}$ of $\mathcal{B}_k$. Here $\mathcal{B}_{k,\boldsymbol{\Lambda}}$ is the partition of the variables corresponding to the block-diagonal structure of the adjacency matrix $\mathbf{E}_{k,\lambda} = \left[\mathbb{I}\left\{\left|[\mathbf{S}_k]_{z,z'}\right| > \lambda\right\}\right]_{z\in[q],z'\in[q]}$, which is based on the thresholded absolute value of the sample covariance matrix $\mathbf{S}_k$ in each cluster

$k \in [K]$. It is important to point out that the class of block-diagonal structures detected by the graphical lasso algorithm when the regularization parameter varies is identical to the block-diagonal structures $\mathcal{B}_{k,\lambda}$ detected by the thresholding of the sample covariance for each cluster $k \in [K]$ (Mazumder & Hastie, 2012).

Finally, given $S_{\mathbf{m}}$ defined as in (1.2.23), our full model collection and random subcollection of PSGaBloME models are defined, respectively, as follows:

$$\mathcal{S} = \{S_{\mathbf{m}} : \mathbf{m} \in \mathcal{M}\}, \mathcal{M} = \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times (\mathcal{B}_k)_{k\in[K]} \times \mathcal{P}([q] \times [p]) \times [\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{D_{\max}K}, \quad (1.2.24)$$

$$\widetilde{\mathcal{S}} = \left\{S_{\mathbf{m}} : \mathbf{m} \in \widetilde{\mathcal{M}}\right\}, \widetilde{\mathcal{M}} = \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times (\mathcal{B}_{k,\mathbf{\Lambda}})_{k\in[K]} \times \mathcal{J} \times \mathcal{R}_{(K,\mathbf{J},D_{\max})}. \quad (1.2.25)$$

1.2.8.5 Weak Oracle inequality

Note that in this thesis, see Section 4.3 for more details, the block-diagonal structures, the relevant variables and rank sparse models are designed, for instance, by the Lasso$+l_2$-Rank procedure in Section 4.3.5. Nevertheless, our finite-sample oracle inequality in Theorem 1.2.5, still holds for any random subcollection of $\mathcal{M}$ which is constructed by some suitable tools in the framework of PSGaBloME regression models. Note that Theorem 1.2.5 is restated as Theorem 4.3.2 with a comprehensive proof in Section 4.3.3. Furthermore, the readers can find the main difficulties regarding the proof of Theorem 1.2.5 in Section 1.2.11.8.

Theorem 1.2.5 (Oracle inequality for PSGaBloME models)**.** *Let $(\mathbf{x}_{[n]}, \mathbf{y}_{[n]})$ be the observations arising from the unknown conditional density s_0. For each $\mathbf{m} \equiv (K, L, D, \mathbf{B}, \mathbf{J}, \mathbf{R}) \in \mathcal{M}$, let $S_{\mathbf{m}}$ be given by (1.2.23). Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$ such that*

$$\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \bar{s}_{\mathbf{m}}) \leq \inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + \frac{\epsilon_{KL}}{n}, \text{ and } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0.$$

Furthermore, we construct a random subcollection $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ as in (1.2.25) and consider the collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of η-log likelihood minimizers defined in (1.2.6). Then, there is a constant C such that for any $\rho \in (0,1)$, and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, for every index $\mathbf{m} \in \mathcal{M}$, $\xi_{\mathbf{m}} \in \mathbb{R}^+$, $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$,

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n)\dim(S_{\mathbf{m}}) + (1 \vee \tau)\xi_{\mathbf{m}}\right], \kappa > \kappa_0,$$

the η'-penalized likelihood estimator $\widehat{s}_{\widehat{\mathbf{m}}}$, defined in (1.2.5) on the subset $\widetilde{\mathcal{M}}$ instead of $\mathcal{M}$, satisfies

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta' + \eta}{n}.$$

1.2.9 An l_1-oracle inequality for the Lasso estimator in SGaME regression models

1.2.9.1 Fixed predictors and number of components with linear Gaussian mean functions

Inspired by the framework in Meynet (2013) and Devijver (2015a), the explanatory variables $\mathbf{x}_i$ and the number of components $K \in \mathbb{N}^\star$ are both fixed. We assume that the observed $\mathbf{x}_i, i \in [n]$, are finite. Without loss of generality, we choose to rescale $\mathbf{x}$, so that $\|\mathbf{x}\|_\infty \leq 1$. Therefore, we can assume that the explanatory variables $\mathbf{x}_i \in \mathcal{X} = [0,1]^p$, for all $i \in [n]$. Note that such a restriction is also used in Devijver (2015a). Under only the assumption of bounded parameters, we provide a lower bound on the Lasso regularization parameter λ, which guarantees an oracle inequality. Note that in this non-random explanatory variables setting, we focus on the Lasso for its l_1-regularization properties rather than as a model selection procedure, as in the case of random explanatory variables and unknown K, as in Montuelle et al. (2014), Nguyen et al. (2021c,b), see also Sections 1.2.6 to 1.2.8.

For simplicity, we consider the case where the means of Gaussian experts are linear functions of the explanatory variables; *i.e.,*

$$\boldsymbol{\Upsilon} = \left\{ \boldsymbol{\upsilon} : \mathcal{X} \mapsto \boldsymbol{\upsilon}_{\boldsymbol{\beta}}(\mathbf{x}) := (\boldsymbol{\beta}_{k0} + \boldsymbol{\beta}_k \mathbf{x})_{k\in[K]} \in (\mathbb{R}^q)^K \,\middle|\, \boldsymbol{\beta} = (\boldsymbol{\beta}_{k0}, \boldsymbol{\beta}_k)_{k\in[K]} \in \mathcal{B} = \left(\mathbb{R}^{q\times(p+1)}\right)^K \right\},$$

where $\boldsymbol{\beta}_{k0}$ and $\boldsymbol{\beta}_k$ are respectively the $q \times 1$ vector of bias and the $q \times p$ regression coefficients matrix for the kth expert.

In summary, we wish to estimate s_0 via conditional densities belonging to the class:

$$\{(\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) \,|\, \boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \boldsymbol{\Psi}\}, \tag{1.2.26}$$

where $\boldsymbol{\Psi} = \boldsymbol{\Gamma} \times \boldsymbol{\Xi}$, and $\boldsymbol{\Xi} = \mathcal{B} \times \boldsymbol{V}$.

From hereon in, for a vector $\mathbf{x} \in \mathbb{R}^p$, we assume that $\mathbf{x} = (x_1, \ldots, x_p)$ is in the column form. Similarly, the parameter of the entire model, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, is also a column vector, where we consider any matrix as a vector produced using vec$(\cdot)$: the vectorization operator that stacks the columns of a matrix into a vector.

1.2.9.2 Boundedness assumption on the softmax gating and Gaussian parameters

We shall restrict our study to estimate s_0 by conditional PDFs belonging to the model class $\mathcal{S}$, which has boundedness assumptions on the softmax gating and Gaussian expert parameters. Specifically, we assume that there exists deterministic constants $A_{\boldsymbol{\gamma}}, A_{\boldsymbol{\beta}}, a_{\boldsymbol{\Sigma}}, A_{\boldsymbol{\Sigma}} > 0$, such that $\boldsymbol{\psi} \in \widetilde{\boldsymbol{\Psi}}$, where

$$\begin{aligned}
\widetilde{\boldsymbol{\Gamma}} &= \left\{ \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \mid \forall k \in [K], \sup_{\mathbf{x}\in\mathcal{X}} \left(|\gamma_{k0}| + \left| \boldsymbol{\gamma}_k^\top \mathbf{x} \right| \right) \leq A_{\boldsymbol{\gamma}} \right\}, \\
\widetilde{\boldsymbol{\Xi}} &= \left\{ \xi \in \boldsymbol{\Xi} \mid \forall k \in [K], \max_{z\in\{1,\ldots,q\}} \sup_{\mathbf{x}\in\mathcal{X}} \left(|[\boldsymbol{\beta}_{k0}]_z| + |[\boldsymbol{\beta}_k \mathbf{x}]_z| \right) \leq A_{\boldsymbol{\beta}}, a_{\boldsymbol{\Sigma}} \leq m\left(\boldsymbol{\Sigma}_k^{-1}\right) \leq M\left(\boldsymbol{\Sigma}_k^{-1}\right) \leq A_{\boldsymbol{\Sigma}} \right\}, \\
\widetilde{\boldsymbol{\Psi}} &= \widetilde{\boldsymbol{\Gamma}} \times \widetilde{\boldsymbol{\Xi}}.
\end{aligned} \tag{1.2.27}$$

Since

$$a_G := \frac{\exp\left(-A_{\boldsymbol{\gamma}}\right)}{\sum_{l=1}^K \exp\left(A_{\boldsymbol{\gamma}}\right)} \leq \sup_{\mathbf{x}\in\mathcal{X}, \boldsymbol{\gamma}\in\widetilde{\boldsymbol{\Gamma}}} \frac{\exp\left(\gamma_{k0} + \boldsymbol{\gamma}_k^\top \mathbf{x}\right)}{\sum_{l=1}^K \exp\left(\gamma_{l0} + \boldsymbol{\gamma}_l^\top \mathbf{x}\right)} \leq \frac{\exp\left(A_{\boldsymbol{\gamma}}\right)}{\sum_{l=1}^K \exp\left(-A_{\boldsymbol{\gamma}}\right)} =: A_G,$$

there exists deterministic positive constants a_G, A_G, such that

$$a_G \leq \sup_{\mathbf{x}\in\mathcal{X}, \boldsymbol{\gamma}\in\widetilde{\boldsymbol{\Gamma}}} g_k\left(\mathbf{x}; \boldsymbol{\gamma}\right) \leq A_G. \tag{1.2.28}$$

We wish to use the model class $\mathcal{S}$ of conditional PDFs to estimate s_0, where

$$\mathcal{S} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) \,\middle|\, \boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Psi}} \right\}. \tag{1.2.29}$$

To simplify the proofs, we shall assume that the true density s_0 belongs to $\mathcal{S}$. That is to say, there exists $\boldsymbol{\psi}_0 = (\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \in \widetilde{\boldsymbol{\Psi}}$, such that $s_0 = s_{\boldsymbol{\psi}_0}$.

Since we are working with conditional PDFs and not with classical densities, we define the following adapted Kullback–Leibler information, that takes into account the structure of conditional PDFs. For fixed explanatory variables $(\mathbf{x}_i)_{1\leq i\leq n}$, we consider the average loss function

$$\mathrm{KL}_n(s, t) = \frac{1}{n} \sum_{i=1}^n \mathrm{KL}\left(s\left(\cdot|\mathbf{x}_i\right), t\left(\cdot|\mathbf{x}_i\right)\right) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^q} \ln\left(\frac{s\left(\mathbf{y}|\mathbf{x}_i\right)}{t\left(\mathbf{y}|\mathbf{x}_i\right)}\right) s\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y}. \tag{1.2.30}$$

However, since we want to handle high-dimensional data, we have to regularize the maximum likelihood estimator (MLE) in order to obtain reasonable estimates. Here, we shall consider l_1-regularization and the associated so-called Lasso estimator, which is the l_1-norm penalized MLE defined as follows:

$$\widehat{s}^{\,\mathrm{Lasso}}(\lambda) := \arg\min_{s_{\boldsymbol{\psi}}\in\mathcal{S}} \left\{ -\frac{1}{n} \sum_{i=1}^n \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \mathrm{pen}_\lambda(\boldsymbol{\psi}) \right\}, \tag{1.2.31}$$

where $\lambda \geq 0$ is a regularization parameter to be tuned, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ and

$$\text{pen}_\lambda(\boldsymbol{\psi}) = \lambda \left\| \boldsymbol{\psi}^{[1,2]} \right\|_1 := \lambda \left(\left\| \boldsymbol{\psi}^{[1]} \right\|_1 + \left\| \boldsymbol{\psi}^{[2]} \right\|_1 \right), \tag{1.2.32}$$

$$\left\| \boldsymbol{\psi}^{[1]} \right\|_1 = \|\boldsymbol{\gamma}\|_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} |\gamma_{kj}|, \tag{1.2.33}$$

$$\left\| \boldsymbol{\psi}^{[2]} \right\|_1 = \|\text{vec}(\boldsymbol{\beta})\|_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{z=1}^{q} \left| [\boldsymbol{\beta}_k]_{z,j} \right|. \tag{1.2.34}$$

From now on, we denote $\|\boldsymbol{\beta}\|_p$ ($p \in \{1, 2, \infty\}$) by the induced p-norm of a matrix , which differs from $\|\text{vec}(\boldsymbol{\beta})\|_p$.

Note that $\text{pen}_\lambda(\boldsymbol{\psi})$ is a Lasso regularization term encouraging sparsity for both the gating and expert parameters. Recall that this penalty is also studied in Khalili (2010), Chamroukhi & Huynh (2018), and Chamroukhi & Huynh (2019), in which the authors studied the univariate case: $\mathbf{Y} \in \mathbb{R}$. Notice that, without considering the l_2-norm, the penalty function considered in such frameworks belongs to our framework and the l_1-oracle inequality from Theorem 1.2.7 can be obtained for it. Indeed, by considering $\lambda = \min\left\{\lambda_1^{[1]}, \ldots, \lambda_K^{[1]}, \lambda_1^{[2]}, \ldots, \lambda_K^{[2]}, \frac{\lambda^{[3]}}{2}\right\}$, the condition for a regularization parameter's lower bound, (1.2.36) from Theorem 1.2.7, can also be applied to model their models , which leads to an l_1-oracle inequality.

1.2.9.3 An l_1-oracle inequality for the Lasso estimator

In this section, Theorem 1.2.7 provides an l_1-oracle inequality for SGaME regression models. It is one the contributions of this thesis and is motivated by the problem studied in Meynet (2013) and Devijver (2015a).

Firstly, we aim to prove that the negative of differential entropy (see its definition, *e.g.,* from Mansuripur (1987, Chapter 9)) of the true unknown conditional density $s_0 \in S$, defined in (1.2.26), is finite, see more in Lemma 1.2.6, which is restated in Lemma 4.2.1 and is proved in Section 4.2.4.3.

Lemma 1.2.6 (Differential entropy of SGaME regression model with boundedness assumptions on parameter spaces)**.** *There exist a nonnegative constant* $H_{s_0} = \max\{0, \ln C_{s_0}\}$, $C_{s_0} = (2\pi)^{-q/2} \left(2A_\Sigma^{-1}\right)^{-q/2}$, *such that*

$$\max\left\{0, \sup_{x \in \mathcal{X}} \int_{\mathbb{R}^q} \ln\left(s_0\left(y|x\right)\right) s_0\left(y|x\right) dy\right\} \leq H_{s_0} < \infty. \tag{1.2.35}$$

Note that Theorem 1.2.7 is restated as Theorem 4.2.2. Furthermore, it is briefly and fully proved in Sections 1.2.12 and 4.2.3.3, respectively, see also Nguyen et al. (2020c).

Theorem 1.2.7 (l_1-oracle inequality for SGaME regression models)**.** *We observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right) \in ([0,1]^p \times \mathbb{R}^q)$, *coming from the unknown conditional mixture of Gaussian experts regression models* $s_0 := s_{\boldsymbol{\psi}_0} \in \mathcal{S}$, *cf.* (1.2.29). *We define the Lasso estimator* $\widehat{s}^{\text{Lasso}}(\lambda)$, *by* (1.2.31), *where* $\lambda \geq 0$ *is a regularization parameter to be tuned. Then, if*

$$\lambda \geq \kappa \frac{K B_n'}{\sqrt{n}} \left(q \ln n \sqrt{\ln(2p+1)} + 1 \right), \tag{1.2.36}$$

$$B_n' = \max\left(A_{\boldsymbol{\Sigma}}, 1 + K A_G\right) \left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}} \left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right), \tag{1.2.37}$$

for some absolute constants $\kappa \geq 148$*, the estimator* $\widehat{s}^{\text{Lasso}}(\lambda)$ *satisfies the following* l_1*-oracle inequality:*

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}^{\text{Lasso}}(\lambda)\right)\right] \leq & \left(1+\kappa^{-1}\right) \inf_{s_{\boldsymbol{\psi}} \in \mathcal{S}} \left(\text{KL}_n\left(s_0, s_{\boldsymbol{\psi}}\right) + \lambda \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right) + \lambda \\ & + \sqrt{\frac{K}{n}} \left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}} + H_{s_0}\right) \sqrt{2qA_{\boldsymbol{\gamma}}} \\ & + 302q\sqrt{\frac{K}{n}} \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}}\ln n\right)\right) \\ & \times K\left(1+\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right). \end{aligned} \tag{1.2.38}$$

$$\tag{1.2.39}$$

Next, we state the following Theorem 1.2.8, which is an l_1-ball MoE regression model selection theorem for l_1-penalized maximum conditional likelihood estimation in the Gaussian mixture framework. Note that Theorem 1.2.7 is an immediate consequence of Theorem 1.2.8. Furthermore, Theorem 1.2.8 is restated as Theorem 4.2.3 and is proved in Section 4.2.3.4, see also Nguyen et al. (2020c).

Theorem 1.2.8. *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$ *with unknown conditional Gaussian mixture PDF* s_0*. For all* $m \in \mathbb{N}^\star$*, consider the* l_1*-ball*

$$S_m = \left\{s_{\boldsymbol{\psi}} \in \mathcal{S}, \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 \leq m\right\}, \tag{1.2.40}$$

and let $\widehat{s}_m$ *be a* η_m*-*ln*-likelihood minimizer in* S_m *for some* $\eta_m \geq 0$*:*

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) \leq \inf_{s_m \in S_m} \left(-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)\right) + \eta_m. \tag{1.2.41}$$

Assume that, for all $m \in \mathbb{N}^\star$*, the penalty function satisfies* $\text{pen}(m) = \lambda m$*, where* λ *is defined later. Then, we define the penalized likelihood estimator* $\widehat{s}_{\widehat{m}}$*, where* $\widehat{m}$ *is defined via the satisfaction of the inequality*

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_{\widehat{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}(\widehat{m}) \leq \inf_{m \in \mathbb{N}^\star} \left(-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}(m)\right) + \eta, \tag{1.2.42}$$

for some $\eta \geq 0$*. Then, if*

$$\lambda \geq \kappa \frac{KB_n'}{\sqrt{n}} \left(q \ln n \sqrt{\ln(2p+1)} + 1\right), \tag{1.2.43}$$

$$B_n' = \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}}\ln n\right)\right), \tag{1.2.44}$$

for some absolute constants $\kappa \geq 148$*, then*

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\right] \leq & \left(1+\kappa^{-1}\right) \inf_{m \in \mathbb{N}^\star} \left(\inf_{s_m \in S_m} \text{KL}_n\left(s_0, s_m\right) + \text{pen}(m) + \eta_m\right) + \eta \\ & + \sqrt{\frac{K}{n}} \left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}} + H_{s_0}\right) \sqrt{2qA_{\boldsymbol{\gamma}}} \\ & + 302q\sqrt{\frac{K}{n}} \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}}\ln n\right)\right) \\ & \times K\left(1+\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right). \end{aligned} \tag{1.2.45}$$

1.2.10 Our contributions for weak oracle inequalities in deterministic collection of MoE models via Theorem 1.2.2

The deterministic collection of MoE models include GLLiM, GLoME, SGaME and LinBoSGaME models where weak oracle inequalities were well studied for last two models via a general conditional density model selection theorem of Cohen & Le Pennec (2011, Theorem 2), see also Cohen & Le Pennec (2013, Theorem 2.2.) and Montuelle et al. (2014, Theorem 2). We first summarize this general model selection theorem and the techniques that Montuelle et al. (2014) used to control the bracketing entropy of LinBoSGaME models with softmax gating networks in Sections 1.2.10.1 and 1.2.10.2, respectively. Then, we explain why such techniques can not be directly applied to our collection of GLoME models via Section 1.2.10.3 to highlight the main challenges and our contributions.

1.2.10.1 A general conditional density model selection theorem for deterministic collection of models

Before stating a general model selection for conditional density, we have to present some regularity assumptions.

First, we need an information theory type assumption to control the complexity of our collection. We assume the existence of a Kraft-type inequality for the collection (Massart, 2007, Barron et al., 2008).

Assumption 1.2.1 (K). *There is a family* $(z_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *of non-negative numbers and a real number* Ξ *such that*

$$\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < +\infty.$$

For technical reasons, a seperability assumption always satisfied in the setting of this paper, is also required. It is a mild condition, which is classical in empirical process theory (Van Der Vaart & Wellner, 1996, van de Geer, 2000). This assumption allows us to work with a countable subset.

Assumption 1.2.2 (Sep). *For every model* $S_{\mathbf{m}}$ *in the collection* $\mathcal{S}$*, there exists some countable subset* $S'_{\mathbf{m}}$ *of* $S_{\mathbf{m}}$ *and a set* $\mathcal{X}'_{\mathbf{m}}$ *with* $\iota\left(\mathcal{X}\setminus\mathcal{X}'_{\mathbf{m}}\right)=0$*, where* ι *denotes Lebesgue measure, such that for every* $t\in S_{\mathbf{m}}$*, there exists some sequence* $(t_k)_{k\geq 1}$ *of elements of* $S'_{\mathbf{m}}$*, such that for every* $\mathbf{y}\in\mathcal{Y}$ *and every* $\mathbf{x}\in\mathcal{X}'_{\mathbf{m}}, \ln\left(t_k\left(\mathbf{x}|\mathbf{y}\right)\right)\xrightarrow{k\to+\infty}\ln\left(t\left(\mathbf{x}|\mathbf{y}\right)\right)$.

Next, recall that the bracketing entropy of a set S with respect to any distance d, denoted by $\mathcal{H}_{[\cdot],d}((\delta,S))$, is defined as the logarithm of the minimal number $\mathcal{N}_{[\cdot],d}\left(\delta,S\right)$ of brackets $[t^-,t^+]$ covering S, such that $d(t^-,t^+)\leq\delta$. That is,

$$N_{[\cdot],d}\left(\delta,S\right):=\min\left\{n\in\mathbb{N}^\star:\exists t_1^-,t_1^+,\ldots,t_n^-,t_n^+\text{ s.t }d(t_k^-,t_k^+)\leq\delta, S\subset\bigcup_{k=1}^{n}\left[t_k^-,t_k^+\right]\right\},\tag{1.2.46}$$

where the bracket $s\in\left[t_k^-,t_k^+\right]$ is defined by $t_k^-(\mathbf{x},\mathbf{y})\leq s(\mathbf{x},\mathbf{y})\leq t_k^+(\mathbf{x},\mathbf{y}),\ \forall(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$.

We also need the following important assumption on Dudley-type integral of these bracketing entropies, which is utilized often in empirical process theory (Van Der Vaart & Wellner, 1996, van de Geer, 2000, Kosorok, 2007).

Assumption 1.2.3 (H). *For every model* $S_{\mathbf{m}}$ *in the collection* $\mathcal{S}$*, there is a non-decreasing function* $\phi_{\mathbf{m}}$ *such that* $\delta\mapsto\frac{1}{\delta}\phi_{\mathbf{m}}(\delta)$ *is non-increasing on* $(0,\infty)$ *and for every* $\delta\in\mathbb{R}^+$*,*

$$\int_0^\delta\sqrt{\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes\mathrm{n}}}\left(\delta,S_{\mathbf{m}}\left(\widetilde{s},\delta\right)\right)}d\delta\leq\phi_{\mathbf{m}}(\delta),$$

where $S_{\mathbf{m}}\left(\widetilde{s},\delta\right)=\{s_{\mathbf{m}}\in S_{\mathbf{m}}:\mathrm{d}^{\otimes\mathrm{n}}\left(\widetilde{s},s_{\mathbf{m}}\right)\leq\delta\}$*. The model complexity of* $S_{\mathbf{m}}$ *is then defined as* $\mathcal{D}_{\mathbf{m}}$ $=n\delta_{\mathbf{m}}^2$*, where* $\delta_{\mathbf{m}}$ *is the unique root of* $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta)=\sqrt{n}\delta$*.*

Observe that the model complexity does not depend on the bracketing entropies of the global models $S_{\mathbf{m}}$, but rather on those of smaller localized sets $S_{\mathbf{m}}(\widetilde{s}, \delta)$. Now we are able to state an important weak oracle inequality, Theorem 1.2.9, from Cohen & Le Pennec (2011).

Theorem 1.2.9 (Theorem 2 from Cohen & Le Pennec 2011)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, arising from an unknown conditional density* s_0*. Let* $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *be an at most countable conditional density model collection. Assume that Assumption 3.2.1 (K), Assumption 3.2.2 (Sep), and Assumption 3.2.3 (H) hold for every model* $S_{\mathbf{m}} \in \mathcal{S}$*. Then, for any* $\rho \in (0,1)$ *and any* $C_1 > 1$*, there is a constant* κ_0 *depending only on* ρ *and* C_1*, such that for every index* $\mathbf{m} \in \mathcal{M}$*,*

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left(n\delta_{\mathbf{m}}^2 + z_{\mathbf{m}}\right)$$

with $\kappa > \kappa_0$ *and* $\delta_{\mathbf{m}}$ *is the unique root of* $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$*, such that the* η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$ *satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right) + \frac{\kappa_0 C_1 \Xi}{n} + \frac{\eta + \eta'}{n}. \tag{1.2.47}$$

For the sake of generality, this Theorem 1.2.9 is relatively abstract. Since the assumptions of the previous Theorem 1.2.9 are as general as possible. But from the practical point of view, a natural question is the existence of interesting model collections that satisfy these assumptions. We will sketch of the proof fo LinBoSGaME models of Montuelle et al. (2014, Theorem 1) and show that their result can not directly applicable to the GLoME setting. The main reason is that the technique for handling the linear combination of bounded functions for the weight functions of logistic schemes of Montuelle et al. (2014) is not valid for the Gaussian gating parameters in GLoME models. Therefore, we propose a *reparameterization trick*[3] to bound the metric entropy of the Gaussian gating parameters space; see Equation (1.2.50) for more details.

1.2.10.2 Sketch of the proof for LinBoSGaME models

To prove the main conditional density model selection theorem for LinBoSGaME models, Montuelle et al. (2014, Theorem 1), the authors have to make use of Theorem 1.2.9. Then, they need to prove that their collection of LinBoSGaME models have to satisfy Assumption 1.2.1 (K), Assumption 1.2.2 (Sep), and Assumption 1.2.3 (H). However, they did not prove Assumption 1.2.1 (K) and Assumption 1.2.2 (Sep) and considered them as assumptions on their LinBoSGaME models because of the complexity of LinBoSGaME models and technical reasons. Therefore, the main difficulty remains on verifying Assumption 1.2.3 (H) via several bracketing entropy controls of the linear combination of bounded functions for the weight functions of logistic schemes.

Firstly, they define the following distance over conditional densities:

$$\sup_{\mathbf{y}} d_{\mathbf{x}}(s,t) = \sup_{\mathbf{y}\in\mathcal{Y}}\left(\int_{\mathcal{X}}\left(\sqrt{s(\mathbf{x}|\mathbf{y})} - \sqrt{t(\mathbf{x}|\mathbf{y})}\right)^2 d\mathbf{x}\right)^{1/2}.$$

This leads straightforwardly to $\mathrm{d}^{2\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}^2(s,t)$. Then, they also define

$$\sup_{\mathbf{y}} d_k\left(g, g'\right) = \sup_{\mathbf{y}\in\mathcal{Y}}\left(\sum_{k=1}^{K}\left(\sqrt{g_k(\mathbf{y})} - \sqrt{g_k'(\mathbf{y})}\right)^2\right)^{1/2},$$

for any gating functions g and g'. To this end, given any densities s and t over $\mathcal{X}$, the following distance, depending on $\mathbf{y}$, is constructed as follows:

$$\sup_{\mathbf{y}}\max_{k} d_{\mathbf{x}}(s,t) = \sup_{\mathbf{y}\in\mathcal{Y}}\max_{k\in[K]} d_{\mathbf{x}}\left(s_k(\cdot,\mathbf{y}), t_k(\cdot,\mathbf{y})\right) = \sup_{\mathbf{y}\in\mathcal{Y}}\max_{k\in[K]}\left(\int_{\mathcal{X}}\left(\sqrt{s_k(\mathbf{x},\mathbf{y})} - \sqrt{t_k(\mathbf{x},\mathbf{y})}\right)^2 d\mathbf{x}\right)^{1/2}.$$

[3]Recall that we only use this nomenclature to perform a change of variables of the Gaussian gating parameters space of GLoME models via the logistic weights of SGaME models. This reparameterization trick does not stand for the well-known one of Variational Autoencoders (VAEs) in the deep learning literature (see Kingma & Welling, 2013, for more details).

Then, they prove that definition of complexity of model $S_{\mathbf{m}}$ in Assumption 1.2.3 (H) is related to an classical entropy dimension with respect to a Hellinger type divergence $\mathrm{d}^{\otimes \mathrm{n}}$, due to Proposition 1.2.10.

Proposition 1.2.10 (Proposition 2 from Cohen & Le Pennec 2011)**.** *For any $\delta \in (0, \sqrt{2}]$, such that $\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right)$, the function*

$$\phi_{\mathbf{m}}(\delta) = \delta \sqrt{\dim(S_{\mathbf{m}})} \left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min(\delta, 1)}\right)} \right)$$

satisfies Assumption 1.2.3 (H). Furthermore, the unique solution $\delta_{\mathbf{m}}$ of $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$, satisfies

$$n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}}) \left(2\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 + \left(\ln \frac{n}{\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 \dim(S_{\mathbf{m}})} \right)_+ \right).$$

Therefore, Proposition 1.2.10 implies that Assumption 1.2.3 (H) can be proved via Lemma 1.2.11.

Lemma 1.2.11. *For any $\delta \in (0, \sqrt{2}]$, the collection of LinBoSGaME models, $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m} \in \mathcal{M}}$, satisfies*

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right) \right).$$

Lemma 1.2.11 is then obtained by decomposing the entropy terms between the softmax gating functions and the Gaussian experts. Note that both LinBoSGaME and GLoME models share the same structures of Gaussian experts mean, recall Figures 1.4 and 1.6 for more details. Therefore, we only highlight our contributions regarding the control of bracketing entropy for the parameter of gating network compared to Montuelle et al. (2014). To do that, the author rewrite the softmax gating parameters' space in (1.1.16) as follows:

$$\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}^{K-1}, \mathbf{W} = \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \sum_{d=1}^{d_{\mathbf{W}}} \omega_d \theta_{\mathbf{W},d}(\mathbf{y}) \in \mathbb{R} : \max_{d \in [d_{\mathbf{W}}]} |\omega_d| \leq T_{\mathbf{W}} \right\}, \tag{1.2.48}$$

$$\mathcal{P}_K = \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^K e^{\mathbf{w}_l(\mathbf{y})}} \right)_{k \in [K]} =: \left(g_{\mathbf{w},k}(\mathbf{y})\right)_{k \in [K]}, \mathbf{w} \in \mathbf{W}_{K,d_{\mathbf{W}}} \right\}.$$

Then, they also require the definition of metric entropy of the set $\mathbf{W}_K$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_K)$, which measures the logarithm of the minimal number of balls of radius at most δ, according to a distance $d_{\|\sup\|_\infty}$, needed to cover $\mathbf{W}_K$ where

$$d_{\|\sup\|_\infty}\left((s_k)_{k \in [K]}, (t_k)_{k \in [K]}\right) = \max_{k \in [K]} \sup_{\mathbf{y} \in \mathcal{Y}} \|s_k(\mathbf{y}) - t_k(\mathbf{y})\|_2, \tag{1.2.49}$$

for any K-tuples of functions $(s_k)_{k \in [K]}$, $(t_k)_{k \in [K]}$ and $\|s_k(\mathbf{y}) - t_k(\mathbf{y})\|_2$ is the Euclidean distance in $\mathbb{R}^L$. By using Lemma 5 and Proposition 2 from Montuelle et al. 2014, see also Lemma 3.2.9 for more detail, Lemma 1.2.11 holds true if we can prove Lemma 1.2.12. Note that the first inequality of Lemma 1.2.12 comes from Montuelle et al. (2014, Lemma 4) and describes relationship between the bracketing entropy of $\mathcal{P}_K$ and the entropy of $\mathbf{W}_K$.

Lemma 1.2.12. *For all $\delta \in (0, \sqrt{2}]$, there exists a constant $C_{\mathbf{W}_K}$ such that*

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) \leq \mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, \mathbf{W}_K\right) \leq \dim(\mathbf{W}_K)\left(C_{\mathbf{W}_K} + \ln\left(\frac{20\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right).$$

By the nice linear property from the construction of linear combination of a finite set of bounded functions whose coefficients belong to a compact set, in the argument from Montuelle et al. (2014, Proof of Part 1 of Lemma 1, Page 1689), the second inequality of Lemma 1.2.12 is then easily established as follows. However, this will be not the case for our Lemma 1.2.13, which is used for controlling the bracketing entropy not only for many standard MoE regression models with Gaussian gating networks, see Section 1.2.10.3 for more details.

Proof of the second inequality of Lemma 1.2.12. Note that for all

$$\mathbf{w} = (0, w_k)_{k\in[K-1]} \in \mathbf{W}_{K,d_{\mathbf{W}}}, \quad \mathbf{v} = (0, v_k)_{k\in[K-1]} \in \mathbf{W}_{K,d_{\mathbf{W}}},$$

it holds that

$$\begin{aligned} d_{\|\sup\|_\infty}(\mathbf{w}-\mathbf{v}) &= \max_{k\in[K-1]} \|\mathbf{w}_k - \mathbf{v}_k\|_\infty = \max_{k\in[K-1]} \sup_{\mathbf{y}\in\mathbf{Y}} \left| \sum_{i=1}^{d_{\mathbf{W}}} \omega_{k,i}^{\mathbf{w}} \theta_{\mathbf{W},i}(\mathbf{y}) - \sum_{i=1}^{d_{\mathbf{W}}} \omega_{k,i}^{\mathbf{v}} \theta_{\mathbf{W},i}(\mathbf{y}) \right| \\ &\leq \max_{k\in[K-1]} \sum_{i=1}^{d_{\mathbf{W}}} \left|\omega_{k,i}^{\mathbf{w}} - \omega_{k,i}^{\mathbf{v}}\right| \underbrace{\sup_{\mathbf{y}\in\mathbf{Y}} |\theta_{\mathbf{W},i}(\mathbf{y})|}_{\leq 1} \leq d_{\mathbf{W}} \max_{k\in[K-1],i\in[d_{\mathbf{W}}]} \left|\omega_{k,i}^{\mathbf{w}} - \omega_{k,i}^{\mathbf{v}}\right|. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} \mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, \mathbf{W}_K\right) &\leq \mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}d_{\mathbf{W}}}, \left\{\boldsymbol{\omega} \in \mathbb{R}^{(K-1)d_{\mathbf{W}}} : \|\boldsymbol{\omega}\|_\infty \leq T_{\mathbf{W}}\right\}\right) \\ &\leq (K-1)d_{\mathbf{W}} \ln\left(1 + \frac{20\sqrt{K-1}d_{\mathbf{W}}T_{\mathbf{W}}}{3\sqrt{3}\delta}\right) \\ &\leq (K-1)d_{\mathbf{W}} \ln\left(\sqrt{2} + \frac{20\sqrt{K-1}d_{\mathbf{W}}T_{\mathbf{W}}}{3\sqrt{3}} + \ln\left(\frac{1}{\delta}\right)\right). \end{aligned}$$

□

1.2.10.3 Our contributions on the proof for GLoME models

To prove Theorem 1.2.2, we also need to make use of Theorem 1.2.9 from Cohen & Le Pennec (2011, 2013). Then, our model collection has to satisfy Assumption 1.2.1 (K), Assumption 1.2.2 (Sep), and Assumption 1.2.3 (H).

As in the proof of LinBoSGaME models, Assumption 1.2.1 (K) and Assumption 1.2.2 (Sep) can be easily verified. In our proof for GLoME models, as a complementary to the proof for LinBoSGaME models, we consider an explicit example where the model is defined by $\mathcal{M} = \mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} = [K_{\max}] \times [d_{\max}]$, $K_{\max}, d_{\max} \in \mathbb{N}^\star$, leads to the Assumption 3.2.1 (K) is always satisfied. It is interesting to find the optimal family $(z_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ satisfying Assumption 1.2.1 (K). To the best of our knowledge, this question is only partially answered in some special cases of MoE regression models, *e.g.,* Gaussian finite mixture model as in Figure 1.3 (a) Maugis & Michel (2011b), finite mixture of Gaussian regression models as in Figure 1.3 (b). However, for the standard MoE regression models as in Figure 1.3 (d), *e.g.,* LinBoSGaME and GLoME models, such question still remains open due to the complexity of models. We hope to resolve this important and interesting problem in our future work. Furthermore, remark that the Assumption 1.2.2 (Sep) is true when we consider Gaussian densities Massart (2007).

Therefore, our model has only to satisfy the remaining Assumption 1.2.3 (H). Following the same strategy as in the proof of LinBoSGaME models, the main task for GLoME models is to prove Lemma 1.2.13, which is similar but much more difficult compared to Lemma 1.2.12.

Another important contribution lies on the numerical experiments in Section 3.2.3. Note that our main objective here is to investigate how well the empirical tensorized Kullback–Leibler divergence between the true model (s_0^*) and the selected model $\widehat{s}^*_{\widehat{\mathbf{m}}}$ follows the finite-sample oracle inequality of Theorem 1.2.2, as well as the rate of convergence of the error term. Therefore, we focus on 1-dimensional data sets, that is, with $L = D = 1$. Beyond the statistical estimation and model selection objectives considered here, the dimensionality reduction capability of GLLiM in high-dimensional regression data, typically $D \gg L$, can be found in (Deleforge et al., 2015c, Section 6).

For the Gaussian gating parameters, to make use of the first inequality from Lemma 1.2.12 of Montuelle et al. (2014), we propose the following *reparameterization trick* of the Gaussian gating

space, which is defined in (1.1.5), via the logistics scheme $\mathcal{P}_K$ and the nonlinear space $\mathcal{W}_K$ as follows:

$$\mathcal{W}_K = \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \left(\ln\left(\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right)\right)_{k\in[K]} =: \left(\mathbf{w}_k(\mathbf{y};\boldsymbol{\omega})\right)_{k\in[K]} = \mathbf{w}\left(\mathbf{y};\boldsymbol{\omega}\right) : \boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}_K \right\}, \tag{1.2.50}$$

$$\mathcal{P}_K = \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^K e^{\mathbf{w}_l(\mathbf{y})}} \right)_{k\in[K]} =: \left(g_{\mathbf{w},k}\left(\mathbf{y}\right)\right)_{k\in[K]}, \mathbf{w} \in \mathcal{W}_K \right\}. \tag{1.2.51}$$

We aim to provide the following important upper bound for metric entropy of nonlinear space Lemma 1.2.13, which play a key step for controlling the bracketing entropy not only for GLoME models but also for any standard MoE regression models with Gaussian gating networks, *e.g.,* BLoME models, see again Figure 1.3 (d) and Figure 1.4 for comprehensive descriptions of this general class.

Lemma 1.2.13. *For all $\delta \in (0, \sqrt{2}]$, there exists a constant $C_{\mathcal{W}_K}$ such that*

$$\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, \mathcal{W}_K\right) \leq \dim\left(\mathcal{W}_K\right)\left(C_{\mathcal{W}_K} + \ln\left(\frac{20\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right).$$

Proof of Lemma 1.2.13. Note that Lemma 1.2.13 is obtained by proving that there exists a constant $C_{\mathcal{W}_K}$ such that $\forall \delta \in (0, 2]$,

$$\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\delta, \mathcal{W}_K\right) \leq \dim\left(\mathcal{W}_K\right)\left(C_{\mathcal{W}_K} + \ln\left(\frac{1}{\delta}\right)\right), \tag{1.2.52}$$

where $\dim\left(\mathcal{W}_K\right) = K - 1 + KL + K\frac{L(L+1)}{2}$.

In order to establish the proof for (1.2.52), we have to construct firstly the $\delta_{\boldsymbol{\pi}}$-covering $\mathbf{\Pi}_{K-1,\boldsymbol{\omega}}$ of $\mathbf{\Pi}_{K-1}$ via Lemma 1.2.14, which is proved in Section 1.2.10.4.

Lemma 1.2.14 (Covering number of probability simplex with maximum norm)**.** *Given any $\delta_{\boldsymbol{\pi}} > 0$, any $\boldsymbol{\pi} \in \mathbf{\Pi}_{K-1}$, we can choose $\widehat{\boldsymbol{\pi}} \in \mathbf{\Pi}_{K-1,\boldsymbol{\omega}}$, an $\delta_{\boldsymbol{\pi}}$-covering of $\mathbf{\Pi}_{K-1}$, so that $\max_{k\in[K]}|\pi_k - \widehat{\pi}_k| \leq \delta_{\boldsymbol{\pi}}$. Furthermore, it holds that*

$$\mathcal{N}\left(\delta_{\boldsymbol{\pi}}, \mathbf{\Pi}_{K-1}, \|\cdot\|_\infty\right) \leq \frac{K\left(2\pi e\right)^{K/2}}{\delta_{\boldsymbol{\pi}}^{K-1}}. \tag{1.2.53}$$

Then, by definition of the covering number, (1.2.52) is obtained immediately via Lemma 1.2.15, which controls the covering number of $\mathcal{W}_K$ and is proved in Section 1.2.10.5.

Lemma 1.2.15. *Given a bounded set $\mathcal{Y}$ in $\mathbb{R}^L$ such that $\mathcal{Y} = \left\{\mathbf{y} \in \mathbb{R}^L : \|\mathbf{y}\|_\infty \leq C_{\mathcal{Y}}\right\}$, it holds that $\mathcal{W}_K$ has a covering number satisfied $\mathcal{N}\left(\delta, \mathcal{W}_K, \mathrm{d}_{\|\sup\|_\infty}\right) \leq C\delta^{-\dim(\mathcal{W}_K)}$, for some constant C.*

Indeed, Lemma 1.2.15 implies the desired result by noting that

$$\begin{aligned}\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\delta, \mathcal{W}_K\right) &= \ln \mathcal{N}\left(\delta, \mathcal{W}_K, d_{\|\sup\|_\infty}\right) \leq \ln\left[\frac{C}{\delta^{\dim(\mathcal{W}_K)}}\right] \\ &= \dim\left(\mathcal{W}_K\right)\left[\frac{1}{\dim\left(\mathcal{W}_K\right)}\ln C + \ln\left(\frac{1}{\delta}\right)\right] = \dim\left(\mathcal{W}_K\right)\left(C_{\mathcal{W}_K} + \ln\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

□

1.2.10.4 Proof of Lemma 1.2.14

Note that Genovese & Wasserman (2000, Lemma 2) provide a result for controlling a $\delta_{\boldsymbol{\pi}}$-Hellinger bracketing of $\mathbf{\Pi}_{K-1}$. However, such result can not be applied for our Lemma 1.2.14 since they use $\delta_{\boldsymbol{\pi}}$-Hellinger bracketing entropy while we use $\delta_{\boldsymbol{\pi}}$-covering number for the probability complex with maximum norm.

Given any $\boldsymbol{\pi} = (\pi_k)_{k\in[K]} \in \mathbf{\Pi}_{K-1}$, let $\boldsymbol{\xi} = (\boldsymbol{\xi}_k)_{k\in[K]}$ where $\boldsymbol{\xi}_k = \sqrt{\pi_k}, \forall k \in [K]$. Then $\boldsymbol{\pi} \in \mathbf{\Pi}_{K-1}$ if and only if $\boldsymbol{\xi} \in Q^+ \cap U$, where U is the surface of the unit sphere and Q^+ is the positive quadrant of

$\mathbb{R}^K$. Next, we divide the unit cube in $\mathbb{R}^K$ into disjoint cubes with sides parallel to the axes and sides of length $\delta_{\boldsymbol{\pi}}/\sqrt{K}$. Let $(\mathbf{C}_j)_{j\in[N]}$ is the subset of these cubes that have non-empty intersection with $Q^+ \cap U$. For any $j \in [N]$, let $\boldsymbol{\nu}_j = (\boldsymbol{\nu}_{j,k})_{k\in[K]}$ be the center of the cube $\mathbf{C}_j$ and $\boldsymbol{\nu}_j^2 = \left(\boldsymbol{\nu}_{j,k}^2\right)_{k\in[K]}$.

Then $\{\boldsymbol{\nu}_j\}_{j\in[N]}$ is a $\delta_{\boldsymbol{\pi}}/\left(2\sqrt{K}\right)$-covering of $Q^+\cap U$, since we have for any $\boldsymbol{\xi} = (\boldsymbol{\xi}_k)_{k\in[K]} \in Q^+\cap U$, there exists $j_0 \in [N]$ such that $\boldsymbol{\xi} \in \mathbf{C}_{j_0}$, and

$$\|\boldsymbol{\xi} - \boldsymbol{\nu}_{j_0}\|_\infty = \max_{k\in[K]} |\boldsymbol{\xi}_k - \boldsymbol{\nu}_{j_0,k}| \leq \frac{\delta_{\boldsymbol{\pi}}}{2\sqrt{K}}. \tag{1.2.54}$$

Therefore, it follows that $\boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}} := \left\{\boldsymbol{\nu}_j^2\right\}_{j\in[N]}$ is a $\delta_{\boldsymbol{\pi}}$-covering of $\boldsymbol{\Pi}_{K-1}$, since for any $\boldsymbol{\pi} = (\pi_k)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$, (1.2.54) leads to the existence of $j_0 \in [N]$, such that

$$\left\|\boldsymbol{\pi} - \boldsymbol{\nu}_{j_0}^2\right\|_\infty = \max_{k\in[K]} \left|\boldsymbol{\xi}_k^2 - \boldsymbol{\nu}_{j_0,k}^2\right| = \max_{k\in[K]} \left\{|\boldsymbol{\xi}_k - \boldsymbol{\nu}_{j_0,k}|\,|\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}|\right\} \leq \frac{\delta_{\boldsymbol{\pi}}}{2\sqrt{K}} \max_{k\in[K]} |\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}| \leq \frac{\delta_{\boldsymbol{\pi}}}{\sqrt{K}} \leq \delta_{\boldsymbol{\pi}},$$

where we used the fact that $\max_{k\in[K]} |\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}| \leq 2$. Now, it remains to count the number of cubes N. Let $\mathcal{T}_a = \{\boldsymbol{z} \in Q^+ : \|\boldsymbol{z}\|_2 \leq a\}$ and let $\mathcal{C} = \bigcup_{j\in[N]} \mathbf{C}_j$. Note that $\mathcal{C} \subset \mathcal{T}_{1+\delta_{\boldsymbol{\pi}}} - \mathcal{T}_{1-\delta_{\boldsymbol{\pi}}} \equiv \mathcal{T}$, and so

$$\text{Volume}(\mathcal{T}) \geq \text{Volume}(\mathcal{C}) = N\left(\frac{\delta_{\boldsymbol{\pi}}}{\sqrt{K}}\right)^K.$$

Note that here we use the notation π for the Archimedes' constant, which differs from $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$ for the mixing proportion of the GLoME model. Then, we define $\mathcal{V}_K(a) = a^K\pi^{K/2}$ as the volume of a sphere of radius a. Since $z! \geq z^z e^{-z}$ and $(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K = K\int_{1-\delta_{\boldsymbol{\pi}}}^{1+\delta_{\boldsymbol{\pi}}} z^{K-1}dz \leq 2\delta_{\boldsymbol{\pi}} K(1+\delta_{\boldsymbol{\pi}})^{K-1}$, it follows that

$$\begin{aligned}
\mathcal{N}(\delta_{\boldsymbol{\pi}}, \boldsymbol{\Pi}_{K-1}, \|\cdot\|_\infty) \leq N &\leq \frac{\text{Volume}(\mathcal{C})}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K} = \frac{1}{2^K}\frac{\mathcal{V}_K(1+\delta_{\boldsymbol{\pi}}) - \mathcal{V}_K(1-\delta_{\boldsymbol{\pi}})}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K} \\
&= \frac{1}{2^K}\frac{\left[(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K\right]}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K}\frac{\pi^{K/2}}{(K/2)!} \leq \left(\frac{\pi e}{2}\right)^{K/2}\frac{\left[(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K\right]}{\delta_{\boldsymbol{\pi}}^K} \\
&\leq \frac{K(2\pi e)^{K/2}}{\delta_{\boldsymbol{\pi}}^{K-1}}.
\end{aligned}$$

1.2.10.5 Proof of Lemma 1.2.15

In order to find an upper bound for a covering number of $\mathcal{W}_K$, we wish to construct a finite δ-covering $\mathbf{W}_{K,\boldsymbol{\omega}}$ of $\mathcal{W}_K$, with respect to the distance $\mathrm{d}_{\|\sup\|_\infty}$. That is, given any $\delta > 0, \mathbf{w}(\cdot;\boldsymbol{\omega}) \in \mathcal{W}_K$, we aim to prove that there exists $\mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}}) \in \mathbf{W}_{K,\boldsymbol{\omega}}$ such that

$$\mathrm{d}_{\|\sup\|_\infty}(\mathbf{w}(\cdot;\boldsymbol{\omega}), \mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}})) = \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} |\mathbf{w}_k(\mathbf{y};\boldsymbol{\omega}) - \mathbf{w}_k(\mathbf{y};\widehat{\boldsymbol{\omega}})| \leq \delta. \tag{1.2.55}$$

In order to accomplish such task, given any positive constants $\delta_{\boldsymbol{c}}, \delta_{\boldsymbol{\Gamma}}, \delta_{\boldsymbol{\pi}}$, and any $k \in [K]$, let us define

$$\begin{aligned}
\mathcal{F} &= \{\mathcal{Y} \ni \mathbf{y} \mapsto \ln(\phi_L(\mathbf{y};\boldsymbol{c},\boldsymbol{\Gamma})) : \|\boldsymbol{c}\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}) \leq M(\boldsymbol{\Gamma}) \leq A_{\boldsymbol{\Gamma}}\}, \\
\mathcal{F}_{\mathbf{c}_k} &= \Big\{\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) : \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}, \\
&\qquad \mathbf{c}_{k,j} \in \{-C_{\mathcal{Y}} + l\delta_{\boldsymbol{c}}/L : l = 0,\ldots,\lceil 2C_{\mathcal{Y}}L/\delta_{\boldsymbol{c}}\rceil\}, j \in [L]\Big\}, \qquad (1.2.56) \\
\mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k} &= \Bigg\{\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) : \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k}, \\
&\qquad [\mathrm{vec}(\boldsymbol{\Gamma}_k)]_{i,j} = \gamma_{i,j}\frac{\delta_{\boldsymbol{\Gamma}}}{L^2}; \gamma_{i,j} = \gamma_{j,i} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{L^2A_{\boldsymbol{\Gamma}}}{\delta_{\boldsymbol{\Gamma}}}\right\rfloor, \left\lfloor\frac{L^2A_{\boldsymbol{\Gamma}}}{\delta_{\boldsymbol{\Gamma}}}\right\rfloor\right], i \in [L], j \in [L]\Bigg\}, \qquad (1.2.57) \\
\mathbf{W}_{K,\boldsymbol{\omega}} &= \{\mathbf{w}(\cdot;\boldsymbol{\omega}) : \mathbf{w}(\cdot;\boldsymbol{\omega}) \in \mathcal{W}_K, \forall k \in [K], \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}, \boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}\}. \qquad (1.2.58)
\end{aligned}$$

Here, $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are ceiling and floor functions, respectively, and $\mathrm{vec}(\cdot)$ is an operator that stacks matrix columns into a column vector. In particular, we denote $\boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$ as a $\delta_{\boldsymbol{\pi}}$-covering of $\boldsymbol{\Pi}_{K-1}$, which is defined in Lemma 1.2.14. By the previous definition, it holds that $\forall k \in [K]$, $\mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k} \subset \mathcal{F}_{\mathbf{c}_k} \subset \mathcal{F}$, and $\mathbf{W}_{K,\boldsymbol{\omega}} \subset \mathcal{W}_K$.

Next, we claim that $\mathbf{W}_{K,\boldsymbol{\omega}}$ is a finite δ-covering of $\mathcal{W}_K$ with respect to the distance $\mathrm{d}_{\|\sup\|_\infty}$. To do this, for any $\mathbf{w}(\cdot;\boldsymbol{\omega}) = (\ln(\pi_k \phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)))_{k\in[K]} \in \mathcal{W}_K$, $\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}$, $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$, and for any $k \in [K]$, by (3.2.43), we first choose a function $\ln(\phi_L(\cdot;\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k}$ so that

$$\|\widehat{\boldsymbol{c}}_k - \mathbf{c}_k\|_1 = \sum_{j=1}^{L} |\widehat{\boldsymbol{c}}_{k,j} - \mathbf{c}_{k,j}| \leq L\frac{\delta_{\boldsymbol{c}}}{L} = \delta_{\boldsymbol{c}}.$$

Furthermore, by (1.2.57), we can obtain a result to construct the covariance matrix lattice. That is, any $\ln(\phi_L(\cdot;\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k}$ can be approximated by $\ln\left(\phi_L\left(\cdot;\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) \in \mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}$ such that

$$\left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) - \mathrm{vec}(\boldsymbol{\Gamma}_k)\right\|_1 \equiv \left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right\|_1 = \sum_{i=1}^{L}\sum_{j=1}^{L}\left|\left[\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right]_{i,j}\right| \leq \frac{L^2\delta_{\boldsymbol{\Gamma}}}{L^2} = \delta_{\boldsymbol{\Gamma}}.$$

Note that since for any $k \in [K]$, $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))$ is differentiable, it is also continuous w.r.t. $\mathbf{y}$ and its parameters $\mathbf{c}_k$ and $\boldsymbol{\Gamma}_k$. Thus, for every fixed $\mathbf{y} \in \mathcal{Y}$, for every $\widehat{\boldsymbol{c}}_k, \mathbf{c}_k \in \mathcal{X}$ with $\widehat{\boldsymbol{c}}_k \leq \mathbf{c}_k$, and for every $\widehat{\boldsymbol{\Gamma}}_k$, $\boldsymbol{\Gamma}_k$, where $\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) \leq \mathrm{vec}(\boldsymbol{\Gamma}_k)$, we can apply the mean value theorem (see Duistermaat & Kolk 2004, Lemma 2.5.1) to $\ln(\phi_L(\mathbf{y};\cdot,\boldsymbol{\Gamma}_k))$ and $\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\cdot))$ on the intervals $[\widehat{\boldsymbol{c}}_k,\mathbf{c}_k]$ and $\left[\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right),\mathrm{vec}(\boldsymbol{\Gamma}_k)\right]$ for some $z_{\mathbf{c}_k} \in (\widehat{\boldsymbol{c}}_k,\mathbf{c}_k)$ and $z_{\boldsymbol{\Gamma}_k} \in \left(\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right),\mathrm{vec}(\boldsymbol{\Gamma}_k)\right)$, respectively, to get

$$\begin{aligned}
\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)) &= (\widehat{\boldsymbol{c}}_k - \mathbf{c}_k)^\top \nabla_{\mathbf{c}_k} \ln(\phi_L(\mathbf{y};z_{\mathbf{c}_k},\boldsymbol{\Gamma}_k)),\\
\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) &= \left(\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) - \mathrm{vec}(\boldsymbol{\Gamma}_k)\right)^\top \nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)} \ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,z_{\boldsymbol{\Gamma}_k})).
\end{aligned}$$

Moreover, $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \nabla_{\mathbf{c}_k}\ln(\phi_L(\mathbf{y};z_{\mathbf{c}_k},\boldsymbol{\Gamma}_k))$ and $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)}\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,z_{\boldsymbol{\Gamma}_k}))$ are continuous functions on the compact set $\mathcal{U} := \mathcal{Y}\times\mathcal{Y}\times[a_{\boldsymbol{\Gamma}},A_{\boldsymbol{\Gamma}}]^{L^2}$ leads to they attain minimum and maximum values (see Duistermaat & Kolk 2004, Theorem 1.8.8). That is, we can set

$$\begin{aligned}
\mathbf{0} < (C_{\boldsymbol{c}})_{1,\ldots,L}^\top &:= \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k))\in\mathcal{U}} |\nabla_{\mathbf{c}_k}\ln|\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)|| < \infty,\\
\mathbf{0} < (C_{\boldsymbol{\Gamma}})_{1,\ldots,L^2}^\top &:= \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k))\in\mathcal{U}} \left|\nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)}\ln|\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)(x)|\right| < \infty.
\end{aligned}$$

Therefore, by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sup_{\mathbf{y}\in\mathcal{Y}} |\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))| &\leq |\widehat{\boldsymbol{c}}_k - \mathbf{c}_k|^\top (C_{\boldsymbol{c}})_{1,\ldots,L}^\top = C_{\boldsymbol{c}}\|\widehat{\boldsymbol{c}}_k - \mathbf{c}_k\|_1 \leq C_{\boldsymbol{c}}\delta_{\boldsymbol{\pi}},\\
\sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k))\right| &\leq C_{\boldsymbol{\Gamma}}\left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right\|_1 \leq C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}},
\end{aligned}$$

and by using the triangle inequality, it follows that

$$\max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))\right| \leq C_{\boldsymbol{c}}\delta_{\boldsymbol{c}} + C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}}. \tag{1.2.59}$$

Moreover, for every $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$, Lemma 1.2.14 implies that we can choose $\widehat{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$ so that $\max_{k\in[K]}|\pi_k - \widehat{\boldsymbol{\pi}}_k| \leq \delta_{\boldsymbol{\pi}}$. Notice that $[a_{\boldsymbol{\pi}},\infty) \ni t \mapsto \ln(t), a_{\boldsymbol{\pi}} > 0$ is a Lipschitz continuous function on $[a_{\boldsymbol{\pi}},\infty]$. Indeed, by the mean value theorem, it holds that there exists $c \in (t_1,t_2)$, such that

$$|\ln(t_1) - \ln(t_2)| = \ln'(c)|t_1 - t_2| \leq \frac{1}{a_{\boldsymbol{\pi}}}|t_1 - t_2|, \text{ for all } t_1,t_2 \in [a_{\boldsymbol{\pi}},\infty). \tag{1.2.60}$$

Therefore, (1.2.55) can be obtained by the following evaluation

$$\begin{aligned}
\mathrm{d}_{\|\sup\|_\infty}\left(\mathbf{w}(\cdot;\boldsymbol{\omega}),\mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}})\right) &= \max_{k\in[K]}\sup_{\mathbf{y}\in\mathcal{Y}}\left|\ln\left(\pi_k\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)\right)-\ln\left(\widehat{\boldsymbol{\pi}}_k\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&= \max_{k\in[K]}\sup_{\mathbf{y}\in\mathcal{Y}}\left|\ln\left(\pi_k\right)-\ln\left(\widehat{\boldsymbol{\pi}}_k\right)+\ln\left(\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)\right)-\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&\leq \max_{k\in[K]}\left|\ln\left(\pi_k\right)-\ln\left(\widehat{\boldsymbol{\pi}}_k\right)\right|+\max_{k\in[K]}\sup_{\mathbf{y}\in\mathcal{Y}}\left|\ln\left(\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)\right)-\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&\leq \frac{1}{a_{\boldsymbol{\pi}}}\max_{k\in[K]}\left|\pi_k-\widehat{\boldsymbol{\pi}}_k\right|+C_{\boldsymbol{c}}\delta_{\boldsymbol{c}}+C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}}\,(\text{using (3.2.47) and (1.2.59)}) \\
&\leq \frac{\delta_{\boldsymbol{\pi}}}{a_{\boldsymbol{\pi}}}+C_{\boldsymbol{c}}\delta_{\boldsymbol{c}}+C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}}\,(\text{using Lemma 1.2.14})\leq\frac{\delta}{3}+\frac{\delta}{3}+\frac{\delta}{3}=\delta,
\end{aligned}$$

where we choose $\delta_{\boldsymbol{\pi}}=\frac{\delta a_{\boldsymbol{\pi}}}{3},\delta_{\boldsymbol{c}}=\frac{\delta}{3C_\mu},\delta_{\boldsymbol{\Gamma}}=\frac{\delta}{3C_{\boldsymbol{\Gamma}}}$. Finally, we get the covering number

$$\begin{aligned}
\mathcal{N}\left(\delta,\mathcal{W}_K,\mathrm{d}_{\|\sup\|_\infty}\right) &\leq \mathrm{card}\left(\mathbf{W}_{K,\boldsymbol{\omega}}\right)=\left[\frac{2C_{\mathcal{Y}}L}{\delta_{\boldsymbol{c}}}\right]^{KL}\left[\frac{2A_{\boldsymbol{\Gamma}}L^2}{\delta_{\boldsymbol{\Gamma}}}\right]^{\frac{L(L+1)}{2}K}\mathcal{N}\left(\delta_{\boldsymbol{\pi}},\boldsymbol{\Pi}_{K-1},\|\cdot\|_\infty\right) \\
&= \frac{C}{\delta^{KL+K\frac{L(L+1)}{2}+K-1}}=\frac{C}{\delta^{\dim(\mathcal{W}_K)}},
\end{aligned}$$

where

$$C=\left(6C_{\boldsymbol{c}}C_{\mathcal{Y}}L\right)^{KL}\left(6C_{\boldsymbol{\Gamma}}A_{\boldsymbol{\Gamma}}L^2\right)^{\frac{L(L+1)}{2}K}\left(\frac{3}{a_{\boldsymbol{\pi}}}\right)^{K-1}K\left(2\pi e\right)^{K/2}.$$

1.2.11 Our contributions for weak oracle inequalities in random subcollection of MoE models via Theorems 1.2.3 and 1.2.5

The deterministic collection of MoE models include BLLiM, BLoME, PSGaBloME and LinBoSGaBloME models where weak oracle inequalities were only well studied for finite mixture of Gaussian regression models via a model selection theorem for MLE among a random subcollection of models in regression framework of Devijver (2015b, Theorem 5.1), see also Devijver & Gallopin (2018, Theorem 7.3). This is an extension of a whole collection of conditional densities from Cohen & Le Pennec (2011, Theorem 2), and of Massart (2007, Theorem 7.11), working only for density estimation. In Section 1.2.11.1, we first summarize this theorem and the techniques that Devijver (2017a) used to control the bracketing entropy of finite mixture of Gaussian regression models with joint rank and variable selection for parsimonious estimation in a high-dimensional framework. Then, we explain why such techniques can not be directly applied to our collection of BLLiM, BLoME, PSGaBloME and LinBoSGaBloME models to highlight the main challenges and our contributions.

1.2.11.1 A model selection theorem for MLE among a random subcollection

We can now state the main result of (Devijver, 2015b, Theorem 5.1) for the model selection theorem for MLE among a random subcollection.

Theorem 1.2.16 (Theorem 5.1 from Devijver (2015b))**.** *Let $\left(\mathbf{x}_{[n]},\mathbf{y}_{[n]}\right)$ be observations coming from an unknown conditional density s_0. Let the model collection $\mathcal{S}=\left(S_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ be an at most countable collection of conditional density sets. Assume that Assumption 1.2.1 (K), Assumption 1.2.2 (Sep), and Assumption 1.2.3 (H) hold for every $\mathbf{m}\in\mathcal{M}$. Let $\epsilon_{KL}>0$, and $\bar{s}_{\mathbf{m}}\in S_{\mathbf{m}}$, such that*

$$\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,\bar{s}_{\mathbf{m}}\right)\leq\inf_{t\in S_{\mathbf{m}}}\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,t\right)+\frac{\epsilon_{KL}}{n};$$

and let $\tau>0$, such that

$$\bar{s}_{\mathbf{m}}\geq e^{-\tau}s_0. \tag{1.2.61}$$

Introduce $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$*, a random subcollection of* $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$*. Consider the collection* $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *of* η*-log likelihood minimizer satisfying* (1.2.6) *for all* $\mathbf{m}\in\widetilde{\mathcal{M}}$*. Then, for any* $\rho\in(0,1)$*, and any* $C_1>1$*, there are two constants* κ_0 *and* C_2 *depending only on* ρ *and* C_1*, such that, for every index* $\mathbf{m}\in\mathcal{M}$*,*

$$\text{pen}(\mathbf{m})\geq\kappa\left(\mathcal{D}_{\mathbf{m}}+(1\vee\tau)\xi_{\mathbf{m}}\right),$$

with $\kappa>\kappa_0$*, and where the model complexity* $\mathcal{D}_{\mathbf{m}}$ *is defined in Assumption 1.2.3 (H), the* η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined as in* (1.2.5) *on the subset* $\widetilde{\mathcal{M}}\subset\mathcal{M}$*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\text{JKL}_{\rho}^{\otimes\text{n}}\left(s_0,\widehat{s}_{\widehat{\mathbf{m}}}\right)\right]\leq C_1\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}}\text{KL}^{\otimes\text{n}}\left(s_0,t\right)+2\frac{\text{pen}(\mathbf{m})}{n}\right)\right]+C_2(1\vee\tau)\frac{\Xi^2}{n}+\frac{\eta'+\eta}{n}.$$

1.2.11.2 Sketches of the proofs for our BLoME and PSGaBloME models

To work with conditional density estimation in the BLoME and PSGaBloME regression models, it is natural to make use of Theorem 1.2.9. However, it is worth mentioning that, because the model collection constructed by the BLLiM Devijver et al. (2017) and our Lasso+l_2-MLE and Lasso+l_2-Rank procedures, see Section 4.3.5 for more details, are both random, we have to use a model selection theorem for MLE among a random subcollection (cf. Devijver, 2015b, Theorem 5.1 and Devijver & Gallopin, 2018, Theorem 7.3). This is the extension of Cohen & Le Pennec (2011, Theorem 2), which dealt with conditional density estimation but not with random subcollection, and of Massart (2007, Theorem 7.11), working only for density estimation.

Then, we explain how we use Theorem 1.2.16 to get the oracle inequalities, Theorems 1.2.3 and 1.2.5. To this end, our model collections of BLoME and PSGaBloME models have to satisfy some regularity assumptions, which are briefly and fully proved in Sections 1.2.11.3, 1.2.11.8, 3.3.3 and 4.3.3, respectively. For BLoME models, the main difficulty in proving our oracle inequality lies in bounding the bracketing entropy of the Gaussian gating functions and Gaussian experts with block-diagonal covariance matrices. To overcome the former issue, we follow a reparameterization trick of the Gaussian gating parameters space (Nguyen et al., 2021c). For the second one, we utilize the recent novel result on block-diagonal covariance matrices in Devijver & Gallopin (2018). While to work with PSGaBloME models, we have to control the bracketing entropy of the weights and means restricted on relevant variables as well as rank sparse models, and in particular with block-diagonal covariance matrices for PSGaBloME model. To overcome the former issue, we need to extend the strategies from Montuelle et al. (2014), Devijver (2017a). For the second one, we need to extend the result on block-diagonal covariance matrices from Gaussian graphical models in Devijver & Gallopin (2018) to standard MoE regression models, based on some ideas of Gaussian mixture models from Genovese & Wasserman (2000), Maugis & Michel (2011b).

1.2.11.3 Our contributions on BLoME models

It should be stressed that all we need is to verify that Assumption 1.2.3 (H), Assumption 1.2.2 (Sep) and Assumption 1.2.1 (K) hold for every $\mathbf{m}\in\mathcal{M}$. According to the result from Devijver (2015b, Section 5.3), Assumption 1.2.2 (Sep) holds when we consider Gaussian densities and the assumption defined by (1.2.61) is true if we assume further that the true conditional density s_0 is bounded and compactly supported. Furthermore, since we restricted d and K to $\mathcal{D}_{\boldsymbol{\Upsilon}}=[d_{\max}]$ and $\mathcal{K}=[K_{\max}]$, respectively, it is true that there exists a family $(\xi_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ and $\Xi>0$ such that, Assumption 1.2.1 (K) is satisfied. Therefore, the proof for the remaining Assumption 1.2.3 (H) is our contribution. Next, we explain that to the best of our knowledge, there are no results that can be directly applied to Assumption 1.2.3 (H) for BLoME models due to their complexity with the Gaussian gating functions and Gaussian experts with block-diagonal covariance matrices. This highlights our contributions with this challenge problem compared to the works of Genovese & Wasserman (2000), Maugis & Michel (2011b), Devijver (2015b, 2017a), Devijver & Gallopin (2018), Montuelle et al. (2014).

By using Lemma 1.2.11, Assumption 1.2.3 (H) holds true if we can prove that for any $\delta\in(0,\sqrt{2}]$,

the collection of LinBoSGaME models, $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, satisfies

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{1.2.62}$$

Proof of 1.2.62. Note that (1.2.62) can be established by first decomposing the entropy term between the Gaussian gating functions and the Gaussian experts. Motivated by our *reparameterization trick* of the Gaussian gating space in Section 1.2.10.3, we define for $\mathcal{P}_K$ via $\mathcal{W}_k$ and Gaussian experts $\mathcal{G}_{K,d,\mathbf{B}}$ as follows.

$$\begin{aligned}
\mathcal{W}_K &= \left\{\mathcal{Y} \ni \mathbf{y} \mapsto \left(\ln\left(\pi_k \phi\left(\mathbf{y}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right)\right)_{k\in[K]} =: \left(\mathbf{w}_k(\mathbf{y}; \boldsymbol{\omega})\right)_{k\in[K]} = \mathbf{w}\left(\mathbf{y}; \boldsymbol{\omega}\right) : \boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}_K\right\}, \\
\mathcal{P}_K &= \left\{\mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^K e^{\mathbf{w}_l(\mathbf{y})}}\right)_{k\in[K]} =: \left(g_k\left(\mathbf{y}; \mathbf{w}\right)\right)_{k\in[K]}, \mathbf{w} \in \mathcal{W}_K\right\}, \text{ and} \\
\mathcal{G}_{K,d,\mathbf{B}} &= \left\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x}, \mathbf{y}) \mapsto \left(\phi\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right)\right)_{k\in[K]} : \boldsymbol{\upsilon}_d \in \mathbf{\Upsilon}_{K,d}, \mathbf{\Sigma}(\mathbf{B}) \in \mathbf{V}_K(\mathbf{B})\right\}.
\end{aligned}$$

There are two possible ways to decompose the bracketing entropy of $S_{\mathbf{m}}$ based on different distances. For the first approach, we can use Lemma 1.2.17 (Montuelle et al., 2014, Lemma 5):

Lemma 1.2.17. *For all $\delta \in (0, \sqrt{2}]$ and $\mathbf{m} \in \mathcal{M}$,*

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} \max_k d_{\mathbf{x}}}\left(\frac{\delta}{5}, \mathcal{G}_{K,d,\mathbf{B}}\right).$$

As mentioning in Appendix B.2.1 from Montuelle et al. (2014), Lemma 1.2.17 boils down to assuming that $\mathbf{Y}$ is bounded. Furthermore, they also claim that this boundedness assumption can be relaxed when using smaller distance $\mathrm{d}^{\otimes \mathrm{n}}$ but bounding the corresponding bracketing entropy becomes much more challenging. We successful weaken such boundedness assumption via utilizing the smaller distance: $\mathrm{d}^{\otimes \mathrm{n}}$, for the bracketing entropy of $S_{\mathbf{m}}$ although bounding such bracketing entropy for $\mathcal{W}_K$ and $\mathcal{G}_{K,\mathcal{B}}$ becomes much more challenging. This is one of our contributions regarding the controlling bracketing entropy of BLoME models. Consequently, this leads to the second approach via Lemma 1.2.18 (Montuelle et al., 2014, Lemma 6).

Lemma 1.2.18. *For all $\delta \in (0, \sqrt{2}]$,*

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],d_{\mathcal{P}_K}}\left(\frac{\delta}{2}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right),$$

where

$$\begin{aligned}
d^2_{\mathcal{P}_K}\left(g^+, g^-\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n d_k^2\left(g^+\left(\mathbf{Y}_i\right), g^-(\mathbf{Y}_i)\right)\right] = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \left(\sqrt{g_k^+\left(\mathbf{Y}_i\right)} - \sqrt{g_k^-\left(\mathbf{Y}_i\right)}\right)^2\right], \\
d^2_{\mathcal{G}_{K,d,\mathbf{B}}}\left(\phi^+, \phi^-\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K d_{\mathbf{x}}^2\left(\phi_k^+\left(\cdot, \mathbf{Y}_i\right), \phi_k^-\left(\cdot, \mathbf{Y}_i\right)\right)\right] \\
&= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \int_{\mathcal{X}} \left(\sqrt{\phi_k^+\left(\mathbf{x}, \mathbf{Y}_i\right)} - \sqrt{\phi_k^+\left(\mathbf{x}, \mathbf{Y}_i\right)}\right)^2 d\mathbf{x}\right].
\end{aligned}$$

Next, we make use of Lemma 1.2.19, which is proved in Section 1.2.11.4, to provide an upper bound on the bracketing entropy of $S_{\mathbf{m}}$ and $\mathcal{P}_K$ on the corresponding distances $\mathrm{d}^{\otimes \mathrm{n}}$ and $d_{\mathcal{P}_K}$, respectively.

Lemma 1.2.19. *It holds that*

$$\mathrm{d}^{\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t), \text{ and } \mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}(\delta, S_{\mathbf{m}}), \tag{1.2.63}$$

$$d_{\mathcal{P}_K}\left(g^+, g^-\right) \leq \sup_{\mathbf{y}} d_k(g^+, g^-), \text{ and } \mathcal{H}_{[\cdot],d_{\mathcal{P}_K}}\left(\frac{\delta}{2}, \mathcal{P}_K\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right). \tag{1.2.64}$$

Lemmas 1.2.18 and 1.2.19 imply that

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right).$$

Based on this metric, one can first relate the bracketing entropy of $\mathcal{P}_K$ to $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathcal{W}_K)$, and then obtain the upper bound for its entropy via Lemma 1.2.13.

Then, we present our main contribution for BLoME models via Lemma 1.2.20. This lemma allows us to construct the Gaussian brackets to handle the metric entropy for Gaussian experts, which is established in Section 1.2.11.5.

Lemma 1.2.20.

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \leq \dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{1.2.65}$$

Finally, (3.3.8) can easily proved via Lemmas 1.2.13 and 1.2.20. □

1.2.11.4 Proof of Lemma 1.2.19

We first aim to prove that $\mathrm{d}^{2\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d^2_{\mathbf{x}}(s,t)$. Indeed, by definition, it follows that

$$\begin{aligned}
\mathrm{d}^{2\otimes \mathrm{n}}\left(s,t\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{y}\right), t\left(\cdot|\mathbf{y}\right)\right) s_{\mathbf{x},0}(\mathbf{y})d\mathbf{y} \leq \sup_{\mathbf{y}} d^2_{\mathbf{x}}\left(s,t\right)\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} s_{\mathbf{x},0}(\mathbf{y})d\mathbf{y} = \sup_{\mathbf{y}} d^2_{\mathbf{x}}\left(s,t\right),
\end{aligned}$$

where $s_{\mathbf{x},0}$ denotes that marginal PDF of s_0, w.r.t. $\mathbf{x}$. Consequently, it holds that $\mathrm{d}^{\otimes \mathrm{n}}(s,t) = \sqrt{\mathrm{d}^{2\otimes \mathrm{n}}(s,t)} \leq \sqrt{\sup_{\mathbf{y}} d^2_{\mathbf{x}}(s,t)} = \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)$. To prove that

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right),$$

it is sufficient to check that

$$\mathcal{N}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{N}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right).$$

By using the definition of bracketing entropy in (1.2.46) and $\mathrm{d}^{\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)$, given

$$\begin{aligned}
A &= \left\{n \in \mathbb{N}^\star : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t } \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)\left(t_k^-, t_k^+\right) \leq \delta, S_{\mathbf{m}} \subset \bigcup_{k=1}^{n}\left[t_k^-, t_k^+\right]\right\}, \\
B &= \left\{n \in \mathbb{N}^\star : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t } \mathrm{d}^{\otimes \mathrm{n}}\left(t_k^-, t_k^+\right) \leq \delta, S_{\mathbf{m}} \subset \bigcup_{k=1}^{n}\left[t_k^-, t_k^+\right]\right\},
\end{aligned}$$

it leads to that $A \subset B$ and then (1.2.63) follows, since

$$\mathcal{N}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)}\left(\delta, S_{\mathbf{m}}\right) = \min A \geq \min B = \mathcal{N}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right).$$

With the similar argument as in the proof of (1.2.63), it holds that $d_{\mathcal{P}_K}\left(g^+, g^-\right) \leq \sup_{\mathbf{y}} d_k(g^+, g^-)$ and (1.2.64) is proved.

1.2.11.5 Proof of Lemma 1.2.20

It is worth mentioning that without any structures on covariance matrices of Gaussian experts from the collection $\mathcal{M}$, Lemma 1.2.20 can be proved using Proposition 2 from Montuelle et al. (2014) and Montuelle et al. (2014, Appendix B.2.3), for constructing of Gaussian brackets to deal with the Gaussian experts. However, dealing with block-diagonal covariance matrices with random subcollection is

much more challenging. We have to establish more constructive bracketing entropies in the spirits of Maugis & Michel (2011b), Devijver (2015b), Devijver & Gallopin (2018).

Given any $k \in [K]$, by defining

$$\mathcal{G}_{d,\mathbf{B}_k} = \{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, \mathbf{y}) \mapsto \phi(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)) =: \phi_k : \boldsymbol{\upsilon}_{k,d} \in \boldsymbol{\Upsilon}_{k,d}, \boldsymbol{\Sigma}_k(\mathbf{B}_k) \in \mathbf{V}_k(\mathbf{B}_k)\}, \tag{1.2.66}$$

it follows that $\mathcal{G}_{K,d,\mathbf{B}} = \prod_{k=1}^K \mathcal{G}_{d,\mathbf{B}_k}$, where $\prod$ stands for the cartesian product. By using Lemma 1.2.21, which is proved in Section 1.2.11.6, it follows that

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \leq \sum_{k=1}^K \mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}}, \mathcal{G}_{d,\mathbf{B}_k}\right). \tag{1.2.67}$$

Lemma 1.2.21. *Given $\mathcal{G}_{K,d,\mathbf{B}} = \prod_{k=1}^K \mathcal{G}_{d,\mathbf{B}_k}$, where $\mathcal{G}_{d,\mathbf{B}_k}$ is defined in (1.2.66), it holds that*

$$\mathcal{N}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \leq \prod_{k=1}^K \mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}}, \mathcal{G}_{d,\mathbf{B}_k}\right),$$

where for any $\phi^+, \phi^- \in \mathcal{G}_{K,d,\mathbf{B}}$ and any $\phi_k^+, \phi_k^- \in \mathcal{G}_{d,\mathbf{B}_k}, k \in [K]$,

$$d^2_{\mathcal{G}_{K,d,\mathbf{B}}}(\phi^+, \phi^-) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \mathrm{d}^2\left(\phi_k^+(\cdot, \mathbf{Y}_i), \phi_k^-(\cdot, \mathbf{Y}_i)\right)\right],$$
$$d^2_{\mathcal{G}_{d,\mathbf{B}_k}}(\phi_k^+, \phi_k^-) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \mathrm{d}^2\left(\phi_k^+(\cdot, \mathbf{Y}_i), \phi_k^-(\cdot, \mathbf{Y}_i)\right)\right].$$

Lemma 1.2.20 is proved via (1.2.67) and Lemma 1.2.22, which is proved in Section 1.2.11.7.

Lemma 1.2.22. *By defining $\mathcal{G}_{d,\mathbf{B}_k}$ as in (1.2.66), for all $\delta \in (0, \sqrt{2}]$, it holds that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2}, \mathcal{G}_{d,\mathbf{B}_k}\right) \leq \dim(\mathcal{G}_{d,\mathbf{B}_k})\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ where} \tag{1.2.68}$$
$$D_{\mathbf{B}_k} = \sum_{g=1}^{G_k} \frac{\mathrm{card}\left(d_k^{[g]}\right)\left(\mathrm{card}\left(d_k^{[g]}\right) - 1\right)}{2},$$
$$C_{\mathcal{G}_{d,\mathbf{B}_k}} = \frac{D_{\mathbf{B}_k} \ln\left(\frac{6\sqrt{6}\lambda_M D^2(D-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim(\boldsymbol{\Upsilon}_{k,d}) \ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)}{\dim(\mathcal{G}_{d,\mathbf{B}_k})}.$$

Indeed, (1.2.67) and (1.2.68) lead to

$$\begin{aligned}
\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) &\leq \sum_{k=1}^K \mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}}, \mathcal{G}_{d,\mathbf{B}_k}\right) \\
&\leq \sum_{k=1}^K \dim(\mathcal{G}_{d,\mathbf{B}_k})\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\sqrt{K}\right) + \ln\left(\frac{1}{\delta}\right)\right) \\
&\leq \dim(\mathcal{G}_{K,d,\mathbf{B}})\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

Here, $C_{\mathcal{G}_{K,d,\mathbf{B}}} = \sum_{k=1}^K C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\sqrt{K}\right)$ and note that $\dim(\mathcal{G}_{K,d,\mathbf{B}}) = \sum_{k=1}^K \dim(\mathcal{G}_{d,\mathbf{B}_k})$, $\dim(\mathcal{G}_{d,\mathbf{B}_k}) = D_{\mathbf{B}_k} + \dim(\boldsymbol{\Upsilon}_{k,d})$, $\dim(\boldsymbol{\Upsilon}_{k,d}) = Dd_{\boldsymbol{\Upsilon}_{k,d}}, C_{\boldsymbol{\Upsilon}_{k,d}} = \sqrt{D}d_{\boldsymbol{\Upsilon}_{k,d}}T_{\boldsymbol{\Upsilon}_{k,d}}$ (in cases where linear combination of bounded functions are used for means, *i.e.*, $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_b$) or $\dim(\boldsymbol{\Upsilon}_{k,d}) = D\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}$, $C_{\boldsymbol{\Upsilon}_{k,d}} = \sqrt{D}\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}T_{\boldsymbol{\Upsilon}_{k,d}}$ (in cases where we use polynomial means, *i.e.*, $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_p$).

1.2.11.6 Proof of Lemma 1.2.21

By the definition of the bracketing entropy in (1.2.46), for each $k \in [K]$, let $\left\{\left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ be a minimal covering of δ_k brackets for $d_{\mathcal{G}_{d,\mathbf{B}_k}}$ of $\mathcal{G}_{d,\mathbf{B}_k}$, with cardinality $\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}$. This leads to

$$\forall l \in \left[\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}\right], d_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l,-}, \phi_k^{l,+}\right) \leq \delta_k.$$

Therefore, we claim that the set $\left\{\prod_{k=1}^K \left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ is a covering of $\frac{\delta}{2}$-bracket for $d_{\mathcal{G}_{K,d,\mathbf{B}}}$ of $\mathcal{G}_{K,d,\mathbf{B}}$ with cardinality $\prod_{k=1}^K \mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}(\delta_k, \mathcal{G}_{d,\mathbf{B}_k})$. Indeed, let any $\phi = (\phi_k)_{k\in[K]} \in \mathcal{G}_{K,d,\mathbf{B}}$. Consequently, for each $k \in [K], \phi_k \in \mathcal{G}_{d,\mathbf{B}_k}$, there exists $l(k) \in \left[\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}\right]$, such that

$$\phi_k^{l(k),-} \leq \phi_k \leq \phi_k^{l(k),+}, d^2_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l(k),+}, \phi_k^{l(k),-}\right) \leq (\delta_k)^2.$$

Then, it follows that $\phi \in [\phi^-, \phi^+] \in \left\{\prod_{k=1}^K \left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$, with $\phi^- = \left(\phi_k^{l(k),-}\right)_{k\in[K]}, \phi^+ = \left(\phi_k^{l(k),+}\right)_{k\in[K]}$, which implies that $\left\{\prod_{k=1}^K \left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ is a bracket covering of $\mathcal{G}_{K,d,\mathbf{B}}$.

Now, we want to verify that the size of this bracket is $\delta/2$ by choosing $\delta_k = \frac{\delta}{2\sqrt{K}}, \forall k \in [K]$. It follows that

$$\begin{aligned} d^2_{\mathcal{G}_{K,d,\mathbf{B}}}\left(\phi^-, \phi^+\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \mathrm{d}^2\left(\phi_k^{l(k),-}(\cdot, \mathbf{Y}_i), \phi_k^{l(k),+}(\cdot, \mathbf{Y}_i)\right)\right] \\ &= \sum_{k=1}^K \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \mathrm{d}^2\left(\phi_k^{l(k),-}(\cdot, \mathbf{Y}_i), \phi_k^{l(k),+}(\cdot, \mathbf{Y}_i)\right)\right] \\ &= \sum_{k=1}^K d^2_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l(k),-}, \phi_k^{l(k),+}\right) \leq K\left(\frac{\delta}{2\sqrt{K}}\right)^2 = \left(\frac{\delta}{2}\right)^2. \end{aligned}$$

To this end, by definition of a minimal $\frac{\delta}{2}$-bracket covering number for $\mathcal{G}_{K,d,\mathbf{B}}$, Lemma 1.2.21 is proved.

1.2.11.7 Proof of Lemma 1.2.22

To provide the upper bound of the bracketing entropy in (1.2.68), our technique is adapted from the work of Genovese & Wasserman (2000) for unidimensional Gaussian mixture families, which is recently generalized to multidimensional case by Maugis & Michel (2011b) for Gaussian mixture models. Furthermore, we make use of the results from Devijver & Gallopin (2018) to deal with block-diagonal covariance matrices, $\mathbf{V}_k(\mathbf{B}_k), k \in [K]$, and from Montuelle et al. (2014) to handle the means of Gaussian experts $\boldsymbol{\Upsilon}_{k,d}, k \in [K]$.

The main idea in our approach is to define firstly a net over the parameter spaces of Gaussian experts, $\boldsymbol{\Upsilon}_{k,d} \times \mathbf{V}_k(\mathbf{B}_k), k \in [K]$, and to construct a bracket covering of $\mathcal{G}_{d,\mathbf{B}_k}$ according to the tensorized Hellinger distance. Note that $\dim(\mathcal{G}_{d,\mathbf{B}_k}) = \dim(\boldsymbol{\Upsilon}_{k,d}) + \dim(\mathbf{V}_k(\mathbf{B}_k))$.

Step 1: Construction of a net for the block-diagonal covariance matrices. Firstly, for $k \in [K]$, we denote by $\mathrm{Adj}(\boldsymbol{\Sigma}_k(\mathbf{B}_k))$ the adjacency matrix associated to the covariance matrix $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$. Note that this matrix of size D^2 can be defined by a vector of concatenated upper triangular vectors. We are going to make use of the result from Devijver & Gallopin (2018) to handle the block-diagonal covariance matrices $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$, via its corresponding adjacency matrix. To do this, we need to construct a discrete space for $\{0,1\}^{D(D-1)/2}$, which is a one-to-one correspondence (bijection) with

$$\mathcal{A}_{\mathbf{B}_k} = \left\{\mathbf{A}_{\mathbf{B}_k} \in \mathcal{S}_D(\{0,1\}) : \exists \boldsymbol{\Sigma}_k(\mathbf{B}_k) \in \mathbf{V}_k(\mathbf{B}_k) \text{ s.t } \mathrm{Adj}(\boldsymbol{\Sigma}_k(\mathbf{B}_k)) = \mathbf{A}_{\mathbf{B}_k}\right\},$$

where $\mathcal{S}_D(\{0,1\})$ is the set of symmetric matrices of size D taking values on $\{0,1\}$.

Then, we want to deduce a discretization of the set of covariance matrices. Let h denotes Hamming distance on $\{0,1\}^{D(D-1)/2}$ defined by

$$d(z,z') = \sum_{i=1}^{n} \mathbb{I}\left\{z \neq z'\right\}, \text{ for all } z, z' \in \{0,1\}^{D(D-1)/2}.$$

Let $\{0,1\}^{D(D-1)/2}_{\mathbf{B}_k}$ be the subset of $\{0,1\}^{D(D-1)/2}$ of vectors for which the corresponding graph has structure $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$. Corollary 1 and Proposition 2 from Supplementary Material A of Devijver & Gallopin (2018) imply that there exists some subset $\mathcal{R}$ of $\{0,1\}^{D(D-1)/2}$, as well as its equivalent $\mathcal{A}^{\text{disc}}_{\mathbf{B}_k}$ for adjacency matrices such that, given $\epsilon > 0$, and

$$\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon) = \left\{ \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathcal{S}_D^{++}(\mathbb{R}) : \text{Adj}\left(\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right) \in \mathcal{A}^{\text{disc}}_{\mathbf{B}_k}, \left[\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right]_{i,j} = \sigma_{i,j}\epsilon, \sigma_{i,j} \in \left[\frac{-\lambda_M}{\epsilon}, \frac{\lambda_M}{\epsilon}\right] \bigcap \mathbb{Z} \right\},$$

it holds that

$$\begin{aligned}
\left\| \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) - \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right) \right\|_2^2 \leq \frac{D_{\mathbf{B}_k}}{2} \wedge \epsilon^2, \forall \left(\mathbf{\Sigma}_k\left(\mathbf{B}_k\right), \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right) \in \left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right)^2 \; s.t. \; \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \neq \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right), \\
\text{card}\left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right) \leq \left(\left\lfloor \frac{2\lambda_M}{\epsilon} \right\rfloor \frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}} \right)^{D_{\mathbf{B}_k}}, \qquad (1.2.69) \\
D_{\mathbf{B}_k} = \dim\left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right) = \sum_{g=1}^{G_k} \frac{\text{card}\left(d_k^{[g]}\right)\left(\text{card}\left(d_k^{[g]}\right) - 1\right)}{2}. \qquad (1.2.70)
\end{aligned}$$

By choosing $\epsilon^2 \leq \frac{D_{\mathbf{B}_k}}{2}$, given $\mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k\left(\mathbf{B}_k\right)$, then there exists $\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right) \in \tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)$, such that

$$\left\| \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) - \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right) \right\|_2^2 \leq \epsilon^2. \qquad (1.2.71)$$

Step 2: Construction of a net for the mean functions. Based on $\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)$, we can construct the following bracket covering of $\mathcal{G}_{d,\mathbf{B}_k}$ by defining the nets for the means of Gaussian experts. The proof of Lemma 1, page 1693, from Montuelle et al. (2014) implies that

$$\mathcal{N}_{[\cdot],\sup_{\mathbf{y}}\|\cdot\|_2}\left(\delta_{\mathbf{\Upsilon}_{k,d}}, \mathbf{\Upsilon}_{k,d}\right) \leq \left(\frac{\exp\left(C_{\mathbf{\Upsilon}_{k,d}}\right)}{\delta_{\mathbf{\Upsilon}_{k,d}}}\right)^{\dim\left(\mathbf{\Upsilon}_{k,d}\right)}.$$

Here $\dim\left(\mathbf{\Upsilon}_{k,d}\right) = Dd_{\mathbf{\Upsilon}_{k,d}}$, and $C_{\mathbf{\Upsilon}_{k,d}} = \sqrt{D}d_{\mathbf{\Upsilon}_{k,d}}T_{\mathbf{\Upsilon}_{k,d}}$ in the general case or $\dim\left(\mathbf{\Upsilon}_{k,d}\right) = D\binom{d_{\mathbf{\Upsilon}_{k,d}}+L}{L}$, and $C_{\mathbf{\Upsilon}_{k,d}} = \sqrt{D}\binom{d_{\mathbf{\Upsilon}_{k,d}}+L}{L}T_{\mathbf{\Upsilon}_{k,d}}$ in the special case of polynomial means. Then, by the definition of bracketing entropy in (1.2.46), for any minimal $\delta_{\mathbf{\Upsilon}_{k,d}}$-bracketing covering of the means from Gaussian experts, denoted by $G_{\mathbf{\Upsilon}_{k,d}}\left(\delta_{\mathbf{\Upsilon}_{k,d}}\right)$, it is true that

$$\text{card}\left(G_{\mathbf{\Upsilon}_{k,d}}\left(\delta_{\mathbf{\Upsilon}_{k,d}}\right)\right) \leq \left(\frac{\exp\left(C_{\mathbf{\Upsilon}_{k,d}}\right)}{\delta_{\mathbf{\Upsilon}_{k,d}}}\right)^{\dim\left(\mathbf{\Upsilon}_{k,d}\right)}. \qquad (1.2.72)$$

Therefore, given $\alpha > 0$, which is specified later, we claim that the set

$$\left\{ [l,u] \;\middle|\; \begin{array}{l} l(\mathbf{x},\mathbf{y}) = (1+2\alpha)^{-D}\,\phi\left(\mathbf{x}; \widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}), (1+\alpha)^{-1}\,\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right), \\ u(\mathbf{x},\mathbf{y}) = (1+2\alpha)^{D}\,\phi\left(\mathbf{x}; \widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}), (1+\alpha)\,\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right), \\ \widetilde{\boldsymbol{\upsilon}}_{k,d} \in G_{\mathbf{\Upsilon}_{k,d}}\left(\delta_{\mathbf{\Upsilon}_{k,d}}\right), \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right) \in \tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon) \end{array} \right\},$$

is a $\delta_{\boldsymbol{\Upsilon}_{k,d}}$-brackets set over $\mathcal{G}_{d,\mathbf{B}_k}$. Indeed, let $\mathcal{X}\times\mathcal{Y}\ni(\mathbf{x},\mathbf{y})\mapsto f(\mathbf{x},\mathbf{y})=\phi\left(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)$ be a function of $\mathcal{G}_{d,\mathbf{B}_k}$, where $\boldsymbol{\upsilon}_{k,d}\in\boldsymbol{\Upsilon}_{k,d}$ and $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\in\mathbf{V}_k\left(\mathbf{B}_k\right)$. According to (1.2.71), there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\in\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)$, such that

$$\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2\leq\epsilon^2.$$

By definition of $G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$, there exists $\widetilde{\boldsymbol{\upsilon}}_{k,d}\in G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$, such that

$$\sup_{\mathbf{y}\in\mathcal{Y}}\|\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y})-\boldsymbol{\upsilon}_{k,d}(\mathbf{y})\|_2^2\leq\delta^2_{\boldsymbol{\Upsilon}_{k,d}}. \tag{1.2.73}$$

Step 3: Upper bound of the number of the bracketing entropy. Next, we wish to make use of Lemma 1.2.23 to evaluate the ratio of two Gaussian densities.

Lemma 1.2.23 (Proposition C.1 from Maugis & Michel (2011b)). *Let $\phi\left(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)$ and $\phi\left(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)$ be two Gaussian densities. If $\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1$ is a positive definite matrix then for all $\mathbf{x}\in\mathbb{R}^D$,*

$$\frac{\phi\left(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)}{\phi\left(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)}\leq\sqrt{\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}}\exp\left[\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^\top\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right)^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\right].$$

The following Lemma 1.2.24 allows us to fulfill the assumptions of Lemma 1.2.23.

Lemma 1.2.24 (Similar to Lemma B.8 from Maugis & Michel (2011b)). *Assume that $0<\epsilon<\lambda_m^2/9$, and set $\alpha=3\sqrt{\epsilon}/\lambda_m$. Then, for every $k\in[K]$, $(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)-\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)$ and $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-(1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)$ are both positive definite matrices. Moreover, for all $\mathbf{x}\in\mathbb{R}^D$,*

$$\mathbf{x}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)-\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}\geq\epsilon\,\|\mathbf{x}\|_2^2,\quad\mathbf{x}^\top\left[\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-(1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}\geq\epsilon\,\|\mathbf{x}\|_2^2.$$

Proof of Lemma 1.2.24. For all $\mathbf{x}\neq\mathbf{0}$, since $\sup_{\lambda\in\text{vp}\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right)}|\lambda|=\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2\leq\epsilon$, where vp denotes the spectrum of matrix, $-\epsilon\geq-\lambda_m/3$, and $\alpha=3\epsilon/\lambda_m$, it follow that

$$\begin{aligned}
\mathbf{x}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)-\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}&=(1+\alpha)\,\mathbf{x}^\top\left[\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)-\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}+\alpha\mathbf{x}^\top\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\mathbf{x}\\
&\geq-(1+\alpha)\left\|\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)-\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right\|_2\|\mathbf{x}\|_2^2+\alpha\lambda_m\|\mathbf{x}\|_2^2\\
&\geq(\alpha\lambda_m-(1+\alpha)\,\epsilon)\,\|\mathbf{x}\|_2^2=(\alpha\lambda_m-\alpha\epsilon-\epsilon)\,\|\mathbf{x}\|_2^2\\
&\geq\left(\frac{2}{3}\alpha\lambda_m-\epsilon\right)\|\mathbf{x}\|_2^2=\epsilon\,\|\mathbf{x}\|_2^2>0,\ \text{and}\\
\mathbf{x}^\top\left[\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-(1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}&=(1+\alpha)^{-1}\,\mathbf{x}^\top\left[\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right]\mathbf{x}+\left(1-(1+\alpha)^{-1}\right)\mathbf{x}^\top\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\mathbf{x}\\
&\geq\left(\frac{\alpha\lambda_m-\epsilon}{1+\alpha}\right)\|\mathbf{x}\|_2^2=\frac{2\epsilon}{1+\alpha}\|\mathbf{x}\|_2^2\geq\epsilon\,\|\mathbf{x}\|_2^2>0\,(\text{ since }0<\alpha<1).
\end{aligned}$$

□

By Lemma 1.2.23 and the same argument as in the proof of Lemma B.9 from Maugis & Michel (2011b), given $0<\epsilon<\lambda_m/3$, where ϵ is chosen later, and $\alpha=3\epsilon/\lambda_m$, we obtain

$$\max\left\{\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})},\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right\}\leq(1+2\alpha)^{-\frac{D}{2}}\exp\left(\frac{\|\boldsymbol{\upsilon}_{k,d}(\mathbf{y})-\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y})\|_2^2}{2\epsilon}\right). \tag{1.2.74}$$

Because $\ln(\cdot)$ is a non-decreasing function, $\ln(1+2\alpha)\geq\alpha,\forall\alpha\in[0,1]$. Combined with (1.2.73) where $\delta^2_{\boldsymbol{\Upsilon}_{k,d}}=D\alpha\epsilon$, we conclude that

$$\max\left\{\ln\left(\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right),\ln\left(\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right)\right\}\leq-\frac{D}{2}\ln(1+2\alpha)+\frac{\delta^2_{\boldsymbol{\Upsilon}_{k,d}}}{2\epsilon}\leq-\frac{D}{2}\alpha+\frac{\delta^2_{\boldsymbol{\Upsilon}_{k,d}}}{2\epsilon}=0.$$

This means that $l(\mathbf{x},\mathbf{y})\leq f(\mathbf{x},\mathbf{y})\leq u(\mathbf{x},\mathbf{y}),\forall(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$. Hence, it remains to bound the size of bracket $[l,u]$ w.r.t. $d_{\mathcal{G}_{d,\mathbf{B}_k}}$. To this end, we aim to verify that $d^2_{\mathcal{G}_{d,\mathbf{B}_k}}(l,u)\leq\frac{\delta}{2}$. To do that, we make use of the following Lemma 1.2.25.

Lemma 1.2.25 (Proposition C.3 from Maugis & Michel (2011b)). *Let $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two Gaussian densities with full rank covariance. It holds that*

$$d^2\left(\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right) = 2\left\{1-2^{D/2}\left|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right|^{-1/4}\left|\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right|^{-1/2}\exp\left[-\frac{1}{4}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^\top\left(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right)^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\right]\right\}.$$

Therefore, using the fact that $\cosh(t)=\frac{e^{-t}+e^t}{2}$, Lemma 1.2.25 leads to, for all $\mathbf{y}\in\mathcal{Y}$:

$$\begin{aligned}
\mathrm{d}^2(l(\cdot,\mathbf{y}),u(\cdot,\mathbf{y})) &= \int_{\mathcal{X}}\left[l(\mathbf{x},\mathbf{y})+u(\mathbf{x},\mathbf{y})-2\sqrt{l(\mathbf{x},\mathbf{y})u(\mathbf{x},\mathbf{y})}\right]d\mathbf{x}\\
&= (1+2\alpha)^{-D}+(1+2\alpha)^{D}-2\\
&+\mathrm{d}^2\left(\phi\left(\cdot;\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)^{-1}\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right),\phi\left(\cdot;\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right)\right)\\
&= 2\cosh\left[D\ln(1+2\alpha)\right]-2\\
&\quad+2\left[1-2^{D/2}\left[(1+\alpha)^{-1}+(1+\alpha)\right]^{-D/2}\left|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right|^{-1/2}\left|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right|^{1/2}\right]\\
&= 2\cosh\left[D\ln(1+2\alpha)\right]-2+2-2\left[\cosh\left(\ln(1+\alpha)\right)\right]^{-D/2}\\
&= 2g\left(D\ln(1+2\alpha)\right)+2h\left(\ln(1+\alpha)\right),
\end{aligned}$$

where $g(t)=\cosh(t)-1=\frac{e^{-t}+e^t}{2}-1$, and $h(t)=1-\cosh(t)^{-D/2}$. The upper bounds of terms g and h separately imply that, for all $\mathbf{y}\in\mathcal{Y}$,

$$\mathrm{d}^2(l(\cdot,\mathbf{y}),u(\cdot,\mathbf{y}))\leq 2\left(2\cosh\left(\frac{1}{\sqrt{6}}\right)\alpha^2D^2+\frac{1}{4}\alpha^2D^2\right)\leq 6\alpha^2D^2=\frac{\delta^2}{4},$$

where we choose $\alpha=\frac{3\epsilon}{\lambda_m}$, $\epsilon=\frac{\delta\lambda_m}{6\sqrt{6}D}$, $\forall\delta\in(0,1]$, $D\in\mathbb{N}^\star$, $\lambda_m>0$, which appears in (1.2.74) and satisfies $\alpha=\frac{\delta}{2\sqrt{6}D}$ and $0<\epsilon<\frac{\lambda_m}{3}$. Indeed, studying functions g and h yields

$$\begin{aligned}
g'(t)&=\sinh(t),g''(t)=\cosh(t)\leq\cosh(c),\forall t\in[0,c],c\in\mathbb{R}_+,\\
h'(t)&=\frac{D}{2}\cosh(t)^{-D/2-1}\sinh(t),\\
h''(t)&=\frac{D}{2}\left(-\frac{D}{2}-1\right)\cosh(t)^{-D/2-2}\sinh^2(t)+\frac{D}{2}\cosh(t)^{-D/2}\\
&=\frac{D}{2}\left(1-\left(\frac{D}{2}+1\right)\left(\frac{\sinh(t)}{\cosh(t)}\right)^2\right)\cosh(t)^{-D/2}\leq\frac{D}{2},
\end{aligned}$$

where we used the fact that $\cosh(t)\geq 1$. Then, since $g(0)=0,g'(0)=0,h(0)=0,h'(0)=0$, by applying Taylor's Theorem, it is true that

$$\begin{aligned}
g(t)&=g(t)-g(0)-g'(0)t=R_{0,1}(t)\leq\cosh(c)\frac{t^2}{2},\forall t\in[0,c],\\
h(t)&=h(t)-h(0)-h'(0)t=R_{0,1}(t)\leq\frac{D}{2}\frac{t^2}{2}\leq\frac{D^2}{2}\frac{t^2}{2},\forall t\geq 0.
\end{aligned}$$

We wish to find an upper bound for $t=D\ln(1+2\alpha)$, $D\in\mathbb{N}^\star$, $\alpha=\frac{\delta}{2\sqrt{6}D}$, $\delta\in(0,1]$. Since ln is an increasing function, then we have

$$t=D\ln\left(1+\frac{\delta}{\sqrt{6}D}\right)\leq D\ln\left(1+\frac{1}{\sqrt{6}D}\right)\leq D\frac{1}{\sqrt{6}D}=\frac{1}{\sqrt{6}},\forall\delta\in(0,1],$$

since $\ln\left(1+\frac{1}{\sqrt{6}D}\right)\leq\frac{1}{\sqrt{6}D}$, $\forall D\in\mathbb{N}^\star$. Then, since $\ln(1+2\alpha)\leq 2\alpha,\forall\alpha\geq 0$,

$$\begin{aligned}
g\left(D\ln(1+2\alpha)\right)&\leq\cosh\left(\frac{1}{\sqrt{6}}\right)\frac{\left(D\ln(1+2\alpha)\right)^2}{2}\leq\cosh\left(\frac{1}{\sqrt{6}}\right)\frac{D^2}{2}4\alpha^2,\\
h\left(\ln(1+\alpha)\right)&\leq\frac{D^2}{2}\frac{\left(\ln(1+\alpha)\right)^2}{2}\leq\frac{D^2\alpha^2}{4}.
\end{aligned}$$

Note that the set of $\delta/2$-brackets $[l, u]$ over $\mathcal{G}_{d,\mathbf{B}_k}$ is totally defined by the parameter spaces $\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)$ and $G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$. This leads to an upper bound of the $\delta/2$-bracketing entropy of $\mathcal{G}_{d,\mathbf{B}_k}$ evaluated from an upper bound of the two set cardinalities. Hence, given any $\delta > 0$, by choosing $\epsilon = \frac{\delta\lambda_m}{6\sqrt{6}D}$, $\alpha = \frac{3\epsilon}{\lambda_m} = \frac{\delta}{2\sqrt{6}D}$, and $\delta^2_{\boldsymbol{\Upsilon}_{k,d}} = D\alpha\epsilon = D\frac{\delta}{2\sqrt{6}D}\frac{\delta\lambda_m}{6\sqrt{6}D} = \frac{\delta^2\lambda_m}{72D}$, it holds that

$$
\begin{aligned}
\mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2}, \mathcal{G}_{d,\mathbf{B}_k}\right) &\leq \text{card}\left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right) \times \text{card}\left(G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)\right) \\
&\leq \left(\left\lfloor \frac{2\lambda_M}{\epsilon} \right\rfloor \frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta_{\boldsymbol{\Upsilon}_{k,d}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \quad \text{(using (1.2.70) and (1.2.72))} \\
&\leq \left(\frac{2\lambda_M 6\sqrt{6}D}{\delta\lambda_m}\frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \\
&= \left(\frac{6\sqrt{6}\lambda_M D^2\left(D-1\right)}{\lambda_m D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \left(\frac{1}{\delta}\right)^{D_{\mathbf{B}_k}+\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)}.
\end{aligned}
$$

Finally, by definition of bracketing entropy in (1.2.46), we obtain

$$
\begin{aligned}
\mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2}, \mathcal{G}_{d,\mathbf{B}_k}\right) &\leq D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M D^2\left(D-1\right)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right) \\
&\quad + \left(D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\right)\ln\left(\frac{1}{\delta}\right) = \dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\frac{1}{\delta}\right)\right),
\end{aligned}
$$

where $\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right) = D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)$ and

$$
C_{\mathcal{G}_{d,\mathbf{B}_k}} = \frac{D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M D^2(D-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)}{\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)}.
$$

1.2.11.8 Our contributions on PSGaBloME models

Note that the proof of Theorem 1.2.5 follows the same idea from Section 1.2.11.3 for which we constructed to prove weak oracle inequality of BLoME models. The only different is that one need to take into account the joint rank and variable selection for parsimonious estimation in a high-dimensional framework. Therefore, we only highlight the main different results which must be adapted to such frameworks for controlling the bracketing entropy of PSGaBloME models, see more in Section 1.2.11.9. This is our first contribution for PSGaBloME models. Another important contribution lies on our constructions for Lasso+l_2-MLE and Lasso+l_2-Rank procedures to deal with high-dimensional data in PSGaBloME models. Such procedures are inspired by the ideas from Khalili (2010), Stadler et al. (2010), Devijver (2015b, 2017a,b).

These procedures are decomposed into three main steps. First, we construct a model collection, with models more or less sparse, with more or less mixture components and with more or less terms in polynomial of weights and means. Second, we refit estimations with the MLE or estimate the parameters by MLE under rank constraint on the restricted set of relevant columns. To this end, a model is selected thanks to the slope heuristic, which is a data-driven criterion based on non-asymptotic theory. In particular, this leads to a classification or clustering according to the MAP principle on the selected model. It is important to emphasize that since we have to deal with multivariate responses in PSGaBloME regression models, we propose new penalty functions in (1.2.81) and the corresponding generalized EM algorithm in Section 1.2.11.11, see also Section 4.3.6 for a comprehensive detail. Finally, Sections 1.2.11.10 and 1.2.11.11 are devoted in our contribution regarding the practical point of view of PSGaBloME regression models via the previous Lasso+l_2-MLE and Lasso+l_2-Rank procedures.

1.2.11.9 Controlling the bracketing entropy of PSGaBloME models

The bracketing entropy of PSGaBloME models can be controlled via the following definitions:

$$\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto g_{\mathbf{w},k}\left(\mathbf{x}\right) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)}, \mathbf{w} = \left(w_k\right)_{k\in[K]} \in \mathbf{W}_{K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}}} \right\},$$

$$\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})} = \{0\} \otimes \mathbf{W}_{\mathbf{J}_{\boldsymbol{\omega}}}^{K-1}, \mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]} = \left\{ \boldsymbol{\alpha} = \left(\alpha_t\right)_{t\in[p]} \in \mathcal{A} : \boldsymbol{\alpha}_j > 0, j \in [\mathbf{J}_{\boldsymbol{\omega}}] \right\},$$

$$\mathbf{W}_{\mathbf{J}_{\boldsymbol{\omega}}} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto w\left(\mathbf{x}\right) = \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \max_{\boldsymbol{\alpha}\in\mathcal{A}} |\boldsymbol{\omega}_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}} \right\},$$

$$\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} = \left\{ \mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \left(\phi_q\left(\mathbf{x};\boldsymbol{\upsilon}_k(\mathbf{y}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)\right)_{k\in[K]} : \boldsymbol{\upsilon} \in \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}, \boldsymbol{\Sigma}(\mathbf{B}) \in \mathbf{V}_K(\mathbf{B}) \right\}.$$

As in Section 1.2.11.3, the most difficult task lies on proving the following Lemma 1.2.26, allowing us to construct the Gaussian brackets to handle with the entropy metric for Gaussian experts, which is comprehensively established in Section 4.3.4.2.

Lemma 1.2.26. *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right).$$

To prove Lemma 1.2.26, step 2 is the main different result compared to the previous 3 steps for proof of Lemma 1.2.22, which requires more work to prove. Before presenting one of our main contributions, we need to define Given any $k \in [K]$, we first define the following set and its corresponding distance:

$$\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)} = \left\{ \mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \phi_q\left(\mathbf{y};\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) : \boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)} \in \boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}, \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k(\mathbf{B}_k) \right\},$$

$$d^2_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\left(\phi_k^+,\phi_k^-\right) = \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{d}^2\left(\phi_k^+\left(\mathbf{X}_i,\cdot\right),\phi_k^-\left(\mathbf{X}_i,\cdot\right)\right)\right].$$

Step 2: Construction of a net for the Gaussian expert mean functions. We claim that given any $\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} > 0$, any $\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)} \in \boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}$, there exist a minimal covering of δ_k-bracket $G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$ and a function $\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)} \in G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \left\|\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) - \boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})\right\|_2^2 \leq \delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}, \tag{1.2.75}$$

$$\operatorname{card}\left(G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)\right) \leq \left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}. \tag{1.2.76}$$

To accomplish this, we use the singular value decomposition of $\boldsymbol{\beta}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} \left[\sigma_{kd}\right]_r \left[\mathbf{u}_{kd}\right]_{\bullet,r} \left[\mathbf{v}_{kd}^{\top}\right]_{r,\bullet}$, $k \in [K], d \in [D]$, with $\left[\sigma_{kd}\right]_r, r \in [R_{kd}]$, denote the singular values of $\boldsymbol{\beta}_{kd}^{R_{kd}}$, with corresponding orthogonal unit vectors $\left(\left[\mathbf{u}_{kd}\right]_{\bullet,r}\right)_{r\in[R_{kd}]}$ and $\left(\left[\mathbf{v}_{kd}^{\top}\right]_{r,\bullet}\right)_{r\in[R_{kd}]}$. Then, we construct $\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) = \widetilde{\boldsymbol{\beta}}_{k0} + \sum_{d=1}^{D} \widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} \mathbf{x}^d$, where $\widetilde{\boldsymbol{\beta}}_{k0}$ and $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} \left[\widetilde{\sigma}_{kd}\right]_r \left[\widetilde{\mathbf{u}}_{kd}\right]_{\bullet,r} \left[\widetilde{\mathbf{v}}_{kd}^{\top}\right]_{r,\bullet}$, $k \in [K], d \in [D]$, are determined so that (1.2.75) and (1.2.76) are satisfied. Note that for each $k \in [K], d \in [D]$, it holds

that

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})-\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})\right\|_2 &= \left\|\widetilde{\boldsymbol{\beta}}_{k0}-\boldsymbol{\beta}_{k0}+\sum_{d=1}^{D}\left(\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}-\boldsymbol{\beta}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2\\ &\leq \left\|\widetilde{\boldsymbol{\beta}}_{k0}-\boldsymbol{\beta}_{k0}\right\|_2+\sum_{d=1}^{D}\left\|\left(\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}-\boldsymbol{\beta}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2\\ &\leq \sqrt{q}\left\|\widetilde{\boldsymbol{\beta}}_{k0}-\boldsymbol{\beta}_{k0}\right\|_\infty+p\sqrt{q}\sum_{d=1}^{D}\left\|\left\|\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}-\boldsymbol{\beta}_{kd}^{R_{kd}}\right\|\right\|_\infty\left\|\mathbf{x}^d\right\|_\infty\\ &\leq \sqrt{q}\left\|\widetilde{\boldsymbol{\beta}}_{k0}-\boldsymbol{\beta}_{k0}\right\|_\infty+p\sqrt{q}\sum_{d=1}^{D}\left\|\left\|\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}-\boldsymbol{\beta}_{kd}^{R_{kd}}\right\|\right\|_\infty,\end{aligned}$$

where we used the fact that for all $d\in[D]$, $\mathbf{x}\in\mathcal{X}$, $\left\|\mathbf{x}^d\right\|_\infty\leq 1$ as $\mathcal{X}=[0,1]^p$. Thus, (1.2.75) is immediately followed if we now choose $\widetilde{\boldsymbol{\beta}}_{k0}$ and $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}$ such that

$$\sqrt{q}\left\|\boldsymbol{\beta}_{k0}-\widetilde{\boldsymbol{\beta}}_{k0}\right\|_\infty\leq\frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2},\tag{1.2.77}$$

$$\left\|\left\|\boldsymbol{\beta}_{kd}^{R_{kd}}-\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}\right\|\right\|_\infty\leq\frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2Dp\sqrt{q}}.\tag{1.2.78}$$

Let us now see how to construct $\widetilde{\boldsymbol{\beta}}_{k0}$ to get (1.2.77). This task can be accomplished if for all $k\in[K]$, $z\in[q]$, we set

$$\begin{aligned}B&=\mathbb{Z}\cap\left[\left\lfloor-A_{\mathbf{u},\mathbf{v}}\frac{2\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right\rfloor,\left\lfloor A_{\mathbf{u},\mathbf{v}}\frac{2\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right\rfloor\right],\\ \left[\widetilde{\boldsymbol{\beta}}_{k0}\right]_z&=\arg\min_{b\in B}\left|\left[\boldsymbol{\beta}_{k0}\right]_z-\frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2\sqrt{q}}b\right|.\end{aligned}$$

Next, let us now see how to construct $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}$ to get (1.2.78). The boundedness assumption in (4.3.6) implies that

$$\begin{aligned}\left\|\left\|\boldsymbol{\beta}_{kd}^{R_{kd}}-\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}\right\|\right\|_\infty &= \max_{z\in[q],j\in[p]}\left|\sum_{r=1}^{R_{kd}}\left(\left[\sigma_{kd}\right]_r\left[\mathbf{u}_{kd}\right]_{z,r}\left[\mathbf{v}_{kd}^\top\right]_{r,j}-\left[\widetilde{\sigma}_{kd}\right]_r\left[\widetilde{\mathbf{u}}_{kd}\right]_{z,r}\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right)\right|\\ &= \max_{z\in[q],j\in[p]}\left|\sum_{r=1}^{R_{kd}}\left(\left(\left[\sigma_{kd}\right]_r-\left[\widetilde{\sigma}_{kd}\right]_r\right)\left[\mathbf{u}_{kd}\right]_{z,r}\left[\mathbf{v}_{kd}^\top\right]_{r,j}\right.\right.\\ &\qquad-\left[\widetilde{\sigma}_{kd}\right]_r\left(\left[\widetilde{\mathbf{u}}_{kd}\right]_{z,r}-\left[\mathbf{u}_{kd}\right]_{z,r}\right)\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\\ &\qquad\left.\left.-\left[\widetilde{\sigma}_{kd}\right]_r\left[\mathbf{u}_{kd}\right]_{z,r}\left(\left[\mathbf{v}_{kd}^\top\right]_{r,j}-\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right)\right)\right|\\ &\leq \max_{r\in[R_{kd}]}\left|\left[\sigma_{kd}\right]_r-\left[\widetilde{\sigma}_{kd}\right]_r\right|\max_{z\in[q],j\in[p]}\sum_{r=1}^{R_{kd}}\left|\left[\mathbf{u}_{kd}\right]_{z,r}\left[\mathbf{v}_{kd}^\top\right]_{r,j}\right|\\ &\quad+\max_{z\in[q],r\in[R_{kd}]}\left|\left[\widetilde{\mathbf{u}}_{kd}\right]_{z,r}-\left[\mathbf{u}_{kd}\right]_{z,r}\right|\max_{j\in[p]}\sum_{r=1}^{R_{kd}}\left|\left[\widetilde{\sigma}_{kd}\right]_r\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right|\\ &\quad+\max_{r\in[R_{kd}],j\in[p]}\left|\left[\mathbf{v}_{kd}^\top\right]_{r,j}-\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right|\max_{z\in[q]}\sum_{r=1}^{R_{kd}}\left|\left[\widetilde{\sigma}_{kd}\right]_r\left[\mathbf{u}_{kd}\right]_{z,r}\right|\\ &\leq R_{kd}A_{\mathbf{u},\mathbf{v}}^2\max_{r\in[R_{kd}]}\left|\left[\sigma_{kd}\right]_r-\left[\widetilde{\sigma}_{kd}\right]_r\right|\\ &\quad+R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma\left(\max_{z\in[q],r\in[R_{kd}]}\left|\left[\widetilde{\mathbf{u}}_{kd}\right]_{z,r}-\left[\mathbf{u}_{kd}\right]_{z,r}\right|+\max_{r\in[R_{kd}],j\in[p]}\left|\left[\mathbf{v}_{kd}^\top\right]_{r,j}-\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right|\right).\end{aligned}$$

Therefore, (1.2.78) is immediately implied if we now choose $[\widetilde{\sigma}_{kd}]_r$, $[\widetilde{\mathbf{u}}_{kd}]_{z,r}$ and $\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}$ such that

$$\begin{aligned}
\max_{r\in[R_{kd}]} |[\sigma_{kd}]_r - [\widetilde{\sigma}_{kd}]_r| &\leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}},\\
\max_{z\in[q],r\in[R_{kd}]} \left|[\widetilde{\mathbf{u}}_{kd}]_{z,r} - [\mathbf{u}_{kd}]_{z,r}\right| &\leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}},\\
\max_{r\in[R_{kd}],j\in[p]} \left|\left[\mathbf{v}_{kd}^\top\right]_{r,j} - \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}\right| &\leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}.
\end{aligned}$$

This task can be accomplished as follows: for all $r \in [R_{kd}]$, $j \in [p]$, $z \in [q]$, set

$$\begin{aligned}
S &= \mathbb{Z} \cap \left[0, \left\lfloor A_\sigma \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor\right],\\
[\widetilde{\sigma}_{kd}]_r &= \arg\min_{\zeta\in S} \left|[\sigma_{kd}]_r - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}\zeta\right|,\\
U &= \mathbb{Z} \cap \left[\left\lfloor -A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor, \left\lfloor A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor\right],\\
[\widetilde{\mathbf{u}}_{kd}]_{z,r} &= \arg\min_{\mu\in U} \left|[\mathbf{u}_{kd}]_{z,r} - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}\mu\right|,\\
\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} &= \arg\min_{\upsilon\in U} \left|\left[\mathbf{v}_{kd}^\top\right]_{r,j} - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}\upsilon\right|.
\end{aligned}$$

Note that, according to Strang (2019, I.8), we only need to determine the vectors $\left(\left([\widetilde{\mathbf{u}}_{kd}]_{z,r}\right)_{z\in[q-r]}\right)_{r\in[R_{kd}]}$ and $\left(\left([\widetilde{\mathbf{v}}_{kd}]_{r,j}\right)_{j\in[\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})-r]}\right)_{r\in[R_{kd}]}$ since the remaining elements of such vectors belong to the the nullspace of $\boldsymbol{\beta}_{kd}^{R_{kd}}$ and $\boldsymbol{\beta}_{kd}^{R_{kd}\top}$. The number of total free parameters in the previous two vectors are

$$\begin{aligned}
\sum_{r=1}^{R_{kd}} (q-r) &= R_{kd}\left(\frac{2q - R_{kd} - 1}{2}\right),\\
\sum_{r=1}^{R_{kd}} (\mathrm{card}\,(\mathbf{J}_{\boldsymbol{\omega}}) - r) &= R_{kd}\left(\frac{2\,\mathrm{card}\,(\mathbf{J}_{\boldsymbol{\omega}}) - R_{kd} - 1}{2}\right).
\end{aligned}$$

To this end, for all $k \in [K]$, $d \in [D]$, and $z \in [q]$, we let

$$\left[\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}\right]_{z,j} = \begin{cases} \sum_{r=1}^{R_{kd}} [\widetilde{\sigma}_{kd}]_r [\widetilde{\mathbf{u}}_{kd}]_{z,r} \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} & \text{if } j \in \mathbf{J}_{\boldsymbol{\omega}},\\ 0 & \text{if } j \in \mathbf{J}_{\boldsymbol{\omega}}^C. \end{cases}$$

In particular, (1.2.76) is proved by the following entropy controlling

$$\begin{aligned}
&\operatorname{card}\left(G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)\right)\\
&\leq \left[\frac{4A_{\mathbf{u},\mathbf{v}}\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^q \prod_{d=1}^{D}\left[\frac{6R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{R_{kd}}\left[\frac{12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{R_{kd}(q+\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})-R_{kd}-1)}\\
&= \left[\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}, \text{ where}\\
&\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right) = q + \sum_{d=1}^{D} R_{kd}\left(q+\operatorname{card}\left(\mathbf{J}_{\boldsymbol{\omega}}\right)-R_{kd}\right),\ C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} = \frac{\ln\left(C_{(D,\mathbf{J},\mathbf{R}_k)}\right)}{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)},\\
&\text{and } C_{(D,\mathbf{J},\mathbf{R}_k)} = \left[4A_{\mathbf{u},\mathbf{v}}\sqrt{q}\right]^q\left[12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}\right]^{\sum_{d=1}^{D}R_{kd}(q+\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})-R_{kd})}2^{-\sum_{d=1}^{D}R_{kd}}.
\end{aligned}$$

1.2.11.10 Lasso+l_2-MLE and Lasso+l_2-Rank procedures

For the sake of simplicity, both the weights of softmax gating networks and the means of Gaussian experts are defined as the following simple polynomial functions:

$$\begin{aligned}
\mathbf{W}_{K,d_{\mathbf{W}}} &= \{0\}\otimes\mathbf{W}_{K-1},\\
\mathbf{W}_{K-1} &= \left\{\mathcal{X}\ni\mathbf{x}\mapsto w_k(\mathbf{x}) = \omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^\top\mathbf{x}^l, \forall k\in[K-1]:\max_{l\in[L]}|\boldsymbol{\omega}_{kl}|\leq T_{\mathbf{W}}\right\},\\
\boldsymbol{\Upsilon}_{K,D} &= \left\{\mathcal{X}\ni\mathbf{x}\mapsto\left(\boldsymbol{\beta}_{k0}+\sum_{d=1}^{D}\boldsymbol{\beta}_{kd}\mathbf{x}^d\right)_{k\in[K]}:\max\left\{\left\|\boldsymbol{\beta}_{kd}\right\|_\infty:k\in[K],d\in(\{0\}\cup[D])\right\}\leq T_{\boldsymbol{\Upsilon}}\right\}.
\end{aligned}$$

However, our finite-sample oracle inequality still holds for a more general case when we utilize general polynomials, defined in (1.2.22), for weights of the gating networks.

Model collection construction of PSGaBloME regression models

We firstly fix $K\in\mathcal{K}$, $L\in\mathcal{L}$ and $D\in\mathcal{D}$. To detect the relevant indices and construct the set $\mathbf{J}\in\mathcal{J}$, by generalizing the idea from Khalili (2010), Stadler et al. (2010), Devijver (2015b, 2017a,b), we utilize an l_2-penalized log-likelihood functions instead of the log-likelihood and combine with two l_1-penalties on the terms of polynomials from weights and the means. It is worth mentioning that in order to deal with PSGaBloME model, we must extend the results from Khalili (2010), Stadler et al. (2010) to multivariate response $\mathbf{Y}\in\mathbb{R}^q$ and the results from Devijver (2015b, 2017a,b) to mixture of polynomial experts with any arbitrarily degree of weights and mean functions. More precisely, we consider

$$\widehat{\boldsymbol{\psi}}^{\text{Lasso}+l_2}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}_{(K,L,D,\mathbf{J},\mathbf{R})}}{\arg\min}\left\{-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\operatorname{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi})\right\}, \tag{1.2.79}$$

$$s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K}g_k\left(\mathbf{x};\boldsymbol{\omega}\right)\phi_q\left(\mathbf{y};\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}),\boldsymbol{\Sigma}_k\right),\ \boldsymbol{\psi} = \left(\omega_{k0},\left(\boldsymbol{\omega}_{kl}\right)_{l\in[L]},\boldsymbol{\beta}_{k0},\left(\boldsymbol{\beta}_{kd}\right)_{d\in[D]},\boldsymbol{\Sigma}_k\right)_{k\in[K]}, \tag{1.2.80}$$

$$\operatorname{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{kl}^{[1]}\left\|\boldsymbol{\omega}_{kl}\right\|_1+\sum_{k=1}^{K}\sum_{d=1}^{D}\lambda_{kd}^{[2]}\left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1+\frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\boldsymbol{\omega}_{kl}\right\|_2^2,\quad \mathbf{Q}_k^\top\mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1}. \tag{1.2.81}$$

Here, $\boldsymbol{\lambda} = \left(\left(\lambda_{kl}^{[1]}\right)_{k\in[K],l\in[L]},\left(\lambda_{kd}^{[2]}\right)_{k\in[K],d\in[D]},\frac{\lambda^{[3]}}{2}\right)$ is a vector of non-negative regularization parameters, for any $k\in[K]$, $l\in[L]$, $d\in[D]$, $\left\|\boldsymbol{\omega}_{kl}\right\|_1 = \sum_{j=1}^{p}\left|\left[\boldsymbol{\omega}_{kl}\right]_j\right|$, $\left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1 = \sum_{j=1}^{p}\sum_{z=1}^{q}\left|\left[\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right]_{z,j}\right|$,

$\|\boldsymbol{\omega}_{kl}\|_2^2 = \sum_{j=1}^{p} [\boldsymbol{\omega}_{kl}]_j^2$ is the Euclidean norm in $\mathbb{R}^p$, and the Cholesky decomposition $\boldsymbol{\Sigma}_k^{-1} = \mathbf{Q}_k^\top \mathbf{Q}_k$ defines $\mathbf{Q}_k$ for all $k \in [K]$. Remark that the first two terms from (1.2.81) are the usual l_1-estimator, called the Lasso estimator, while the l_2 penalty function for the gating network is added to avoid wildly large positive and negative estimates of the regression coefficients corresponding to the mixing proportions. This behavior can be observed in logistic/multinomial regression when the number of potential features is large and highly correlated (*e.g.,* Park & Hastie (2008), Bunea et al. (2008)). However, this also affects the sparsity of the regularization model, which is confirmed from numerical experiments from Chamroukhi & Huynh (2018), Chamroukhi & Huynh (2019).

Computing those estimators leads to construct the relevant variables set. For a fixed number of mixture components $K \in \mathcal{K}$, fixed degrees $L \in \mathcal{L}$ and $D \in \mathcal{D}$ of polynomials from mean and weight functions, denote by $\mathbf{G}_{K,L,D}$ a candidate of grid of regularization parameters. Fixing a regularization parameter $\boldsymbol{\lambda} \in \mathbf{G}_{K,L,D}$, we could then use a generalized EM algorithm which is originally introduced by Dempster et al. (1977) and is extended for PSGaBloME models with univariate response, *e.g.,* Jordan & Jacobs (1994), Khalili (2010), Chamroukhi & Huynh (2018), Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019), to compute the Lasso $+l_2$ estimator, and construct the set of relevant variables $\mathbf{J}_{(K,L,D,\boldsymbol{\lambda})}$, saying the non-zero coefficients. We denote by $\mathcal{J}$ the random collection of all these sets,

$$\mathcal{J} = \bigcup_{K\in\mathcal{K}} \bigcup_{L\in\mathcal{L}} \bigcup_{D\in\mathcal{D}} \bigcup_{\boldsymbol{\lambda}\in\mathbf{G}_{K,L,D}} \mathbf{J}_{(K,L,D,\boldsymbol{\lambda})}. \tag{1.2.82}$$

Refitting

The Lasso+l_2-MLE procedure

The second step consists of approximating the MLE

$$\widehat{s}^{(K,L,D,\mathbf{J})} = \arg \min_{t\in\mathcal{S}_{(K,L,D,\mathbf{J})}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(t \left(\mathbf{y}_i | \mathbf{x}_i \right) \right) \right\}, \tag{1.2.83}$$

which can be accomplished by using an EM algorithm for each model $(K, L, D, \mathbf{J}) \in \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times \mathcal{J}$. Remark that we estimate all parameters, to reduce bias induced by the Lasso $+l_2$ estimator. The reason why we need to refit the Lasso $+l_2$ estimator can be referred to Devijver (2015b, Section 2.3).

The Lasso+l_2-Rank procedure

We use the generalized EM algorithm to estimate the parameters by MLE under rank constraint on the restricted set of relevant columns.

Model selection

The third step is devoted to model selection. We follow the framework from Devijver (2017b, Section 3) to select the refitted model rather than selecting the regularization parameter. Instead of using an asymptotic criterion, such as BIC or AIC, we use the slope heuristic, originally introduced by Birgé & Massart (2007) and recently reviewed by Baudry et al. (2012) and Arlot (2019), which is a data-driven non-asymptotic criterion for selecting a model among a collection of models. For an oracle inequality to only justify the penalty shape when using slope heuristic used here, see Theorem 1.2.5 for more details.

1.2.11.11 Generalized EM algorithm

Note that we will not present in an exhaustive way the generalized EM algorithm here. The readers can find its description in more details in Section 4.3.6. However, we would like to highlight our contributions as follows. The EM algorithm (Dempster et al., 1977, McLachlan & Krishnan, 1997) is most commonly known as a technique to produce MLEs in settings where the data under study is incomplete or when optimization of the likelihood would be simplified if an additional set of variables

were known. The iterative EM algorithm consists of an expectation (E) step followed by a maximization (M) step. Generally, during the E step the conditional expectation of the complete (i.e. observed and unobserved) data log-likelihood is computed, given the data and current parameter values. In the M step the expected log-likelihood is maximized with respect to the model parameters. The imputation of latent variables often makes maximization of the expected log-likelihood more feasible. The log-likelihood function of the PSGaBloME model is

$$\mathrm{L}\left(\boldsymbol{\psi}\right) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} g_k\left(\mathbf{x}_i;\boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i),\boldsymbol{\Sigma}_k\right) \right].$$

It is difficult to directly obtain MLEs from this likelihood. In the EM framework, to alleviate this, the data are augmented by imputing for each incomplete observed-data vector $(\mathbf{x}_i,\mathbf{y}_i)_{i\in[n]}$, the K-dimensional binary random variable $\mathbf{z}_i = (z_{ik})_{k\in[K]}$ (which is also called the latent (unobserved) random variable or the allocation variable in the mixture model context). This latent variable has a 1-of-K representation in which a particular element z_{ik} is equal to 1 and all other elements are equal to 0. More precisely, for any $i \in [n]$, $k \in [K]$, z_{ik} is an indicator binary-valued variable such that $z_{ik} = 1$ if the ith pair $(\mathbf{x}_i,\mathbf{y}_i)$ is generated from the kth expert component and $z_{ik} = 0$ otherwise. Here, for any $i \in [n]$, given the predictor $\mathbf{x}_i$, $\mathbf{z}_i$ are unobserved i.i.d. random variables following a multinomial distribution:

$$\mathbf{z}_i|\mathbf{x}_i \sim \mathrm{Mult}\left(1, \left(g_k\left(\mathbf{x}_i;\boldsymbol{\omega}\right)\right)_{k\in[K]}\right).$$

The EM algorithm for solving (1.2.83) firstly requires the construction of the penalized complete-data log-likelihood

$$\mathrm{PL_c}\left(\boldsymbol{\psi},\mathbf{z}\right) = \mathrm{L_c}\left(\boldsymbol{\psi},\mathbf{z}\right) - \mathrm{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}), \tag{1.2.84}$$

$$\mathrm{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \lambda_{kl}^{[1]} \left\|\boldsymbol{\omega}_{kl}\right\|_1 + \sum_{k=1}^{K}\sum_{d=1}^{D} \lambda_{kd}^{[2]} \left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1 + \frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L} \left\|\boldsymbol{\omega}_{kl}\right\|_2^2, \quad \mathbf{Q}_k^\top\mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1},$$

via the standard complete-data log-likelihood

$$\mathrm{L_c}\left(\boldsymbol{\psi},\mathbf{z}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln\left[g_k\left(\mathbf{x}_i;\boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i),\boldsymbol{\Sigma}_k\right)\right].$$

The generalized EM, or GEM, algorithm addresses the problem of an intractable M-step. Instead of aiming to maximize the conditional expectation of $\mathrm{PL_c}\left(\boldsymbol{\psi}\right)$ with respect to $\boldsymbol{\psi}$, it seeks instead to change the parameters in such a way as to increase its value. Then, the GEM algorithm for the PSGaBloME model in its general form runs as follows. After starting with an initial solution $\boldsymbol{\psi}^{(0)}$, it alternates between the following steps until convergence (*e.g.,* when there is no longer significant change in the relative variation of the regularized log-likelihood).

It is important to emphasize that one of the main difficulty in the generalized EM algorithm for mixture model is an optimization problem in a generalized M-step. Motivated by the recent novel works from Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019) for SGaME model with linear mean Gaussian experts and scalar responses, we propose and compare three approaches for maximizing objective function in Generalized M-step based on a majorization–minimization (MM) algorithm, a coordinate ascent algorithm and proximal Newton-type method. These approaches have some advantages since they do not use any approximate for the penalty function, and have a separate structure which avoid matrix inversion. Note that we extend the work from Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019) to devise a novel MM algorithm for the PSGaBloME model with polynomial mean of Gaussian functions and multivariate responses.

1.2.12 Our contributions for l_1-oracle inequality for the Lasso estimator via Theorem 1.2.8

1.2.12.1 Sketch of the proof.

Motivated by the idea from Meynet (2013) and Devijver (2015a), we study the Lasso as the solution of a penalized maximum likelihood model selection procedure over countable collections of models in

an l_1-ball. Therefore, the main Theorem 1.2.7 is an immediate consequence of Theorem 1.2.8, which is an l_1-ball MoE regression model selection theorem for l_1-penalized maximum conditional likelihood estimation in the Gaussian mixture framework. The proof of Theorem 1.2.8 can be deduced from Proposition 4.2.4 and Proposition 4.2.5, which address the cases for large and small values of $\mathbf{Y}$.

Note that as the same contribution of the works of Devijver (2015a), which extended the Meynet (2013) to multivariate, Proposition 4.2.4 constitutes our main technical contribution, mainly on multivariate calculation. Its proof follows the arguments developed in the proof of a more general model selection theorem for maximum likelihood estimators: Massart (2007, Theorem 7.11). More precisely, the proof of Proposition 4.2.4 is in the spirit of Vapnik's method of structural risk minimization, which is established initially in Vapnik (1982) and briefly summarized in Section 8.2 in Massart (2007). In particularly, to obtain an upper bound of the empirical process in expectation, we shall use concentration inequalities combined with symmetrization arguments.

1.2.12.2 Our contributions

Our first contribution is Lemma 1.2.6, which helps to relax an assumption on the true unknown conditional density s_0. In fact both Meynet (2013, Page 660, Equation (4.28)) and Devijver (2015a, Page 661, the last inequality at the bottom) used the following assumption $s_0 \leq 1$.

Our second contribution lies in controlling the deviation $\sup_{f_m \in F_m} |\nu_n(-f_m)|$ appearing in Lemma 1.2.27 and from (1.2.85). More precisely, let $m \in \mathbb{N}^\star$, we have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f_m(\mathbf{Y}_i|\mathbf{x}_i) - \mathbb{E}_{\mathbf{Y}_{[n]}} \left[f_m(\mathbf{Y}_i|\mathbf{x}_i) \right] \right) \right|. \tag{1.2.85}$$

Lemma 1.2.27. *Let $M_n > 0$. Consider the event*

$$\mathcal{T} = \left\{ \max_{i=1,\dots,n} \|\mathbf{Y}_i\|_\infty = \max_{i=1,\dots,n} \max_{z \in \{1,\dots,q\}} |[\mathbf{Y}_i]_z| \leq M_n \right\},$$

and set

$$B_n = \max(A_{\boldsymbol{\Sigma}}, 1 + KA_G) \left(1 + q\sqrt{q} (M_n + A_{\boldsymbol{\beta}})^2 A_{\boldsymbol{\Sigma}} \right), \text{ and} \tag{1.2.86}$$

$$\Delta_{m'} = m' \sqrt{\ln(2p+1)} \ln n + 2\sqrt{K} \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}} \right). \tag{1.2.87}$$

Then, on the event $\mathcal{T}$, for all $m' \in \mathbb{N}^\star$, and for all $t > 0$, with probability greater than $1 - e^{-t}$,

$$\sup_{f_{m'} \in F_{m'}} |\nu_n(-f_{m'})| \mathbb{1}_{\mathcal{T}} \leq \frac{4KB_n}{\sqrt{n}} \left[37q\Delta_{m'} + \sqrt{2} \left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_\Sigma} \right) \sqrt{t} \right]. \tag{1.2.88}$$

To control the deviation of (1.2.85), we shall use concentration and symmetrization arguments. To do that, we have to provide new upper bounds with some adjustments compared to the works of Meynet (2013) and Devijver (2015a) via Lemmas 1.2.28 to 1.2.30. To do that, we let $M_n > 0$ and consider the event

$$\mathcal{T} = \left\{ \max_{i=1,\dots,n} \|\mathbf{Y}_i\|_\infty = \max_{i=1,\dots,n} \max_{z \in \{1,\dots,q\}} |[\mathbf{Y}_i]_z| \leq M_n \right\},$$

and put $B_n = \max(A_{\boldsymbol{\Sigma}}, 1 + KA_G) \left(1 + q\sqrt{q} (M_n + A_{\boldsymbol{\beta}})^2 A_{\boldsymbol{\Sigma}} \right)$.

Lemma 1.2.28. *On the event $\mathcal{T}$, for all $m \in \mathbb{N}^\star$,*

$$\sup_{f_m \in F_m} \|f_m\|_n \mathbb{1}_{\mathcal{T}} \leq 2KB_n \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}} \right) =: R_n. \tag{1.2.89}$$

Lemma 1.2.29. *Let $\delta > 0$ and $m \in \mathbb{N}^\star$. On the event $\mathcal{T}$, we have the following upper bound of the δ-packing number of the set of functions F_m, equipped with the metric induced by the norm $\|\cdot\|_n$:*

$$M\left(\delta, F_m, \|\cdot\|_n\right) \leq (2p+1)^{\frac{72B_n^2q^2K^2m^2}{\delta^2}} \left(1+\frac{18B_nKqA_{\boldsymbol{\beta}}}{\delta}\right)^K \left(1+\frac{18B_nKA_{\boldsymbol{\gamma}}}{\delta}\right)^K \left(1+\frac{18B_nKq\sqrt{q}}{a_{\boldsymbol{\Sigma}}\delta}\right)^K.$$

Via the upper bounds provided in Lemmas 1.2.28 and 1.2.29, we can apply Lemma 6.1 in Massart, 2007, see also Lemma 4.2.9 for more details, to get an upper bound on $\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m\in\mathcal{F}_m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(\mathbf{Y}_i|\mathbf{x}_i)\right|\right]$. We thus obtain the following results.

Lemma 1.2.30. *Let $m \in \mathbb{N}^\star$, consider $(\epsilon_1, \ldots, \epsilon_n)$, a Rademacher sequence independent of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Then, on the event $\mathcal{T}$,*

$$\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m\in\mathcal{F}_m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(\mathbf{Y}_i|\mathbf{x}_i)\right|\right] \leq \frac{74KB_nq}{\sqrt{n}}\Delta_m, \tag{1.2.90}$$

$$\Delta_m := m\sqrt{\ln(2p+1)}\ln n + 2\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right). \tag{1.2.91}$$

In particular, the proofs of Lemmas 1.2.28 and 1.2.29, require an upper bound on the uniform norm of the gradient of $\ln s_{\boldsymbol{\psi}}$, for $s_{\boldsymbol{\psi}} \in \mathcal{S}$ from SGaME models, which is more complex and difficult compared to the class of finite mixture of Gaussian regression models. More precisely, we provide the following Lemma 1.2.31, where the upper bound is used to modify the upper bound in Lemmas 1.2.28 and 1.2.29.

Lemma 1.2.31. *Given $s_{\boldsymbol{\psi}}$, as described in (1.2.28), it holds that*

$$\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}}\left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\cdot|\mathbf{x})\right)}{\partial\boldsymbol{\psi}}\right\|_\infty \leq G(\cdot),$$

$$G : \mathbb{R}^q \ni \mathbf{y} \mapsto G(\mathbf{y}) = \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right). \tag{1.2.92}$$

Finally, we also correct some errors regarding the upper bounds from l_1-oracle inequalities from Meynet (2013), Devijver (2015a).

1.2.12.3 Discussions and comparisons

Theorem 1.2.7 is non-asymptotic: the number n of observations is fixed while the number of covariates p can grow with respect to n, and in fact can be much larger than n. Note that, as in Khalili (2010), the true order K of the MoE model (the true number of experts in our model) is assumed to be known. From a pragmatic perspective, one may estimate it via the AIC of Akaike (1974), the BIC of Schwarz et al. (1978), or slope heuristic of Birgé & Massart (2007), see also Section 1.2.4. Our result follows directly the lineage of research of Meynet (2013) and Devijver (2015a). In fact, our theorem combined Vapnik's structural risk minimization paradigm (*e.g.,* Vapnik, 1982) and theory of model selection for conditional density estimation (*e.g.,* Cohen & Le Pennec, 2011), which is an extended version of the density estimation results from Massart (2007).

A great amount of attention has been paid to obtaining a lower bound for λ, cf. (1.2.36), in the oracle inequality, with optimal dependence on p and q, which are the only parameters not to be fixed and which can grow the possibility that $p \gg n$. In fact, the condition that $\kappa \geq 148$ is implied by Theorem 1.2.8, namely, the following conditions:

$$\lambda \geq \kappa\frac{KB_n'}{\sqrt{n}}\left(q\ln n\sqrt{\ln(2p+1)}+1\right),$$

$$B_n' = \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}}\ln n\right)\right),$$

for some absolute constants $\kappa \geq 148$. Such conditions are obtained from the proof of Theorem 1.2.8, which requires the results of Proposition 4.2.4. In the proof of Proposition 4.2.4, the original condition for κ is as follows: let $\kappa \geq 1$ and assume that $\text{pen}(m) = \lambda m$, for all $m \in \mathbb{N}^\star$ with

$$\lambda \geq \kappa \frac{4KB_n}{\sqrt{n}} \left(37q \ln n \sqrt{\ln(2p+1)} + 1\right).$$

Then, to simplify the constant on the lower bound for λ, we replaced $4 \times 37\kappa, \kappa \geq 1$ by $\kappa' = 4 \times 37\kappa \geq 148$.

Note that, we recover the same dependence of form $\sqrt{\ln(2p+1)}$ as for the homogeneous linear regression in Stadler et al. (2010) and of form $\sqrt{\ln(2p+1)}\frac{(\ln n)^2}{\sqrt{n}}$ for the mixture Gaussian regression models in Meynet (2013). On the contrary, the dependence on q for the mixture of multivariate Gaussian regression models in Devijver (2015a) was of form $q^2 + q$, while we obtain the form $q^2\sqrt{q}$, here. The main reason is that we need to control the larger class, S, of the finite mixture of experts models with softmax gating functions and Gaussian experts, and we use a different technique to evaluate the upper bound on the uniform norm of the gradient for each element in S. Furthermore, the dependence on n for the homogeneous linear regression in Stadler et al. (2010) was of order $\frac{1}{\sqrt{n}}$, while we have an extra $(\ln n)^2$ factor, here. In fact, the same situation can be found in the l_1-oracle inequalities of Meynet (2013), and Devijver (2015a). As explained in Meynet (2013), using a non-linear Kullback–Leibler information leads to a scenario where the linearity arguments developed in Stadler et al. (2010) with the quadratic loss function can not be exploited. Instead, we need to use the entropy arguments to handle our model, which leads to an extra $(\ln n)^2$ factor. Motivated by the frameworks from Meynet (2013), Devijver (2015a), we have paid attention to giving an explicit dependence not only on n, p and q, but also on the number of mixture components K as well as on the regressors and $A_{\boldsymbol{\beta}}, A_{\boldsymbol{\Sigma}}, A_G$–all the quantities bounding the parameters of the model. However, we should be aware of the fact that these dependences may not be optimal. In our lower bound, we obtain the factor K^2 dependence instead of K, as in Meynet (2013), Devijver (2015a). This can be explained via the fact that we used another technique to handle the more complex model when dealing with the upper bound on the uniform norm of the gradient of $\ln s_{\boldsymbol{\psi}}$, for $s_{\boldsymbol{\psi}} \in S$, in Lemma 4.2.14. We refer to Meynet (2013, Remark 5.8) for some data sets for which the dependence on K might be reduced to an order of $\sqrt{K}$ for the mixture Gaussian regression models. Establishing the optimal rates for such problems is still open.

We further note that our theorem ensures that there exists a λ sufficiently large for which the estimate has good properties, but does not give an explicit value for λ. However, in Theorem 1.2.7, we provide at least the lower bound for the value of λ via the bound $\lambda \geq \kappa C(p, q, n, K)$, where $\kappa \geq 148$, even though this value is obviously overly pessimistic. A possible solution for calibrating penalties from the data are the AIC and BIC approaches, motivated by asymptotic arguments. Another method is the slope heuristic introduced first by Birgé & Massart (2007) and further discussed in Baudry et al. (2012), which is a non-asymptotic criterion for selecting a model among a collection of models. Such strategy, however, will not be further considered here as the implementations will result in an overextension of the length and scope of the manuscript. Furthermore, the technical developments required for such methods is non-trivial and constitutes a significant research direction that we hope to pursue in the future. We refer the reader to the numerical experiments from Montuelle et al. (2014) (using the slope heuristic), and Chamroukhi & Huynh (2018), and Chamroukhi & Huynh (2019) (using BIC), for the practical implementation of penalization methods in the framework of MoE regression models.

As in Devijver (2015a), we suppose that the regressors belong to $\mathcal{X} = [0, 1]^p$, for simplicity. However, the arguments in our proof are valid for covariates of any scale.

To the best of our knowledge, we are the first to prove the non-asymptotic l_1-oracle inequality of Theorem 1.2.7, for the mixture of Gaussian experts regression models with l_1-regularization. Note that by extending the theoretical developments for mixture of linear regression models in Khalili & Chen (2007), a standard asymptotic theory for MoE models is established in Khalili (2010). Therefore, our non-asymptotic result in Theorem 1.2.7 can be considered as complementary to such asymptotic results for SGaME regression models.

Theorem 1.2.7 is also complementary to Theorem 1.2.5 and Theorem 1 of Montuelle et al. (2014),

who also considered SGaME models. Notice that they focused on model selection and obtained a *weak oracle inequality* for the penalized MLE, while we aimed to study the l_1-regularization properties of the Lasso estimators. However, we can compare their procedure with Theorem 1.2.8.

The main reason explaining their result being considered a *weak oracle inequality* is that we can see that Theorem 1.2.5 and Theorem 1 of Montuelle et al. (2014) use difference divergence on the left (the $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$, tensorized Jensen–Kullback–Leibler divergence), and on the right (the $\mathrm{KL}^{\otimes \mathrm{n}}$, tensorized Kullback–Leibler divergence). However, under a strong assumption, the two divergences are *equivalent* for the conditional PDFs considered. This strong assumption is nevertheless satisfied, if we assume that $\mathcal{X}$ is compact, as is the case of $\mathcal{X} = [0,1]^p$ in Theorem 1.2.8, s_0 is compactly supported, and the regression functions are uniformly bounded, and there is a uniform lower bound on the eigenvalues of the covariance matrices.

To illustrate the strictness of the compactness assumption for s_0, we only need to consider s_0 as a univariate Gaussian PDF, which obviously does not satisfy such a hypothesis. Therefore, in such case, Theorem 1.2.5 and Theorem 1 in Montuelle et al. (2014) are actually weaker than Theorem 1.2.8, with respect to the compact support assumption on the true conditional PDF s_0. On the contrary, the only assumption used to establish Theorem 1.2.8 is the boundedness of the parameters of the mixtures, which is also assumed in Theorem 1.2.5 and in Montuelle et al. (2014, Theorem 1).

Note that the constant $1 + \kappa^{-1}$ from the upper bound in Theorem 1.2.8 and C_1 from Theorem 1.2.5 can not be taken to be equal to 1. This fact is consequential when s_0 does not belong to the approximation class, *i.e.,* when the model is misspecified. This problem also occurred in the l_1-oracle inequalities from Meynet (2013) and Devijver (2015a). Deriving an oracle inequality such that $1 + \kappa^{-1} = 1$, for the Kullback–Leibler loss, is still an open problem. However, one way to handle this difficulty is to use the approximation capacity of the class MoE regression models (cf. Nguyen et al. 2021a, 2019, 2020d,b). Indeed, if we take a large enough number of mixture components, K, we could approximate well a wide class of densities, then the term on the first right-hand side, $\left(1 + \kappa^{-1}\right) \inf_{s_{\boldsymbol{\psi}} \in S} \left(\mathrm{KL}_n\left(s_0, s_{\boldsymbol{\psi}}\right) + \lambda \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right)$, in which S depends on K, is small for K well-chosen. In the next Section 1.3, we study the approximation capabilities of mixtures of experts models in a variety of contexts, including conditional density approximation and approximate Bayesian computation, after providing improvements upon approximation results in the context of unconditional mixture distributions.

1.3 Approximation capabilities of the mixtures of experts models

1.3.1 Finite mixture models

Define $(\mathbb{U}, \|\cdot\|_{\mathbb{U}})$ to be a normed vector space (NVS), and let $x \in \left(\mathbb{R}^d, \|\cdot\|_2\right)$, for some $d \in \mathbb{N}^\star$, where $\|\cdot\|_2$ is the Euclidean norm. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying $f \geq 0$ and $\int f \mathrm{d}\lambda = 1$, where λ is the Lebesgue measure. We say that f is a probability density function (PDF) on the domain $\mathbb{R}^d$ (which we will omit for brevity, from hereon in). Let $g : \mathbb{R}^d \to \mathbb{R}$ be another PDF and define the functional class $\mathcal{M}^g = \bigcup_{m \in \mathbb{N}^\star} \mathcal{M}_m^g$, where

$$\mathcal{M}_m^g = \left\{ h_m^g : h_m^g(\cdot) = \sum_{i=1}^{m} \frac{c_i}{\sigma_i^d} g\left(\frac{\cdot - \mu_i}{\sigma_i}\right), \mu_i \in \mathbb{R}^d, \sigma_i \in \mathbb{R}_+, c \in \Pi_{m-1}, i \in [m] \right\},$$

$c^\top = (c_1, \ldots, c_m)$, $\mathbb{R}_+ = (0, \infty)$, a probability simplex defined by

$$\Pi_{m-1} = \left\{ (\pi_i)_{i \in [m]} \in \mathbb{R}^m | \forall i \in [m], \pi_i > 0, \sum_{i=1}^{m} \pi_i = 1 \right\}, \tag{1.3.1}$$

$[m] = \{1, \ldots, m\}, m \in \mathbb{N}^\star$, and $(\cdot)^\top$ is the matrix transposition operator. We say that any $h_m^g \in \mathcal{M}_m^g$ is a m-component location-scale finite mixture of the PDF g.

The study of PDFs in the class $\mathcal{M}_m^g$ is an evergreen area of applied and technical research, in statistics. We point the interested reader to the many comprehensive books on the topic, such as Everitt & Hand (1981), Titterington et al. (1985), McLachlan & Basford (1988), Lindsay (1995),

McLachlan & Peel (2000), Frühwirth-Schnatter (2006), Schlattmann (2009), Mengersen et al. (2011), and Fruhwirth-Schnatter et al. (2019).

Much of the popularity of finite mixture models stem from the folk theorem, which states that for any density f, there exists an $h \in \mathcal{M}_m^g$, for some sufficiently large number of components $m \in \mathbb{N}^\star$, such that h approximates f arbitrarily closely, in some sense. Examples of this folk theorem come in statements such as: "provided the number of component densities is not bounded above, certain forms of mixture can be used to provide arbitrarily close approximation to a given probability distribution" (Titterington et al., 1985, p. 50), "the [mixture] model forms can fit any distribution and significantly increase model fit" (Walker & Ben-Akiva, 2011, p. 173), and "a mixture model can approximate almost any distribution" (Yona, 2010, p. 500). Other statements conveying the same sentiment are reported in Nguyen & McLachlan (2019). There is a sense of vagary in the reported statements, and little is ever made clear regarding the technical nature of the folk theorem.

In order to proceed, we require the following definitions. We say that f is compactly supported on $\mathbb{K} \subset \mathbb{R}^d$, if $\mathbb{K}$ is compact and if $\mathbf{1}_{\mathbb{K}^\complement} f = 0$, where $\mathbf{1}_{\mathbb{X}}$ is the indicator function that takes value 1 when $x \in \mathbb{X}$ and 0, elsewhere, and $(\cdot)^\complement$ is the set complement operator (i.e., $\mathbb{X}^\complement = \mathbb{R}^d \backslash \mathbb{X}$). Here, $\mathbb{X}$ is a generic subset of $\mathbb{R}^d$. Furthermore, we say that $f \in \mathcal{L}_p(\mathbb{X})$ for any $1 \leq p < \infty$, if

$$\|f\|_{\mathcal{L}_p(\mathbb{X})} = \left(\int |\mathbf{1}_{\mathbb{X}} f|^p \, \mathrm{d}\lambda \right)^{1/p} < \infty,$$

and for $p = \infty$, if

$$\|f\|_{\mathcal{L}_\infty(\mathbb{X})} = \inf \{a \geq 0 : \lambda(\{x \in \mathbb{X} : |f(x)| > a\}) = 0\} < \infty,$$

where we call $\|\cdot\|_{\mathcal{L}_p(\mathbb{X})}$ the $\mathcal{L}_p$-norm on $\mathbb{X}$. When $\mathbb{X} = \mathbb{R}^d$, we shall write $\|\cdot\|_{\mathcal{L}_p(\mathbb{R}^d)} = \|\cdot\|_{\mathcal{L}_p}$. Denote the class of all bounded functions on $\mathbb{X}$ by

$$\mathcal{B}(\mathbb{X}) = \{f \in \mathcal{L}_\infty(\mathbb{X}) : \exists a \in [0, \infty), \text{ such that } |f(x)| \leq a, \forall x \in \mathbb{X}\}$$

and write

$$\|f\|_{\mathcal{B}(\mathbb{X})} = \sup_{x \in \mathbb{X}} |f(x)|.$$

For brevity, we shall write $\mathcal{B}(\mathbb{R}^d) = \mathcal{B}$, and $\|f\|_{\mathcal{B}(\mathbb{R}^d)} = \|f\|_{\mathcal{B}}$.

In addition, we define the so-called Kullback–Leibler divergence, see Kullback & Leibler (1951), between any two PDFs f and g on $\mathbb{X}$ as

$$\mathrm{KL}_{\mathbb{X}}(f, g) = \int \mathbf{1}_{\mathbb{X}} f \log\left(\frac{f}{g}\right) \mathrm{d}\lambda.$$

In Nguyen & McLachlan (2019), the approximation of PDFs f by the class $\mathcal{M}_m^g$ was explored in a restrictive setting. Let $\{h_m^g\}$ be a sequence of functions that draw elements from the nested sequence of sets $\{\mathcal{M}_m^g\}$ (i.e., $h_1^g \in \mathcal{M}_1^g, h_2^g \in \mathcal{M}_2^g, \ldots$). The following result of Zeevi & Meir (1997) was presented in Nguyen & McLachlan (2019), along with a collection of its implications, such as the results of from Li & Barron (1999) and Rakhlin et al. (2005).

Theorem 1.3.1 (Zeevi & Meir, 1997)**.** *If*

$$f \in \{g : \mathbf{1}_{\mathbb{K}} g \geq \beta, \, \beta > 0\} \cap \mathcal{L}_2(\mathbb{K})$$

and g are PDFs and $\mathbb{K}$ is compact, then there exists a sequence $\{h_m^g\}$ such that

$$\lim_{m \to \infty} \|f - h_m^g\|_{\mathcal{L}_2(\mathbb{K})} = 0 \text{ and } \lim_{m \to \infty} KL_{\mathbb{K}}(f, h_m^g) = 0.$$

Although powerful, this result is restrictive in the sense that it only permits approximation in the $\mathcal{L}_2$ norm on compact sets $\mathbb{K}$, and that the result only allows for approximation of functions f that are strictly positive on $\mathbb{K}$. In general, other modes of approximation are desirable, in particular approximation in $\mathcal{L}_p$-norm for $p = 1$ or $p = \infty$ are of interest, where the latter case is generally referred to

as uniform approximation. Furthermore, the strict-positivity assumption, and the restriction on compact sets limits the scope of applicability of Theorem 1.3.1. An example of an interesting application of extensions beyond Theorem 1.3.1 is within the $\mathcal{L}_1$-norm approximation framework of Devroye & Lugosi (2001).

Let $g : \mathbb{R}^d \to \mathbb{R}$ again be a PDF. Then, for each $m \in \mathbb{N}^\star$, we define

$$\mathcal{N}_m^g = \left\{ h : h(x) = \sum_{i=1}^{m} c_i \frac{1}{\sigma_i^d} g\left(\frac{x - \mu_i}{\sigma_i} \right), \mu_i \in \mathbb{R}^d, \sigma_i \in \mathbb{R}_+, c_i \in \mathbb{R}, i \in [m] \right\},$$

which we call the set of m-component location-scale linear combinations of the PDF g. In the past, results regarding approximations of PDFs f via functions $\eta \in \mathcal{N}_m^g$ have been more forthcoming. For example, in the case of $g = \phi$, where

$$\phi(x) = (2\pi)^{-n/2} \exp\left(-\|x\|_2^2 / 2 \right), \tag{1.3.2}$$

is the standard normal PDF. We denote the class of continuous functions and uniformly continuous functions by $\mathcal{C}$ and $\mathcal{C}^u$, respectively. The classes of bounded continuous shall be denoted by $\mathcal{C}_b = \mathcal{C} \cap \mathcal{B}$.

We have the result that for every PDF f, compact set $\mathbb{K} \subset \mathbb{R}^d$, and $\epsilon > 0$, there exists an $m \in \mathbb{N}^\star$ and $h \in \mathcal{N}_m^\phi$, such that $\|f - h\|_{\mathcal{L}_\infty(\mathbb{K})} < \epsilon$ (Sandberg, 2001, Lem. 1). Furthermore, upon defining the set of continuous functions that vanish at infinity by

$$\mathcal{C}_0 = \left\{ f \in \mathcal{C} : \forall \epsilon > 0, \exists \text{ a compact } \mathbb{K} \subset \mathbb{R}^d, \text{ such that } \|f\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} < \epsilon \right\},$$

we also have the result: for every PDF $f \in \mathcal{C}_0$ and $\epsilon > 0$, there exists an $m \in \mathbb{N}^\star$ and $h \in \mathcal{N}_m^\phi$, such that $\|f - h\|_{\mathcal{L}_\infty} < \epsilon$ (Sandberg, 2001, Thm. 2). Both of the results from Sandberg (2001) are simple implications of the famous Stone–Weierstrass theorem (cf. Stone (1948) and De Branges (1959)).

To the best of our knowledge, the strongest available claim that is made regarding the folk theorem, within a probabilistic or statistical context, is that of (DasGupta, 2008, Thm. 33.2). Let $\{\eta_m^g\}$ be a sequence of functions that draw elements from the nested sequence of sets $\{\mathcal{N}_m^g\}$, in the same manner as $\{h_m^g\}$. We paraphrase the claim without loss of fidelity, as follows.

Claim 1.3.2. If $f, g \in \mathcal{C}$ are PDFs and $\mathbb{K} \subset \mathbb{R}^d$ is compact, then there exists a sequence $\{\eta_m^g\}$, such that

$$\lim_{m \to \infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

Unfortunately, the proof of Claim 1.3.2 is not provided within DasGupta (2008). The only reference of the result is to an undisclosed location in Cheney & Light (2000), which, upon investigation, can be inferred to be Theorem 5 of (Cheney & Light, 2000, Ch. 20). It is further notable that there is no proof provided for the theorem. Instead, it is stated that the proof is similar to that of Theorem 1 in (Cheney & Light, 2000, Ch. 24), which is a reproduction of the proof for (Xu et al., 1993, Lem. 3.1).

There is a major problem in applying the proof technique of (Xu et al., 1993, Lem. 3.1) in order to prove Claim 1.3.2. The proof of (Xu et al., 1993, Lem. 3.1) critically depends upon the statement that "there is no loss of generality in assuming that $f(x) = 0$ for $x \in \mathbb{R}^d \backslash 2\mathbb{K}$". Here, for $a \in \mathbb{R}_+$, $a\mathbb{K} = \{x \in \mathbb{R}^d : x = ay, y \in \mathbb{K}\}$. The assumption is necessary in order to write any convolution with f and an arbitrary continuous function as an integral over a compact domain, and then to use a Riemann sum to approximate such an integral. Subsequently, such a proof technique does not work outside the class of continuous functions that are compactly supported on $a\mathbb{K}$. Thus, one cannot verify Claim 1.3.2 from the materials of Xu et al. (1993), Cheney & Light (2000), and DasGupta (2008), alone.

Some recent results in the spirit of Claim 1.3.2 have been obtained by Nestoridis & Stefanopoulos (2007) and Nestoridis et al. (2011), using methods from the study of universal series (see for example in Nestoridis & Papadimitropoulos (2005)).

Let

$$\mathcal{W} = \left\{ f \in \mathcal{C}_0 : \sum_{y \in \mathbb{Z}^d} \sup_{x \in [0,1]^d} |f(x + y)| < \infty \right\}$$

denote the so-called Wiener's algebra (see, e.g., Feichtinger (1977)) and let

$$\mathcal{V} = \left\{ f \in \mathcal{C}_0 : \forall x \in \mathbb{R}^d,\ |f(x)| \leq \beta \left(1 + \|x\|_2\right)^{-d-\theta},\ \beta, \theta \in \mathbb{R}_+ \right\}$$

be a class of functions with tails decaying at a faster rate than $o\left(\|x\|_2^d\right)$. In Nestoridis et al. (2011), it is noted that $\mathcal{V} \subset \mathcal{W}$. Further, let

$$\mathcal{C}_c = \left\{ f \in \mathcal{C} : \exists \text{ a compact set } \mathbb{K}, \text{ such that } \mathbf{1}_{\mathbb{K}^\complement} f = 0 \right\},$$

denote the set of compactly supported continuous functions. The following Theorem 1.3.3 was proved in Nestoridis & Stefanopoulos (2007).

Theorem 1.3.3 (Nestoridis & Stefanopoulos, 2007, Thm. 3.2)**.** *If $g \in \mathcal{V}$, then the following statements hold.*

(a) For any $f \in \mathcal{C}_c$, there exists a sequence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), such that

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_1} + \|f - \eta_m^g\|_{\mathcal{L}_\infty} = 0.$$

(b) For any $f \in \mathcal{C}_0$, there exists a sequence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), such that

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty} = 0.$$

(c) For any $1 \leq p < \infty$ and $f \in \mathcal{L}_p$, there exists a sequence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), such that

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_p} = 0.$$

(d) For any measurable f, there exists a sequence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), such that

$$\lim_{m\to\infty} \eta_m^g = f, \text{ almost everywhere.}$$

(e) If ν is a σ-finite Borel measure on $\mathbb{R}^d$, then for any ν-measurable f, there exists a sequence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), such that

$$\lim_{m\to\infty} \eta_m^g = f,$$

almost everywhere, with respect to ν.

The result was then improved upon, in Nestoridis et al. (2011), whereupon the more general space $\mathcal{W}$ was taken as a replacement for $\mathcal{V}$, in Theorem 1.3.3. Denote the class of bounded continuous functions by $\mathcal{C}_b = \mathcal{C} \cap \mathcal{L}_\infty$. The following theorem was proved in Nestoridis et al. (2011).

Theorem 1.3.4 (Nestoridis et al., 2011, Thm. 3.2)**.** *If $g \in \mathcal{W}$, then the following statements are true.*

(a) The conclusion of Theorem 1.3.3(a) holds, with $\mathcal{C}_c$ replaced by $\mathcal{C}_0 \cap \mathcal{L}_1$.

(b) The conclusions of Theorem 1.3.3(b)–(e) hold.

(c) For any $f \in \mathcal{C}_b$ and compact $\mathbb{K} \subset \mathbb{R}^d$, there exists a sequence $\{\eta_m^g\}$, such that

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

Utilizing the techniques from Nestoridis & Stefanopoulos (2007), Bacharoglou (2010) proved a similar set of results to Theorem 1.3.3, under the restriction that f is a non-negative function with support $\mathbb{R}$, using $g = \phi$ (i.e. g has form (1.3.2), where $d = 1$) and taking $\left\{h_m^\phi\right\}$ as the approximating sequence, instead of $\{\eta_m^g\}$. That is, the following result is obtained.

Theorem 1.3.5 (Bacharoglou, 2010, Cor. 2.5)**.** *If $f : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$, then the following statements are true.*

(a) For any PDF $f \in \mathcal{C}_c$, there exists a sequence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, such that

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_1} + \left\|f - h_m^\phi\right\|_{\mathcal{L}_\infty} = 0.$$

(b) For any $f \in \mathcal{C}_0$, such that $\|f\|_{\mathcal{L}_1} \leq 1$, there exists a sequence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, such that

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_\infty} = 0.$$

(c) For any $1 < p < \infty$ and $f \in \mathcal{C} \cap \mathcal{L}_p$, such that $\|f\|_{\mathcal{L}_1} \leq 1$, there exists a sequence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, such that

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_p} = 0.$$

(d) For any measurable f, there exists a sequence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, such that

$$\lim_{m\to\infty} h_m^\phi = f, \text{ almost everywhere.}$$

(e) For any PDF $f \in \mathcal{C}$, there exists a sequence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, such that

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_1} = 0.$$

Theorem 1.3.5 is restrictive in two ways. First, it does not permit characterization of approximation via the class $\mathcal{M}_m^g$ for any g except the normal PDF ϕ. Although ϕ is traditionally the most common choice for g in practice, the modern mixture model literature has seen the use of many more exotic component PDFs, such as the student-t PDF and its skew and modified variants (see, e.g., Peel & McLachlan, 2000, Forbes & Wraith, 2014, and Lee & McLachlan, 2016). Thus, its use is somewhat limited in the modern context. Furthermore, modern applications tend to call for $d > 1$, further restricting the impact of the result as a theoretical bulwark for finite mixture modeling in practice. A remark in Bacharoglou (2010) states that the result can generalized to the case where $g \in \mathcal{V}$ instead of $g = \phi$. However, no suggestions were proposed, regarding the generalization of Theorem 1.3.5 to the case of $d > 1$.

In Section 2.1, we prove a novel set of results that largely generalize Theorem 1.3.5. Using techniques inspired by Donahue et al. (1997) and Cheney & Light (2000), we are able to obtain a set of results regarding the approximation capability of the class of m-component mixture models $\mathcal{M}_m^g$, when $g \in \mathcal{C}_0$ or $g \in \mathcal{V}$, and for any $d \in \mathbb{N}^\star$. By definition of $\mathcal{V}$, the majority of our results extend beyond the proposed possible generalizations of Theorem 1.3.5.

Motivated by the incomplete proofs of Xu et al. (1993, Lem 3.1) and Theorem 5 from Cheney & Light (2000, Chapter 20), as well as the restricted results of Nestoridis & Stefanopoulos (2007), Bacharoglou (2010), and Nestoridis et al. (2011), in Section 2.1, see also in Nguyen et al. (2020d), we establish and prove Theorem 1.3.6 regarding sequences of PDFs $\{h_m^g\}$ from $\mathcal{M}^g$. Note that Theorem 1.3.6 is restated as Theorem 2.1.1 and proved in Section 2.1.

Theorem 1.3.6 (Nguyen et al., 2020d, Theorem 5)**.** *If we assume that f and g are PDFs and that $g \in \mathcal{C}_0$, then the following statements are true.*

(a) For any $f \in \mathcal{C}_0$, there exists a sequence $\{h_m^g\}$ $(h_m^g \in \mathcal{M}_m^g)$, such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty} = 0.$$

(b) For any $f \in \mathcal{C}_b$ and compact $\mathbb{K} \subset \mathbb{R}^d$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

(c) For any $1 < p < \infty$ and $f \in \mathcal{L}_p$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

(d) For any measurable f, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} h_m^g = f, \text{ almost everywhere.}$$

(e) If ν is a σ-finite Borel measure on $\mathbb{R}^d$, then for any ν-measurable f, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} h_m^g = f,$$

almost everywhere, with respect to ν.

If we assume instead that $g \in \mathcal{V}$, then the following statement is also true.

(f) For any $f \in \mathcal{C}$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_1} = 0.$$

In particular, in Section 2.2, see also in Nguyen et al. (2020b), we establish Theorem 1.3.7 which improves upon Theorem 1.3.6 in a number of ways. More specifically, while statements (a), (d), and (e) still hold under the same assumptions as in Theorem 1.3.6; statement (b) from Theorem 1.3.6 is improved by relaxing the assumption on the PDF g, for further details see statement (a) of Theorem 1.3.7; and statement (c) and (f) from Theorem 1.3.6 is drastically improved to apply to any $f \in \mathcal{L}_1$ and $g \in \mathcal{L}_\infty$, see more in statement (b) of Theorem 1.3.7. The goal of Section 2.2 is to seek the weakest set of assumptions in order to establish approximation theoretical results over the widest class of probability density problems. We note in particular that our improvement with respect to statement (b) from Theorem 1.3.6 yields exactly the result of Theorem 5 from Cheney & Light (2000, Chapter 20), which was incorrectly proved (see also DasGupta, 2008, Theorem 33.2). Moreover, it is good to point out that extending Theorem 1.3.4 (c) from $\mathcal{C}_b$ to $\mathcal{C}$ for f can immediately established as follows: on a compact K, since f is a continuous function, it is always bounded, therefore $\tilde{f} = \min\{f, \sup_K f\}$ is a bounded continuous function which coincides with f on K. So it suffices to apply the result to $\tilde{f}$ to deduce that it is still true with f. The same argument applies to conclusion (b) of Theorem 1.3.6, which is later restated as Theorem 2.1.1. Then, the interest of such a remark is to better demonstrate that what is difficult here to obtain our new Theorem 1.3.7, which is restated as Theorem 2.2.1, it is to relax the assumptions made on g, and not the assumption made on f.

Theorem 1.3.7 (Nguyen et al., 2020b, Theorem 2)**.** *Let $h_m^g \in \mathcal{M}_m^g$ denote an m-component location finite mixture PDF. If we assume that f and g are PDFs, then the following statements are true.*

(a) If $f, g \in \mathcal{C}$ and $\mathbb{K} \subset \mathbb{R}^d$ is a compact set, then there exists a sequence $\{h_m^g\}_{m=1}^\infty \subset \mathcal{M}^g$, such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{B}(\mathbb{K})} = 0.$$

(b) For $p \in [1, \infty)$, if $f \in \mathcal{L}_p$ and $g \in \mathcal{L}_\infty$, then there exists a sequence $\{h_m^g\}_{m=1}^\infty \subset \mathcal{M}^g$, such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

Note that Theorem 1.3.7 is restated as Theorem 2.2.1 and proved in Section 2.2.

1.3.2 Mixture of experts models

Let $\mathbb{W} = \mathbb{Y} \times \mathbb{X}$, where $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{Y} \subseteq \mathbb{R}^q$, for $d, q \in \mathbb{N}^\star$. Suppose that the input and output random variables, $\boldsymbol{X} \in \mathbb{X}$ and $\boldsymbol{Y} \in \mathbb{Y}$, are related via the conditional PDF in the functional class:

$$\mathcal{F} = \left\{ f : \mathbb{W} \to [0, \infty) \,|\, \int_{\mathbb{Y}} f_{|\boldsymbol{X}} (\boldsymbol{y}, \boldsymbol{x}) \, \mathrm{d}\lambda (\boldsymbol{y}) = 1, \forall \boldsymbol{x} \in \mathbb{X} \right\},$$

where λ denotes the Lebesgue measure. The MoE approach seeks to approximate the unknown target conditional PDF f by a function of the MoE form:

$$m_{|\boldsymbol{X}} (\boldsymbol{y}, \boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{Gate}_k (\boldsymbol{x}) \, \mathrm{Expert}_k (\boldsymbol{y}),$$

where $\mathbf{Gate} = (\mathrm{Gate}_k)_{k \in [K]} \in \mathcal{G}^K$ ($[K] = \{1, \ldots, K\}$), $\mathrm{Expert}_1, \ldots, \mathrm{Expert}_K \in \mathcal{E}$, and $K \in \mathbb{N}^\star$. Here, we say that m is a K-component MoE model with gates arising from the class $\mathcal{G}^K$ and experts arising from $\mathcal{E}$, where $\mathcal{E}$ is a class of PDFs with support $\mathbb{Y}$.

The most popular choices for $\mathcal{G}^K$ are the parametric softmax and Gaussian gating classes:

$$\mathcal{G}_S^K = \left\{ \mathbf{Gate} = (\mathrm{Gate}_k (\cdot; \boldsymbol{\gamma}))_{k \in [K]} \,|\forall k \in [K], \mathrm{Gate}_k (\cdot; \boldsymbol{\gamma}) = \frac{\exp \left(a_k + \boldsymbol{b}_k^\top \cdot \right)}{\sum_{l=1}^K \exp \left(a_l + \boldsymbol{b}_l^\top \cdot \right)}, \boldsymbol{\gamma} \in \mathbb{G}_S^K \right\}$$

and

$$\mathcal{G}_G^K = \left\{ \mathbf{Gate} = (\mathrm{Gate}_k (\cdot; \boldsymbol{\gamma}))_{k \in [K]} \,|\forall k \in [K], \mathrm{Gate}_k (\cdot; \boldsymbol{\gamma}) = \frac{\pi_k \phi (\cdot; \boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \phi (\cdot; \boldsymbol{\nu}_l, \boldsymbol{\Sigma}_l)}, \boldsymbol{\gamma} \in \mathbb{G}_G^K \right\},$$

respectively, where

$$\mathbb{G}_S^K = \left\{ \boldsymbol{\gamma} = (a_1, \ldots, a_K, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K) \in \mathbb{R}^K \times \left(\mathbb{R}^d \right)^K \right\}$$

and

$$\mathbb{G}_G^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K) \in \Pi_{K-1} \times \left(\mathbb{R}^d \right)^K \times \mathbb{S}_d^K \right\}.$$

Here,

$$\phi (\cdot; \boldsymbol{\nu}, \boldsymbol{\Sigma}) = |2\pi \boldsymbol{\Sigma}|^{-1/2} \exp \left[-\frac{1}{2} (\cdot - \boldsymbol{\nu})^\top \boldsymbol{\Sigma}^{-1} (\cdot - \boldsymbol{\nu}) \right]$$

is the multivariate normal density function with mean vector $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_K)$ is a vector of weights in the $K-1$ probability simplex Π_{K-1}, defined in (1.3.1), and $\mathbb{S}_d$ is the class of $d \times d$ symmetric positive definite matrices. The softmax and Gaussian gating classes were first introduced by Jacobs et al. (1991) and Xu et al. (1995), respectively. Typically, one chooses experts that arise from some location-scale class:

$$\mathcal{E}_\psi = \left\{ g_\psi (\cdot; \boldsymbol{\mu}, \sigma) : \mathbb{Y} \to [0, \infty) \,|\, g_\psi (\cdot; \boldsymbol{\mu}, \sigma) = \frac{1}{\sigma^q} \psi \left(\frac{\cdot - \boldsymbol{\mu}}{\sigma} \right), \boldsymbol{\mu} \in \mathbb{R}^q, \sigma \in (0, \infty) \right\},$$

where ψ is a PDF, with respect to $\mathbb{R}^q$ in the sense that $\psi : \mathbb{R}^q \to [0, \infty)$ and $\int_{\mathbb{R}^q} \psi (\boldsymbol{y}) \, \mathrm{d}\lambda (\boldsymbol{y}) = 1$.

We shall say that $f \in \mathcal{L}_p (\mathbb{W})$ for any $p \in [1, \infty)$ if

$$\|f\|_{p, \mathbb{W}} = \left(\int_{\mathbb{W}} |\mathbf{1}_{\mathbb{W}} f|^p \, \mathrm{d}\lambda (\boldsymbol{z}) \right)^{1/p} < \infty,$$

where $\mathbf{1}_{\mathbb{W}}$ is the indicator function that takes value 1 when $\boldsymbol{z} \in \mathbb{W}$, and 0 otherwise. Further, we say that $f \in \mathcal{L}_\infty (\mathbb{W})$ if

$$\|f\|_{\infty, \mathbb{W}} = \inf \{ a \geq 0 | \lambda (\{ \boldsymbol{z} \in \mathbb{W} | \, |f (\boldsymbol{z})| > a \}) = 0 \} < \infty.$$

We shall refer to $\|\cdot\|_{p,\mathbb{W}}$ as the $\mathcal{L}_p$ norm on $\mathbb{W}$, for $p \in [0,\infty]$, and where the context is obvious, we shall drop the reference to $\mathbb{W}$.

Suppose that the target conditional PDF f is in the class $\mathcal{F}_p = \mathcal{F} \cap \mathcal{L}_p$. We address the problem of approximating f, with respect to the $\mathcal{L}_p$ norm, using MoE models in the softmax and Gaussian gated classes,

$$\mathcal{M}_S^{\psi} = \Bigg\{ m_K^{\psi} : \mathbb{W} \to [0,\infty) \,|\, m_K^{\psi}(\boldsymbol{y},\boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{Gate}_k(\boldsymbol{x})\, g_{\psi}(\boldsymbol{y};\boldsymbol{\mu}_k,\sigma_k), \\ g_{\psi} \in \mathcal{E}_{\psi} \cap \mathcal{L}_{\infty}, \mathbf{Gate} \in \mathcal{G}_S^K, \boldsymbol{\mu}_k \in \mathbb{Y}, \sigma_k \in (0,\infty), k \in [K], K \in \mathbb{N}^{\star} \Bigg\}, \tag{1.3.3}$$

and

$$\mathcal{M}_G^{\psi} = \Bigg\{ m_K^{\psi} : \mathbb{W} \to [0,\infty) \,|\, m_K^{\psi}(\boldsymbol{y},\boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{Gate}_k(\boldsymbol{x})\, g_{\psi}(\boldsymbol{y};\boldsymbol{\mu}_k,\sigma_k), \\ g_{\psi} \in \mathcal{E}_{\psi} \cap \mathcal{L}_{\infty}, \mathbf{Gate} \in \mathcal{G}_G^K, \boldsymbol{\mu}_k \in \mathbb{Y}, \sigma_k \in (0,\infty), k \in [K], K \in \mathbb{N}^{\star} \Bigg\}, \tag{1.3.4}$$

by showing that both $\mathcal{M}_S^{\psi}$ and $\mathcal{M}_G^{\psi}$ are dense in the class $\mathcal{F}_p$, when $\mathbb{X} = [0,1]^d$ and $\mathbb{Y}$ is a compact subset of $\mathbb{R}^q$. Our denseness results are enabled by the indicator function approximation result of Jiang & Tanner (1999b), and the finite mixture model denseness theorems of Nguyen et al. (2020b) and Nguyen et al. (2020d).

Our Theorems 1.3.8 and 1.3.9, Lemma 1.3.10, and Corollary 1.3.11 contribute to an enduring continuity of sustained interest in the approximation capabilities of MoE models. Related to our results are contributions regarding the approximation capabilities of the conditional expectation function of the classes $\mathcal{M}_S^{\psi}$ and $\mathcal{M}_G^{\psi}$, see definitions in (1.3.3) and (1.3.4), respectively, (Jiang & Tanner, 1999b, Krzyzak & Schafer, 2005, Mendes & Jiang, 2012, Nguyen et al., 2016, 2019, Wang & Mendel, 1992, Zeevi et al., 1998) and the approximation capabilities of subclasses of $\mathcal{M}_S^{\psi}$ and $\mathcal{M}_G^{\psi}$, with respect to the Kullback–Leibler divergence (Jiang & Tanner, 1999a, Norets et al., 2010, Norets & Pelenis, 2014). Our results can be seen as complements to the Kullback–Leibler approximation theorems of Norets et al. (2010) and Norets & Pelenis (2014), by the relationship between the Kullback–Leibler divergence and the $\mathcal{L}_2$ norm (Zeevi & Meir, 1997). That is, when $f > 1/\kappa$, for all $(\boldsymbol{y},\boldsymbol{x}) \in \mathbb{W}$ and some constant $\kappa > 0$, we have that the integrated conditional Kullback–Leibler divergence considered by Norets & Pelenis (2014):

$$\int_{\mathbb{X}} D\left(f_{|\boldsymbol{X}}(\cdot,\boldsymbol{x}) \,\|\, m_K^{\psi}(\cdot,\boldsymbol{x})\right) \mathrm{d}\lambda(\boldsymbol{x}) = \int_{\mathbb{X}} \int_{\mathbb{Y}} f_{|\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) \log \frac{f_{|\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x})}{m_K^{\psi}(\boldsymbol{y},\boldsymbol{x})} \mathrm{d}\lambda(\boldsymbol{y})\, \mathrm{d}\lambda(\boldsymbol{x})$$

satisfies

$$\int_{\mathbb{X}} D\left(f_{|\boldsymbol{X}}(\cdot,\boldsymbol{x}) \,\|\, m_K^{\psi}(\cdot,\boldsymbol{x})\right) \mathrm{d}\lambda(\boldsymbol{x}) \leq \kappa^2 \left\| f - m_K^{\psi} \right\|_{2,\mathbb{W}}^2,$$

and thus a good approximation in the integrated Kullback–Leibler divergence is guaranteed if one can find a good approximation in the $\mathcal{L}_2$ norm, which is guaranteed by our main results. Note that Theorem 1.3.8 is restated as Theorem 2.3.1 and proved in Section 2.3.3.1.

Theorem 1.3.8. *Assume that $\mathbb{X} = [0,1]^d$ for $d \in \mathbb{N}^{\star}$ and $\mathbb{Y}$ is a compact subset in $\mathbb{R}^q$, $q \in \mathbb{N}^{\star}$. For any $f \in \mathcal{F} \cap \mathcal{C}$, any $p \in [1,\infty)$, there exists a sequence $\left\{m_K^{\psi}\right\}_{K \in \mathbb{N}^{\star}} \subset \mathcal{M}_S^{\psi}$, where $\psi \in \mathcal{C}(\mathbb{R}^q)$ is a PDF on support $\mathbb{R}^q$, such that $\lim_{K\to\infty} \left\| f - m_K^{\psi} \right\|_p = 0$.*

Since convergence in Lebesgue spaces does not imply point-wise modes of convergence, the following result is also useful and interesting in some restricted scenarios. Here, we note that the mode of convergence is almost uniform, which implies almost everywhere convergence and convergence in measure (cf. Bartle 1995, Lem 7.10 and Thm. 7.11). The almost uniform convergence of $\left\{m_K^{\psi}\right\}_{K \in \mathbb{N}^{\star}}$

to f in the following result is to be understood in the sense of Bartle (1995, Def. 7.9). That is, for every $\delta > 0$, there exists a set $\mathbb{U}_\delta \subset \mathbb{W}$ with $\lambda(\mathbb{W}) < \delta$, such that $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star}$ converges to f, uniformly on $\mathbb{W}\backslash\mathbb{U}_\delta$. Note that Theorem 1.3.9 is restated as Theorem 2.3.2 and proved in Section 2.3.3.2.

Theorem 1.3.9. *Assume that $\mathbb{X} = [0,1]$ and $\mathbb{Y}$ is a compact subset in $\mathbb{R}^q$, $q \in \mathbb{N}^\star$. For any $f \in \mathcal{F}\cap\mathcal{C}$, there exists a sequence $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star} \subset \mathcal{M}_S^\psi$, where $\psi \in \mathcal{C}(\mathbb{R}^q)$ is a PDF on support $\mathbb{R}^q$, such that $\lim_{K\to\infty} m_K^\psi = f$, almost uniformly.*

The following Lemma 1.3.10 establishes the connection between the gating classes $\mathcal{G}_S^K$ and $\mathcal{G}_G^K$. Note that Lemma 1.3.10 is restated as Lemma 2.3.3 and proved in Section 2.3.4.1.

Lemma 1.3.10. *For each $K \in \mathbb{N}^\star$, $\mathcal{G}_S^K \subset \mathcal{G}_G^K$. Further, if we define the class of Gaussian gating vectors with equal covariance matrices:*

$$\mathcal{G}_E^K = \left\{ \textbf{Gate} = (\text{Gate}_k(\cdot;\boldsymbol{\gamma}))_{k\in[K]} \,|\forall k \in [K], \text{Gate}_k(\cdot;\boldsymbol{\gamma}) = \frac{\pi_k\phi(\cdot;\boldsymbol{\nu}_k,\boldsymbol{\Sigma})}{\sum_{l=1}^K \pi_l\phi(\cdot;\boldsymbol{\nu}_l,\boldsymbol{\Sigma})}, \boldsymbol{\gamma} \in \mathbb{G}_E^K \right\},$$

where

$$\mathbb{G}_E^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi},\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_K,\boldsymbol{\Sigma}) \in \Pi_{K-1} \times \left(\mathbb{R}^d\right)^K \times \mathbb{S}_d \right\},$$

then $\mathcal{G}_E^K \subset \mathcal{G}_S^K$.

By using Lemma 1.3.10, Theorems 1.3.8 and 1.3.9 imply the following Corollary 1.3.11, regarding the approximation capability of the class $\mathcal{M}_G^\psi$. Note that Corollary 1.3.11 is restated as Corollary 2.3.4 in Section 2.3.

Corollary 1.3.11. *Theorems 1.3.8 and 1.3.9 hold when $\mathcal{M}_S^\psi$ is replaced by $\mathcal{M}_G^\psi$ in their statements.*

1.4 Universal approximation for mixture of experts models in approximate Bayesian computation

Approximate Bayesian computation (ABC) (see, *e.g.,* Sisson et al. 2018) appears as a natural candidate for addressing problems, where there is a lack of availability or tractability of the likelihood. Such cases occur when the direct model or data generating process is not available, but is available as a simulation procedure; *e.g.,* when the data generating process is characterized as a series of ordinary differential equations, as in Mesejo et al. (2016), Hovorka et al. (2004). In addition, typical features or constraints that can occur in practice are that: (1) the observations $\mathbf{y}$ are high-dimensional, because they represent signals in time or spectra, as in Schmidt & Fernando (2015), Bernard-Michel et al. (2009), Ma et al. (2013); and (2) the parameter $\boldsymbol{\theta}$, to be estimated, is itself multi-dimensional with correlated dimensions so that independently predicting its components is sub-optimal; *e.g.,* when there are known constraints such as when the parameter elements are concentrations or probabilities that sum to one (Deleforge et al., 2015a, Lemasson et al., 2016, Bernard-Michel et al., 2009).

The fundamental idea of ABC is to generate parameter proposals $\boldsymbol{\theta}$ in a parameter space Θ using a prior distribution $\pi(\boldsymbol{\theta})$ and accept a proposal if the simulated data $\mathbf{z}$ for that proposal is similar to the observed data $\mathbf{y}$, both in an observation space $\mathcal{Y}$. This similarity is usually measured using a distance or discriminative measure D and a simulated sample $\mathbf{z}$ is retained if $D(\mathbf{z},\mathbf{y})$ is smaller than a given threshold ϵ. In this simple form, the procedure is generally referred to as rejection ABC. Other variants are possible and often recommended, for instance using MCMC or sequential procedures (*e.g.*, Del Moral et al., 2012, Buchholz & Chopin, 2019), but we will focus on the rejection version for the purpose of Section 2.4.

In the case of a rejection algorithm, selected samples are drawn from the so-called ABC quasi-posterior, which is an approximation to the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. Under conditions similar to

that of Bernton et al. (2019), regarding the existence of a PDF $f_{\boldsymbol{\theta}}(\mathbf{z})$ for the likelihood, the ABC quasi-posterior depends on D and on a threshold ϵ, and can be written as

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}\,. \tag{1.4.1}$$

More specifically, the similarity between $\mathbf{z}$ and $\mathbf{y}$ is generally evaluated based on two components: the choice of summary statistics $s(\cdot)$ to account for the data in a more robust manner, and the choice of a distance to compare the summary statistics. That is, $D(\mathbf{y},\mathbf{z})$ in (1.4.1) should then be replaced by $D(s(\mathbf{y}), s(\mathbf{z}))$, whereupon we overload D to also denote the distance between summary statistics $s(\cdot)$.

However, there is no general rule for constructing good summary statistics for complex models and if a summary statistic does not capture important characteristics of the data, the ABC algorithm is likely to yield samples from an incorrect posterior (Blum et al., 2013, Fearnhead & Prangle, 2012, Gutmann et al., 2018). Great insight has been gained through the work of Fearnhead & Prangle (2012), who introduced the *semi-automatic* ABC framework and showed that under a quadratic loss, the optimal choice for the summary statistic of $\mathbf{y}$ was the true posterior mean of the parameter: $s(\mathbf{y}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$. This conditional expectation cannot be calculated analytically but can be estimated by regression using a learning data set prior to the ABC procedure itself.

In Fearnhead & Prangle (2012), it is suggested that a simple regression model may be enough to approximate $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$, but this has since been contradicted, for instance by Jiang et al. (2017) and Wiqvist et al. (2019), who show that the quality of the approximation can matter in practice. Still focusing on posterior means as summary statistics, they use deep neural networks that capture complex non-linear relationships and exhibit much better results than standard regression approaches. However, deep neural networks remain very computationally costly tools, both in terms of the required size of training data and number of parameters and hyperparameters to be estimated and tuned.

In Section 2.4, see also Forbes et al. (2021), our first contribution is to investigate an alternative efficient way to construct summary statistics, in the same vein as semi-automatic ABC, but based on posterior moments, not restricted to the posterior means. Although this natural extension was already proposed in Jiang et al. (2017), it requires the availability of a flexible and tractable regression model, able to capture complex non-linear relationships and to provide posterior moments, straightforwardly. As such, Jiang et al. (2017) did not consider an implementation of the procedure. For this purpose, the GLLiM method (Deleforge et al., 2015c), that we recall in Section 2.4.1, appears as a good candidate, with properties that balance between the computationally expensive neural networks and the simple standard regression techniques. In contrast to most regression methods that provide only pointwise predictions, GLLiM provides, at low cost, a parametric estimation of the full true posterior distributions. In particular, we prove universal theorems that the quasi-posterior distribution resulting from ABC with surrogate posteriors built from GLLiM converges to the true one, under standard conditions, see more in Section 2.4.3. Using a learning set of parameters and observations couples, GLLiM learns a family of finite Gaussian mixtures whose parameters depend analytically on the observation to be inverted. For any observed data, the true posterior can be approximated as a Gaussian mixture, whose moments are easily computed in closed form and turned into summary statistics for subsequent ABC sample selection.

More precisely, we provide two types of results, below. In the first result (Theorem 1.4.1), the true posterior is used to compare samples $\mathbf{y}$ and $\mathbf{z}$. This result aims at providing insights on the proposed quasi-posterior formulation and at illustrating its potential advantages. In the second result (Theorem 1.4.2), a surrogate posterior is learned and used to compare samples. Conditions are specified under which the resulting ABC quasi-posterior converges to the true posterior.

1.4.1 Convergence of the ABC quasi-posterior

In this section, we assume a fixed given observed $\mathbf{y}$ and the dependence on $\mathbf{y}$ is omitted from the notation, when there is no confusion.

Let us first recall the standard form of the ABC quasi-posterior, omitting summary statistics from

the notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z}\ . \tag{1.4.2}$$

If D is a distance and $D(\mathbf{y},\mathbf{z})$ is continuous in $\mathbf{z}$, the ABC posterior in (1.4.2) can be shown to have the desirable property of converging to the true posterior when ϵ tends to 0 (see Prangle et al., 2018).

The proof is based on the fact that when ϵ tends to 0, due to the property of the distance D, the set $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y},\mathbf{z}) \leq \epsilon\}$, defining the indicator function in (1.4.2), tends to the singleton $\{\mathbf{y}\}$ so that consequently $\mathbf{z}$ in the likelihood can be replaced by the observed $\mathbf{y}$, which then leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta})f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore to the true posterior as desired (see also Rubio & Johansen, 2013, Bernton et al., 2019). It is interesting to note that this proof is based on working on the term under the integral only and is using the equality, at convergence, of $\mathbf{z}$ to $\mathbf{y}$, which is actually a stronger than necessary assumption for the result to hold. Alternatively, if we first rewrite (1.4.2) using Bayes' theorem, it follows that

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta})\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z}\ \propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}\ . \tag{1.4.3}$$

That is, when accounting for the normalizing constant:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\mathbf{z})\ d\mathbf{z}}\ . \tag{1.4.4}$$

Using this equivalent formulation, we can then replace $D(\mathbf{y},\mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, with D now denoting a distance on densities, and obtain the same convergence result when ϵ tends to 0. More specifically, we can show the following general result. Let us define our ABC quasi-posterior as,

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z},$$

which can be written as

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} \pi\left(\mathbf{z}\right) d\mathbf{z}}\ . \tag{1.4.5}$$

The following Theorem 1.4.1 shows that $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ converges to $\pi\left(\cdot \mid \mathbf{y}\right)$ in total variation, for fixed $\mathbf{y}$. Note that Theorem 1.4.1 is restated as Theorem 2.4.1 and proved in Section 2.4.6.1.

Theorem 1.4.1. *For every $\epsilon > 0$, let $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D\left(\pi\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{z}\right)\right) \leq \epsilon\}$. Assume the following:*

(A1) $\pi\left(\boldsymbol{\theta} \mid \cdot\right)$ is continuous for all $\boldsymbol{\theta} \in \Theta$, and $\sup_{\boldsymbol{\theta}\in\Theta} \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right) < \infty$;

(A2) There exists a $\gamma > 0$ such that $\sup_{\boldsymbol{\theta}\in\Theta} \sup_{\mathbf{z}\in A_\gamma} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) < \infty$;

(A3) $D\left(\cdot,\cdot\right) : \Pi \times \Pi \to \mathbb{R}_+$ is a metric on the functional class $\Pi = \{\pi\left(\cdot \mid \mathbf{y}\right) : \mathbf{y} \in \mathcal{Y}\}$;

(A4) $D\left(\pi\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{z}\right)\right)$ is continuous, with respect to $\mathbf{z}$.

Under (A1)–(A4), $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ in (1.4.5) converges in total variation to $\pi\left(\cdot \mid \mathbf{y}\right)$, for fixed $\mathbf{y}$, as $\epsilon \to 0$.

It appears that what is important is not to select $\mathbf{z}$'s that are close (and at the limit equal) to the observed $\mathbf{y}$ but to choose $\mathbf{z}$'s so that the posterior $\pi(\cdot \mid \mathbf{z})$ (the term appearing in the integral in (1.4.3)) is close (and at the limit equal) to $\pi(\cdot \mid \mathbf{y})$. And this last property is less demanding than $\mathbf{z} = \mathbf{y}$. Potentially, there may be several $\mathbf{z}$'s satisfying $\pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})$, but this is not problematic when using (1.4.3), while it is problematic when following the standard proof as in Bernton et al. (2019).

1.4.2 Convergence of the ABC quasi-posterior with surrogate posteriors

In most ABC settings, based on data discrepancy or summary statistics, the above consideration and result are not useful because the true posterior is unknown by construction and cannot be used to compare samples. However this principle becomes useful in our setting, which is based on surrogate posteriors. While the previous result can be seen as an oracle of sorts, it is more interesting in practice to investigate whether a similar result holds when using surrogate posteriors in the ABC likelihood. This is the goal of Theorem 1.4.2 below, which we prove for a restricted class of target distribution and of surrogate posteriors that are learned as mixtures.

We now assume that $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and consider the following class $\mathcal{H}_{\mathcal{X}}$ of distributions on $\mathcal{X}$, $\mathcal{H}_{\mathcal{X}} = \{g_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \Psi\}$, with constraints on the parameters, Ψ being a bounded parameter set. In addition the densities in $\mathcal{H}_{\mathcal{X}}$ are assumed to satisfy for any $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$, there exist arbitrary positive scalars a, b and B such that

$$\text{for all } \mathbf{x} \in \mathcal{X},\ a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \text{ and } \sup_{\mathbf{x} \in \mathcal{X}} |\log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x})| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1 .$$

We denote by p^K a K-component mixture of distributions from $\mathcal{H}_{\mathcal{X}}$ and defined for all $\mathbf{y} \in \mathcal{Y}$, $p^{K,N}(\cdot \mid \mathbf{y})$ as follows:

$$\forall \boldsymbol{\theta} \in \Theta, \quad p^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p^K\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right),$$

with $\boldsymbol{\phi}^*_{K,N}$ the maximum likelihood estimator (MLE) for the data set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$, generated from the true joint distribution $\pi(\cdot, \cdot)$:

$$\boldsymbol{\phi}^*_{K,N} = \arg\max_{\boldsymbol{\phi} \in \Phi} \sum_{n=1}^{N} \log\left(p^K(\boldsymbol{\theta}_n, \mathbf{y}_n; \boldsymbol{\phi})\right).$$

In addition, for every $\epsilon > 0$, let $A^{K,N}_{\epsilon,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) \leq \epsilon\right\}$ and $q^{K,N}_{\epsilon}$ denote the ABC quasi-posterior defined with $p^{K,N}$ by

$$q^{K,N}_{\epsilon}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z} . \tag{1.4.6}$$

Note that Theorem 1.4.2 is restated as Theorem 2.4.2 and proved in Section 2.4.6.2.

Theorem 1.4.2. *Assume the following: $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and*

(B1) For joint density π, there exists G_π a probability measure on Ψ such that, with $g_{\boldsymbol{\varphi}} \in \mathcal{H}_{\mathcal{X}}$, $\pi(\mathbf{x}) = \int_\Psi g_{\boldsymbol{\varphi}}(\mathbf{x})\ G_\pi(d\boldsymbol{\varphi})$;

(B2) The true posterior density $\pi(\cdot \mid \cdot)$ is continuous both with respect to $\boldsymbol{\theta}$ and $\mathbf{y}$;

(B3) $D(\cdot, \cdot) : \Pi \times \Pi \to \mathbb{R}_+ \cup \{0\}$ is a metric on a functional class Π, which contains the class $\left\{p^{K,N}(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}^\star, N \in \mathbb{N}^\star\right\}$. In particular, $D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) = 0$, if and only if $p^{K,N}(\cdot \mid \mathbf{y}) = p^{K,N}(\cdot \mid \mathbf{z})$;

(B4) For every $\mathbf{y} \in \mathcal{Y}$, $\mathbf{z} \mapsto D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right)$ is a continuous function on $\mathcal{Y}$.

Then, under (B1)–(B4), the Hellinger distance $D_{\mathrm{H}}\left(q^{K,N}_{\epsilon}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converges to 0 in some measure λ, with respect to $\mathbf{y} \in \mathcal{Y}$ and in probability, with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$. That is, for any $\alpha > 0, \beta > 0$, it holds that

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D^2_{\mathrm{H}}\left(q^{K,N}_{\epsilon}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1. \tag{1.4.7}$$

In Section 2.4, our second contribution is to propose to compare directly the full surrogate posterior distributions provided by GLLiM, without reducing them to their moments. So doing, we introduce the idea of functional summary statistics, which also requires a different notion of the usual distances

or discrepancy measures to compare them. Recent developments in optimal transport-based distances designed for Gaussian mixtures (Delon & Desolneux, 2020, Chen et al., 2019) match perfectly this need via the so-called Mixture-Wasserstein distance as referred to in Delon & Desolneux (2020), and denoted throughout the text as MW_2. There exist other distances between mixtures that are tractable, and among them the L_2 distance is also considered in this work.

As an alternative to semi-automatic ABC, in the works of Nguyen et al. (2020a), Jiang et al. (2018), Bernton et al. (2019), Park et al. (2016), Gutmann et al. (2018), the difficulties associated with finding efficient summary statistics were bypassed by adopting, respectively, the Energy Distance, a Kullback–Leibler divergence estimator, the Wasserstein distance, the Maximum Mean Discrepancy (MMD), and classification accuracy to provide a data discrepancy measure. Such approaches compare simulated data and observed data by looking at them as *i.i.d.* samples from distributions, respectively linked to the simulated and true parameter, except for Bernton et al. (2019) and Gutmann et al. (2018) who proposed solutions to also handle time series. We suspect that to be effective, these methods require that the observed and simulated data contain each a moderately large number of samples. Typically, they cannot be applied if we observe only one limited sample related to the parameter to be recovered. This is a major difference with the approach that we propose.

We propose not to compare samples from distributions, but comparing directly the distributions by their surrogates using distances between distributions. It is always possible to use the previous data discrepancies by simulating first samples from the distributions to be compared but this is likely to be computationally sub-optimal. We can instead use the same Wasserstein, Kullback–Leibler divergence, *etc.*, but in their *population* versions rather than in their empirical versions. As an example, a Wasserstein-based distance can be computed between Mixtures of Gaussians, thanks to the recent work of Delon & Desolneux (2020) and Chen et al. (2019). Closed form expressions also exist for the L_2 distance, for the MMD with a Gaussian RBF kernel, or a polynomial kernel (see Sriperumbudur et al., 2010, Muandet et al., 2012) and for the Jensen–Rényi divergence of degree two (see Wang et al., 2009). Kristan et al. (2011) also proposed an algorithm based on the so-called inscented transform in order to compute the Hellinger distance between two Gaussian mixtures, although it is unclear what the complexity of this algorithm is.

To emphasize the difference to more standard summaries, we refer to our surrogate posteriors as functional summary statistics. The term has already been used by Soubeyrand et al. (2013) in the ABC context in their attempts to characterize spatial structures (*e.g.* spatial point processes) using statistics that are functions (*e.g.* correlograms or variograms). Their approach is different in spirit in that it does not address the issue of choosing the summary statistics. Given some functional statistics whose definition and nature may change for each considered model, their goal was to optimize the distances to compare them so as to extract the best information on the parameters of interest. Soubeyrand et al. (2013) propose a weighted L_2 distance to compare such statistics. In our proposal, the functional statistics are probability distributions. They arise as a way to bypass the summary statistics choice, but in this work, we make use of existing metrics to compare them, without optimization.

1.5 Outline and Contributions

The rest of the manuscript is organized as follows. In Chapter 2, we present our first main contributions by establishing theoretical approximation results of MoE models over the widest class of PDFs and conditional PDFs, under the weakest set of assumptions, from the works:

(C1) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *Approximation by finite mixtures of continuous density functions that vanish at infinity.* Cogent Mathematics & Statistics, volume 7, page 1750861. Cogent OA, 2020.
Link: https://www.tandfonline.com/doi/full/10.1080/25742558.2020.1750861
(Nguyen et al., 2020d).

(C2) Hien Duy Nguyen, **TrungTin Nguyen**, Faicel Chamroukhi, and Geoffrey McLachlan. *Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models.* Journal of Statistical Distributions and Applications, 8(1), 13, 2021.

Link: https://doi.org/10.1186/s40488-021-00125-0
(Nguyen et al., 2021a).

(C3) **TrungTin Nguyen**, Faicel Chamroukhi, Hien D Nguyen, and Geoffrey J McLachlan. *Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces.* arXiv preprint arXiv:2008.09787. To appear, Communications in Statistics - Theory and Methods, 2021.
Link: https://arxiv.org/pdf/2008.09787.pdf
(Nguyen et al., 2020b).

We established universal approximation theorems for mixture distributions as well as MoE models. More precisely, we proved that to an arbitrary degree of accuracy, location-scale mixtures of a continuous PDF can approximate any continuous PDF, uniformly, on a compact set; and for any finite $p \geq 1$, location-scale mixtures of an essentially-bounded PDF can approximate any PDF, in the L_p norm. Furthermore, we demonstrated the richness of the class of MoE models by proving denseness results in Lebesgue spaces for conditional PDFs, when the input and output variables are both compactly supported. In another contribution of this thesis subject at large, we considered MoE models in the Bayesian framework. Then, we proved that the quasi-posterior distribution resulting from approximate Bayesian computation (ABC) with surrogate posteriors built from finite Gaussian mixtures using an inverse regression approach, converges to the true one, under standard conditions via the following work:

(C4) Florence Forbes, Hien Duy Nguyen, **TrungTin Nguyen**, and Julyan Arbel. *Approximate Bayesian computation with surrogate posteriors.* hal-03139256. 2021. Link: https://hal.archives-ouvertes.fr/hal-03139256v2/document
(Forbes et al., 2021).

In Chapters 3 and 4, we establish non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds on the penalties hold true, in high-dimensional regression scenarios for a variety of MoE regression models, including Gaussian-gated and softmax-gated Gaussian MoE, based on an inverse regression strategy or a Lasso penalization, respectively. In particular, our oracle inequalities show that the performance in Jensen–Kullback–Leibler type loss of our penalized maximum likelihood estimators are roughly comparable to that of oracle models if we take large enough the constants in front of the penalties, whose forms are only known up to multiplicative constants and proportional to the dimensions of models. These motivate us to make use of the slope heuristic criterion to select several hyperparameters, including the number of mixture components, the amount of sparsity (the coefficients and ranks sparsity levels), the degree of polynomial mean functions, and the potential hidden block-diagonal structures of the covariance matrices of the multivariate predictor or response variable. To support our theoretical results and the statistical study of non-asymptotic model selection in a variety of MoE models, we perform numerical studies by considering simulated and real data, which highlight the performance of our finite-sample oracle inequality results.

In particular, the works from Chapter 3 constitute our second main contributions for the non-asymptotic model selection in a GLoME regression model and a BLoME regression model from the works:

(C5) **TrungTin Nguyen**, Hien Duy Nguyen, Faicel Chamroukhi, and Florence Forbes. *A non-asymptotic penalization criterion for model selection in mixture of experts models.* arXiv preprint arXiv:2104.02640. 2021. Link: https://arxiv.org/pdf/2104.02640.pdf
(Nguyen et al., 2021c).

(C6) **TrungTin Nguyen**, Faicel Chamroukhi, Hien Duy Nguyen, and Florence Forbes. *Non-asymptotic model selection in block-diagonal mixture of polynomial experts models.* arXiv preprint arXiv:2104.08959. 2021. Link: https://arxiv.org/pdf/2104.08959.pdf
(Nguyen et al., 2021b).

Last but not least, our third main contributions for the non-asymptotic joint rank and variable selection results in a PSGaBloME regression models are provided via Chapter 4 from the works:

(C7) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *An l_1-oracle inequality for the lasso in mixture of experts regression models.* arXiv preprint arXiv:2009.10622. 2020. Link: https://arxiv.org/pdf/2009.10622.pdf (Nguyen et al., 2020c).

(C8) *Joint rank and variable selection by a non-asymptotic model selection in mixture of polynomial experts models.* Ongoing work.

Note that these models are useful for high-dimensional heterogeneous data, where the number of explanatory variables can be much larger than the sample size and there exist potential hidden graph-structured interactions between variables.

Finally, Chapter 5 concludes the manuscript and discusses perspectives. In particular, we suggest several conjectures and open problems as future research directions.

Chapter 2

Approximation capabilities of the mixtures of experts models

Chapter 2 is based on the following works:

(C1) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *Approximation by finite mixtures of continuous density functions that vanish at infinity.* Cogent Mathematics & Statistics, volume 7, page 1750861. Cogent OA, 2020.
Link: https://www.tandfonline.com/doi/full/10.1080/25742558.2020.1750861
(Nguyen et al., 2020d).

(C2) Hien Duy Nguyen, **TrungTin Nguyen**, Faicel Chamroukhi, and Geoffrey McLachlan. *Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models.* Journal of Statistical Distributions and Applications, 8(1), 13, 2021. Link: https://doi.org/10.1186/s40488-021-00125-0
(Nguyen et al., 2021a).

(C3) **TrungTin Nguyen**, Faicel Chamroukhi, Hien D Nguyen, and Geoffrey J McLachlan. *Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces.* arXiv preprint arXiv:2008.09787. To appear, Communications in Statistics - Theory and Methods, 2021. Link: https://arxiv.org/pdf/2008.09787.pdf
(Nguyen et al., 2020b).

(C4) Florence Forbes, Hien Duy Nguyen, **TrungTin Nguyen**, and Julyan Arbel. *Approximate Bayesian computation with surrogate posteriors.* hal-03139256. 2021. Link: https://hal.archives-ouvertes.fr/hal-03139256v2/document
(Forbes et al., 2021).

Contents

In Chapter 1, we introduced the necessary concepts and main results of the approximation and model selection capabilities of the MoE models. In Chapter 2, we present our first contributions of the approximation capabilities within these models in more details compared to Section 1.3. More precisely, we aim to prove that to an arbitrary degree of accuracy, location-scale mixtures of a continuous PDF can approximate any continuous PDF, uniformly, on a compact set; and for any finite $p \geq 1$, location-scale mixtures of an essentially-bounded PDF can approximate any PDF, in the L_p norm. Moreover, given input and output variables are both compactly supported, we demonstrate the richness of the class of MoE models by proving denseness results in Lebesgue spaces for conditional PDFs. In another contribution of this Ph.D. subject at large, we considered MoE models in the Bayesian framework. Then, we proved that the quasi-posterior distribution resulting from approximate Bayesian computation (ABC) with surrogate posteriors built from finite Gaussian mixtures using an inverse regression approach, converges to the true one, under standard conditions.

Chapter 2 proceeds as follows. Our main theorems regarding the approximation by finite mixtures of continuous density functions that vanish at infinity and location-scale finite mixtures in Lebesgue spaces are stated and proved in the Section 2.1 and Section 2.2, respectively. In particular, all the proofs of Theorem 2.1.1 (or equivalently Theorem 1.3.6)(a)+(d)+(e), which are later strictly improved by Theorem 2.2.1 (or equivalently Theorem 1.3.7), are not presented in Section 2.1 and can be found in Nguyen et al. (2020d). These universal approximation results are then extended to conditional PDFs via MoE models in Lebesgue spaces in Section 2.3. Universal approximation for MoE in approximate Bayesian computation is presented in Section 2.4.

2.1 Approximation by finite mixtures of continuous density functions that vanish at infinity

The remainder of Section 2.1 is devoted to proving the following Theorem 2.1.1.

Theorem 2.1.1 (Nguyen et al., 2020d, Theorem 5)**.** *If we assume that f and g are PDFs and that $g \in \mathcal{C}_0$, then the following statements are true.*

(a) For any $f \in \mathcal{C}_0$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty} = 0.$$

(b) For any $f \in \mathcal{C}_b$ and compact $\mathbb{K} \subset \mathbb{R}^d$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

(c) For any $1 < p < \infty$ and $f \in \mathcal{L}_p$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

(d) For any measurable f, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} h_m^g = f, \text{ almost everywhere.}$$

(e) If ν is a σ-finite Borel measure on $\mathbb{R}^d$, then for any ν-measurable f, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} h_m^g = f,$$

almost everywhere, with respect to ν.

If we assume instead that $g \in \mathcal{V}$, then the following statement is also true.

(f) For any $f \in \mathcal{C}$, there exists a sequence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_1} = 0.$$

2.1.1 Technical preliminaries

Before we begin to prove the main theorem, we establish some technical results regarding our class of component densities $\mathcal{C}_0$. Let $f, g \in \mathcal{L}_1$ and denote the convolution of f and g by $f \star g = g \star f$. Further, we denote the sequence of dilates of g by $\left\{g_k : g_k(x) = k^d g(kx), k \in \mathbb{N}^\star\right\}$. The following result is an alternative to Lemma 2.1.6 and Corollary 2.1.7. Here, we replace a boundedness assumption on the approximand, in the aforementioned theorem by a vanishing at infinity assumption, instead.

Lemma 2.1.2. *Let g be a PDF and $f \in \mathcal{C}_0$. Then,*

$$\lim_{k\to\infty} \|g_k \star f - f\|_{\mathcal{L}_\infty} = 0.$$

Proof. It suffices to show that for any $\epsilon > 0$, there exists a $k(\epsilon) \in \mathbb{N}^\star$, such that $\|g_k \star f - f\|_{\mathcal{L}_\infty} < \epsilon$, for all $k \geq k(\epsilon)$. By Lemma 2.1.8, $f \in \mathcal{C}_b$, and thus $\|f\|_{\mathcal{L}_\infty} < \infty$. By making the substitution $z = kx$, we obtain for each k

$$\int g_k(x)\,\mathrm{d}\lambda = \int k^d g(kx)\,\mathrm{d}\lambda = \int g(z)\,\mathrm{d}\lambda = 1.$$

By Corollary 2.1.7, we obtain $\lim_{k\to\infty} \int \mathbf{1}_{\{x:\|x\|_2>\delta\}} g_k \mathrm{d}\lambda = 0$ and thus we can choose a $k(\epsilon)$, such that

$$\int \mathbf{1}_{\{x:\|x\|_2>\delta\}} g_k \mathrm{d}\lambda < \frac{\epsilon}{4\|f\|_{\mathcal{L}_\infty}}.$$

Since g is a PDF, we have

$$\begin{aligned}|(g_k \star f)(x) - f(x)| &= \left|\int g_k(y)\left[f(x-y) - f(x)\right]\mathrm{d}\lambda(y)\right| \\ &\leq \int g_k(y)\,|f(x-y) - f(x)|\,\mathrm{d}\lambda(y).\end{aligned}$$

By uniform continuity, for any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that $|f(x-y) - f(x)| < \epsilon/2$, for any $x, y \in \mathbb{R}^d$, such that $\|y\|_2 < \delta(\epsilon)$ (Lemma 2.1.8). Thus, on the one hand, for any $\delta(\epsilon)$, we can pick a $k(\epsilon)$ such that

$$\begin{aligned}\int \mathbf{1}_{\{y:\|y\|_2>\delta(\epsilon)\}} g_k(y)\,|f(x-y) - f(x)|\,\mathrm{d}\lambda(y) &\leq 2\|f\|_{\mathcal{L}_\infty} \int \mathbf{1}_{\{y:\|y\|_2>\delta(\epsilon)\}} g_k \mathrm{d}\lambda \\ &\leq 2\|f\|_{\mathcal{L}_\infty} \times \frac{\epsilon}{4\|f\|_{\mathcal{L}_\infty}} = \frac{\epsilon}{2},\end{aligned} \tag{2.1.1}$$

and on the other hand

$$\begin{aligned}\int \mathbf{1}_{\{y:\|y\|_2\leq\delta(\epsilon)\}} g_k(y)\,|f(x-y)-f(x)|\,\mathrm{d}\lambda(y) &\leq \frac{\epsilon}{2}\int \mathbf{1}_{\{y:\|y\|_2\leq\delta(\epsilon)\}} g_k \mathrm{d}\lambda \\ &\leq \frac{\epsilon}{2}\times 1 = \frac{\epsilon}{2}. \end{aligned} \tag{2.1.2}$$

The proof is completed by summing (2.1.1) and (2.1.2). □

Lemma 2.1.3. *If $f \in \mathcal{C}_0$ is such that $f \geq 0$, and $\epsilon > 0$, then there exists a $h \in \mathcal{C}_c$, such that $0 \leq h \leq f$, and*

$$\|f-h\|_{\mathcal{L}_\infty} < \epsilon$$

Proof. Since $f \in \mathcal{C}_0$, there exists a compact $\mathbb{K} \subset \mathbb{R}^d$ such that $\|f\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} < \epsilon/2$. By Lemma 2.1.9, there exists some $g \in \mathcal{C}_c$, such that $0 \leq g \leq 1$ and $\mathbf{1}_{\mathbb{K}} g = 1$. Let $h = gf$, which implies that $h \geq 0$ and $0 \leq h \leq f$. Furthermore, notice that $\mathbf{1}_{\mathbb{K}}(f-h) = 0$ and $\|h\|_{\mathcal{L}_\infty} \leq \|f\|_{\mathcal{L}_\infty}$, by construction. The proof is completed by observing that

$$\begin{aligned}\|f-h\|_{\mathcal{L}_\infty} &= \|f-h\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} \\ &\leq \|f\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} + \|h\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} \\ &\leq 2\|f\|_{\mathcal{L}_\infty(\mathbb{K}^\complement)} < \epsilon. \end{aligned}$$

□

For any $\delta > 0$, uniformly continuous function f, let

$$w(f,\delta) = \sup_{\{x,y\in\mathbb{R}^d:\|x-y\|_2\leq\delta\}} |f(x)-f(y)|$$

denote the modulus of continuity of f. Furthermore, define the diameter of a set $\mathbb{X} \subset \mathbb{R}^d$ by $\mathrm{diam}(\mathbb{X}) = \sup_{x,y\in\mathbb{X}} \|x-y\|_2$ and denote an open ball, centered at $x \in \mathbb{R}^d$ with radius $r > 0$ by $\mathbb{B}(x,r) = \{y \in \mathbb{R}^d : \|x-y\|_2 < r\}$.

Notice that the class $\mathcal{M}_m^g$ can be parameterized as

$$\mathcal{M}_m^g = \left\{ h : h(x) = \sum_{i=1}^m c_i k_i^d g(k_i x - z_i),\, z_i \in \mathbb{R}^d,\, k_i \in \mathbb{R}_+,\, c \in \Pi_{m-1},\, i \in [m] \right\},$$

where $k_i = 1/\sigma_i$ and $z_i = \mu_i/\sigma_i$. The following result is the primary mechanism that permits us to construct finite mixture approximations for convolutions of form $g_k \star f$. The argument motivated by the approaches taken in (Cheney & Light, 2000, Thm 1, Ch. 24), (Nestoridis & Stefanopoulos, 2007, Lem. 3.1), and (Nestoridis et al., 2011, Thm. 3.1).

Lemma 2.1.4. *Let $f \in \mathcal{C}$ and $g \in \mathcal{C}_0$ be PDFs. Furthermore, let $\mathbb{K} \subset \mathbb{R}^d$ be compact and $h \in \mathcal{C}_c$, where $\mathbf{1}_{\mathbb{K}^\complement} h = 0$ and $0 \leq h \leq f$. Then for any $k \in \mathbb{N}^\star$, there exists a sequence $\{h_m^g\}$, such that*

$$\lim_{m\to\infty} \|g_k \star h - h_m^g\|_{\mathcal{L}_\infty} = 0.$$

Proof. It suffices to show that for any $k \in \mathbb{N}^\star$ and $\epsilon > 0$, there exists a sufficiently large enough $m(\epsilon) \in \mathbb{N}^\star$ so that for all $m \geq m(\epsilon), h_m^g \in \mathcal{M}_m^g$ such that

$$\|g_k \star h - h_m^g\|_{\mathcal{L}_\infty} < \epsilon. \tag{2.1.3}$$

For any $k \in \mathbb{N}^\star$, we can write

$$\begin{aligned}(g_k \star h)(x) &= \int g_k(x-y)\,h(y)\,\mathrm{d}\lambda(y) \\ &= \int \mathbf{1}_{\{y:y\in\mathbb{K}\}} g_k(x-y)\,h(y)\,\mathrm{d}\lambda(y) \\ &= \int \mathbf{1}_{\{y:y\in\mathbb{K}\}} k^d g(kx-ky)\,h(y)\,\mathrm{d}\lambda(y) \\ &= \int \mathbf{1}_{\{z:z\in k\mathbb{K}\}} g(kx-z)\,h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z). \end{aligned}$$

Here, $k\mathbb{K}$ is continuous image of a compact set, and hence is compact (cf. (Rudin, 1976, Thm. 4.14)). By Lemma 2.1.10, for any $\delta > 0$, there exists $\kappa_i \in \mathbb{R}^d$ $(i \in [m-1],\ m \in \mathbb{N}^\star)$, such that $k\mathbb{K} \subset \bigcup_{i=1}^{m-1} \mathbb{B}(\kappa_i, \delta/2)$. Further, if $\mathbb{B}_i^\delta = k\mathbb{K} \cap \mathbb{B}(\kappa_i, \delta/2)$, then we have $k\mathbb{K} = \bigcup_{i=1}^{m-1} \mathbb{B}_i^\delta$. We can obtain a disjoint covering of $k\mathbb{K}$ by taking $\mathbb{A}_1^\delta = \mathbb{B}_1$ and $\mathbb{A}_i^\delta = \mathbb{B}_i^\delta \backslash \bigcup_{j=1}^{i-1} \mathbb{B}_j^\delta$ $(i \in [m-1])$ and noting that $k\mathbb{K} = \bigcup_{i=1}^{m-1} \mathbb{A}_i^\delta$, by construction (cf. (Cheney & Light, 2000, Ch. 24)). Furthermore, each $\mathbb{A}_i^\delta$ is a Borel set and $\operatorname{diam}\left(\mathbb{A}_i^\delta\right) \leq \delta$.

For convenience, let $\Pi_m^\delta = \left\{\mathbb{A}_i^\delta : i \in [m-1]\right\}$ denote the disjoint covering, or partition, of $k\mathbb{K}$. We seek to show that there exists an $m \in \mathbb{N}^\star$ and Π_m^δ, such that

$$\left\| g_k \star h - \sum_{i=1}^{m} c_i k_i^d g\left(k_i x - z_i\right) \right\|_{\mathcal{L}_\infty} < \epsilon,$$

where $k_i = k$,

$$c_i = k^{-d} \int \mathbf{1}_{\left\{z : z \in \mathbb{A}_i^\delta\right\}} h\left(z/k\right) \mathrm{d}\lambda(z),$$

and $z_i \in \mathbb{A}_i^\delta$, for $i \in [m-1]$.

Further, $z_m \in \mathbb{A}_{m-1}^\delta$ and $c_m = 1 - \sum_{i=1}^{m-1} c_i$, with k_m chosen as follows. By Lemma 2.1.8, $g \leq C < \infty$ for some positive C. Then, $\left\| c_m k_m^d g\left(k_m x - z_m\right) \right\|_{\mathcal{L}_\infty} \leq c_m k_m^d C$. We may choose k_m so that $k_m^d = \epsilon / \left(2 c_m C\right)$, so that

$$\left\| c_m k_m^d g\left(k_m x - z_m\right) \right\|_{\mathcal{L}_\infty} \leq \frac{\epsilon}{2}.$$

Since $0 \leq h \leq f$, the sum of c_i $(i \in [m-1])$ satisfies the inequality

$$\begin{aligned} \sum_{i=1}^{m-1} c_i &= k^{-d} \sum_{i=1}^{m-1} \int \mathbf{1}_{\left\{z : z \in \mathbb{A}_i^\delta\right\}} h\left(\frac{z}{k}\right) \mathrm{d}\lambda = k^{-d} \int \mathbf{1}_{\{z : z \in k\mathbb{K}\}} h\left(\frac{z}{k}\right) \mathrm{d}\lambda \\ &= \int \mathbf{1}_{\{x : x \in \mathbb{K}\}} h \mathrm{d}\lambda \leq \int \mathbf{1}_{\{x : x \in \mathbb{K}\}} f \mathrm{d}\lambda \leq \int f \mathrm{d}\lambda = 1. \end{aligned}$$

Thus, $0 \leq c_m \leq 1$, and our construction implies that $h_m^g \in \mathcal{M}_m^g$, where

$$h_m^g(x) = \sum_{i=1}^{m} c_i k_i^d g\left(k_i x - z_i\right) \forall x \in \mathbb{R}^n.$$

We can bound the left-hand side of (2.1.3) as follows:

$$\begin{aligned} \left\| g_k \star h - h_g^m \right\|_{\mathcal{L}_\infty} &\leq \left\| \left(g_k \star h\right)(x) - \sum_{i=1}^{m-1} c_i k_i^d g\left(k_i x - z_i\right) \right\|_{\mathcal{L}_\infty} + \left\| c_m k_m^d g\left(k_m x - z_m\right) \right\|_{\mathcal{L}_\infty} \\ &\leq \left\| \left(g_k \star h\right)(x) - \sum_{i=1}^{m-1} c_i k_i^d g\left(k_i x - z_i\right) \right\|_{\mathcal{L}_\infty} + \frac{\epsilon}{2} \\ &= \left\| \int \mathbf{1}_{\{z : z \in k\mathbb{K}\}} g\left(kx - z\right) h\left(\frac{z}{k}\right) \mathrm{d}\lambda(z) \right. \\ &\quad \left. - \sum_{i=1}^{m-1} \int \mathbf{1}_{\left\{z : z \in \mathbb{A}_i^\delta\right\}} g\left(kx - z_i\right) h\left(\frac{z}{k}\right) \mathrm{d}\lambda(z) \right\|_{\mathcal{L}_\infty} + \frac{\epsilon}{2} \\ &\leq \sum_{i=1}^{m-1} \int \mathbf{1}_{\left\{z : z \in \mathbb{A}_i^\delta\right\}} \left\| g\left(kx - z\right) - g\left(kx - z_i\right) \right\|_{\mathcal{L}_\infty} h\left(\frac{z}{k}\right) \mathrm{d}\lambda(z) + \frac{\epsilon}{2}. \end{aligned} \tag{2.1.4}$$

Since

$$\left\| kx - z - \left(kx - z_i\right) \right\|_2 = \left\| z - z_i \right\|_2 \leq \operatorname{diam}\left(\mathbb{A}_i^\delta\right) \leq \delta,$$

we have $\left| g\left(kx - z\right) - g\left(kx - z_i\right) \right| \leq w(g, \delta)$, for each $i \in [m-1]$. Since $\lim_{\delta \to 0} w(g, \delta) = 0$ (cf. (Makarov & Podkorytov, 2013, Thm. 4.7.3)), we may choose a $\delta(\epsilon) > 0$ so that $w\left(g, \delta(\epsilon)\right) < \epsilon / \left(2k^d\right)$.

We may proceed from (2.1.4) as follows:

$$\begin{aligned}\left\|g_k \star h - h_g^m\right\|_{\mathcal{L}_\infty} &\leq w\left(g, \delta\left(\epsilon\right)\right) \int \mathbf{1}_{\{z: z \in k\mathbb{K}\}} h\left(\frac{z}{k}\right) \mathrm{d}\lambda + \frac{\epsilon}{2} \\ &= w\left(g, \delta\left(\epsilon\right)\right) k^d \int h \mathrm{d}\lambda + \frac{\epsilon}{2} \\ &\leq w\left(g, \delta\left(\epsilon\right)\right) k^d + \frac{\epsilon}{2} \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned} \tag{2.1.5}$$

To conclude the proof, it suffices to choose an appropriate sequence of partitions $\Pi_m^{\delta(\epsilon)}, m \geq m(\epsilon)$, for some large but finite $m(\epsilon)$, so that (2.1.4) and (2.1.5) hold, which is possible by Lemma 2.1.10. □

For any $r \in \mathbb{N}^\star$, let $\bar{\mathbb{B}}_r = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq r\right\}$ be a closed ball of radius r, centered at the origin.

Lemma 2.1.5. *If $f \in \mathcal{L}_1$, such that $f \geq 0$, then*

$$\lim_{r \to \infty} \left\|f - \mathbf{1}_{\bar{\mathbb{B}}_r} f\right\|_{\mathcal{L}_1} = 0.$$

Proof. By construction, each element of the sequence $\left\{\mathbf{1}_{\bar{\mathbb{B}}_r} f\right\}$ $(r \in \mathbb{N}^\star)$ is measurable, $0 \leq \mathbf{1}_{\bar{\mathbb{B}}_r} f \leq f$, and

$$\lim_{r \to \infty} \mathbf{1}_{\bar{\mathbb{B}}_r} f = f,$$

point-wise. We obtain our conclusion via the Lebesgue dominated convergence theorem. □

2.1.2 Proof of Theorem 2.1.1(a)

We now proceed to prove each of the parts of Theorem 2.1.1. To prove Theorem 2.1.1(a) it suffices to show that for every $\epsilon > 0$, there exists a $h_m^g \in \mathcal{M}_m^g$, such that $\|f - h_m^g\|_{\mathcal{L}_\infty} < \epsilon$.

Start by applying Lemma 2.1.3 to obtain $h \in \mathcal{C}_c$, such that $0 \leq h \leq f$ and $\|f - h\|_{\mathcal{L}_\infty} < \epsilon/2$. Then, we have

$$\begin{aligned}\|f - h_m^g\|_{\mathcal{L}_\infty} &\leq \|f - h\|_{\mathcal{L}_\infty} + \|h - h_m^g\|_{\mathcal{L}_\infty} \\ &< \frac{\epsilon}{2} + \|h - h_m^g\|_{\mathcal{L}_\infty}. \end{aligned} \tag{2.1.6}$$

The goal is to find a h_m^g, such that $\|h - h_m^g\|_{\mathcal{L}_\infty} < \epsilon/2$. Since $h \in \mathcal{C}_c$, we may find a compact $\mathbb{K} \subset \mathbb{R}^d$ such that $\|h\|_{\mathcal{L}_\infty\left(\mathbb{K}^\complement\right)} = 0$. Apply Lemma 2.1.2 to show the existence of a $k\left(\epsilon\right)$, such that

$$\|h - g_k \star h\|_{\mathcal{L}_\infty} < \frac{\epsilon}{4},$$

for all $k \geq k\left(\epsilon\right)$. With a fixed $k = k\left(\epsilon\right)$, apply Lemma 2.1.4 to show that there exists a $h_m^g \in \mathcal{M}_m^g$, such that

$$\left\|g_{k(\epsilon)} \star h - h_m^g\right\|_{\mathcal{L}_\infty} < \frac{\epsilon}{4}.$$

By the triangle inequality, we have

$$\begin{aligned}\|h - h_m^g\|_{\mathcal{L}_\infty} &\leq \left\|h - g_{k(\epsilon)} \star h\right\|_{\mathcal{L}_\infty} + \left\|g_{k(\epsilon)} \star h - h_m^g\right\|_{\mathcal{L}_\infty} \\ &< \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}. \end{aligned} \tag{2.1.7}$$

The proof is complete by substitution of (2.1.7) into (2.1.6).

2.1.3 Proof of Theorem 2.1.1(d) and Theorem 2.1.1(e)

A combination of Lusin's theorem and Urysohn's lemma, see more detail in Lemma 2.1.9, renders a sequence of continuous functions with compact support $\left(g_k\right)_{k \in \mathbb{N}^\star}$ converging to f almost everywhere.

By Theorem 2.1.1(a), for each $k \in \mathbb{N}^\star$, there exists a sequence $\{h_{m_k}^g\}$ that uniformly converges to g_k, that is, $\|h_{m_k}^g - g_k\|_\infty < 1/k, \forall k \in \mathbb{N}^\star$. Thus, by Lemma 2.1.14, $\{h_{m_k}^g\}$ almost uniformly converges to g_k and also converges almost everywhere, to g_k, with respect to any measure ν. We prove Theorem 2.1.1(d) by setting $\nu = \lambda$, and we prove Theorem 2.1.1(e) by not specifying ν.

2.1.4 Comments and discussion

2.1.4.1 Relationship to Theorem 1.3.1

In the proof of Theorem 1.3.1, the famous Hilbert space approximation result of Jones (1992) and Barron (1993) was used to bound the $\mathcal{L}_2$ norm between any approximand $f \in \mathcal{L}_2$ and a convex combination of bounded functions in $\mathcal{L}_2$. This approximation theorem is exactly the $p = 2$ case of the more general theorem of Donahue et al. (1997), as presented in Lemma 2.1.13. Thus, one can view Theorem 2.1.1(c) as the $p \in (1, \infty)$ generalization of Theorem 1.3.1.

2.1.4.2 The class $\mathcal{W}$ is a proper subset of the class $\mathcal{C}_0$

Here, we comment on the nature of class $\mathcal{W}$, which was investigated by Bacharoglou (2010) and Nestoridis et al. (2011). We recall that Bacharoglou (2010) conjectured that Theorem 1.3.5 generalizes from $g = \phi$ to $g \in \mathcal{V}$. In Theorem 2.1.1(a)–(e), we assume that $g \in \mathcal{C}_0$. We can demonstrate that $g \in \mathcal{C}_0$ is a strictly weaker condition than $g \in \mathcal{V}$ or $g \in \mathcal{W}$.

For example, consider the function in $g : \mathbb{R} \to \mathbb{R}$ such that $g(x) = 0$ if $x < 0$ and

$$g(x) = \sum_{i=1}^{\infty} \frac{2^{2i}}{i} \left[(x - i + 1)^{2i} \mathbf{1}_{\{i-1 \leq x < i-1/2\}} + (x - i)^{2i} \mathbf{1}_{\{i-1/2 \leq x < i\}} \right] \text{ if } x \geq 0,$$

and note that

$$\int \mathbf{1}_{(-1/2,1/2)} \frac{(2x)^{2i}}{i} \mathrm{d}\lambda = \frac{1}{2i^2 + i} < \frac{1}{i^2}.$$

Since $\sum_{i=1}^{\infty} (1/i^2) = \pi^2/6$, $g \in \mathcal{L}_1$. Furthermore, g is continuous since all stationary points of g are continuous. In $\mathbb{R}$, $g \in \mathcal{C}_0$ if

$$\lim_{x \to \pm\infty} g(x) = 0.$$

For $x \leq 0$, we observe that $g = 0$ and thus the left limit is satisfied. On the right, for any $1/\epsilon > 0$, we have $x(\epsilon) \geq \lceil \epsilon \rceil - 1/2$, so that $g(x) < 1/\epsilon$, for all $x > x(\epsilon)$, where $\lceil \cdot \rceil$ is the ceiling operator. Therefore, $g \in \mathcal{C}_0$.

Within each interval $i - 1 \leq x < i$, we observe that g is locally maximized at $x = i - 1/2$. The local maximum corresponding to each of these points is $1/i$. Thus $g \notin \mathcal{W}$, since

$$\sum_{i=1}^{\infty} \frac{1}{i} < \sum_{y \in \mathbb{Z}} \sup_{x \in [0,1]} |g(x + y)|,$$

where $\sum_{i=1}^{\infty} (1/i) = \infty$. Furthermore, $g \notin \mathcal{V}$ since $\mathcal{V} \subset \mathcal{W}$.

2.1.4.3 Convergence in measure

Along with the conclusions of Theorem 2.1.1(d) and (e), Lemma 2.1.14 also implies convergence in measure. That is, if ν is a σ-finite Borel measure on $\mathbb{R}^d$, then for any ν-measurable f, there exists a sequence $\{h_m^g\}$, such that for any $\epsilon > 0$,

$$\lim_{m \to \infty} \upsilon \left(\left\{ x \in \mathbb{R}^d : |f(x) - h_m^g(x)| \geq \epsilon \right\} \right) = 0.$$

2.1.5 Technical results

Throughout Section 2.1, we utilize a number of established technical results. For the convenience of the reader, we append these results within Section 2.1.5. Sources from which we draw the unproved results are provided in Section 2.1.6.

Lemma 2.1.6. *Let $\{g_k\}$ be a sequence of PDFs in $\mathcal{L}_1$ and for every $\delta > 0$*

$$\lim_{k\to\infty} \int \mathbf{1}_{\{x:\|x\|_2>\delta\}} g_k \, d\lambda = 0.$$

Then, for all $f \in \mathcal{L}_p$ and $1 \leq p < \infty$,

$$\lim_{k\to\infty} \|g_k \star f - f\|_{\mathcal{L}_p} = 0.$$

Furthermore, for all $f \in \mathcal{C}_b$ and any compact $\mathbb{K} \subset \mathbb{R}^d$,

$$\lim_{k\to\infty} \|g_k \star f - f\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

The sequences $\{g_k\}$ from Lemma 2.1.6 are often called approximate identities or approximations of the identity. A simple construction of approximate identities is by taking dilations $g_k(x) = k^d g(kx)$, which yields the following corollary.

Corollary 2.1.7. *Let g be a PDF. Then the sequence of dilations $\{g_k : g_k(x) = k^d g(kx)\}$, satisfies the hypothesis of Lemma 2.1.6 and hence permits its conclusion.*

Lemma 2.1.8. *The class $\mathcal{C}_0$ is a subset of $\mathcal{C}_b$. Furthermore, if $f \in \mathcal{C}_0$, then f is uniformly continuous.*

Lemma 2.1.9 (Urysohn's Lemma)**.** *If $\mathbb{K} \subset \mathbb{R}^d$ is compact, then there exists some $g \in \mathcal{C}_c$, such that $0 \leq g \leq 1$ and $\mathbf{1}_{\mathbb{K}} g = 1$.*

Lemma 2.1.10. *If $\mathbb{X} \subset \mathbb{R}^d$ is bounded, then for any $r > 0$, $\mathbb{X}$ can be covered by $\bigcup_{i=1}^m \mathbb{B}(x_i, r)$ for some finite $m \in \mathbb{N}^\star$, where $x_i \in \mathbb{R}^d$ and $i \in [m]$.*

Lemma 2.1.11. *If $0 < p < q < r \leq \infty$, then $\mathcal{L}_p \cap \mathcal{L}_r \subset \mathcal{L}_q$.*

Let $\Gamma : \mathbb{R} \to \mathbb{R}$ be the usual gamma function, defined as $\Gamma(z) = \int \mathbf{1}_{(0,\infty)} x^{z-1} \exp(-x) \, \mathrm{d}\lambda$.

Lemma 2.1.12. *If $f \in \mathcal{L}_p$ and $g \in \mathcal{L}_1$, for $1 \leq p \leq \infty$, then $f \star g$ exists and we have $\|f \star g\|_{\mathcal{L}_p} \leq \|g\|_{\mathcal{L}_1} \|f\|_{\mathcal{L}_p}$.*

Lemma 2.1.13. *Let $\mathcal{G} \subset \mathcal{L}_p$, for some $1 \leq p < \infty$, and let $f \in \overline{Conv}(\mathcal{G})$. For any $K > 0$, such that $\|f - \alpha\|_{\mathcal{L}_p} < K$, for all $\alpha \in \mathcal{G}$, there exists a $h_m \in Conv_m(\mathcal{G})$, such that*

$$\|f - h_m\|_{\mathcal{L}_p} \leq \frac{C_p K}{m^{1-1/\alpha}},$$

where $\alpha = \min\{p, 2\}$, and

$$C_p = \begin{cases} 1 & \text{if } 1 \leq p \leq 2, \\ \sqrt{2}\left[\sqrt{\pi}\Gamma\left(\frac{p+1}{2}\right)\right]^{1/p} & \text{if } p > 2. \end{cases}$$

Lemma 2.1.14. *In any measure ν, uniform convergence implies almost uniform convergence, and almost uniform convergence implies almost everywhere convergence and convergence in measure, with respect to ν.*

2.1.6 Sources of results

Lemma 2.1.6 is reported as Theorem 9.3.3 in Makarov & Podkorytov (2013) (see also Theorem 2 of (Cheney & Light, 2000, Ch. 20)). The proof of Corollary 2.1.7 can be taken from that of Theorem 4 of (Cheney & Light, 2000, Ch. 20). Lemma 2.1.8 appears in Conway (2012), as Proposition 1.4.5. Lemma 2.1.9 is taken from Corollary 1.2.9 of Conway (2012). Lemma 2.1.10 appears as Theorem 1.2.2 in Conway (2012). Lemma 2.1.11 can be found in (Folland, 1999, Prop. 6.10). Lemma 2.1.12 can be found in (Makarov & Podkorytov, 2013, Thm. 9.3.1). Lemma 2.1.13 appears as Corollary 2.6 in Donahue et al. (1997). Lemma 2.1.14 can be obtained from the definition of almost uniform convergence, Lemma 7.10, and Theorem 7.11 of Bartle (1995).

2.2 Universal approximation theorems for location-scale finite mixtures in Lebesgue spaces

Section 2.2 progresses as follows. The main result of this Section 2.2 is stated in Section 2.2.1. Technical preliminaries to the proof of the main result are presented in Section 2.2.2. The proof is then established in Section 2.2.3. Additional technical results required throughout the paper are reported in the Section 2.2.4.

2.2.1 Main result

Theorem 2.2.1 (Nguyen et al., 2020b, Theorem 2)**.** *Let $h_m^g \in \mathcal{M}^g$ denote an m-component location finite mixture PDF. If we assume that f and g are PDFs, then the following statements are true.*

(a) If $f, g \in \mathcal{C}$ and $\mathbb{K} \subset \mathbb{R}^d$ is a compact set, then there exists a sequence $\{h_m^g\}_{m=1}^{\infty} \subset \mathcal{M}^g$, such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{B}(\mathbb{K})} = 0.$$

(b) For $p \in [1, \infty)$, if $f \in \mathcal{L}_p$ and $g \in \mathcal{L}_\infty$, then there exists a sequence $\{h_m^g\}_{m=1}^{\infty} \subset \mathcal{M}^g$, such that

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

2.2.2 Technical preliminaries

Notice that $\mathcal{M}_m^g$ can be parameterized via dilates. That is, we can write

$$\mathcal{M}_m^g = \left\{ h_m^g : h_m^g(\cdot) = \sum_{i=1}^{m} c_i k_i^d g(k_i \times \cdot - k_i \mu_i), \mu_i \in \mathbb{R}^d, k_i \in \mathbb{R}_+, c \in \Pi_{m-1}, i \in [m] \right\},$$

where $k_i = 1/\sigma_i$.

Define $(\mathbb{U}, \|\cdot\|_{\mathbb{U}})$ to be a normed vector space (NVS) and let $\mathbb{F}$ be a subset of $\mathbb{U}$, and denote the convex hull of $\mathbb{F}$ by $\text{conv}(\mathbb{F})$ is the smallest convex subset in $\mathbb{U}$ that contains $\mathbb{F}$ (cf. Brezis, 2010, Chapter 1). By definition, we may write

$$\text{conv}(\mathbb{F}) = \left\{ \sum_{i\in[m]} \alpha_i f_i : f_i \in \mathbb{F}, \alpha \in \Pi_{m-1}, i \in [m], m \in \mathbb{N}^\star \right\},$$

where $\alpha^\top = (\alpha_1, \ldots, \alpha_m)$.

Define the class of "basic" densities, which will serve as the approximation building blocks, as follows

$$\mathcal{G}^g = \left\{ k^d g(k \times \cdot - k\mu), \mu \in \mathbb{R}^d, k \in \mathbb{R}_+ \right\},$$

and suppose that we can choose a suitable NVS $(\mathbb{U}, \|\cdot\|_{\mathbb{U}})$, such that $\mathcal{G}^g \subset \mathcal{M}^g \subset \mathbb{U}$. Then, by definition, it holds that $\mathcal{M}^g$ is a convex hull of $\mathcal{G}^g$.

For $u \in \mathbb{U}$ and $r > 0$, we define the open and closed balls of radius r, centered around u, by:

$$\mathbb{B}(u, r) = \{v \in \mathbb{U} : \|u - v\|_{\mathbb{U}} < r\},$$

and

$$\overline{\mathbb{B}}(u, r) = \{v \in \mathbb{U} : \|u - v\|_{\mathbb{U}} \leq r\},$$

respectively. For brevity, we also write $\mathbb{B}_r = \mathbb{B}(0, r)$ and $\overline{\mathbb{B}}_r = \overline{\mathbb{B}}(0, r)$. A set $\mathbb{F} \subset \mathbb{U}$ is open, if for every $u \in \mathbb{F}$, there exists an $r > 0$, such that $\mathbb{B}(u, r) \subset \mathbb{F}$. We say that $\mathbb{F}$ is closed if its complement is open, and by definition, we say that $\mathbb{U}$ and the empty set are both closed and open.

We call the smallest closed set containing $\mathbb{F}$ its closure, and we denote it by $\overline{\mathbb{F}}$. A sequence $\{u_m\} \subset \mathbb{U}$ converges to $u \in \mathbb{U}$, if $\lim_{m\to\infty} \|u_m - u\|_{\mathbb{U}} = 0$, and we denote it symbolically by $\lim_{m\to\infty} u_m = u$. That is, for every $\epsilon > 0$, there exists an $N(\epsilon) \in \mathbb{N}^\star$, such that $m \geq N(\epsilon)$ implies that $\|u_m - u\|_{\mathbb{U}} < \epsilon$.

By Lemma 2.2.6, we can write the closure of $\mathbb{F}$ as

$$\overline{\mathbb{F}} = \left\{ u \in \mathbb{U} : u = \lim_{m\to\infty} u_m, u_m \in \mathbb{F} \right\}$$

and hence

$$\overline{\mathcal{M}^g} = \left\{ h \in \mathbb{U} : h = \lim_{m\to\infty} h_m^g, h_m^g \in \mathcal{M}^g \right\}.$$

Thus, by definition, it holds that $\overline{\mathcal{M}^g}$ is a closed and convex subset of $\mathbb{U}$.

If $f \in \mathcal{C}$ is a PDF on $\mathbb{R}^d$, we denote its support by

$$\operatorname{supp} f = \left\{ x \in \mathbb{R}^d : f(x) \neq 0 \right\}$$

and furthermore, we denote the set of compactly supported continuous functions by

$$\mathcal{C}_c = \{ f \in \mathcal{C} : \operatorname{supp} f \text{ is compact} \}.$$

For open sets $\mathbb{V} \subset \mathbb{R}^d$, we will write $f \prec \mathbb{V}$ as shorthand for $f \in \mathcal{C}_c$, $0 \leq f \leq 1$, and $\operatorname{supp} f \subset \mathbb{V}$.

The following lemmas permit us to prove the primary technical mechanism that is used to prove our main result presented in Theorem 2.2.1.

Lemma 2.2.2. *Let $f \in \mathcal{C}$ be a PDF. Then, for every compact $\mathbb{K} \subset \mathbb{R}^d$, we can choose $h \in \mathcal{C}_c$, such that $\operatorname{supp} h \subset \mathbb{B}_r$, $0 \leq h \leq f$, and $h = f$ on $\mathbb{K}$, for some $r \in \mathbb{R}_+$.*

Proof. Since $\mathbb{K}$ is bounded, there exists some $r \in \mathbb{R}_+$, such that $\mathbb{K} \subset \mathbb{B}_r$. Lemma 2.2.10 implies that there exists a function $u \prec \mathbb{B}_r$, such that $u(x) = 1$, for all $x \in \mathbb{K}$. We can then set $h = uf$ to obtain the desired result of Lemma 2.2.2. □

Lemma 2.2.3. *Let $h \in \mathcal{C}_c$, such that $\operatorname{supp} h \subset \mathbb{B}_r$, $0 \leq h$, and $\int h d\lambda \leq 1$, and let $g \in \mathcal{C}$ be a PDF. Then, for any $k \in \mathbb{R}_+$, there exists a sequence $\{h_m^g\}_{m=1}^{\infty} \subset \mathcal{M}^g$, so that*

$$\lim_{m\to\infty} \|g_k \star h - h_m^g\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} = 0. \tag{2.2.1}$$

Furthermore, if $g \in \mathcal{C}_b^u$, we have the stronger result that

$$\lim_{m\to\infty} \|g_k \star h - h_m^g\|_{\mathcal{B}} = 0. \tag{2.2.2}$$

Proof. It suffices to show that given any $r, k, \epsilon \in \mathbb{R}_+$, there exists a sufficiently large $m(\epsilon, r, k) \in \mathbb{N}^\star$ such that for all $m \geq m(\epsilon, r, k)$, there exists a $h_m^g \in \mathcal{M}_m^g$ satisfying

$$\|g_k \star h - h_m^g\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} < \epsilon. \tag{2.2.3}$$

First, write

$$\begin{aligned}(g_k \star h)(x) &= \int g_k(x-y) h(y) \, \mathrm{d}\lambda(y) = \int \mathbf{1}_{\{y : y \in \overline{\mathbb{B}}_r\}} g_k(x-y) h(y) \, \mathrm{d}\lambda(y) \\ &= \int \mathbf{1}_{\{y : y \in \overline{\mathbb{B}}_r\}} k^d g(kx - ky) h(y) \, \mathrm{d}\lambda(y) = \int \mathbf{1}_{\{z : z \in \overline{\mathbb{B}}_{rk}\}} g(kx - z) h\left(\frac{z}{y}\right) \mathrm{d}\lambda(z),\end{aligned}$$

where $\overline{\mathbb{B}}_{rk}$ is a continuous image of a compact set, and hence is also compact (cf. Rudin, 1976, Theorem 4.14). By Lemma 2.1.10, for any $\delta > 0$, there exist $\kappa_i \in \mathbb{R}^d$ ($i \in [m-1]$, for some $m \in \mathbb{N}^\star$), such that $\overline{\mathbb{B}}_{rk} \subset \bigcup_{i=1}^{m-1} \mathbb{B}(\kappa_i, \delta/2)$. Further, if $\mathbb{B}_i^\delta = \mathbb{B}_{rk}^\delta = \overline{\mathbb{B}}_{rk} \cap \mathbb{B}(\kappa_i, \delta/2)$, then $\overline{\mathbb{B}}_{rk} = \bigcup_{i=1}^{m-1} \mathbb{B}_i^\delta$. We can hence obtain a disjoint covering of $\overline{\mathbb{B}}_{rk}$ by taking $\mathbb{A}_1^\delta = \mathbb{B}_1^\delta$, and $\mathbb{A}_i^\delta = \mathbb{B}_i^\delta \setminus \bigcup_{j=1}^{i-1} \mathbb{B}_j^\delta$ ($i \in [m-1]$) (cf. Cheney & Light, 2000, Chapter 24). Notice that $\overline{\mathbb{B}}_{rk} = \bigcup_{i=1}^{m-1} \mathbb{A}_i^\delta$, each $\mathbb{A}_i^\delta$ is a Borel set, and $\operatorname{diam}(\mathbb{A}_i^\delta) \leq \delta$, by construction.

We shall denote the disjoint cover of $\overline{\mathbb{B}}_{rk}$ by $\Pi_m^\delta = \{\mathbb{A}_i^\delta\}_{i=1}^{m-1}$. We seek to show that there exists an $m \in \mathbb{N}^\star$ and Π_m^δ, such that

$$\left\| g_k \star h - \sum_{i=1}^{m} c_i k_i^d g(k_i x - z_i) \right\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} < \epsilon,$$

where $k_i = k$, $c_i = k^{-d} \int \mathbf{1}_{\{z:z\in\mathbb{A}_i^\delta\}} h(z/k)\,\mathrm{d}\lambda(z)$, and $z_i \in \mathbb{A}_i^\delta$, for $i \in [m-1]$. We then set $z_m = 0$ and $c_m = 1 - \sum_{i=1}^{m-1} c_i$. Here, c_m depends only on r and ϵ. Suppose that $c_m > 0$. Then, since $g \neq 0$, there exists some $s \in \mathbb{R}_+$ such that $C_s = \sup_{w\in\overline{\mathbb{B}}_s} g(w) > 0$. We can choose

$$k_m = \min\left\{\frac{s}{r}, \left(\frac{\epsilon}{2c_m C_s}\right)^{1/d}\right\},$$

so that $\|g(k_m \times \cdot)\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} \leq C_s$ and

$$\|g(k_m \times \cdot)\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} \leq \frac{c_m \epsilon C_s}{2c_m C_s} = \epsilon/2.$$

Moreover, if we assume that $g \in \mathcal{C}_b^u$, then there exists a constant $C \in (0,\infty)$ such that $\|g\|_{\mathcal{B}} \leq C$. In this case, we can choose $k_m^d = \epsilon/(2c_m C)$ to obtain

$$\left\|c_m k_m^d g(k_m \times \cdot - z_m)\right\|_{\mathcal{B}} \leq \epsilon/2.$$

Since $0 \leq h$ and $\int h\mathrm{d}\lambda \in [0,1]$, the sum $\sum_{i=1}^{m-1} c_i$ satisfies the inequalities:

$$\begin{aligned} 0 \leq \sum_{i=1}^{m-1} c_i &= k^{-d} \sum_{i=1}^{m-1} \int \mathbf{1}_{\{z:z\in\mathbb{A}_i^\delta\}} h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z) \\ &= k^{-d} \int \mathbf{1}_{\{z:z\in k\mathbb{K}\}} h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z) = \int \mathbf{1}_{\{x:x\in\mathbb{K}\}} h\mathrm{d}\lambda \leq 1. \end{aligned}$$

Thus, $c_m \in [0,1]$, and our construction of h_m^g implies that $h_m^g = \sum_{i=1}^m c_i k_i^d g(k_i x - z_i) \in \mathcal{M}_m^g$.

We can then bound the left-hand side of (2.2.3) as follows:

$$\begin{aligned} \|g_k \star h - h_m^g\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} &\leq \left\|g_k \star h - \sum_{i=1}^{m-1} c_i k_i^d g(k_i \times \cdot - z_i)\right\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} + \left\|c_m k_m^d g(k_m \times \cdot - z_m)\right\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} \\ &\leq \left\|g_k \star h - \sum_{i=1}^{m-1} c_i k_i^d g(k_i \times \cdot - z_i)\right\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} + \frac{\epsilon}{2} \\ &= \left\|\int \mathbf{1}_{\{z:z\in\overline{\mathbb{B}}_{rk}\}} g(kx - z)\, h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z) - \sum_{i=1}^{m-1} \int \mathbf{1}_{\{z:z\in\mathbb{A}_i^\delta\}} g(kx - z_i)\, h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z)\right\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} \\ &\quad + \frac{\epsilon}{2} \\ &\leq \sum_{i=1}^{m-1} \int \mathbf{1}_{\{z:z\in\mathbb{A}_i^\delta\}} |g(kx - z) - g(kx - z_i)|\, h\left(\frac{z}{k}\right)\mathrm{d}\lambda(z) + \frac{\epsilon}{2}. \end{aligned} \tag{2.2.4}$$

Since $x \in \overline{\mathbb{B}}_r$, $z \in \mathbb{A}_i^\delta$, and $z_i \in \overline{\mathbb{B}}_{rk}$, it holds that $\|kx - z_i\|_2 = \|kx - z\|_2 \leq 2rk$, and

$$\|kx - z - (kx - z_i)\|_2 = \|z - z_i\|_2 \leq \operatorname{diam}\left(\mathbb{A}_i^\delta\right) \leq \delta.$$

Note that $g \in \mathcal{C}$, and thus g is uniformly continuous on the compact set $\overline{\mathbb{B}}_{2rk}$, implying that

$$|g(kx - z) - g(kx - z_i)| \leq w(g, 2rk, \delta),$$

for each $i \in [m-1]$, where

$$w(g, r, \delta) = \sup\left\{|g(x) - g(y)| : \|x - y\|_2 \leq \delta \text{ and } x, y \in \overline{\mathbb{B}}_r\right\}$$

denotes a modulus of continuity. Since $\lim_{\delta\to 0} w(g, 2rk, \delta) = 0$ (cf. Makarov & Podkorytov, 2013, Theorem 4.7.3), we may choose a $\delta(\epsilon, r, k) > 0$, such that

$$w(g, 2rk, \delta(\epsilon, r, k)) < \frac{\epsilon}{2k^d}.$$

We then proceed from (2.2.4) as follows:

$$\begin{aligned}
\|g_k \star h - h_m^g\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} &\leq w(g, 2rk, \delta(\epsilon, r, k)) \int \mathbf{1}_{\{z: z\in\overline{\mathbb{B}}_{rk}\}} h\left(\frac{z}{k}\right) \mathrm{d}\lambda(z) + \frac{\epsilon}{2} \\
&= w(g, 2rk, \delta(\epsilon, r, k)) k^d \int h \mathrm{d}\lambda + \frac{\epsilon}{2} \\
&\leq w(g, 2rk, \delta(\epsilon, r, k)) k^d + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \qquad (2.2.5)
\end{aligned}$$

To conclude the proof of (2.2.1), it suffices to choose an appropriate sequence of partitions $\Pi_m^{\delta(\epsilon,r,k)}$, such that $m \geq m(\epsilon, r, k)$, for some sufficiently large $m(\epsilon, r, k)$, so that (2.2.4) and (2.2.5) hold. This is possible via Lemma 2.1.10. When $g \in \mathcal{C}_b^u$, we notice that (2.2.4) and (2.2.5) both hold for all $x \in \mathbb{R}^d$. Thus, we have the stronger result of (2.2.2). □

We present the primary tools for proving Theorem 2.2.1 in the following pair of lemmas. The first one in Lemma 2.2.4 permits the approximation of convolutions of the form $g_k \star f$ in the $\mathcal{L}_1$ functional space, and the second presented in Lemma 2.2.5 generalizes this first result to the spaces $\mathcal{L}_p$, where $p \in [1, \infty)$, under an essentially bounded assumption.

Lemma 2.2.4. *If f and g are PDFs in the NVS $(\mathcal{L}_1, \|\cdot\|_{\mathcal{L}_1})$, then $\mathcal{M}^g \subset \mathcal{L}_1$ and $g_k \star f \in \mathcal{L}_1$, for every $k \in \mathbb{R}_+$. Furthermore, there exists a sequence $\{h_m^g\}_{m=1}^{\infty} \subset \mathcal{M}^g$, such that*

$$\lim_{m\to\infty} \|g_k \star f - h_m^g\|_{\mathcal{L}_1} = 0.$$

Proof. For any $k \in \mathbb{R}_+$, we can show that $g_k \in \mathcal{L}_1$, since

$$\|g_k\|_{\mathcal{L}_1} = \int g_k \mathrm{d}\lambda = \int k^d g(kx) \mathrm{d}\lambda(x) = \int g \mathrm{d}\lambda = 1.$$

If $h_m^g \in \mathcal{M}_m^g$, then $h_m^g \in \mathcal{L}_1$, since it is a finite sum of functions in $\mathcal{L}_1$, and thus, $\mathcal{M}^g \subset \mathcal{L}_1$. Note that since f is a PDF, we have $f \in \mathcal{L}_1$, and by Lemma 2.2.11, we also have that $g_k \star f \in \mathcal{L}_1$. By Lemma 2.2.12, it then follows that

$$\begin{aligned}
\|g_k \star f\|_{\mathcal{L}_1} &= \int g_k \star f \mathrm{d}\lambda \\
&= \int \left[\int g_k(x - y) f(y) \mathrm{d}\lambda(y)\right] \mathrm{d}\lambda(x) \\
&= \int \left[\int g_k(x - y) \mathrm{d}\lambda(x)\right] f(y) \mathrm{d}\lambda(y) \\
&= \|g_k\|_{\mathcal{L}_1} \|f\|_{\mathcal{L}_1} = 1
\end{aligned}$$

By definition of of the closure of $\mathcal{M}^g$ in $\mathcal{L}_1$, it suffices to show that for any $k \in \mathbb{R}_+$, $g_k \star f \in \overline{\mathcal{M}^g}$. We seek a contradiction by assuming that $g_k \star f \notin \overline{\mathcal{M}^g}$. Then, we can choose $\mathbb{A} = \overline{\mathcal{M}^g}$ and $\mathbb{B} = \{g_k \star f\}$ so that $\mathbb{A}, \mathbb{B} \subset \mathcal{L}_1$ are nonempty convex subsets, such that $\mathbb{A} \cap \mathbb{B} = \emptyset$. Furthermore, $\mathbb{A}$ is closed and $\mathbb{B}$ is compact. By Lemma 2.2.7, there exists a continuous linear functional $\phi \in \mathcal{L}_1^*$, such that $\phi(v) < \alpha < \phi(w)$, for all $v \in \mathbb{A}$ and $w \in \mathbb{B}$. By definition of $\mathbb{B}$, for all $v \in \overline{\mathcal{M}^g} \subset \mathcal{L}_1$ we have

$$\phi(v) < \alpha < \phi(g_k \star f).$$

By Lemma 2.2.9, with $\phi \in \mathcal{L}_1^*$, there exists a unique function $u \in \mathcal{L}_\infty$, such that, for all $v \in \mathcal{L}_1$,

$$\phi(v) = \int u(x) v(x) \mathrm{d}\lambda(x).$$

If we let $v = g_k(\cdot - \mu) \in \overline{\mathcal{M}^g} \subset \mathcal{L}_1$, then we obtain the inequalities

$$\sup_{\mu \in \mathbb{R}^d} \int u(x)\, g_k(x-\mu)\, \mathrm{d}\lambda(x) < \alpha < \int u(x)\, (g_k \star f)(x)\, \mathrm{d}\lambda(x).$$

The left-hand inequality can be reduced as follows:

$$\begin{aligned}
\alpha &< \int u(x)\, (g_k \star f)(x)\, \mathrm{d}\lambda(x) \\
&= \int u(x) \left[\int g_k(x-\mu)\, f(\mu)\, \mathrm{d}\lambda(\mu) \right] \mathrm{d}\lambda(x) \\
&= \int f(\mu) \left[\int u(x)\, g_k(x-\mu)\, \mathrm{d}\lambda(x) \right] \mathrm{d}\lambda(\mu) \\
&< \alpha \int f(\mu)\, \mathrm{d}\lambda(\mu) = \alpha,
\end{aligned}$$

where the third line is due to Lemma 2.2.12 and the final equality is because f is a PDF. This yields the sought contradiction. □

Lemma 2.2.5. *If $f, g \in \mathcal{L}_\infty$ are PDFs in the NVS $\left(\mathcal{L}_\infty, \|\cdot\|_{\mathcal{L}_p}\right)$, for $p \in [1, \infty)$, then, $\mathcal{M}^g \subset \mathcal{L}_p$ and $g_k \star f \in \mathcal{L}_p$, for any $k \in \mathbb{R}_+$. Furthermore, there exists a sequence $\{h_m^g\}_{m=1}^\infty \subset \mathcal{M}^g$, such that*

$$\lim_{m \to \infty} \|g_k \star f - h_m^g\|_{\mathcal{L}_p} = 0.$$

Proof. We obtain the result for $p = 1$ via Lemma 2.2.4. Otherwise, since $g \in \mathcal{L}_1 \cap \mathcal{L}_\infty$, we know that $g \in \mathcal{L}_p$ and $g_k \in \mathcal{L}_p$, for each $k \in \mathbb{R}_+$, via Lemma 2.1.11. For any $h_m^g \in \mathcal{M}_m^g$, we then have $h_m^g \in \mathcal{L}_p$ via finite summation, and hence $\mathcal{M}^g \in \mathcal{L}_p$. Since $f \in \mathcal{L}_1$, Lemma 2.2.11 implies that $g_k \star f \in \mathcal{L}_p$. By definition of the closure of $\mathcal{M}^g$, it suffices to show that $g_k \star f \in \overline{\mathcal{M}^g}$, for any $k \in \mathbb{R}_+$. This can be achieved by seeking a contradiction under the assumption that $g_k \star f \notin \overline{\mathcal{M}^g}$ and using Lemma 2.2.8 in the same manner as Lemma 2.2.9 is used in the proof of Lemma 2.2.4. □

2.2.3 Proof of the main result

2.2.3.1 Proof of Theorem 2.2.1 (a)

To prove the statement (a) of Theorem 2.2.1, it suffices to show that there exists a sufficiently large $m(\epsilon, \mathbb{K}) \in \mathbb{N}^\star$, such that for all $m \geq m(\epsilon, \mathbb{K})$, there exists a $h_m^g \in \mathcal{M}_m^g$, such that $\|f - h_m^g\|_{\mathcal{B}(\mathbb{K})} < \epsilon$, for any $\epsilon > 0$ and compact set $\mathbb{K} \subset \mathbb{R}^d$.

First, Lemma 2.2.2 implies that we can choose a $h \in \mathcal{C}_c$, such that $\operatorname{supp} h \subset \overline{\mathbb{B}}_r$, $0 \leq h \leq f$, and $h = f$ on $\mathbb{K}$, for some $r > 0$, where $\mathbb{K} \subset \overline{\mathbb{B}}_r$. We then have $\|f - h\|_{\mathcal{B}(\mathbb{K})} = 0$.

Since $h \in \mathcal{C}_c \subset \mathcal{C}_b^u$, Lemma 2.1.6 and Corollary 2.1.7 then imply that there exists a $k(\epsilon) \in \mathbb{R}_+$, such that for all $k \geq k(\epsilon)$, $\|h - g_k \star h\|_{\mathcal{B}(\mathbb{K})} < \epsilon/2$. We shall assume that $k \geq k(\epsilon)$, from hereon in.

Lemma 2.2.3 then implies that there exists an $m(\epsilon, r, k) \in \mathbb{N}^\star$, such that for any $m \geq m(\epsilon, r, k)$, there exists a $h_m^g \in \mathcal{M}_m^g$, such that $\|g_k \star h - h_m^g\|_{\mathcal{B}(\mathbb{K})} < \|g_k \star h - h_m^g\|_{\mathcal{B}(\overline{\mathbb{B}}_r)} < \epsilon/2$. The triangle inequality then completes the proof.

2.2.3.2 Proof of Theorem 2.2.1 (b)

To prove the statement (a) of Theorem 2.2.1, it suffices to show that there exists a sufficiently large $m(\epsilon) \in \mathbb{N}^\star$, such that for all $m \geq m(\epsilon)$, there exists a $h_m^g \in \mathcal{M}_m^g$, such that $\|f - h_m^g\|_{\mathcal{L}_p} < \epsilon$, for any $\epsilon > 0$.

First, Lemma 2.1.6 and Corollary 2.1.7 imply that there exists a $k(\epsilon) \in \mathbb{R}_+$, such that for any $k \geq k(\epsilon)$, it follows that $\|f - g_k \star f\|_{\mathcal{L}_p} < \epsilon/2$. We shall assume $k \geq k(\epsilon)$, from hereon in.

Lemma 2.2.4 and 2.2.5 imply that there exists an $m(\epsilon) \in \mathbb{N}^\star$, such that for all $m \geq m(\epsilon)$, there exists a $h_m^g \in \mathcal{M}_m^g$, such that $\|g_k \star f - h_m^g\|_{\mathcal{L}_p} < \epsilon/2$. The triangle inequality then completes the proof.

2.2.4 Technical results

We state a number of technical results that are used throughout Section 2.2 in Section 2.2.4.

Lemma 2.2.6 (Folland, 1999, Proposition 0.22). *Let $(\mathbb{U}, \|\cdot\|_{\mathbb{U}})$ be an NVS, and let $\mathbb{F} \subset \mathbb{U}$ and $u \in \mathbb{U}$. Then the following statements are equivalent: (a) $u \in \overline{\mathbb{F}}$; (b) $\mathbb{B}(u, r) \cap \mathbb{F} \neq \emptyset$, for all $r > 0$; and (c) there exists a sequence $\{u_m\} \subset \mathbb{F}$ that converges to u.*

Let $\mathbb{U}$ be a locally convex linear topological space over $\mathbb{R}$ and recall that a functional is a function defined on $\mathbb{U}$ (or some subspace of $\mathbb{U}$), with values in $\mathbb{R}$. We denote the due space of $\mathbb{U}$ (the space of all continuous linear functions on $\mathbb{U}$) by $\mathbb{U}^*$.

Lemma 2.2.7 (Second geometric form of the Hahn-Banach theorem, see, *e.g.*, Brezis, 2010, Theorem 1.7). *Let $\mathbb{A}, \mathbb{B} \subset \mathbb{U}$ be two nonempty convex subsets, such that $\mathbb{A} \cap \mathbb{B} \neq \emptyset$. Assume that $\mathbb{A}$ is closed and that $\mathbb{B}$ is compact. Then, there exists a continuous linear functional $\phi \in \mathbb{U}^*$, such that its corresponding hyperplane $H = \{u \in \mathbb{U} : \phi(u) = \alpha\}$ $(\alpha \in \mathbb{R})$ strictly separates $\mathbb{A}$ and $\mathbb{B}$. That is, there exists some $\epsilon > 0$, such that $\phi(u) \leq \alpha - \epsilon$ and $\phi(v) \geq \alpha + \epsilon$, for all $u \in \mathbb{A}$ and $v \in \mathbb{B}$. Or, in other words, $\sup_{u \in \mathbb{A}} \phi(u) < \inf_{v \in \mathbb{B}} \phi(v)$.*

Lemma 2.2.8 (Riesz representation theorem for $\mathcal{L}_p$, $p > 1$, see, *e.g.*, Brezis, 2010, Theorem 4.11). *If $p \in \mathbb{R}_+$, and $\phi \in (\mathcal{L}_p)^*$, then, there exists a unique function $u \in \mathcal{L}_q$, such that for all $v \in \mathcal{L}_q$,*

$$\phi(v) = \int u(x)\, v(x)\, d\lambda(x),$$

where $1/p + 1/q = 1$.

Lemma 2.2.9 (Riesz representation theorem for $\mathcal{L}_1$, see, *e.g.*, Brezis, 2010, Theorem 4.14). *If $\phi \in (\mathcal{L}_1)^*$, then there exists a unique $u \in \mathcal{L}_\infty$, such that for all $v \in \mathcal{L}_1$,*

$$\phi(v) = \int u(x)\, v(x)\, d\lambda(x).$$

Lemma 2.2.10 (Rudin, 1987, Theorem 2.13). *Let $\mathbb{V}_1, \ldots, \mathbb{V}_n$ be open subsets of $\mathbb{R}^d$, and let $\mathbb{K}$ be a compact set, such that $\mathbb{K} \subset \bigcup_{i=1}^n \mathbb{V}_i$. Then, there exists functions $h_i \prec \mathbb{V}_i$ $(i \in [n])$, such that $\sum_{i=1}^n h_i(x) = 1$, for all $x \in \mathbb{K}$. The set $\{h_i\}$ is referred to as the partition of unity on $\mathbb{K}$, subordinated to the cover $\{\mathbb{V}_i\}$.*

Lemma 2.2.11 (Folland, 1999, Proposition 8.8). *If $f \in \mathcal{L}_p$ $(1 \leq p \leq \infty)$ and $g \in \mathcal{L}_1$, then $f \star g$ exists and we have $\|f \star g\|_{\mathcal{L}_p} \leq \|f\|_{\mathcal{L}_p} \|f\|_{\mathcal{L}_1}$. Furthermore, if p and q are such that $1/p + 1/q = 1$, then $f \in \mathcal{L}_p$ and $g \in \mathcal{L}_q$, then $f \star g$ exists, is bounded and uniformly continuous, and $\|f \star g\|_{\mathcal{L}_\infty} \leq \|f\|_{\mathcal{L}_p} \|f\|_{\mathcal{L}_q}$. In particular, if $p \in \mathbb{R}_+$, then $f \star g \in \mathcal{C}_0$.*

Lemma 2.2.12 (Fubini's Theorem, see, *e.g.*, Rudin, 1987, Theorem 8.8). *Let $(\mathbb{X}, \mathcal{X}, \nu_1)$ and $(\mathbb{Y}, \mathcal{Y}, \nu_2)$ be σ-finite measure spaces, and assume that f is a $(\mathcal{X} \times \mathcal{Y})$-measurable function on $\mathbb{X} \times \mathbb{Y}$. If*

$$\int_{\mathbb{X}} \left[\int_{\mathbb{Y}} |f(x, y)|\, d\nu_1(x) \right] d\nu_2(y) < \infty,$$

then

$$\int_{\mathbb{X} \times \mathbb{Y}} |f|\, d(\nu_1 \times \nu_2) = \int_{\mathbb{X}} \left[\int_{\mathbb{Y}} |f(x, y)|\, d\nu_1(x) \right] d\nu_2(y) = \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} |f(x, y)|\, d\nu_2(y) \right] d\nu_1(x) < \infty.$$

2.3 Universal approximation theorems for mixture of experts models in Lebesgue spaces

Section 2.3 proceeds as follows. The main result is presented in Section 2.3.1. Technical lemmas are provided in Section 2.3.2. The proofs of our results are then presented in Section 2.3.3. Proofs of required lemmas that do not appear elsewhere are provided in Section 2.3.4.

2.3.1 Main results

Theorem 2.3.1. *Assume that* $\mathbb{X} = [0,1]^d$ *for* $d \in \mathbb{N}^\star$ *and* $\mathbb{Y}$ *is a compact subset in* $\mathbb{R}^q$, $q \in \mathbb{N}^\star$. *For any* $f \in \mathcal{F} \cap \mathcal{C}$, *any* $p \in [1, \infty)$, *there exists a sequence* $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star} \subset \mathcal{M}_S^\psi$, *where* $\psi \in \mathcal{C}(\mathbb{R}^q)$ *is a PDF on support* $\mathbb{R}^q$, *such that* $\lim_{K\to\infty} \left\| f - m_K^\psi \right\|_p = 0$.

Since convergence in Lebesgue spaces does not imply point-wise modes of convergence, the following result is also useful and interesting in some restricted scenarios. Here, we note that the mode of convergence is almost uniform, which implies almost everywhere convergence and convergence in measure (cf. Bartle 1995, Lem 7.10 and Thm. 7.11). The almost uniform convergence of $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star}$ to f in the following result is to be understood in the sense of Bartle (1995, Def. 7.9). That is, for every $\delta > 0$, there exists a set $\mathbb{U}_\delta \subset \mathbb{W}$ with $\lambda(\mathbb{W}) < \delta$, such that $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star}$ converges to f, uniformly on $\mathbb{W}\backslash\mathbb{U}_\delta$.

Theorem 2.3.2. *Assume that* $\mathbb{X} = [0,1]$ *and* $\mathbb{Y}$ *is a compact subset in* $\mathbb{R}^q$, $q \in \mathbb{N}^\star$. *For any* $f \in \mathcal{F} \cap \mathcal{C}$, *there exists a sequence* $\left\{m_K^\psi\right\}_{K\in\mathbb{N}^\star} \subset \mathcal{M}_S^\psi$, *where* $\psi \in \mathcal{C}(\mathbb{R}^q)$ *is a PDF on support* $\mathbb{R}^q$, *such that* $\lim_{K\to\infty} m_K^\psi = f$, *almost uniformly.*

The following result establishes the connection between the gating classes $\mathcal{G}_S^K$ and $\mathcal{G}_G^K$.

Lemma 2.3.3. *For each* $K \in \mathbb{N}^\star$, $\mathcal{G}_S^K \subset \mathcal{G}_G^K$. *Further, if we define the class of Gaussian gating vectors with equal covariance matrices:*

$$\mathcal{G}_E^K = \left\{ \textbf{Gate} = \left(\text{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right)\right)_{k\in[K]} | \forall k \in [K], \text{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right) = \frac{\pi_k \phi\left(\cdot;\boldsymbol{\nu}_k,\boldsymbol{\Sigma}\right)}{\sum_{l=1}^K \pi_l \phi\left(\cdot;\boldsymbol{\nu}_l,\boldsymbol{\Sigma}\right)}, \boldsymbol{\gamma} \in \mathbb{G}_E^K \right\},$$

where

$$\mathbb{G}_E^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \boldsymbol{\Sigma}) \in \Pi_{K-1} \times \left(\mathbb{R}^d\right)^K \times \mathbb{S}_d \right\},$$

then $\mathcal{G}_E^K \subset \mathcal{G}_S^K$.

By applying Lemma 2.3.3, Theorems 2.3.1 and 2.3.2 imply the following Corollary 2.3.4, regarding the approximation capability of the class $\mathcal{M}_G^\psi$.

Corollary 2.3.4. *Theorems 2.3.1 and 2.3.2 hold when* $\mathcal{M}_S^\psi$ *is replaced by* $\mathcal{M}_G^\psi$ *in their statements.*

Remark 2.3.5. To the best of our knowledge,our Theorems 2.3.1 and 2.3.2 are the first theorems to study universal approximation theorems in Lebesgue and continuous spaces for the approximation capabilities of MoE models with gating networks in subclasses of $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$. This contributes to an enduring continuity of sustained interest in the approximation capabilities of MoE models. Related to our results are contributions regarding the approximation capabilities of the conditional expectation function of the classes $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$, see definitions in (1.3.3) and (1.3.4), respectively, (Jiang & Tanner, 1999b, Krzyzak & Schafer, 2005, Mendes & Jiang, 2012, Nguyen et al., 2016, 2019, Wang & Mendel, 1992, Zeevi et al., 1998) and the approximation capabilities of subclasses of $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$, with respect to the Kullback–Leibler divergence (Jiang & Tanner, 1999a, Norets et al., 2010, Norets & Pelenis, 2014). Our results can be seen as complements to the Kullback–Leibler approximation theorems of Norets et al. (2010) and Norets & Pelenis (2014), by the relationship between the Kullback–Leibler divergence and the $\mathcal{L}_2$ norm (Zeevi & Meir, 1997). That is, when $f > 1/\kappa$, for all $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{W}$ and some constant $\kappa > 0$, we have that the integrated conditional Kullback–Leibler divergence considered by Norets & Pelenis (2014):

$$\int_{\mathbb{X}} D\left(f_{|\boldsymbol{X}}\left(\cdot,\boldsymbol{x}\right) \| m_K^\psi\left(\cdot,\boldsymbol{x}\right)\right) \mathrm{d}\lambda\left(\boldsymbol{x}\right) = \int_{\mathbb{X}}\int_{\mathbb{Y}} f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}\right) \log \frac{f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}\right)}{m_K^\psi\left(\boldsymbol{y},\boldsymbol{x}\right)} \mathrm{d}\lambda\left(\boldsymbol{y}\right) \mathrm{d}\lambda\left(\boldsymbol{x}\right)$$

satisfies

$$\int_{\mathbb{X}} D\left(f_{|\boldsymbol{X}}\left(\cdot,\boldsymbol{x}\right)\|m_K^{\psi}\left(\cdot,\boldsymbol{x}\right)\right)\mathrm{d}\lambda\left(\boldsymbol{x}\right)\leq\kappa^2\left\|f-m_K^{\psi}\right\|_{2,\mathbb{W}}^2,$$

and thus a good approximation in the integrated Kullback–Leibler divergence is guaranteed if one can find a good approximation in the $\mathcal{L}_2$ norm, which is guaranteed by our main results.

2.3.2 Technical lemmas

Let $\mathbb{K}^n=\left\{\left(k_1,\ldots,k_d\right)\in[n]^d\right\}$ and $\kappa:\mathbb{K}^n\to\left[n^d\right]$ be a bijection for each $n\in\mathbb{N}^{\star}$. For each $\left(k_1,\ldots,k_d\right)\in\mathbb{K}^n$ and $k\in\left[n^d\right]$, we define $\mathbb{X}_k^n=\mathbb{X}_{\kappa(k_1,\ldots,k_d)}^n=\prod_{i=1}^d\mathbb{I}_{k_i}^n$, where $\mathbb{I}_{k_i}^n=\left[\left(k_i-1\right)/n,k_i/n\right)$ for $k_i\in[n-1]$, and $\mathbb{I}_n^n=\left[\left(n-1\right)/n,1\right]$.

We call $\left\{\mathbb{X}_k^n\right\}_{k\in\left[n^d\right]}$ a fine partition of $\mathbb{X}$, in the sense that $\mathbb{X}=[0,1]^d=\bigcup_{k=1}^{n^d}\mathbb{X}_k^n$, for each n, and that $\lambda\left(\mathbb{X}_k^n\right)=n^{-d}$ gets smaller, as n increases. The following result from Jiang & Tanner (1999b) establishes the approximation capability of softmax gates.

Lemma 2.3.6 (Jiang & Tanner, 1999b, p. 1189)**.** *For each $n\in\mathbb{N}^{\star}$, $p\in[1,\infty)$ and $\epsilon>0$, there exists a gating function*

$$\mathbf{Gate}=\left(\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right)\right)_{k\in\left[n^d\right]}\in\mathcal{G}_S^{n^d}$$

for some $\boldsymbol{\gamma}\in\mathbb{G}_S^{n^d}$, such that

$$\sup_{k\in\left[n^d\right]}\left\|\mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}}-\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right)\right\|_{p,\mathbb{X}}\leq\epsilon.$$

When, $d=1$, we have also the following almost uniform convergence alternative to Lemma 2.3.6.

Lemma 2.3.7. *Let $\mathbb{X}=[0,1]$. Then, for each $n\in\mathbb{N}^{\star}$, there exists a sequence of gating functions:*

$$\left\{\mathbf{Gate}_l=\left(\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right)\right)_{k\in\left[n^d\right]}\right\}_{l\in\mathbb{N}^{\star}}\subset\mathcal{G}_S^n,$$

defined by $\left\{\boldsymbol{\gamma}_l\right\}_{l\in\mathbb{N}^{\star}}\subset\mathbb{G}_S^n$, such that

$$\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right)\to\mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}},$$

almost uniformly, simultaneously for all $k\in\left[n^d\right]$.

For PDF ψ on support $\mathbb{R}^q$, define the class of finite mixture models by

$$\begin{aligned}\mathcal{H}^{\psi}\quad=\quad&\left\{h_K^{\psi}:\mathbb{R}^q\to[0,\infty)\,|h_K^{\psi}\left(\boldsymbol{y}\right)=\sum_{k=1}^K c_k g_{\psi}\left(\boldsymbol{y};\boldsymbol{\mu}_k,\sigma_k\right),\right.\\&\left.g_{\psi}\in\mathcal{E}_{\psi}\cap\mathcal{L}_{\infty},\left(c_k\right)_{k\in[K]}\in\Pi_{K-1},\boldsymbol{\mu}_k\in\mathbb{Y},\sigma_k\in(0,\infty),k\in[K],K\in\mathbb{N}^{\star}\right\}.\end{aligned}$$

We require the following result, from Nguyen et al. (2020b), regarding the approximation capabilities of $\mathcal{H}^{\psi}$.

Lemma 2.3.8 (Nguyen et al., 2020b, Theorem 2.2.1(b))**.** *If $f\in\mathcal{C}\left(\mathbb{Y}\right)$ is a PDF on $\mathbb{Y}$, $\psi\in\mathcal{C}\left(\mathbb{R}^q\right)$ is a PDF on $\mathbb{R}^q$, and $\mathbb{Y}\subset\mathbb{R}^q$ is compact, then there exists a sequence $\left\{h_K^{\psi}\right\}_{K\in\mathbb{N}^{\star}}\subset\mathcal{H}^{\psi}$, such that $\lim_{K\to\infty}\left\|f-h_K^{\psi}\right\|_{\mathcal{B}(\mathbb{Y})}=0$.*

2.3.3 Proofs of main results

2.3.3.1 Proof of Theorem 2.3.1

To prove the result, it suffices to show that for each $\epsilon > 0$, there exists a $m_K^{\psi} \in \mathcal{M}_S^{\psi}$, such that

$$\left\| f - m_K^{\psi} \right\|_p < \epsilon.$$

The main steps of the proof are as follows. We firstly approximate $f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x})$ by

$$\upsilon_n(\boldsymbol{y}, \boldsymbol{x}) = \sum_{k=1}^{n^d} \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n), \tag{2.3.1}$$

where $\boldsymbol{x}_k^n \in \mathbb{X}_k^n$, for each $k \in \left[n^d\right]$, such that

$$\|f - \upsilon_n\|_p < \frac{\epsilon}{3}, \tag{2.3.2}$$

for all $n \geq N_1(\epsilon)$, for some sufficiently large $N_1(\epsilon) \in \mathbb{N}^{\star}$. Then we approximate $\upsilon_n(\boldsymbol{y}, \boldsymbol{x})$ by

$$\eta_n(\boldsymbol{y}, \boldsymbol{x}) = \sum_{k=1}^{n^d} \text{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n), \tag{2.3.3}$$

where $\boldsymbol{\gamma}_n \in \mathbb{G}_S^{n^d}$ and $\textbf{Gate} = (\text{Gate}_k(\cdot; \boldsymbol{\gamma}_n))_{k \in \left[n^d\right]} \in \mathcal{G}_S^{n^d}$, so that

$$\|\upsilon_n - \eta_n\|_p \leq \sup_{k \in \left[n^d\right]} \left\| \text{Gate}_k(\cdot; \boldsymbol{\gamma}) - \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} \right\|_{p,\mathbb{X}} \sum_{k=1}^{n^d} \left\| f_{|\boldsymbol{X}}(\cdot, \boldsymbol{x}_k^n) \right\|_{p,\mathbb{Y}} < \frac{\epsilon}{3}, \tag{2.3.4}$$

using Lemma 2.3.6.

Finally, we approximate $\eta_n(\boldsymbol{y}, \boldsymbol{x})$ by $m_{K_n}^{\psi}(\boldsymbol{y}, \boldsymbol{x})$, where

$$m_{K_n}^{\psi}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{k=1}^{n^d} \text{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}) h_{n_k}^k(\boldsymbol{y}, \boldsymbol{x}_k^n) \tag{2.3.5}$$

and

$$h_{n_k}^k(\boldsymbol{y}, \boldsymbol{x}_k^n) = \sum_{i=1}^{n_k} c_i^k g_{\psi}\left(\boldsymbol{y}; \boldsymbol{\mu}_i^k, \sigma_i^k\right) \in \mathcal{H}^{\psi} \tag{2.3.6}$$

for $n_k \in \mathbb{N}^{\star}$ ($k \in \left[n^d\right]$), such that $K_n = \sum_{k=1}^{n^d} n_k$. Here, we establish that there exists $N_2(\epsilon, n, \boldsymbol{\gamma}_n) \in \mathbb{N}^{\star}$, so that when $n_k \geq N_2(\epsilon, n, \boldsymbol{\gamma}_n)$,

$$\left\| \eta_n - m_{K_n}^{\psi} \right\|_p \leq \sup_{k \in \left[n^d\right]} \|\text{Gate}_k(\cdot; \boldsymbol{\gamma})\|_{p,\mathbb{X}} \sum_{k=1}^{n^d} \left\| f(\cdot|\boldsymbol{x}_k^n) - h_{n_k}^k(\cdot|\boldsymbol{x}_k^n) \right\|_{p,\mathbb{Y}} < \frac{\epsilon}{3}. \tag{2.3.7}$$

Results (2.3.2)–(2.3.7) then imply that for each $\epsilon > 0$, there exists $N_1(\epsilon)$, $\boldsymbol{\gamma}_n$, and $N_2(\epsilon, n, \boldsymbol{\gamma}_n)$, such that for all $K_n = \sum_{k=1}^{n^d} n_k$, where $n_k \geq N_2(\epsilon, n, \boldsymbol{\gamma}_n)$ (for each $k \in \left[n^d\right]$) and $n \geq N_1(\epsilon)$. The following inequality results from an application of the triangle inequality:

$$\left\| f - m_{K_n}^{\psi} \right\|_p \leq \|f - \upsilon_n\|_p + \|\upsilon_n - \eta_n\|_p + \left\| \eta_n - m_{K_n}^{\psi} \right\|_p < 3 \times \frac{\epsilon}{3} = \epsilon.$$

We now focus our attention to proving each of the results: (2.3.2)–(2.3.7). To prove (2.3.2), we note that since f is uniformly continuous (because $\mathbb{W} = \mathbb{Y} \times \mathbb{X}$ is compact, and $f \in \mathcal{C}$), there exists a function (2.3.1) such that for all $\varepsilon > 0$,

$$\sup_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{W}} \left| f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}) - \upsilon(\boldsymbol{y}, \boldsymbol{x}) \right| < \varepsilon. \tag{2.3.8}$$

We can construct such an approximation by considering the fact that as n increases, the diameter $\delta_n = \sup_{k \in n^d} \operatorname{diam}\left(\mathbb{X}_k^n\right)$ of the fine partition goes to zero. By the uniform continuity of f, for every $\varepsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that if $\|(\boldsymbol{y}_1, \boldsymbol{x}_1) - (\boldsymbol{y}_2, \boldsymbol{x}_2)\| < \delta(\epsilon)$, then $|f(\boldsymbol{y}_1|\boldsymbol{x}_1) - f(\boldsymbol{y}_2|\boldsymbol{x}_2)| < \varepsilon$, for all pairs $(\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2) \in \mathbb{W}$. Here, $\|\cdot\|$ denotes the Euclidean norm. Furthermore, for any $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{W}$, we have

$$\begin{aligned}\left|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}) - \upsilon_n(\boldsymbol{y}, \boldsymbol{x})\right| &= \left|\sum_{k=1}^{n^d} \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} \left[f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}) - f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right]\right| \\ &\leq \sum_{k=1}^{n^d} \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} \left|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}) - f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right|, \end{aligned} \tag{2.3.9}$$

by the triangle inequality.

Since $\boldsymbol{x}_k^n \in \mathbb{X}_k^n$, for each k and n, we have the fact that $\|(\boldsymbol{y}, \boldsymbol{x}) - (\boldsymbol{y}, \boldsymbol{x}_k^n)\| < \delta_n$ for $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{Y} \times \mathbb{X}_k^n$. By uniform continuity, for each ε, we can find a sufficiently small $\delta(\epsilon)$, such that $\left|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}) - f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right| < \varepsilon$, if $\|(\boldsymbol{y}, \boldsymbol{x}) - (\boldsymbol{y}, \boldsymbol{x}_k^n)\| < \delta(\epsilon)$, for all k. The desired result (2.3.8) can be obtained by noting that the right hand side of (2.3.9) consists of only one non-zero summand for any $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{W}$, and by choosing $n \in \mathbb{N}^\star$ sufficiently large, so that $\delta_n < \delta(\epsilon)$.

By (2.3.8), we have the fact that $\upsilon_n \to f$, point-wise. We can bound υ_n as follows:

$$\upsilon_n(\boldsymbol{y}, \boldsymbol{x}) \leq \sum_{i=1}^{n^p} \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} \sup_{\boldsymbol{\zeta} \in \mathbb{Y}, \boldsymbol{\xi} \in \mathbb{X}} f(\boldsymbol{\zeta}, \boldsymbol{\xi}) = \sup_{\boldsymbol{\zeta} \in \mathbb{Y}, \boldsymbol{\xi} \in \mathbb{X}} f(\boldsymbol{\zeta}, \boldsymbol{\xi}), \tag{2.3.10}$$

where the right-hand side is a constant and is therefore in $\mathcal{L}_p$, since $\mathbb{W}$ is compact. An application of the Lebesgue dominated convergence theorem in $\mathcal{L}_p$ then yields (2.3.2).

Next we write

$$\begin{aligned}\|\upsilon_n - \eta_n\|_p &= \left\|\sum_{k=1}^{n^d} \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n) - \sum_{k=1}^{n^d} \operatorname{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right\|_p \\ &\leq \sum_{k=1}^{n^d} \left\|\left[\mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} - \operatorname{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n)\right] f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right\|_p.\end{aligned}$$

Since the norm arguments are separable in $\boldsymbol{x}$ and $\boldsymbol{y}$, we apply Fubini's theorem to get

$$\begin{aligned}\|\upsilon_n - \eta_n\|_p &= \sum_{k=1}^{n^d} \left\|\left[\mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} - \operatorname{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n)\right]\right\|_{p,\mathbb{X}} \left\|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right\|_{p,\mathbb{Y}} \\ &\leq \sup_{k \in \left[n^d\right]} \left\|\left[\mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} - \operatorname{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n)\right]\right\|_{p,\mathbb{X}} \sum_{k=1}^{n^d} \left\|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right\|_{p,\mathbb{Y}}\end{aligned}$$

Because $f \in \mathcal{B}$ and n^d is finite, for any fixed $n \in \mathbb{N}^\star$, we have $C_1(n) = \sum_{k=1}^{n^d} \left\|f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)\right\|_{p,\mathbb{Y}} < \infty$. For each $\epsilon > 0$, we need to choose a $\boldsymbol{\gamma}_n \in \mathbb{G}_S^{n^d}$, such that

$$\sup_{k \in \left[n^d\right]} \left\|\left[\mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} - \operatorname{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n)\right]\right\|_{p,\mathbb{X}} < \frac{\epsilon}{3C_1(n)},$$

which can be achieved via a direct application of Lemma 2.3.6. We have thus shown (2.3.4).

Lastly, we are required to approximate $f_{|\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x}_k^n)$ for each $k \in \left[n^d\right]$, by a function of form (2.3.6). Since $\mathbb{Y}$ is compact and f and ψ are continuous, we can apply of Lemma 2.3.8, directly. Note that over a set of finite measure, convergence in $\|\cdot\|_{\mathcal{B}}$ implies convergence in $\mathcal{L}_p$ norm, for all $p \in [1, \infty]$ (cf. Oden & Demkowicz 2010, Prop. 3.9.3).

We can then write (2.3.5) as

$$
\begin{aligned}
m_{K_n}^{\psi}\left(\boldsymbol{y},\boldsymbol{x}\right) &= \sum_{k=1}^{n^d}\frac{\exp\left(a_{n,k}+\boldsymbol{b}_{n,k}^{\top}\boldsymbol{x}\right)}{\sum_{l=1}^{n^d}\exp\left(a_{n,l}+\boldsymbol{b}_{n,l}^{\top}\boldsymbol{x}\right)}h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\\
&= \sum_{k=1}^{n^d}\sum_{i=1}^{n_k}\frac{\exp\left(a_{n,k}+\boldsymbol{b}_{n,k}^{\top}\boldsymbol{x}\right)}{\sum_{l=1}^{n^d}\exp\left(a_{n,l}+\boldsymbol{b}_{n,l}^{\top}\boldsymbol{x}\right)}\frac{c_i^k}{\sum_{l=1}^{n_k}c_l^k}g_{\psi}\left(\boldsymbol{y};\boldsymbol{\mu}_i^k,\sigma_i^k\right)\\
&= \sum_{k=1}^{n^d}\sum_{i=1}^{n_k}\frac{\exp\left(\log c_i^k+a_{n,k}+\boldsymbol{b}_{n,k}^{\top}\boldsymbol{x}\right)}{\sum_{l=1}^{n^d}\sum_{j=1}^{n_k}\exp\left(\log c_j^k+a_{n,l}+\boldsymbol{b}_{n,l}^{\top}\boldsymbol{x}\right)}g_{\psi}\left(\boldsymbol{y};\boldsymbol{\mu}_i^k,\sigma_i^k\right),
\end{aligned}
\tag{2.3.11}
$$

where $\boldsymbol{\gamma}_n = \left(a_{n,1},\ldots,a_{n,n^d},\boldsymbol{b}_{n,1},\ldots,\boldsymbol{b}_{n,n^d}\right)$. From (2.3.11), we observe that $m_{K_n}^{\psi}\in\mathcal{M}_S^{\psi}$, with $K_n = \sum_{k=1}^{n^d}n_k$.

To obtain (2.3.7), we write

$$
\begin{aligned}
\left\|\eta_n - m_{K_n}^{\psi}\right\|_p &= \left\|\sum_{k=1}^{n^d}\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_n\right)f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right) - \sum_{k=1}^{n^d}\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}\right)h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right\|_p\\
&\leq \sum_{k=1}^{n^d}\left\|\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_n\right)\left[f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right) - h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right]\right\|_p.
\end{aligned}
$$

By separability and Fubini's theorem, we then have

$$
\begin{aligned}
\left\|\eta_n - m_{K_n}^{\psi}\right\| &\leq \sum_{k=1}^{n^d}\left\|\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_n\right)\right\|_{p,\mathbb{X}}\left\|f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right) - h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right\|_{p,\mathbb{Y}}\\
&\leq \sup_{k\in\left[n^d\right]}\left\|\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_n\right)\right\|_{p,\mathbb{X}}\sum_{k=1}^{n^d}\left\|f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right) - h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right\|_{p,\mathbb{Y}}.
\end{aligned}
$$

Let $C_2\left(n,\boldsymbol{\gamma}_n\right) = \sup_{k\in\left[n^d\right]}\left\|\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_n\right)\right\|_{p,\mathbb{X}}$. Then, we apply Lemma 2.3.8 n^d times to establish the existence of a constant $N_2\left(\epsilon,n,\boldsymbol{\gamma}_n\right)\in\mathbb{N}^{\star}$, such that for all $k\in\left[n^d\right]$ and $n_k\geq N_2\left(\epsilon,n,\boldsymbol{\gamma}_n\right)$,

$$
\left\|f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right) - h_{n_k}^{k}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right\|_{p,\mathbb{Y}} \leq \frac{\epsilon}{3C_2\left(n,\boldsymbol{\gamma}_n\right)n^d}.
$$

Thus, we have

$$
\left\|\eta_n - m_{K_n}^{\psi}\right\| \leq C_2\left(n,\boldsymbol{\gamma}_n\right)\times n^d\times\frac{\epsilon}{3C_2\left(n,\boldsymbol{\gamma}_n\right)n^d} = \frac{\epsilon}{3},
$$

which completes our proof.

2.3.3.2 Proof of Theorem 2.3.2

The proof is procedurally similar to that of Theorem 2.3.1 and thus we only seek to highlight the important differences. Firstly, for any $\epsilon>0$, we approximate $f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}\right)$ by $\upsilon_n\left(\boldsymbol{x}|\boldsymbol{y}\right)$ of form (2.3.1), with $d=1$. Result (2.3.2) implies uniform convergence, in the sense that there exists an $N_1\left(\epsilon\right)\in\mathbb{N}^{\star}$, such that for all $n\geq N_1\left(\epsilon\right)$,

$$
\left\|f-\upsilon_n\right\|_{\mathcal{B}} < \frac{\epsilon}{3}. \tag{2.3.12}
$$

We now seek to approximate υ_n by η_n of form (2.3.3), with $\boldsymbol{\gamma}_n=\boldsymbol{\gamma}_l$ for some $l\in\mathbb{N}^{\star}$. Upon application of Lemma 2.3.7, it follows that for each $k\in\left[n^d\right]$ and $\varepsilon>0$, there exists a measurable set $\mathbb{B}_k\left(\varepsilon\right)\subseteq\mathbb{X}$, such that

$$\lambda\left(\mathbb{B}_k\left(\varepsilon\right)\right) < \frac{\varepsilon}{n^d \lambda\left(\mathbb{Y}\right)}$$

and

$$\left\| \mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right) - \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} \right\|_{\mathcal{B}\left(\mathbb{B}_k^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3}$$

for all $l \geq M_k\left(\epsilon, n\right)$, for some $M_k\left(\epsilon, n\right) \in \mathbb{N}^{\star}$. Here, $\left(\cdot\right)^{\complement}$ is the set complement operator.

Since $f \in \mathcal{B}$, we have the bound $C\left(n\right) = \sum_{k=1}^{n^d} \left\| f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})} < \infty$. Write $\mathbb{B}\left(\varepsilon\right) = \bigcup_{k=1}^{n^d} \mathbb{B}_k\left(\varepsilon\right)$. Then, $\mathbb{B}^{\complement}\left(\varepsilon\right) = \bigcap_{k=1}^{n^d} \mathbb{B}_k^{\complement}\left(\varepsilon\right)$,

$$\lambda\left(\mathbb{B}\left(\varepsilon\right)\right) \leq \sum_{k=1}^{n^d} \lambda\left(\mathbb{B}_k^{\complement}\left(\varepsilon\right)\right) < \frac{\varepsilon}{\lambda\left(\mathbb{Y}\right)},$$

and

$$\left\| \mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right) - \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \leq \min_{k\in\left[n^d\right]} \left\| \mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right) - \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} \right\|_{\mathcal{B}\left(\mathbb{B}_k^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3C\left(n\right)},$$

for all $l \geq M\left(\epsilon, n\right) = \max_{k\in\left[n^d\right]} M_k\left(\epsilon, n\right)$. Here we use the fact that the supremum over some intersect of sets is less than or equal to the minimum of the supremum over each individual set.

Upon defining $\mathbb{C}\left(\varepsilon\right) = \mathbb{Y} \times \mathbb{B}\left(\varepsilon\right) \subset \mathbb{W}$, we observe that

$$\lambda\left(\mathbb{C}\left(\varepsilon\right)\right) = \lambda\left(\mathbb{B}\left(\varepsilon\right)\right)\lambda\left(\mathbb{Y}\right) \leq \frac{\varepsilon}{\lambda\left(\mathbb{Y}\right)} \times \lambda\left(\mathbb{Y}\right) = \varepsilon,$$

and $\mathbb{C}\left(\varepsilon\right) \subset \mathbb{Y} \times \mathbb{B}\left(\varepsilon\right)$. Note also that

$$\left(\mathbb{Y} \times \mathbb{B}\left(\varepsilon\right)\right)^{\complement} = \mathbb{W} \backslash \left(\mathbb{Y} \times \mathbb{B}\left(\varepsilon\right)\right) = \mathbb{Y} \times \mathbb{B}^{\complement}\left(\varepsilon\right)$$

and

$$\mathbb{C}^{\complement}\left(\varepsilon\right) = \left(\mathbb{Y} \times \mathbb{B}^{\complement}\left(\varepsilon\right)\right) \cup \left(\mathbb{Y}^{\complement} \times \mathbb{B}\left(\varepsilon\right)\right) \cup \left(\mathbb{Y}^{\complement} \times \mathbb{B}^{\complement}\left(\varepsilon\right)\right).$$

It follows that

$$\left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} \leq \max\left\{ \left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{Y}\times\mathbb{B}^{\complement}(\varepsilon)\right)}, \left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{Y}^{\complement}\times\mathbb{B}(\varepsilon)\right)}, \left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{Y}^{\complement}\times\mathbb{B}^{\complement}(\varepsilon)\right)} \right\}.$$

Since $\mathbb{Y}^{\complement} \times \mathbb{B}\left(\varepsilon\right)$ and $\mathbb{Y}^{\complement} \times \mathbb{B}^{\complement}\left(\varepsilon\right)$ are empty, via separability, we have

$$\begin{aligned}
\left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} &= \left\|\upsilon_n - \eta_n\right\|_{\mathcal{B}\left(\mathbb{Y}\times\mathbb{B}^{\complement}(\varepsilon)\right)} \\
&= \sup_{\boldsymbol{z}\in\mathbb{Y}\times\mathbb{B}^{\complement}(\varepsilon)} \left| \sum_{k=1}^{n^d} \left[\mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} - \mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right) \right] f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right) \right| \\
&\leq \sup_{\boldsymbol{z}\in\mathbb{Y}\times\mathbb{B}^{\complement}(\varepsilon)} \sum_{k=1}^{n^d} \left| \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} - \mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right) \right| f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right) \\
&\leq \sum_{k=1}^{n^d} \sup_{\boldsymbol{z}\in\mathbb{Y}\times\mathbb{B}^{\complement}(\varepsilon)} \left| \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} - \mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right) \right| f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right) \\
&= \sum_{k=1}^{n^d} \left\| \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} - \mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \left\| f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})} \\
&\leq \sup_{k\in[n]} \left\| \mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}} - \mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \sum_{k=1}^{n^d} \left\| f_{|\boldsymbol{X}}\left(\boldsymbol{y}, \boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})}.
\end{aligned}$$

Recall that the $\sum_{k=1}^{n^d}\left\|f_{|\boldsymbol{X}}\left(\boldsymbol{y},\boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})}=C\left(n\right)<\infty$ and that we can choose $l\geq M\left(\epsilon,n\right)$ so that

$$\sup_{k\in[n]}\left\|\mathbf{1}_{\left\{\boldsymbol{x}\in\mathbb{X}_k^n\right\}}-\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right)\right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)}<\frac{\epsilon}{3C\left(n\right)},$$

and thus

$$\left\|\upsilon_n-\eta_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}<\frac{\epsilon}{3C\left(n\right)}\times C\left(n\right)=\frac{\epsilon}{3}, \tag{2.3.13}$$

as required.

Finally, by noting that for each $k\in\left[n^d\right]$, both (2.3.6) and $f\left(\cdot|\boldsymbol{x}_k^n\right)$ are continuous over $\mathbb{Y}$, we apply Lemma 2.3.8 to obtain an $N_2\left(\epsilon,n,l\right)\in\mathbb{N}^{\star}$, such that for any $\epsilon>0$ and $n_k\geq N_2\left(\epsilon,n,l\right)$, we have

$$\left\|f\left(\cdot|\boldsymbol{x}_k^n\right)-h_{n_k}^k\left(\cdot|\boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})}<\frac{\epsilon}{3M_1n}.$$

Here $M_1=\sup_{k\in\left[n^d\right]}\left\|\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}_l\right)\right\|_{\mathcal{B}(\mathbb{X})}<\infty$, since $\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right)$ is continuous in $\boldsymbol{x}$, and $\mathbb{X}$ is compact. Therefore, for all $K_n=\sum_{k=1}^{n^d}n_k$, $n_k\geq N_2\left(\epsilon,n,l\right)$,

$$\begin{aligned}\left\|\eta_n-m_{K_n}^{\psi}\right\|_{\mathcal{B}}&\leq\sup_{k\in\left[n^d\right]}\left\|\mathrm{Gate}_k\left(\boldsymbol{x};\boldsymbol{\gamma}_l\right)\right\|_{\mathcal{B}(\mathbb{X})}\sum_{k=1}^{n^d}\left\|f\left(\cdot|\boldsymbol{x}_k^n\right)-h_{n_k}^k\left(\cdot|\boldsymbol{x}_k^n\right)\right\|_{\mathcal{B}(\mathbb{Y})}\\&=M_1\times n^d\times\frac{\epsilon}{3M_1n^d}=\frac{\epsilon}{3}.\end{aligned} \tag{2.3.14}$$

In summary, via (2.3.12), (2.3.13), and (2.3.14), for each $\epsilon>0$, for any $\varepsilon>0$, there exists a $\mathbb{C}\left(\varepsilon\right)\subset\mathbb{W}$ and constants $N_1\left(\epsilon\right),M\left(\epsilon,n\right),N_2\left(\epsilon,n,l\right)\in\mathbb{N}^{\star}$, such that for all $K_n=\sum_{k=1}^{n^d}n_k$, with $n_k\geq N_2\left(\epsilon,n,l\right)$, $l\geq M\left(\epsilon,n\right)$, and $n\geq N_1\left(\epsilon\right)$, it follows that $\lambda\left(\mathbb{C}\left(\varepsilon\right)\right)<\varepsilon$, and

$$\begin{aligned}\left\|f-m_{K_n}^{\psi}\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}&\leq\left\|f-\upsilon_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}+\left\|\upsilon_n-\eta_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}+\left\|\eta_n-m_{K_n}^{\psi}\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}\\&\leq\left\|f-\upsilon_n\right\|_{\mathcal{B}}+\left\|\upsilon_n-\eta_n\right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}+\left\|\eta_n-m_{K_n}^{\psi}\right\|_{\mathcal{B}}\\&<3\times\frac{\epsilon}{3}=\epsilon.\end{aligned}$$

This completes the proof.

2.3.4 Proofs of lemmas

2.3.4.1 Proof of Lemma 2.3.3

We firstly prove that any gating vector from $\mathcal{G}_S^K$ can be equivalently represented as an element of $\mathcal{G}_G^K$. For any $\boldsymbol{x}\in\mathbb{R}^d$, $d\in\mathbb{N}^{\star}$, $k\in[K]$, $a_k\in\mathbb{R}$, $\boldsymbol{b}_k\in\mathbb{R}^d$, and $K\in\mathbb{N}^{\star}$, choose $\boldsymbol{\nu}_k=\boldsymbol{b}_k$, $\tau_k=a_k+\boldsymbol{b}_k^{\top}\boldsymbol{b}_k/2$ and

$$\pi_k=\exp\left(\tau_k\right)/\sum_{l=1}^{K}\exp\left(\tau_l\right).$$

This implies that $\sum_{l=1}^{K}\pi_l=1$, $\pi_l>0$, for all $l\in[K]$, and

$$\begin{aligned}\frac{\exp\left(a_k+\boldsymbol{b}_k^{\top}\boldsymbol{x}\right)}{\sum_{l=1}^{K}\exp\left(a_k+\boldsymbol{b}_k^{\top}\boldsymbol{x}\right)}&=\frac{\exp\left(\tau_k-\boldsymbol{v}_k^{\top}\boldsymbol{v}_k/2+\boldsymbol{v}_k^{\top}\boldsymbol{x}\right)}{\sum_{l=1}^{K}\exp\left(\tau_l-\boldsymbol{v}_l^{\top}\boldsymbol{v}_l/2+\boldsymbol{v}_l^{\top}\boldsymbol{x}\right)}\\&=\frac{\exp\left(\tau_k\right)\exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)/2\right)}{\sum_{l=1}^{K}\exp\left(\tau_l\right)\exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)/2\right)}\\&=\frac{\pi_k\left(2\pi\right)^{-d/2}\exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)/2\right)}{\sum_{l=1}^{K}\pi_l\left(2\pi\right)^{-d/2}\exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)^{\top}\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)/2\right)}\\&=\frac{\pi_k\phi\left(\boldsymbol{x};\boldsymbol{\nu}_k,\mathbf{I}\right)}{\sum_{l=1}^{K}\pi_l\phi\left(\boldsymbol{x};\boldsymbol{\nu}_l,\mathbf{I}\right)},\end{aligned}$$

where $\mathbf{I}$ is the identity matrix of appropriate size. This proves that $\mathcal{G}_S^K \subset \mathcal{G}_G^K$.

Next, to show that $\mathcal{G}_E^K \subset \mathcal{G}_S^K$, we write

$$\begin{aligned}
&\frac{\pi_k \phi\left(\boldsymbol{x}; \boldsymbol{\nu}_k, \boldsymbol{\Sigma}\right)}{\sum_{l=1}^K \pi_l \phi\left(\boldsymbol{x}; \boldsymbol{\nu}_l, \boldsymbol{\Sigma}\right)} \\
=&\frac{\pi_k \left|2\pi\boldsymbol{\Sigma}\right|^{-1/2} \exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)/2\right)}{\sum_{l=1}^K \pi_l \left|2\pi\boldsymbol{\Sigma}\right|^{-1/2} \exp\left(-\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)/2\right)} \\
=&\frac{1}{\sum_{l=1}^K \exp\left(-\log\left(\pi_l^{-2}/\pi_k^{-2}\right)/2 - \left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)/2 - \left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)/2\right)},
\end{aligned}$$

and note that

$$\begin{aligned}
&\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_l\right) - \left(\boldsymbol{x}-\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\nu}_k\right) \\
=&-2\left(\boldsymbol{\nu}_l-\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \left(\boldsymbol{\nu}_l+\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\nu}_l-\boldsymbol{\nu}_k\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{\pi_k \phi\left(\boldsymbol{x}; \boldsymbol{\nu}_k, \boldsymbol{\Sigma}\right)}{\sum_{l=1}^K \pi_l \phi\left(\boldsymbol{x}; \boldsymbol{\nu}_l, \boldsymbol{\Sigma}\right)} \\
=&\frac{1}{\sum_{l=1}^K \exp\left(-\log\left(\pi_l^{-2}/\pi_k^{-2}\right)/2 - \left(\boldsymbol{\nu}_l+\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\nu}_l-\boldsymbol{\nu}_k\right)/2 - \left(\boldsymbol{\nu}_l-\boldsymbol{\nu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)}.
\end{aligned}$$

Next, notice that we can write

$$\frac{\exp\left(a_k + \boldsymbol{b}_k^\top \boldsymbol{x}\right)}{\sum_{l=1}^K \exp\left(a_l + \boldsymbol{b}_l^\top \boldsymbol{x}\right)} = \frac{1}{\sum_{l=1}^K \exp\left(\alpha_l + \boldsymbol{\beta}_l^\top \boldsymbol{x}\right)},$$

where $\alpha_l = a_l - a_k$ and $\boldsymbol{\beta}_l = \boldsymbol{\beta}_l - \boldsymbol{\beta}_k$. We now choose a_k and $\boldsymbol{b}_k$, such that for every $l \in [K]$,

$$\alpha_l = a_l - a_k = -\frac{1}{2}\log\left(\frac{\pi_l^{-2}}{\pi_k^{-2}}\right) - \frac{1}{2}\left(\boldsymbol{\nu}_l^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}_l - \boldsymbol{\nu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}_k\right),$$

and

$$\boldsymbol{\beta}_l = \boldsymbol{\beta}_l - \boldsymbol{\beta}_k = \boldsymbol{\nu}_l^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{\nu}_k^\top \boldsymbol{\Sigma}^{-1}.$$

To complete the proof, we choose

$$a_k = \log\left(\pi_k\right) - \frac{1}{2}\boldsymbol{\nu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}_k$$

and $b_k = \boldsymbol{\nu}_k^\top \boldsymbol{\Sigma}^{-1}$, for each $k \in [K]$.

2.3.4.2 Proof of Lemma 2.3.7

For $l \in [0, \infty)$, write

$$\mathrm{Gate}_k\left(x, l\right) = \frac{\exp\left(\left[x - c_k\right] lk\right)}{\sum_{i=1}^n \exp\left(\left[x - c_i\right] li\right)},$$

where $x \in \mathbb{X} = [0,1]$, and $c_k = (k-1)/(2k)$. We identify that $\mathbf{Gate} = \left(\mathrm{Gate}_k\left(x, l\right)\right)_{k\in[n]}$ belongs to the class $\mathcal{G}_S^n$. The proof of the Section 4 Proposition from Jiang & Tanner (1999b) reveals that for all $k \in [n]$,

$$\mathrm{Gate}_k\left(x, l\right) \to \mathbf{1}_{\left\{x \in \mathbb{I}_k^n\right\}}$$

almost everywhere in λ, as $l \to \infty$. The result then follows via an application of Egorov's theorem (cf. Folland 1999, Thm. 2.33).

2.4 Universal approximation for mixture of experts models in approximate Bayesian computation

The Gaussian Locally Linear Mapping (GLLiM) model is briefly described in Section 2.4.1. A first exploitation of GLLiM combined with the semi-automatic ABC principle of Fearnhead & Prangle (2012) is presented in Section 2.4.2.1. Our extension, using functional summary statistics, is then described in Section 2.4.2.2. The approach's theoretical universal approximation properties are investigated in Section 2.4.3 and the practical performance is illustrated in Section 2.4.4, both on synthetic and real data. An appendix gathers additional illustration and detailed proofs, while the code can be found at https://github.com/Trung-TinNGUYEN/GLLiM-ABC. Finally, Section 5.2 concludes the Section 2.4 and discusses perspectives.

2.4.1 Parametric posterior approximation with Gaussian mixtures

A learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ is built from the joint distribution that results from the prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the likelihood $f_{\boldsymbol{\theta}}$, where $[N] = \{1, \ldots, N\}$. The idea is to capture the relationship between $\boldsymbol{\theta}$ and $\mathbf{y}$ with a joint probabilistic model for which computing conditional distributions and moments is straightforward. For the choice of the model to fit to $\mathcal{D}_N$, we propose to use the so-called Gaussian Locally Linear Mapping (GLLiM) model (Deleforge et al., 2015c) for its ability to capture non-linear relationships in a tractable manner, based on flexible mixtures of Gaussian distributions. GLLiM can be considered within the class of inverse regression approaches, such as sliced inverse regression (Li, 1991), partial least squares (Cook & Forzani, 2019), mixtures of regressions approaches of different variants, *e.g.* mixtures of experts (Nguyen et al., 2019, 2021a), see also in Section 2.3, cluster weighted models (Ingrassia et al., 2012), and kernel methods (Nataraj et al., 2018). In contrast to deep learning approaches (see Arridge et al. 2019, for a survey), GLLiM provides for each observed $\mathbf{y}$, a full posterior probability distribution within a family of parametric models $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$. To model non-linear relationships, it uses a mixture of K linear models. More specifically, the expression of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi})$ is analytical and available for all $\mathbf{y}$ with $\boldsymbol{\phi}$ being independent of $\mathbf{y}$:

$$p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y})\, \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}_k\mathbf{y} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k), \tag{2.4.1}$$

where $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\eta_k(\mathbf{y}) = \pi_k\mathcal{N}(\mathbf{y}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) / \sum_{j=1}^{K} \pi_j\mathcal{N}(\mathbf{y}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)$, see Section 3.2.1.4 for greater details. This distribution involves a number of parameters:

$$\boldsymbol{\phi} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}.$$

One interesting property of this parametric model is that the mixture setting provides guarantees that, when choosing K large enough, it is possible to approximate any reasonable relationship (Nguyen et al., 2019, 2020b, 2021a). The parameter $\boldsymbol{\phi}$ can be estimated by fitting a GLLiM model to $\mathcal{D}_N$ using an Expectation-Maximization (EM) algorithm. Details on the model and its estimation are provided in Deleforge et al. (2015c).

Fitting a GLLiM model to $\mathcal{D}_N$ therefore results in a set of parametric distributions $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\phi}^*_{K,N}),\ \mathbf{y} \in \mathcal{Y}\}$, which are mixtures of Gaussian distributions and can be seen as a parametric mapping from $\mathbf{y}$ values to posterior pdfs on $\boldsymbol{\theta}$. The parameter $\boldsymbol{\phi}^*_{K,N}$ is the same for all conditional distributions and does not need to be re-estimated for each new instance of $\mathbf{y}$. When required, it is straightforward to compute the expectation and covariance matrix of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$ in (2.4.1):

$$\begin{aligned}\mathbb{E}_G\left[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right] = \int_{\boldsymbol{\Theta}} p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})\boldsymbol{\theta} d\boldsymbol{\theta} &= \sum_{k=1}^{K} \eta_k(\mathbf{y}) \int_{\boldsymbol{\Theta}} \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k, \boldsymbol{\Sigma}^*_k)\boldsymbol{\theta} d\boldsymbol{\theta} \\ &= \sum_{k=1}^{K} \eta_k(\mathbf{y})\left(\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k\right). \end{aligned} \tag{2.4.2}$$

$$\begin{aligned}\mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}] &= \mathbb{E}_G\left[\boldsymbol{\theta}\boldsymbol{\theta}^\top \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right] - \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]\,\mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]^\top \\ &= \sum_{k=1}^{K} \eta_k(\mathbf{y}) \int_{\boldsymbol{\Theta}} \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k, \boldsymbol{\Sigma}^*_k)\boldsymbol{\theta}\boldsymbol{\theta}^\top d\boldsymbol{\theta} - \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]\,\mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]^\top \\ &= \sum_{k=1}^{K} \eta_k(\mathbf{y}) \left[\boldsymbol{\Sigma}^*_k + (\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k)(\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k)^\top\right] - \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]\,\mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]^\top. \end{aligned} \tag{2.4.3}$$

Expression (2.4.2) then provides approximate posterior means and can be directly used in a semi-automatic ABC procedure. In addition, summary statistics extracted from the covariance matrix (2.4.3) can also be included and is likely to improve the ABC selection as illustrated in Section 2.4.4.

2.4.2 Extended semi-automatic ABC

Semi-automatic ABC refers to an approach introduced in Fearnhead & Prangle (2012), which has since then led to various attempts and improvements, see *e.g.* Jiang et al. (2017) and Wiqvist et al. (2019), without dramatic deviation from the original ideas.

2.4.2.1 Extension to extra summary vectors

A natural idea is to use the approximate posterior expectation provided by GLLiM in (2.4.2) as the summary statistic s of data $\mathbf{y}$, $s(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$, and then to apply standard ABC algorithms, *e.g.* a rejection ABC. It provides a first attempt to combine GLLiM and ABC procedures and has the advantage over neural networks of being easier to estimate without the need for huge learning data sets and hyperparameter tuning.

However, one advantage of GLLiM over most regression methods is not to reduce to pointwise predictions and to provide full posteriors as output. The posteriors can then be used to provide other posterior moments as summary statistics. The same standard ABC procedure as before can be applied but now with $s_1(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$ and $s_2(\mathbf{y}) = \mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$, as given by (2.4.3). In Section 2.4.4, we show examples where s_2 is restricted to the posterior log-variances, *i.e.* the logarithms of the diagonal elements of the posterior covariance matrix.

As illustrated in Section 2.4.4, it is easy to construct examples where the posterior expectations, even when well-approximated, do not perform well as summary statistics. Providing a straightforward and tractable way to add other posterior moments is then already an interesting contribution. However, to really make the most of the GLLiM framework, we propose to further exploit the fact that GLLiM provides more than the means, variances or other moments.

2.4.2.2 Extension to functional summary statistics

Instead of comparing simulated $\mathbf{z}$'s to the observed $\mathbf{y}$, or equivalently their summary statistics, we propose to compare the $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$'s to $p_G(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\phi}^*_{K,N})$, as given by (2.4.1). As approximation of the true posteriors, these quantities are likely to capture the main characteristics of $\boldsymbol{\theta}$ without committing to the choice of a particular moment. The comparison requires an appropriate distance that needs to be a mathematical distance between distributions. The equivalent functional distance to the L_2 distance can still be used, as can the Hellinger distance or any other divergence. A natural choice is the Kullback–Leibler divergence, but computing Kullback–Leibler divergences between mixtures is not straightforward. Computing the Energy statistic (*e.g.*, Nguyen et al., 2020a) appears at first to be easier but in the end that would still resort to Monte Carlo sums. Since model (2.4.1) is parametric, we could also compute distances between the parameters of the mixtures that depend on $\mathbf{y}$. That is for $k \in [K]$, between the mixing proportions $\eta^*_k(\mathbf{y}) = \frac{\pi^*_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}^*_k, \boldsymbol{\Gamma}^*_k)}{\sum_{j=1}^{K} \pi^*_j \mathcal{N}(\mathbf{y}; \boldsymbol{c}^*_j, \boldsymbol{\Gamma}^*_j)}$ and conditional means $\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k$. But this may lead us back to the usual issue with distances between summary statistics and also we may have to face the label switching issue, not easily handled within ABC procedures.

Recently, interesting developments regarding the Wasserstein distance and Gaussian mixtures have emerged (Delon & Desolneux, 2020, Chen et al., 2019), introducing an optimal transport-based distance between Gaussian mixtures. We recall the definition of this distance, denoted by MW_2 in Section 2.4.5 and describe our next ABC procedure, referred to as GLLiM-MW2-ABC. The L_2 distance between mixtures is also very straightforward to compute and recalled in Section 2.4.5, leading to another procedure, which we call GLLiM-L2-ABC. Both procedures are referred to as functional GLLiM-ABC procedures. We will often write GLLiM-D-ABC to include both cases and possibly other distances D.

The semi-automatic ABC extensions that we propose are all summarized in Algorithm 1. To be general, Algorithm 1 is presented with two simulated data sets, one for training GLLiM and constructing the surrogate posteriors, and one for the ABC selection procedure itself, but the same data set could be used as in semi-automatic ABC (Fearnhead & Prangle, 2012). For rejection ABC, the selection also requires the user to fix a threshold ϵ. It is common practice to set ϵ to a quantile of the computed distances. GLLiM then requires the choice of K, the number of Gaussians in the mixtures. K can be chosen using asymptotic model selection criteria (see Deleforge et al., 2015c), or non-asymptotic model selection in Chapters 3 and 4, *e.g.,* slope heuristic (Nguyen et al., 2021c,b), but its precise value is not critical, all the more so if GLLiM is not used for prediction, directly. See details in Section 2.4.4.

Algorithm 1 GLLiM-ABC algorithms – Vector and functional variants

1: **Inverse operator learning.** Apply GLLiM on a training set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ to estimate, for any $\mathbf{z} \in \mathcal{Y}$, the K-Gaussian mixture $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$ in (2.4.1) as a first approximation of the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$, where $\boldsymbol{\phi}^*_{K,N}$ does not depend on $\mathbf{z}$.

2: **Distances computation.** Consider another simulated set $\mathcal{E}_M = \{(\boldsymbol{\theta}_m, \mathbf{z}_m), m \in [M]\}$. For a given observed $\mathbf{y}$, do one of the following for $m \in [M]$:

Vector summary statistics. (Section 2.4.2.1)

GLLiM-E-ABC: Compute summary statistics $s_1(\mathbf{z}_m) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$ (2.4.2).

GLLiM-EV-ABC: Compute both $s_1(\mathbf{z}_m)$ and $s_2(\mathbf{z}_m)$ by considering also posterior log-variances derived from (2.4.3).

In both cases, compute standard distances between summary statistics.

Functional summary statistics. (Section 2.4.2.2)

GLLiM-MW2-ABC: Compute $\mathrm{MW}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

GLLiM-L2-ABC: Compute $\mathrm{L}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

3: **Sample selection.** Select the $\boldsymbol{\theta}_m$ values that correspond to distances under an ϵ threshold (rejection ABC) or apply an ABC procedure that can handle distances, directly.

4: **Sample use.** For a given observed $\mathbf{y}$, use the produced sample of $\boldsymbol{\theta}$ values to compute a closer approximation of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

2.4.3 Universal approximation properties

Before illustrating the proposed GLLiM-D-ABC procedures performance, we investigate the theoretical properties of our ABC quasi-posterior defined via surrogate posteriors.

Let $\mathcal{X} = \Theta \times \mathcal{Y}$ and $(\mathcal{X}, \mathcal{F})$ be a measurable space. Let λ be a σ-finite measure on $\mathcal{F}$. Whenever we mention below that a probability measure Pr on $\mathcal{F}$ has a density, we will understand that it has a Radon–Nikodym derivative with respect to λ (λ can typically be chosen as the Lebesgue measure on the Euclidean space). For all $p \in [1, \infty)$ and f, g in appropriate spaces, let $D_p(f, g) = \left(\int |f(\mathbf{x}) - g(\mathbf{x})|^p \, d\lambda(\mathbf{x})\right)^{1/p}$ denote the L_p distance and $D_H^2(f, g) = \int (\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})})^2 d\lambda(\mathbf{x})$ be the squared Hellinger distance. When not specified otherwise, let D be an arbitrary distance on $\mathcal{Y}$ or on densities, depending on the context. We further denote the L_p norm for vectors by $\|\cdot\|_p$.

In a GLLiM-D-ABC procedure, the ABC quasi-posterior is constructed as follows. Let $p_G^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$ be the surrogate conditional distribution of form (2.4.1), learned from a preliminary GLLiM model with K components and using a learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$.

This conditional distribution is a K-component mixture, which depends on a set of learned parameters $\boldsymbol{\phi}^*_{K,N}$, independent of $\mathbf{y}$. The GLLIM-D-ABC quasi-posterior resulting from the GLLiM-D-ABC procedure then depends both on K, N and the tolerance level ϵ and can be written as

$$q^{K,N}_{G,\epsilon}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}_G(\,\cdot\,|\mathbf{y}),\, p^{K,N}_G(\,\cdot\,|\mathbf{z}))\leq\epsilon\}}\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}\,, \tag{2.4.4}$$

where D is a distance on densities such as the MW_2 and L_2 metrics, which are both proper distances (see Section 2.4.5).

We provide two types of results, below. In the first result (Theorem 2.4.1), the true posterior is used to compare samples $\mathbf{y}$ and $\mathbf{z}$. This result aims at providing insights on the proposed quasi-posterior formulation and at illustrating its potential advantages. In the second result (Theorem 2.4.2), a surrogate posterior is learned and used to compare samples. Conditions are specified under which the resulting ABC quasi-posterior converges to the true posterior.

2.4.3.1 Convergence of the ABC quasi-posterior

In this section, we assume a fixed given observed $\mathbf{y}$ and the dependence on $\mathbf{y}$ is omitted from the notation, when there is no confusion.

Let us first recall the standard form of the ABC quasi-posterior, omitting summary statistics from the notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}\,. \tag{2.4.5}$$

If D is a distance and $D(\mathbf{y},\mathbf{z})$ is continuous in $\mathbf{z}$, the ABC posterior in (2.4.5) can be shown to have the desirable property of converging to the true posterior when ϵ tends to 0 (see Prangle et al., 2018).

The proof is based on the fact that when ϵ tends to 0, due to the property of the distance D, the set $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y},\mathbf{z}) \leq \epsilon\}$, defining the indicator function in (2.4.5), tends to the singleton $\{\mathbf{y}\}$ so that consequently $\mathbf{z}$ in the likelihood can be replaced by the observed $\mathbf{y}$, which then leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore to the true posterior as desired (see also Rubio & Johansen, 2013, Bernton et al., 2019). It is interesting to note that this proof is based on working on the term under the integral only and is using the equality, at convergence, of $\mathbf{z}$ to $\mathbf{y}$, which is actually a stronger than necessary assumption for the result to hold. Alternatively, if we first rewrite (2.4.5) using Bayes' theorem, it follows that

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, \pi(\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z} \;\propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, \pi(\mathbf{z})\, d\mathbf{z}\,. \tag{2.4.6}$$

That is, when accounting for the normalizing constant:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, \pi(\mathbf{z})\, d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\, \pi(\mathbf{z})\, d\mathbf{z}}\,. \tag{2.4.7}$$

Using this equivalent formulation, we can then replace $D(\mathbf{y},\mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, with D now denoting a distance on densities, and obtain the same convergence result when ϵ tends to 0. More specifically, we can show the following general result. Let us define our ABC quasi-posterior as,

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}}\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z},$$

which can be written as

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z})\, \pi(\mathbf{z})\, d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} \pi(\mathbf{z})\, d\mathbf{z}}\,. \tag{2.4.8}$$

The following theorem shows that $q_\epsilon(\cdot \mid \mathbf{y})$ converges to $\pi(\cdot \mid \mathbf{y})$ in total variation, for fixed $\mathbf{y}$. The proof is detailed in Section 2.4.6.1.

Theorem 2.4.1. *For every $\epsilon > 0$, let $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})) \leq \epsilon\}$. Assume the following:*

(A1) $\pi(\boldsymbol{\theta} \mid \cdot)$ is continuous for all $\boldsymbol{\theta} \in \Theta$, and $\sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \mathbf{y}) < \infty$;

(A2) There exists a $\gamma > 0$ such that $\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty$;

(A3) $D(\cdot, \cdot) : \Pi \times \Pi \to \mathbb{R}_+$ is a metric on the functional class $\Pi = \{\pi(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}\}$;

(A4) $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$ is continuous, with respect to $\mathbf{z}$.

Under (A1)–(A4), $q_\epsilon(\cdot \mid \mathbf{y})$ in (2.4.8) converges in total variation to $\pi(\cdot \mid \mathbf{y})$, for fixed $\mathbf{y}$, as $\epsilon \to 0$.

It appears that what is important is not to select $\mathbf{z}$'s that are close (and at the limit equal) to the observed $\mathbf{y}$ but to choose $\mathbf{z}$'s so that the posterior $\pi(\cdot \mid \mathbf{z})$ (the term appearing in the integral in (2.4.6)) is close (and at the limit equal) to $\pi(\cdot \mid \mathbf{y})$. And this last property is less demanding than $\mathbf{z} = \mathbf{y}$. Potentially, there may be several $\mathbf{z}$'s satisfying $\pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})$, but this is not problematic when using (2.4.6), while it is problematic when following the standard proof as in Bernton et al. (2019).

2.4.3.2 Convergence of the ABC quasi-posterior with surrogate posteriors

In most ABC settings, based on data discrepancy or summary statistics, the above consideration and result are not useful because the true posterior is unknown by construction and cannot be used to compare samples. However this principle becomes useful in our setting, which is based on surrogate posteriors. While the previous result can be seen as an oracle of sorts, it is more interesting in practice to investigate whether a similar result holds when using surrogate posteriors in the ABC likelihood. This is the goal of Theorem 2.4.2 below, which we prove for a restricted class of target distribution and of surrogate posteriors that are learned as mixtures.

We now assume that $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and consider the following class $\mathcal{H}_\mathcal{X}$ of distributions on $\mathcal{X}$, $\mathcal{H}_\mathcal{X} = \{g_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \Psi\}$, with constraints on the parameters, Ψ being a bounded parameter set. In addition the densities in $\mathcal{H}_\mathcal{X}$ are assumed to satisfy for any $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$, there exist arbitrary positive scalars a, b and B such that

$$\text{for all } \mathbf{x} \in \mathcal{X},\ a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \text{ and } \sup_{\mathbf{x} \in \mathcal{X}} |\log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x})| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1 .$$

We denote by p^K a K-component mixture of distributions from $\mathcal{H}_\mathcal{X}$ and defined for all $\mathbf{y} \in \mathcal{Y}$, $p^{K,N}(\cdot \mid \mathbf{y})$ as follows:

$$\forall \boldsymbol{\theta} \in \Theta, \quad p^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p^K\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right),$$

with $\boldsymbol{\phi}^*_{K,N}$ the maximum likelihood estimate (MLE) for the data set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$, generated from the true joint distribution $\pi(\cdot, \cdot)$:

$$\boldsymbol{\phi}^*_{K,N} = \arg\max_{\boldsymbol{\phi} \in \Phi} \sum_{n=1}^{N} \log\left(p^K(\boldsymbol{\theta}_n, \mathbf{y}_n; \boldsymbol{\phi})\right).$$

In addition, for every $\epsilon > 0$, let $A^{K,N}_{\epsilon,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) \leq \epsilon\right\}$ and $q^{K,N}_\epsilon$ denote the ABC quasi-posterior defined with $p^{K,N}$ by

$$q^{K,N}_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_\mathcal{Y} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z} . \tag{2.4.9}$$

Theorem 2.4.2. *Assume the following: $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and*

(B1) For joint density π, there exists G_π a probability measure on Ψ such that, with $g_{\boldsymbol{\varphi}} \in \mathcal{H}_\mathcal{X}$, $\pi(\mathbf{x}) = \int_\Psi g_{\boldsymbol{\varphi}}(\mathbf{x})\, G_\pi(d\boldsymbol{\varphi})$;

(B2) The true posterior density $\pi(\cdot \mid \cdot)$ is continuous both with respect to $\boldsymbol{\theta}$ and $\mathbf{y}$;

(B3) $D(\cdot,\cdot) : \Pi \times \Pi \to \mathbb{R}_+ \cup \{0\}$ *is a metric on a functional class* Π*, which contains the class* $\left\{p^{K,N}(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}^\star, N \in \mathbb{N}^\star\right\}$*. In particular,* $D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) = 0$*, if and only if* $p^{K,N}(\cdot \mid \mathbf{y}) = p^{K,N}(\cdot \mid \mathbf{z})$*;*

(B4) For every $\mathbf{y} \in \mathcal{Y}$*,* $\mathbf{z} \mapsto D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right)$ *is a continuous function on* $\mathcal{Y}$*.*

Then, under (B1)–(B4), the Hellinger distance $D_{\mathrm{H}}\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ *converges to* 0 *in some measure* λ*, with respect to* $\mathbf{y} \in \mathcal{Y}$ *and in probability, with respect to the sample* $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$*. That is, for any* $\alpha > 0, \beta > 0$*, it holds that*

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1. \tag{2.4.10}$$

Sketch of the proof of Theorem 2.4.2. For all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$, the quasi-posterior (2.4.9) can be written equivalently as

$$q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, d\mathbf{z},$$
$$\text{with } K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) = \frac{\mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z})}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\tilde{\mathbf{z}})) \leq \epsilon\}} \pi(\tilde{\mathbf{z}}) \, d\tilde{\mathbf{z}}},$$

where $K_\epsilon^{K,N}(\cdot; \mathbf{y})$ is a pdf, with respect to $\mathbf{z} \in \mathcal{Y}$, with compact support $A_{\epsilon,\mathbf{y}}^{K,N} \subset \mathcal{Y}$, by definition of $A_{\epsilon,\mathbf{y}}^{K,N}$ and (B4). Using the relationship between Hellinger and L_1 distances (see details in Section 2.4.6.2 relations (2.4.19) and (2.4.20)), it then holds that

$$D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \leq 2 D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{y})\right), \tag{2.4.11}$$

where there exists $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in B_{\epsilon,\mathbf{y}}^{K,N}$ with

$$B_{\epsilon,\mathbf{y}}^{K,N} = \arg\max_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1\left(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})\right).$$

The next step is to bound the right-hand side of (2.4.11) using the triangle inequality with respect to the Hellinger distance D_{H}. Consider the limit point $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ defined as $\mathbf{z}_{0,\mathbf{y}}^{K,N} = \lim_{\epsilon \to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$. Since for each $\epsilon > 0$, $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in A_{\epsilon,\mathbf{y}}^{K,N}$ it holds that $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$, where $A_{0,\mathbf{y}}^{K,N} = \bigcap_{\epsilon \in \mathbb{Q}_+} A_{\epsilon,\mathbf{y}}^{K,N}$. By continuity of D, $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}(\cdot \mid \mathbf{z}), p^{K,N}(\cdot \mid \mathbf{y})\right) = 0\right\}$ and $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}(\cdot \mid \mathbf{z}) = p^{K,N}(\cdot \mid \mathbf{y})\right\}$, using (B3). The distance on the right-hand side of (2.4.11) can then be decomposed in three parts,

$$\begin{aligned} D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{y})\right) \leq\ & D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{y})\right) \\ & + D_{\mathrm{H}}\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right). \end{aligned} \tag{2.4.12}$$

The first term in the right-hand side can be made close to 0 as ϵ goes to 0 independently of K and N. The two other terms are of the same nature, and the definition of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ yields $p^{K,N}(\cdot \mid \mathbf{y}) = p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})$.

Using the fact that $\pi(\cdot \mid \cdot)$ is a uniformly continuous function in $(\boldsymbol{\theta}, \mathbf{y})$ on a compact set $\mathcal{X}$ and taking the limit $\epsilon \to 0$, yields $\lim_{\epsilon \to 0} D_H^2\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) = 0$ in measure λ, with respect to $\mathbf{y} \in \mathcal{Y}$. Since this result is true whatever the data set $\mathcal{D}_N$, it also holds in probability with respect to $\mathcal{D}_N$. That is, given any $\alpha_1 > 0$, $\beta_1 > 0$, there exists $\epsilon(\alpha_1, \beta_1) > 0$ such that for any $0 < \epsilon < \epsilon(\alpha_1, \beta_1)$,

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) \geq \alpha_1\right) = 0.$$

Next, we prove that

$$D_H^2\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{y})\right) \left(\text{equal to } D_H^2\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right)\right) \text{ and } D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$$

both converge to 0 in measure λ, with respect to $\mathbf{y}$ and in probability, with respect to $\mathcal{D}_N$. Such convergence can be obtained via Rakhlin et al. (2005, Corollary 2.2), and Lemma 2.4.5 in Section 2.4.6.3, which provides the guarantee that we can choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N}$. Equation (2.4.10) in Theorem 2.4.2 follows from the triangle inequality (2.4.12). A detailed proof is provided in Section 2.4.6.2.

Remark 2.4.3. The GLLiM model involving multivariate unconstrained Gaussian distributions does not satisfy the conditions of Theorem 2.4.2 so that $p^{K,N}$ cannot be replaced by $p_G^{K,N}$ in the theorem. However as illustrated in Rakhlin et al. (2005), truncated Gaussian distributions with constrained parameters can meet the restrictions imposed in the theorem. We are not aware of any more general result involving the MLE of Gaussian mixtures. The GLLiM model could as well be replaced by another model satisfying the conditions of the theorem but for practical applications, this model would need to have computational properties such as the tractability of the estimation of its parameters and needs to be efficient in multivariate and potentially high-dimensional settings.

2.4.4 Numerical experiments

Most benchmark examples in ABC correspond to unimodal and light tailed posterior distributions. Such settings may not be the most appropriate to show differences and discriminate between the performance of our methods. We therefore consider settings that are simple in terms of dimension and complexity but exhibit posterior distributions with characteristics such as bimodality and heavy tails. A first set of two synthetic examples is considered with parameters in dimensions 1 or 2 and bimodal posterior distributions (Section 2.4.4.1). A third example is derived from a real application in sound source localization, where the posterior distribution has mass on four 1D manifolds (Section 2.4.4.2). All of these examples are run for a single observation in $d = 10$ dimensions.

To circumvent the choice of an arbitrary summary statistic, Fearnhead & Prangle (2012) showed that the best summary statistic, in terms of the minimal quadratic loss, was the posterior mean. This posterior mean is not known and needs to be approximated. In Fearnhead & Prangle (2012) a regression approach is proposed to provide a way to compute summary statistics prior to the ABC rejection sampling, itself. In this section, the transformations used for the regression part are $(1, y, y^2, y^3, y^4)$ following the procedure suggested in the **abctools** package (Nunes & Prangle, 2015). We refer to this procedure as semi-automatic ABC. We did not try to optimise the procedure using other transformations but did not notice systematic improvements when increasing the number of polynomial terms, for instance. This approach using the posterior mean is further developed in Jiang et al. (2017), where a multilayer perceptron deep neural network regression model is employed and replaces the linear regression model of Fearnhead & Prangle (2012). The deep neuronal network with multiple hidden layers considered by Jiang et al. (2017) offers stronger representational power to approximate the posterior mean and hence to learn an informative summary statistic, when compared to linear regression models. Improved results were obtained by Jiang et al. (2017), but we did not compare our approach to their method. As our current examples are of relatively small dimension d, we also did not draw comparisons with discrepancy-based ABC techniques such as WABC (Bernton et al., 2019) or classification ABC (Gutmann et al., 2018), which are designed for a more data-rich setting.

The performance of the four proposed GLLiM-ABC schemes summarized in Algorithm 1 is compared to that of semi-automatic ABC. All reported results are obtained with a simple rejection scheme as per instances implemented in the **abc** R package (Csillery et al., 2012). The other schemes available in the **abc** package have been tested but no notable performance differences were observed. In regards to the final sample thresholding (*i.e.*, choice of ϵ), following common practice, all methods retain samples for which the distance to the observation is under a small (*e.g.* 0.1%) quantile of all computed distances.

The **xLLiM** R package, available on the CRAN, is used to learn a GLLiM model with K components and an isotropic constraint, from a set $\mathcal{D}_N$ of N simulations from the true model. The isotropic GLLiM is simpler than the fully-specified GLLiM and we observed that it provided surrogate posteriors of sufficient quality for the ABC selection scheme. The exact meaning of this constraint can be found in Deleforge et al. (2015c), Perthame et al. (2017). Another set of simulated couples $(\boldsymbol{\theta}, \mathbf{y})$ of size M is generally used for the ABC rejection scheme unless otherwise specified.

2.4.4.1 Non-identifiable models

It is straightforward to construct models that lead to multimodal posteriors by considering likelihoods that are invariant by some transformation.

Ill-posed inverse problems

Here, we consider inverse problems for which the solution is not unique. This setting is quite common in practice and can occur easily when the forward model exhibits some invariance, *e.g.*, when considering the negative of the parameters. A simple way to model this situation consists of assuming that the observation $\mathbf{y}$ is generated as a realization of

$$\mathbf{y} = F(\boldsymbol{\theta}) + \boldsymbol{\varepsilon},$$

where F is a deterministic theoretical model coming from experts and $\boldsymbol{\varepsilon}$ is a random variable expressing the uncertainty both on the theoretical model and on the measurement process. A common assumption is that $\boldsymbol{\varepsilon}$ is distributed as a centered Gaussian noise. Non-identifiability may then arise when $F(-\boldsymbol{\theta}) = F(\boldsymbol{\theta})$. Following this generative approach, a first simple example is constructed with a Student t-distributed noise leading to the likelihood:

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \mathcal{S}_d(\mathbf{y}; \mu^2 \mathrm{I}_d, \sigma^2 \mathrm{I}_d, \nu),$$

where $\mathcal{S}_d(\cdot; \mu^2 \mathrm{I}_d, \sigma^2 \mathrm{I}_d, \nu)$ is the pdf of a d-variate Student t-distribution with a d-dimensional location parameter with all dimensions equal to μ^2, diagonal isotropic scale matrix $\sigma^2\, \mathrm{I}_d$ and degree-of-freedom (dof) parameter ν. Recall that for a Student t-distribution, a diagonal scale matrix is not inducing independent dimensions so that $\mathbf{y}$ is not a set of *i.i.d.* univariate Student t observations. The dof controls the tail heaviness; *i.e.*, the smaller the value of ν, the heavier the tail. In particular, for $\nu \leq 2$, the variance is undefined, while for $\nu \leq 1$ the expectation is also undefined. In this example, we set $\sigma^2 = 2$, $\nu = 2.1$, and μ is the parameter to estimate.

For all compared procedures, we set $d = 10$, $K = 10$, $N = M = 10^5$, and the tolerance level ϵ to the 0.1% quantile of observed distances, so that all selected posterior samples are of size 100. To visualize posterior samples densities, we use a density estimation procedure based on the **ggplot2** R package with a Gaussian kernel.

Figure 2.1 shows the true and the compared ABC posterior distributions for a 10-dimensional observation $\mathbf{y}$, simulated under a process with $\mu = 1$. The true posterior exhibits the expected symmetry with modes close to the values: $\mu = 1$ and $\mu = -1$. The simple rejection ABC procedure based on GLLiM expectations (GLLiM-E-ABC) and the semi-automatic ABC procedure both show over dispersed samples with wrongly located modes. The GLLiM-EV-ABC exhibits two well located modes but does not preserve the symmetry of the true posterior. The distance-based approaches, GLLiM-L2-ABC and GLLiM-MW2-ABC both capture the bimodality. GLLiM-MW2-ABC is the only method to estimate a symmetric posterior distribution with two modes of equal importance. Note, however, that in term of precision, the posterior distribution estimation remains difficult considering an observation of size only $d = 10$.

This simple example shows that the expectation as a summary statistic suffers from the presence of two equivalent modes, while the approaches based on distances are more robust. There is a clear improvement in complementing the summary statistics with the log-variances. Although in this case, this augmentation provides a satisfying bimodal posterior estimate, it lacks the expected symmetry of the two modes. The GLLiM-MW2-ABC procedure has the advantage of exhibiting a symmetric posterior estimate, that is more consistent with the true posterior.

In the following subsection we present another case that cannot be cast as the above generating process but also exhibit a transformation invariant likelihood.

Sum of moving average models of order 2 (MA(2))

Moving average (MA) models are commonly studied in the ABC literature, see *e.g.* Marin et al. (2012), Jiang et al. (2018), Nguyen et al. (2020a). An example using MA(1) processes is provided in

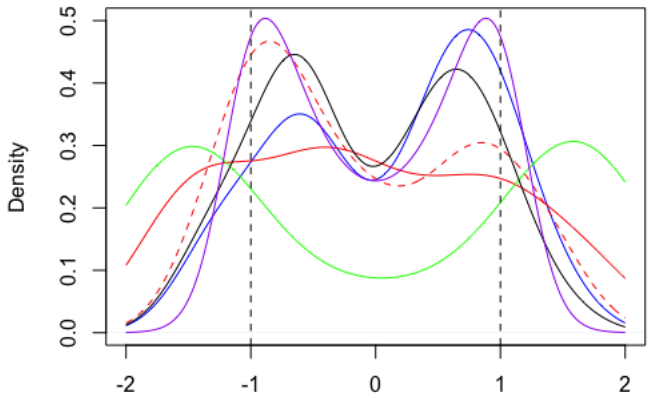


Figure 2.1: Non identifiable Student t-distribution. ABC posterior distributions from the selected samples. GLLiM-L2-ABC in blue, GLLiM-MW2-ABC in black, semi-automatic ABC in green, GLLiM-E-ABC (expectations) in red and GLLiM-EV-ABC (expectations and log-variances) in dotted red line. The true posterior is shown in purple. The dashed lines indicate the μ (equivalent) values used to generate the observation.

Forbes et al. (2021, Section 6.1.2). In this section, we consider MA(2) models. The MA(2) process is a stochastic process $(y'_t)_{t\in\mathbb{N}^\star}$ defined by

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2},$$

where $\{z_t\}$ is an *i.i.d.* sequence, according to a standard normal distribution and θ_1 and θ_2 are scalar parameters. A standard identifiability condition is imposed on this model leading to a prior distribution uniform on the triangle described by the inequalities

$$-2 < \theta_1 < 2, \quad \theta_1 + \theta_2 > -1, \quad \theta_1 - \theta_2 < 1\,.$$

We consider a transformation that consists of taking the opposite sign of θ_1 and keeping θ_2 unchanged. The considered observation corresponds then to a series obtained by summing the two MA models, defined below

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2}, \quad y''_t = z'_t - \theta_1 z'_{t-1} + \theta_2 z'_{t-2}, \quad y_t = y'_t + y''_t,$$

where $\{z_t\}$ and $\{z'_t\}$ are both *i.i.d.* sequences, generated from a standard normal distribution. It follows that a vector of length d, $\mathbf{y} = (y_1, \ldots, y_d)^\top$, is distributed according to a multivariate d-dimensional centered Gaussian distribution with a Toeplitz covariance matrix whose first row is $(2(\theta_1^2 + \theta_2^2 + 1), 0, 2\theta_2, 0, \ldots, 0)$. The likelihood is therefore invariant by the transformation proposed above, and so is the uniform prior over the triangle. It follows that the posterior is also invariant by the same transformation and can then be chosen so as to exhibit two symmetric modes.

For all procedures, we set $K = 80$ and $N = M = 10^5$, and ϵ to the 1% distance quantile, so that all selected posterior samples are of size 1000. An observation of size $d = 10$ is simulated from the model with $\theta_1 = 0.7$ and $\theta_2 = 0.5$. ABC posterior distribution estimates are shown in Figure 2.2.

The level sets of the true posterior can be computed from the exact likelihood and a grid of values for θ_1 and θ_2. For the setting used in this thesis, none of the considered ABC procedures is fully satisfactory, in that the selected samples are all quite dispersed. This is mainly due to the relatively low size of the observation ($d = 10$). This can also be observed in Marin et al. (2012) (Figures 1 and 2), where ABC samples are less dispersed for a size of $d = 100$ and quite spread off when d is reduced to $d = 50$, even when the autocovariance is used as summary statistic.

Despite the relative spread of the parameters accepted after the ABC rejection, the posterior marginals, shown in Figure 2.2, provide an interesting comparison. GLLiM-D-ABC and GLLiM-EV-ABC procedures show symmetric θ_1 values, in accordance with the symmetry and bimodality of the true posterior. The use of the L_2 or MW_2 distances does not lead to significant differences. GLLiM-E-ABC and semi-automatic ABC behave similarly and do not capture the bimodality on θ_1, but the

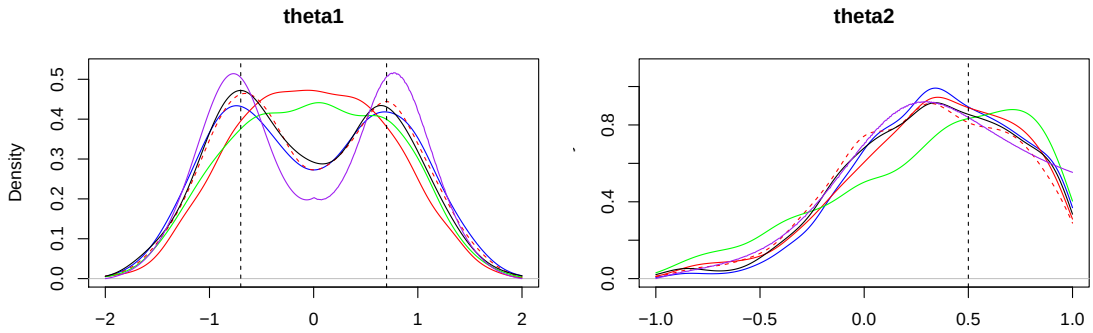


Figure 2.2: Sum of MA(2) models. Posterior marginals from the samples selected with a 1% quantile (1000 values): semi-automatic ABC (green), GLLiM-L2-ABC (blue), GLLiM-MW2-ABC (black), GLLiM-E-ABC (red) and GLLiM-EV-ABC (dotted red). The true marginal posteriors are shown in purple. The dashed lines show the values used to simulate the observation $\theta_1 = 0.7$ and $\theta_2 = 0.5$.

addition of the posterior log-variances in GLLiM-EV-ABC improves on GLLiM-E-ABC. These results suggest that although GLLiM may not provide good approximations of the first posterior moments, it can still provide good enough approximations of the surrogate posteriors in GLLiM-D-ABC. For θ_2, all posteriors are rather close to the true posterior marginal except for semi-automatic ABC which shows a mode at a wrong location when compared to the true posterior.

2.4.4.2 Sound source localization

The next two examples are constructed from a real sound source localization problem in audio processing. Although microphone arrays provide the most accurate sound source localization, setups limited to two microphones, *e.g.* Beal et al. (2003), Hospedales & Vijayakumar (2008), are often considered to mimic binaural hearing that resembles the human head with applications such as autonomous humanoid robot modelling. Binaural localization cues (Wang & Brown, 2006) include interaural time difference (ITD), interaural level difference (ILD) and interaural phase difference (IPD).

Two microphone setup

We first consider an artificial two microphone setup in a 2D scene. The object of interest is a sound source located at an unknown position $\boldsymbol{\theta} = (x, y)$. The two microphones are assumed to be located at known positions, respectively denoted by $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. A good cue for the sound source localization is the interaural time difference (ITD). The ITD is the difference between two times: the time a sound emitted from the source is acquired by microphone 1 at $\boldsymbol{m}_1$ and the time at microphone 2 at $\boldsymbol{m}_2$. ITD values are widely used by auditory scene analysis methods (Wang & Brown, 2006).

The function F that maps a location $\boldsymbol{\theta}$ onto an ITD observation is

$$F(\boldsymbol{\theta}) = \frac{1}{c}(\|\boldsymbol{\theta} - \boldsymbol{m}_1\|_2 - \|\boldsymbol{\theta} - \boldsymbol{m}_2\|_2), \tag{2.4.13}$$

where c is the sound speed in real applications but set to 1 in our example for the purpose of illustration. The important point is that an ITD value does not correspond to a unique point in the scene space, but rather to a whole surface of points. In fact, each isosurface defined by (2.4.13) is represented by one sheet of a two-sheet hyperboloid in 2D. Hence, each ITD observation constrains the location of the auditory source to lie on a 1D manifold. The corresponding hyperboloid is determined by the sign of the ITD. In our example, to create a multimodal posterior, we modify the usual setting by taking the absolute value of the ITD so that solutions can now lie on either of the two hyperboloids. In addition we assume that ITDs are observed with some Student t noise that implies heavy tails and possible outliers. Although the ITD is a univariate measure, we consider a more general d dimensional setting by defining the following Student t likelihood, $\mathbf{y} = (y_1, \ldots, y_d)$ and $\text{ITD}(\boldsymbol{\theta}) = \mid \|\boldsymbol{\theta} - \boldsymbol{m}_1\|_2 - \|\boldsymbol{\theta} - \boldsymbol{m}_2\|_2 \mid$,

where

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \mathcal{S}_d(\mathbf{y}; \mathrm{ITD}(\boldsymbol{\theta})\mathrm{I}_d, \sigma^2\mathrm{I}_d, \nu) \,. \tag{2.4.14}$$

The above likelihood corresponds to a d-variate Student t-distribution with a d-dimensional location parameter with all dimensions equal to $\mathrm{ITD}(\boldsymbol{\theta})$, diagonal isotropic scale matrix equal to $\sigma^2\,\mathrm{I}_d$ and degree-of-freedom (dof) parameter ν.

The parameter space is assumed to be $\Theta = [-2, 2] \times [-2, 2]$ and the prior on $\boldsymbol{\theta}$ is assumed to be uniform on Θ. The microphones' positions are $\boldsymbol{m}_1 = (-0.5, 0)$ and $\boldsymbol{m}_2 = (0.5, 0)$. We assume $\nu = 3$ and $\sigma^2 = 0.01$. The true $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta} = (1.5, 1)$ and we simulate a 10-dimensional $\mathbf{y}$ following model (2.4.14).

The four ABC methods using GLLiM and semi-automatic ABC are compared. Results are reported in Forbes et al. (2021, Section 6.2.1).

Two pairs of microphones setting

We build on the previous example to design a more complex setting. Two pairs of microphones are considered respectively located at $((-0.5, 0), (0.5, 0))$ and $((0, -0.5), (0, 0.5))$. The ITD vectors are assumed to be measured with equal probability either from the first pair or from the second pair. It results a likelihood that is a mixture of two equal weight components both following the previous model but for different microphones locations. The 10-dimensional observation $\mathbf{y}$ is generated from a source at location $(1.5, 1)$. Depending on whether this observation is coming from the first pair or second pair component, it results a true posterior as shown in Figure 2.4 (h) or one with non-intersecting hyperbolas. The contour plot indicates that the observation corresponds to the $((0, -0.5), (0, 0.5))$ pair. A sample obtained using the Metropolis–Hastings algorithm, as implemented in the R package **mcmc** (Geyer & Jonhson, 2020), is shown in Figure 2.4 (d).

The GLLiM model used consists of $K = 20$ Gaussian components with an isotropic constraint. A selected sample of 1000 values is retained by thresholding the distances under the 0.1% quantile. In a first test, semi-automatic ABC and GLLiM use the same data set of size $M = 10^6$ which is also used for the rejection ABC part. Selected samples are shown in Figure 2.3. The mixture provided by GLLiM as an approximation of the true posterior (Figure 2.3 (d)) well captures the main posterior parts. This GLLiM posterior is a 20-component Gaussian mixture of form (2.4.1). The true posterior expectations are all zero and are thus not informative about the location parameters. However, a correct structure can be seen in the GLLiM-E-ABC sample, in contrast to the semi-automatic one that shows no structure as expected. Adding the posterior log-variance estimations has a good impact on the selected sample, which is only marginally different from the GLLiM-D-ABC samples. This suggests that the posterior log-variances are very informative on the location parameters.

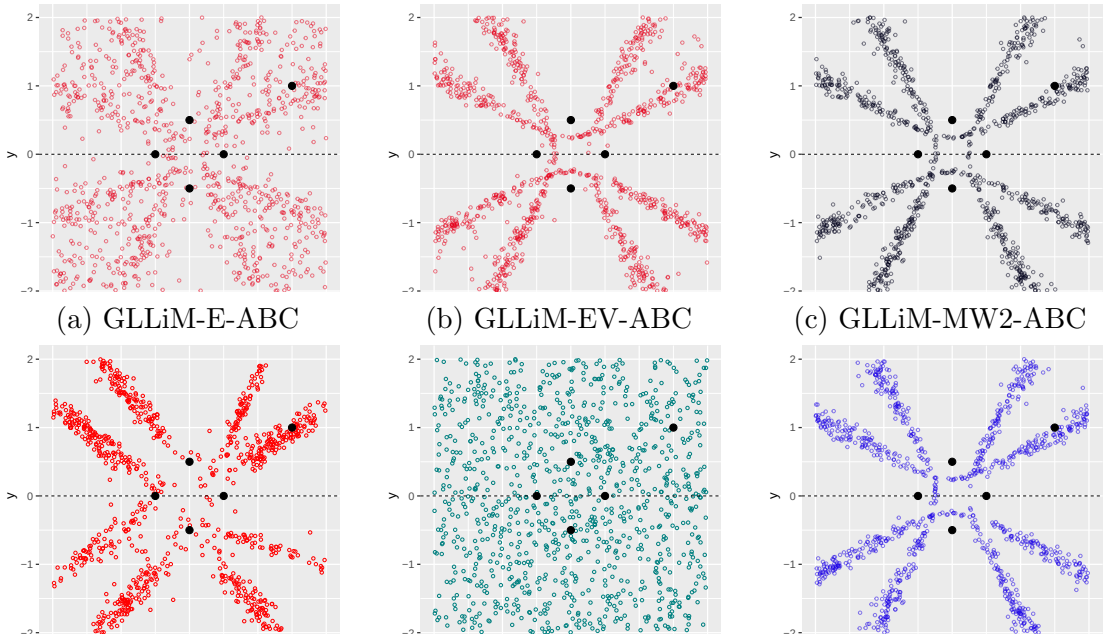
When GLLiM is first learned with a smaller data set of size $N = 10^5$ and different from the rejection ABC data set, results slightly degrade, but not significantly so (Figure 2.4). More badly localized estimations can be seen in the samples of Figure 2.4 (a,b), but the GLLiM-D-ABC samples are well localized and are not really impacted by this difference in the GLLiM learning step. In this case the improvement of GLLiM-D-ABC over GLLiM-EV-ABC is clearer.

2.4.5 Appendix: Distances between Gaussian mixtures

2.4.5.1 Optimal transport-based distance between Gaussian mixtures

Delon & Desolneux (2020), Chen et al. (2019) have introduced a distance specifically designed for Gaussian mixtures based on the Wasserstein distance. In an optimal transport context, by restricting the possible coupling measures (*i.e.*, the optimal transport plan) to a Gaussian mixture, they propose a discrete formulation for this distance. This makes it tractable and suitable for high dimensional problems, while in general using the standard Wasserstein distance between mixtures is problematic. Delon & Desolneux (2020) refer to the proposed new distance as MW_2, for *Mixture Wasserstein*.

The MW_2 definition makes first use of the tractability of the Wasserstein distance between two Gaussians for a quadratic cost. The standard quadratic cost Wasserstein distance between two Gaus-



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC

(d) GLLiM mixture (e) Semi-automatic ABC (f) GLLiM-L2-ABC

Figure 2.3: Sound source localization with a mixture of two microphones pairs. GLLiM is learned with the largest data set of size $M = 10^6$. Selected samples using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) the approximate GLLiM posterior for the observed data, (e) semi-automatic ABC, (f) L_2 distances. Black points on the dotted line are the microphones positions. The fifth black point is the true sound source localization.

sian pdfs $g_1(\cdot) = \mathcal{N}(\cdot\,; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $g_2(\cdot) = \mathcal{N}(\cdot\,; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is (see Delon & Desolneux 2020),

$$\mathrm{W}_2^2(g_1, g_2) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{trace}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2}\right)^{1/2}\right).$$

Section 4 of Delon & Desolneux (2020) shows that the MW_2 distance between two mixtures can be computed by solving the following discrete optimization problem. Let $f_1 = \sum_{k=1}^{K_1} \pi_{1k}\, g_{1k}$ and by $f_2 = \sum_{k=1}^{K_2} \pi_{2k}\, g_{2k}$ be two Gaussian mixtures. Then,

$$\mathrm{MW}_2^2(f_1, f_2) = \min_{\mathbf{w} \in \Pi(\pi_1, \pi_2)} \sum_{k,l} w_{kl}\, \mathrm{W}_2^2(g_{1k}, g_{2l})\,, \tag{2.4.15}$$

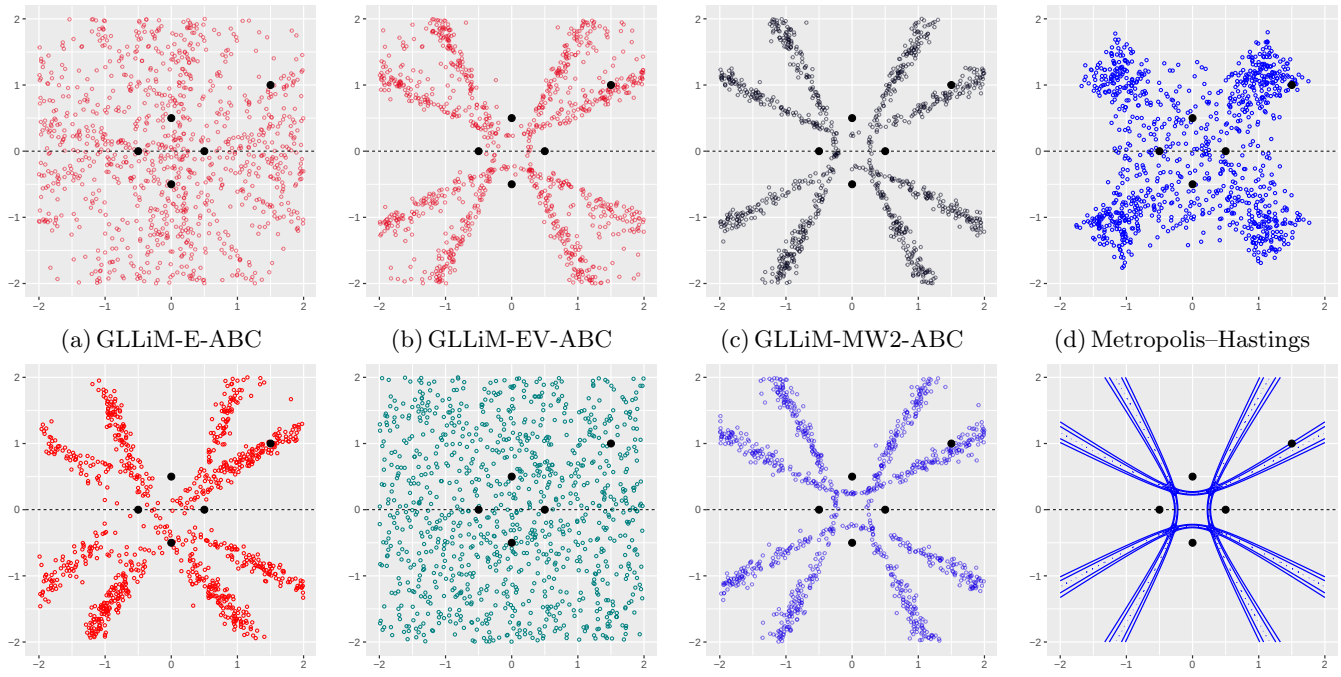
where π_1 and π_2 are the discrete distributions on the simplex defined by the respective weights of the mixtures and $\Pi(\pi_1, \pi_2)$ is the set of discrete joint distributions $\mathbf{w} = (w_{kl}, k \in [K_1], l \in [K_2])$, whose marginals are π_1 and π_2. Finding the minimizer $\mathbf{w}^*$ of (2.4.15) boils down to solving a simple discrete optimal transport problem, where the entries of the $K_1 \times K_2$ dimensional cost matrix are the $\mathrm{W}_2^2(g_{1k}, g_{2l})$ quantities.

As implicitly suggested above, MW_2 is indeed a distance on the space of Gaussian mixtures; see Delon & Desolneux (2020). In particular, for two Gaussian mixtures f_1 and f_2, MW_2 satisfies the equality property according to which $\mathrm{MW}_2(f_1, f_2) = 0$ implies that $f_1 = f_2$. In our experiments, the MW_2 distances were computed using the **transport** R package (Schuhmacher et al., 2020).

2.4.5.2 L2 distance between Gaussian mixtures

The L_2 distance between two Gaussian mixtures is also closed form. Denote by $f_1 = \sum_{k=1}^{K_1} \pi_{1k}\, g_{1k}$ and $f_2 = \sum_{k=1}^{K_2} \pi_{2k}\, g_{2k}$ two Gaussian mixtures,

$$\mathrm{L}_2^2(f_1, f_2) = \sum_{k,l} \pi_{1k}\pi_{1l}\langle g_{1k}, g_{1l}\rangle + \sum_{k,l} \pi_{2k}\pi_{2l}\langle g_{2k}, g_{2l}\rangle - 2\sum_{k,l} \pi_{1k}\pi_{2l}\langle g_{1k}, g_{2l}\rangle, \tag{2.4.16}$$



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC (d) Metropolis–Hastings

(e) GLLiM mixture (f) Semi-automatic ABC (g) GLLiM-L2-ABC (h) True posterior

Figure 2.4: Sound source localization with a mixture of two microphones pairs. GLLiM is learned on a first data set of size $N = 10^5$ while ABC is run using the largest data set of size $M = 10^6$. Selected samples using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) a Metropolis–Hastings algorithm, (e) the approximate GLLiM posterior for the observed data, (f) semi-automatic ABC, (g) L_2 distances, and (h) contours of the true posterior distribution. Black points on the dotted line are the microphones positions. The fifth black point is the true sound source localization.

where $\langle \cdot, \cdot \rangle$ denotes the L_2 scalar product, which is closed form for two Gaussian distributions g_1 and g_2 and given by $\langle g_1, g_2 \rangle = \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. The L_2 distance can be evaluated in $\mathcal{O}(K_1 K_2)$ time. We do not discuss the different properties of the various possible distances but the distance choice has a potential impact on the associated GLLiM-D-ABC procedure. This impact is illustrated in the experimental Section 2.4.4.

2.4.6 Appendix: Proofs

2.4.6.1 Proof of Theorem 2.4.1

We follow steps similar to the proof of Proposition 2 in Bernton et al. (2019). The ABC quasi-posterior can be written as

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Y}} K_\epsilon(\mathbf{z}; \mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z},$$

where $K_\epsilon(\mathbf{z}; \mathbf{y}) \propto \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}}\, \pi(\mathbf{z})$ denotes the density evaluated at some $\mathbf{z}$ of the prior truncated to A_ϵ. $K_\epsilon(\cdot; \mathbf{y})$ is a probability density function (pdf) in $\mathbf{z} \in \mathcal{Y}$ with compact support $A_\epsilon \subset \mathcal{Y}$ by definition of A_ϵ and (A4). It follows that

$$\begin{aligned} |q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| &\leq \int_{\mathcal{Y}} K_\epsilon(\mathbf{z}; \mathbf{y})\, |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|\, d\mathbf{z} \\ &\leq \sup_{\mathbf{z} \in A_\epsilon} |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \\ &= |\pi(\boldsymbol{\theta} \mid \mathbf{z}_\epsilon) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|, \end{aligned}$$

for some $\mathbf{z}_\epsilon \in A_\epsilon$, where the second inequality is due to the fact that $K_\epsilon(\cdot; \mathbf{y})$ is a pdf, and the last equality is due to (A1) and the compacity of A_ϵ.

Since for each $\epsilon > 0$, $\mathbf{z}_\epsilon \in A_\epsilon$, we have $\lim_{\epsilon \to 0} \mathbf{z}_\epsilon \in A_0$, where $A_0 = \bigcap_{\epsilon \in \mathbb{Q}_+} A_\epsilon$. Then, using that by continuity of D, $A_0 = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})) = 0\}$, it follows from the equality property of D, that $A_0 = \{\mathbf{z} \in \mathcal{Y} : \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$. Taking the limit $\epsilon \to 0$ yields

$$|\pi(\boldsymbol{\theta} \mid \mathbf{z}_\epsilon) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \to |\pi(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| = 0$$

and hence $|q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \to 0$, for each $\theta \in \Theta$.

By (A2), we have

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Theta} q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) &= \sup_{\boldsymbol{\theta} \in \Theta} \int_{\mathcal{Y}} K_\epsilon(\mathbf{z}; \mathbf{y}) \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, d\mathbf{z} \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty,\end{aligned}$$

for some γ, so that $\epsilon \leq \gamma$. Finally, by the bounded convergence theorem, we have

$$\lim_{\epsilon \to 0} \int_\Theta |q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \, d\boldsymbol{\theta} = \lim_{\epsilon \to 0} \|q_\epsilon(\cdot \mid \mathbf{y}) - \pi(\cdot \mid \mathbf{y})\|_1 = 0.$$

2.4.6.2 Proof of Theorem 2.4.2

We now provide a detailed proof of Theorem 2.4.2. Given any $\alpha > 0, \beta > 0$, we claim that

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1;$$

or equivalently, for any $\alpha > 0, \beta > 0, \gamma > 0$, we wish to find $\epsilon(\alpha, \beta, \gamma) > 0$, $K(\alpha, \beta, \gamma) \in \mathbb{N}^\star$, and $N(\alpha, \beta, \gamma) \in \mathbb{N}^\star$ so that for all $\epsilon < \epsilon(\alpha, \beta, \gamma), K \geq K(\alpha, \beta, \gamma), N \geq N(\alpha, \beta, \gamma)$:

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) > \alpha\right) \leq \gamma. \tag{2.4.17}$$

To prove (2.4.17), we first recall that we can rewrite $q_\epsilon^{K,N}$ as follows, for all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned}q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) &= \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, d\mathbf{z}, \\ K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) &= \frac{\mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z})}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z}) \, d\mathbf{z}},\end{aligned} \tag{2.4.18}$$

where $K_\epsilon^{K,N}(\cdot; \mathbf{y})$ is a pdf on $\mathbf{z} \in \mathcal{Y}$ with compact support $A_{\epsilon,\mathbf{y}}^{K,N} \subset \mathcal{Y}$ by definition of $A_{\epsilon,\mathbf{y}}^{K,N}$ and (B4).

The Hellinger distance D_{H}, between two densities f and g in appropriate spaces, is related to the L_1 distance D_1 as follows, see Zeevi & Meir (1997, Lemma 1),

$$\left(\frac{1}{2} D_1(f, g)\right)^2 \leq D_H^2(f, g) \leq D_1(f, g). \tag{2.4.19}$$

Applying successively the right-hand-side of (2.4.19), the definition of $q_\epsilon^{K,N}$ and the fact that $K_\epsilon^{K,N}(\cdot; \mathbf{y})$ is a pdf, we can write

$$\begin{aligned}D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) &\leq D_1\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \\ &= \int_\Theta \left|q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})\right| d\lambda(\boldsymbol{\theta}) \\ &\leq \int_\Theta \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) \left|\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})\right| d\lambda(\mathbf{z}) \, d\lambda(\boldsymbol{\theta}) \\ &= \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) \int_\Theta \left|\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})\right| d\lambda(\boldsymbol{\theta}) \, d\lambda(\mathbf{z}) \\ &\leq \sup_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} \int_\Theta \left|\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})\right| d\lambda(\boldsymbol{\theta}) \ .\end{aligned}$$

Then using Makarov & Podkorytov (2013, Corollary 7.1.3) and the continuity of $\pi(\cdot \mid \cdot)$ (B2), it follows that $\mathbf{z} \mapsto D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$ is a continuous function for every $\mathbf{y} \in \mathcal{Y}$. As $A_{\epsilon,\mathbf{y}}^{K,N}$ is compact, since

$$\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in B_{\epsilon,\mathbf{y}}^{K,N} = \arg\max_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right),$$

$$\sup_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) = D_1\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right),$$

and using the left-hand-side of (2.4.19), we finally get that

$$D_{\mathrm{H}}^2\left(q_{\epsilon}^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq 2D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \tag{2.4.20}$$

Consider the limit point $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ defined as $\mathbf{z}_{0,\mathbf{y}}^{K,N} = \lim_{\epsilon \to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$. Since for each $\epsilon > 0$, $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in A_{\epsilon,\mathbf{y}}^{K,N}$ then $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$, where $A_{0,\mathbf{y}}^{K,N} = \bigcap_{\epsilon \in \mathbb{Q}_+} A_{\epsilon,\mathbf{y}}^{K,N}$. By continuity of D, $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}\left(\cdot \mid \mathbf{z}\right), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right)\right.$ and $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}\left(\cdot \mid \mathbf{z}\right) = p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right\}$, using (B3).

The distance on the right-hand side of (2.4.20) can then be bounded by three terms using the triangle inequality for the Hellinger distance D_{H},

$$\begin{aligned} D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq\ & D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right) \\ & + D_{\mathrm{H}}\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \end{aligned} \tag{2.4.21}$$

The first term on the right-hand side can be made close to 0 as ϵ goes to 0 independently of K and N. The two other terms are of the same nature as the definition of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ yields $p^{K,N}\left(\cdot \mid \mathbf{y}\right) = p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})$.

Therefore, we first prove that $\lim_{\epsilon \to 0} D_H^2\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) = 0$ pointwise *i.e.* for each $\mathbf{y}$. Indeed, since $\pi(\cdot \mid \cdot)$ is a uniformly continuous function in $(\boldsymbol{\theta}, \mathbf{y})$, given any $\mathbf{y} \in \mathcal{Y}$, $\alpha_1 > 0$, there exists $\delta(\alpha_1) > 0$ such that for all $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N} \subset \mathcal{Y}$,

$$\sup_{\boldsymbol{\theta} \in \Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \leq \alpha_1, \forall \mathbf{z} \in \mathcal{Y}, \left|\mathbf{z} - \mathbf{z}_{0,\mathbf{y}}^{K,N}\right| < \delta(\alpha_1). \tag{2.4.22}$$

Furthermore, since Θ is a subset of a compact set, $\lambda\left(\Theta\right) < \infty$. Hence, by using the fact that $\lim_{\epsilon \to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} = \mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$ pointwise with respect to $\mathbf{y}$ and choosing $\mathbf{z} = \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$ in (2.4.22), we obtain that given any $\mathbf{y} \in \mathcal{Y}$, and $\alpha_1 > 0$, there exists $\delta(\alpha_1) > 0$, and $\epsilon\left(\delta(\alpha_1)\right) > 0$ such that $\forall 0 < \epsilon < \epsilon\left(\delta(\alpha_1)\right)$, $\left|\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} - \mathbf{z}_{0,\mathbf{y}}^{K,N}\right| < \delta(\alpha_1)$. Using (2.4.19) and (2.4.22), it follows for any ϵ such that $0 < \epsilon < \epsilon\left(\delta(\alpha_1)\right)$,

$$\begin{aligned} D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) &\leq D_1\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \lambda\left(\Theta\right) \\ &\leq \alpha_1 \lambda\left(\Theta\right). \end{aligned} \tag{2.4.23}$$

Such convergence also holds in measure λ. Given any $\alpha_1 > 0$, $\beta_1 > 0$, there exists $\epsilon\left(\alpha_1, \beta_1\right) > 0$ such that for any $0 < \epsilon < \epsilon\left(\alpha_1, \beta_1\right)$,

$$\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) \leq \alpha_1. \tag{2.4.24}$$

Then, since (2.4.24) is true whatever the value of $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$, sampled from the joint $\pi(\cdot, \cdot)$, it also holds, in probability with respect to the data set, that

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) > \alpha_1\right) = 0. \tag{2.4.25}$$

Next, we prove that $D_H^2\Big(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{y})\Big)$, equal to $D_H^2\Big(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\Big)$, and $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\right.$ both converge to 0 in measure λ, with respect to $\mathbf{y}$ and in probability with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$.

We first focus on $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$. Using the monotonicity of the Lebesgue integral and a result from Tsybakov (2008, Lemma 2.4) indicating that the squared Hellinger distance can be bounded by the Kullback–Leibler (KL) divergence, it follows that

$$\int_{\mathcal{Y}} D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) \leq \int_{\mathcal{Y}} \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}).$$

Then since $\pi(\mathbf{y}) \geq a\lambda(\Theta)$

$$\begin{aligned}\int_{\mathcal{Y}} \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) &\leq \frac{1}{a\lambda(\Theta)} \int_{\mathcal{Y}} \pi(\mathbf{y})\ \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) \\ &\leq \frac{1}{a\lambda(\Theta)}\ \mathrm{KL}\left(\pi, p^{K,N}\right)\ , \end{aligned} \tag{2.4.26}$$

where in the last right-hand side, the Kullback–Leibler divergence is on the joint densities π and $p^{K,N}$ and the inequality is coming from a standard relationship between Kullback–Leibler divergences between joint and conditional distributions, *i.e.*

$$\mathrm{KL}\left(\pi, p^{K,N}\right) = \int_{\mathcal{Y}} \pi(\mathbf{y})\ \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) + \int_{\mathcal{Y}} \pi(\mathbf{y})\ \log\left(\frac{\pi(\mathbf{y})}{p^{K,N}(\mathbf{y})}\right)\ d\lambda(\mathbf{y})\ ,$$

with the last integral being a positive Kullback–Leibler divergence. Using Corollary 2.2 in Rakhlin et al. (2005) (see details in Section 2.4.6.3), we can show that $\mathrm{KL}\left(\pi, p^{K,N}\right)$ tends to 0 in probability as K and N tends to infinity. It follows that $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converges to 0 in L_1 distance with respect to $\mathbf{y}$. Using Tao (2011, 1.5. Modes of convergence), $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ also converges to 0 in measure λ with respect to $\mathbf{y}$, and in probability with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ as $K \to \infty, N \to \infty$.
That is, given any $\alpha_2 > 0$, $\beta_2 > 0$, $\gamma_2 > 0$, there exists $K(\alpha_2, \beta_2, \gamma_2) \in \mathbb{N}^\star$, $N(\alpha_2, \beta_2, \gamma_2) \in \mathbb{N}^\star$ such that for any $K \geq K(\alpha_2, \beta_2, \gamma_2)$, $N \geq N(\alpha_2, \beta_2, \gamma_2)$,

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y}, D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{2.4.27}$$

To show that the same as (2.4.27) also holds when replacing $\mathbf{y}$ by $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ in D_H^2, we need to show some measurability property with respect to λ. Lemma 2.4.5, together with its proof in Subsection 2.4.6.3, guaranties first that the map $\mathbf{y} \mapsto \mathbf{z}_0^{K,N}(\mathbf{y}) = \mathbf{z}_{0,\mathbf{y}}^{K,N}$ is measurable. Since $\mathbf{y} \mapsto D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ is a continuous function (using (B4) and Makarov & Podkorytov 2013, Corollary 7.1.3), the measurability of the map implies that
$D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right)$ is also a measurable function (see Tao 2011, 1.3.2. Measurable functions). Consequently Tao (2011, Lemma 1.3.9 Equivalent notions of measurability) the set $\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_2\right\}$ is a measurable set with respect to λ. In addition by the monotonicity of λ and the defintion of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$, the measure of this set satisfies for any $\beta_2 > 0$,

$$\lambda\{\mathbf{y} \in \mathcal{Y} : D_H^2(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})) \geq \beta_2\} \leq \lambda\{\mathbf{y} \in \mathcal{Y} : D_H^2(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})) \geq \beta_2\}.$$

Then (2.4.27) implies that

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{2.4.28}$$

Finally, (2.4.17) can be deduced from (2.4.25), (2.4.27) and (2.4.28) by choosing $\alpha_1 = \alpha_2 = \alpha/3$, $\beta_1 = \beta_2 = \beta^2/36$, $\gamma_2 = \gamma/2$, $\epsilon(\alpha, \beta, \gamma) = \epsilon(\alpha_1, \beta_1)$, $K(\alpha, \beta, \gamma) = K(\alpha_2, \beta_2, \gamma_2)$ and $N(\alpha, \beta, \gamma) = N(\alpha_2, \beta_2, \gamma_2)$.

2.4.6.3 Auxiliary results

Use of Corollary 2.2 of Rakhlin et al. (2005)

In this section, we claim that under the conditions of Theorem 2.4.2, we can prove that $\mathrm{KL}\left(\pi, p^{K,N}\right) \to 0$, in probability as $K \to \infty, N \to \infty$.

To do so we use the following Lemma 2.4.4 coming from Rakhlin et al. (2005). Let us recall that $\mathcal{H}_{\mathcal{X}}$ is a parametric family of pdfs on $\mathcal{X}$, $\mathcal{H}_{\mathcal{X}} = \{g_{\boldsymbol{\varphi}}, \boldsymbol{\varphi} \in \Psi\}$. The set of continuous convex combinations associated with $\mathcal{H}_{\mathcal{X}}$ is defined as

$$\mathcal{C} = \mathrm{conv}\left(\mathcal{H}_{\mathcal{X}}\right) = \left\{ f : f(\mathbf{x}) = \int_{\Psi} g_{\boldsymbol{\varphi}}\left(\mathbf{x}\right) G\left(d\boldsymbol{\varphi}\right), g_{\boldsymbol{\varphi}} \in \mathcal{H}_{\mathcal{X}}, G \text{ is a probability measure on } \Psi \right\}.$$

We write $\mathrm{KL}\left(\pi, \mathcal{C}\right) = \inf_{g \in \mathcal{C}} \mathrm{KL}\left(\pi, g\right)$.
The class of K-component mixtures on $\mathcal{H}_{\mathcal{X}}$ is then defined as

$$\mathcal{C}_K = \mathrm{conv}_K(\mathcal{H}_{\mathcal{X}}) = \left\{ f : f(\mathbf{x}) = \sum_{k=1}^{K} c_k g_{\boldsymbol{\varphi}_k}\left(\mathbf{x}\right), c \in \mathbb{S}^{K-1}, g_{\boldsymbol{\varphi}_k} \in \mathcal{H}_{\mathcal{X}} \right\} \tag{2.4.29}$$

where $\mathbb{S}^{K-1} = \left\{ (c_1, \ldots, c_K) \in \mathbb{R}^K : \sum_{k=1}^{K} c_k = 1, c_k \geq 0, k \in [K] \right\}$.

The result from Rakhlin et al. (2005) is recalled in the following Lemma.

Lemma 2.4.4 (Corollary 2.2. from Rakhlin et al. (2005))**.** *Let $\mathcal{X} = \Theta \times \mathcal{Y}$ be a compact set. Let π be a target density π such that $0 < a \leq \pi(\mathbf{x}) \leq b$, for all $\mathbf{x} \in \mathcal{X}$. Assume that the distributions in $\mathcal{H}_{\mathcal{X}}$ satisfy, for any $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$,*

$$\begin{aligned} &\text{for all } \mathbf{x} \in \mathcal{X},\ 0 < a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \\ &\text{and } \sup_{\mathbf{x} \in \mathcal{X}} \left| \log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x}) \right| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1 , \end{aligned}$$

and that the parameter set Ψ is a cube with side length A with a, b, A, B arbitrary positive scalars. Let $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ be realizations from the joint distribution $\pi(\cdot, \cdot)$ and denote by $p^{K,N}$ the K-component mixture MLE in $\mathcal{C}_K$.
Then, with probability at least $1 - \exp(-t)$,

$$\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \mathrm{KL}\left(\pi, \mathcal{C}\right) + \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3\sqrt{t}}{\sqrt{N}},$$

where c_1, c_2 and c_3 are positive scalars depending only on a, b, A, B and on the dimension of $\mathcal{X}$ (see Rakhlin et al. (2005) for the exact expressions).

Assumption (B1) in Theorem 2.4.2 then implies that $\pi \in \mathcal{C}$ so that $\mathrm{KL}\left(\pi, \mathcal{C}\right) = 0$. Using Lemma 2.4.4, it follows that for all $t > 0$, for all $K \in \mathbb{N}^\star$, and for all $N \in \mathbb{N}^\star$,

$$\Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3\sqrt{t}}{\sqrt{N}} \right) \geq 1 - \exp\left(-t\right). \tag{2.4.30}$$

Choosing $t = N^{1/2}$, (2.4.30) becomes

$$1 - \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3}{N^{1/4}} \right) \leq \exp\left(-N^{1/2}\right). \tag{2.4.31}$$

Therefore, for any $\gamma_1 > 0, \gamma_2 > 0$, there exist $K(\gamma_1, \gamma_2) \in \mathbb{N}^\star$, and $N(\gamma_1, \gamma_2) \in \mathbb{N}^\star$ so that for all $K \geq K(\gamma_1, \gamma_2)$ and $N \geq N(\gamma_1, \gamma_2)$,

$$\begin{aligned} &\frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3}{N^{1/4}} \leq \gamma_1, \\ &\exp\left(-N^{1/2}\right) \leq \gamma_2. \end{aligned}$$

From which we deduce using (2.4.31) that for all $K \geq K(\gamma_1, \gamma_2)$ and all $N \geq N(\gamma_1, \gamma_2)$,

$$1 - \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \gamma_1\right) \leq \gamma_2,$$

that is

$$\lim_{K\to\infty, N\to\infty} \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \gamma_1\right) = 1,$$

which achieves the desired result that $\mathrm{KL}\left(\pi, p^{K,N}\right) \to 0$, in probability as $K \to \infty, N \to \infty$.

Proof of the measurability of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ (Lemma 2.4.5)

We wish to make use of the result from (Aliprantis & Border, 2006, Theorem 18.19 Measurable Maximum Theorem) to prove that we can choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N}$. More specifically this is guarantied by the following Lemma 2.4.5 which is proved below.

Background. The required materials for this lemma and the proof arise from Aliprantis & Border (2006), Chapter 18. The main concepts are recalled below.

Let f be a function on a product space $\mathcal{Y} \times \mathcal{Z}$, such that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$. Assume that $(\mathcal{Y}, \mathcal{F})$ is a measurable space.

The function $f(\mathbf{y}, \mathbf{z})$ is said to be Caratheodory, if f is continuous in $\mathbf{z} \in \mathcal{Z}$ and measurable in $\mathbf{y} \in \mathcal{Y}$.

By definition, a correspondence ζ from a set $\mathcal{Y}$ to a set $\mathcal{Z}$ assigns each $\mathbf{y} \in \mathcal{Y}$ to a subset $\zeta(\mathbf{y}) \in \mathcal{Z}$. We write this relationship as $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$.

A correspondence $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ is measurable (weakly measurable) if $\zeta^{\ell}(F) \in \mathcal{F}$ for each closed (open) subset F of $\mathcal{Z}$, where ζ^{ℓ} is the so-called lower inverse of ζ defined as $\zeta^{\ell}(F) = \{\mathbf{z} \in \mathcal{Z} : \zeta(\mathbf{y}) \cap F \neq \emptyset\}$.

Lemma 18.7 from Aliprantis & Border (2006) states the following: Suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ is Caratheodory, where $(\mathcal{Y}, \mathcal{F})$ is a measurable space, $\mathcal{Z}$ is a metrizable space, and $\mathcal{X}$ is a topological space. For each subset H of $\mathcal{X}$, define the correspondence $\zeta_H : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ by

$$\zeta_H(\mathbf{y}) = \{\mathbf{z} \in \mathcal{Z} : f(\mathbf{y}, \mathbf{z}) \in H\}.$$

If H is open, then ζ_H is a measurable correspondence.

Corollary 18.8 from Aliprantis & Border (2006) states the following: Suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ is Caratheodory, where $(\mathcal{Y}, \mathcal{F})$ is a measurable space, $\mathcal{Z}$ is a metrizable space, and $\mathcal{X}$ is a topological space. Define the correspondence $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ by

$$\zeta(\mathbf{y}) = \{\mathbf{z} \in \mathcal{Z} : f(\mathbf{y}, \mathbf{z}) = 0\}.$$

If $\mathcal{Z}$ is compact, then ζ is a measurable correspondence.

Furthermore, we have the fact that the countable unions of measurable correspondences are also measurable. We say that $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ admits a measurable selector, if there exists a measurable function $f : \mathcal{Y} \to \mathcal{Z}$, such that $f(\mathbf{y}) \in \zeta(\mathbf{y})$, for each $\mathbf{y} \in \mathcal{Y}$.

Theorem 18.19 (Measurable Maximum Theorem) from Aliprantis & Border (2006) then states the following. Let $\mathcal{Z}$ be a separable metrizable space and $(\mathcal{Y}, \mathcal{F})$ be a measurable space. Let $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ be a weakly measurable correspondence with nonempty compact values, and suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is Caratheodory. Define $m : \mathcal{Y} \to \mathbb{R}$ by

$$m(\mathbf{y}) = \max_{\mathbf{z} \in \zeta(\mathbf{y})} f(\mathbf{y}, \mathbf{z}),$$

and define $\mu : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ to be its maximizers:

$$\mu(\mathbf{y}) = \{\mathbf{z} \in \zeta(\mathbf{y}) : f(\mathbf{y}, \mathbf{z}) = m(\mathbf{y})\}.$$

Then 1) the value function m is measurable, 2) the argmax correspondence μ has nonempty and compact values, 3) the argmax correspondence μ is measurable and admits a measurable selector.

In our context, the use of Theorem 18.19 above takes the form of Lemma 2.4.5.

Lemma 2.4.5. *Under the assumptions in Theorem 2.4.2 and with the following definitions,*

$$A_{\epsilon,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), p^{K,N}\left(\cdot \mid \mathbf{z}\right)\right) \leq \epsilon\right\} \text{ and } \quad A_{0,\mathbf{y}}^{K,N} = \bigcap_{\epsilon \in \mathbb{Q}_+} A_{\epsilon,\mathbf{y}}^{K,N},$$

$$B_{\epsilon,\mathbf{y}}^{K,N} = \arg\max_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \text{ and } \quad B_{0,\mathbf{y}}^{K,N} = \bigcap_{\epsilon \in \mathbb{Q}_+} B_{\epsilon,\mathbf{y}}^{K,N},$$

so that $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}(\cdot \mid \mathbf{y}) - p^{K,N}(\cdot \mid \mathbf{z}) = 0\right\}$ *and* $B_{0,\mathbf{y}}^{K,N} = \arg\max_{\mathbf{z} \in A_{0,\mathbf{y}}^{K,N}} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$. *Then, we can always choose an argmax correspondence* $\mathbf{y} \twoheadrightarrow B_{0,\mathbf{y}}^{K,N}$, *which is measurable and admits a measurable selector.*

Proof of Lemma 2.4.5. Let us define the correspondence $\zeta_0^{K,N} : \mathcal{Y} \twoheadrightarrow \mathcal{Y}$ so that $\zeta_0^{K,N}\left(\mathbf{y}\right) = A_{0,\mathbf{y}}^{K,N}$. We claim that this correspondence is a weakly measurable correspondence with nonempty compact values. Indeed, we firstly define the function $f^{K,N}\left(\mathbf{y}, \mathbf{z}\right) = p^{K,N}\left(\cdot \mid \mathbf{y}\right) - p^{K,N}\left(\cdot \mid \mathbf{z}\right)$, and notice that

$$f^{K,N} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

is Caratheodory, since it is a continuous function in $\mathbf{z}$ and measurable in $\mathbf{y}$ by the continuity of $p^{K,N}$. Then, by using the (Aliprantis & Border, 2006, Corollary 18.8) and the fact that $\mathcal{Y}$ is compact, it follows that

$$\zeta_0^{K,N}\left(\mathbf{y}\right) = \left\{\mathbf{z} \in \mathcal{Y} : f^{K,N}\left(\mathbf{y}, \mathbf{z}\right) = 0\right\}$$

is measurable. Then, it is also weakly measurable (see Aliprantis & Border 2006, Lemma 18.2). Furthermore, $\zeta_0^{K,N}$ has nonempty compact values since for any $\mathbf{y} \in \mathcal{Y}$, $\zeta_0^{K,N}(\mathbf{y})$ always contains $\mathbf{y}$, and $\zeta_0^{K,N}(\mathbf{y}) = \left[f^{K,N}\left(\mathbf{y}, \cdot\right)\right]^{-1}\left(\{0\}\right)$ is a compact set since the inverse image of continuous function $f^{K,N}\left(\mathbf{y}, \cdot\right)$ of compact set is also compact.

Then, since we assume that $\left(\mathbf{y}, \mathbf{z}\right) \mapsto D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$ is a continuous function in $\mathbf{z}$ and measurable in $\mathbf{y}$, then it is also a Caratheodory function. We also remark that $B_{0,\mathbf{y}}^{K,N}$ can be written as a argmax correspondence

$$B_{0,\mathbf{y}}^{K,N} = \arg\max_{\mathbf{z} \in \zeta_0^{K,N}(\mathbf{y})} D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right).$$

By using the result from Aliprantis & Border, 2006, Theorem 18.19, Measurable Maximum Theorem, we conclude that the the argmax correspondence $B_{0,\mathbf{y}}^{K,N}$ is measurable and admits a measurable selector, that is, we can always choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N} \in B_{0,\mathbf{y}}^{K,N}$.

Chapter 3

Model selection in the Gaussian-gated localized mixture of experts regression model

Chapter 3 is based on the following works:

(C5) **TrungTin Nguyen**, Hien Duy Nguyen, Faicel Chamroukhi, and Florence Forbes. *A non-asymptotic penalization criterion for model selection in mixture of experts models.* arXiv preprint arXiv:2104.02640. 2021. Link: https://arxiv.org/pdf/2104.02640.pdf (Nguyen et al., 2021c).

(C6) **TrungTin Nguyen**, Faicel Chamroukhi, Hien Duy Nguyen, and Florence Forbes. *Non-asymptotic model selection in block-diagonal mixture of polynomial experts models.* arXiv preprint arXiv:2104.08959. 2021. Link: https://arxiv.org/pdf/2104.08959.pdf (Nguyen et al., 2021b).

Contents

Note that we have already highlighted the main oracle inequalities without detailed proofs regarding non-asymptotic model selection results for GLoME and BLoME models in high-dimensional scenarios based on an inverse regression strategy in Sections 1.2.6 and 1.2.7. In particular, our oracle inequalities show that the performance in Jensen–Kullback–Leibler type loss of our penalized maximum likelihood estimators are roughly comparable to that of oracle models if we take large enough the constants in front of the penalties, whose forms are only known up to multiplicative constants and proportional to the dimensions of models. Such theoretical justifications of the penalty shapes motivate us to make use of the slope heuristic criterion to select several hyperparameters, including the

number of mixture components, the degree of polynomial mean functions, and the potential hidden block-diagonal structures of the covariance matrices of the multivariate predictor or response variable. In Chapter 3, we aim to present such non-asymptotic oracle inequalities in as much detail as possible.

3.1 Introduction

In Chapter 3, we examine MoE models with Gaussian gating functions, first introduced by Xu et al. (1995), for clustering and regression. From hereon in, we refer to these models as the Gaussian-gated localized MoE (GLoME) and the block-diagonal localized mixture of polynomial experts (BLoME) models, to be developed in Sections 3.2 and 3.3, respectively. Furthermore, we refer to MoE models with softmax gating functions as softmax-gated MoE (SGaME). Note that the BLoME model generalizes the GLoME model by enjoying a parsimonious covariance structure, via block-diagonal structures for covariance matrices in the Gaussian experts.

It is worth mentioning that both GLoME and BLoME models have been thoroughly studied in the statistics and machine learning literatures under several contexts: localized MoE (Ramamurti & Ghosh, 1996, 1998, Moerland, 1999, Bouchard, 2003), normalized Gaussian networks (Sato & Ishii, 2000), MoE modeling of priors in Bayesian nonparametric regression (Norets & Pelenis, 2014, Norets & Pati, 2017), cluster-weighted modeling (Ingrassia et al., 2012), supervised Gaussian locally-linear mapping (GLLiM) in inverse regression (Deleforge et al., 2015c), block-diagonal covariance for Gaussian locally-linear mapping (BLLiM) model (Devijver et al., 2017), deep mixture of linear inverse regressions (Lathuilière et al., 2017) and multiple-output Gaussian gated mixture of linear experts (Nguyen et al., 2019). It is also interesting to point out that supervised GLLiM in Deleforge et al. (2015c) is an affine instance of a GLoME model, where linear combination of bounded functions are considered instead of affine for mean functions of Gaussian experts.

One of the main disadvantages of SGaME models is the difficulty of applying an EM algorithm, which requires an internal iterative numerical optimization procedure (*e.g.,* iteratively-reweighted least squares, Newton-Raphson algorithm) to update the softmax parameters. To overcome this problem, we instead use the Gaussian gating network that enables us to link GLoME with finite mixtures of Gaussian models. Then, the maximization with respect to the parameters of the gating network can be solved analytically with the EM algorithm framework, which decreases the computational complexity of the estimation routine. Furthermore, we then can also make use of well established theoretical results for finite mixture models.

In this work, we are interested in controlling and accounting for model complexity when selecting the best data-driven number of mixture components of a model. In general, model selection is often performed using the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Akaike, 1974, Schwarz et al., 1978). An important limitation of these criteria, however, is that they are only valid asymptotically. This implies that there are no finite sample guarantees when using AIC or BIC, for choosing between different levels of complexity. The slope heuristic of Birgé & Massart (2007), supported by a non-asymptotic oracle inequality, is a method that permits finite sample inference in place of AIC and BIC. Recent reviews and practical issues regarding the slope heuristic can be found in Baudry et al. (2012) and Arlot (2019). It should be stressed that a general model selection result, originally established by Massart (2007, Theorem 7.11), guarantees a penalized criterion leads to a good model selection and the penalty being only known up to multiplicative constants and proportional to the dimensions of models. In particular, such multiplicative constants can be calibrated by the slope heuristic approach in a finite sample setting.

Following the concentration inequality-based methods for likelihood penalization of Massart (2007), Massart & Meynet (2011) and Cohen & Le Pennec (2011), a number of finite-sample oracle results have been established for the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and general penalized maximum likelihood estimators (PMLE). These results include the works of Meynet (2013) and Devijver (2015a,b, 2017a) for finite mixture regression models, and Montuelle et al. (2014) and Nguyen et al. (2020c) for SGaME models. However, to the best of our knowledge, in Sections 3.2 and 3.3 (see also in Nguyen et al., 2021c,b), we are the first to provide finite-sample oracle inequalities for PMLE of GLoME and BLoME models, via Theorems 3.2.3 and 3.3.2, respectively.

Note that for the Gaussian gating parameters, the technique for handling the logistic weights in the SGaME models of Montuelle et al. (2014) is not directly applicable to the GLoME or BLoME framework, due to the quadratic form of the canonical link. Therefore, we propose a *reparameterization trick* to bound the metric entropy of the Gaussian gating parameters space; see Equation (3.2.25) and Section 3.2.5.2 for more details. Furthermore, in Nguyen et al. (2021c, Theorem 3.2.3), see also Section 3.2, we extend one of corollaries (in which the authors used linear combination of bounded functions for the functions in softmax gating networks) from Montuelle et al. (2014, Theorem 1) to the quadratic form of the canonical link from gating networks, see more details in Equation (3.2.25) and Lemma 3.2.10.

The main contribution of Chapter 3 is a theoretical result: a non-asymptotic oracle bound on the risk that provides the lower bound on the regularization parameters that ensures non-asymptotic control of the estimator Kullback–Leibler loss for the GLoME or BLoME model.

The goal of this chapter is to study the conditions on penalty functions that guarantee the weak oracle inequality for GLoME and BLoME models. As such, the rest of Chapter 3 is organized as follows. In Sections 3.2.1 and 3.3.1, we introduce the notations and frameworks for GLoME and BLoME models with corresponding special cases, GLLiM and BLLiM models, respectively. In Sections 3.2.2.2 and 3.3.2, we state the main results of Chapter 3: weak oracle inequalities satisfied by the PMLEs. Our results are then illustrated via numerical experiments in Section 3.2.3. Sections 3.2.4 and 3.3.3 are devoted to the proofs of the main results, based on a general model selection theorem. The proofs of technical lemmas can be found in Sections 3.2.5 and 3.3.4.

3.2 A non-asymptotic model selection in the Gaussian-gated localized mixture of experts regression model

3.2.1 Notation and framework

We consider a regression framework and aim at capturing the potential nonlinear relationship between the multivariate response $\mathbf{Y} = (\mathbf{Y}_j)_{j\in[L]}, [L] = \{1, \ldots, L\}$, and the set of covariates $\mathbf{X} = (\mathbf{X}_j)_{j\in[D]}$. Let $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n \subset \left(\mathbb{R}^D \times \mathbb{R}^L\right)^n$ be a random sample, and let $\mathbf{x}$ and $\mathbf{y}$ denote the observed values of the random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively.

3.2.1.1 GLoME models

We consider an extension of the MoE model of Xu et al. (1995), which extended the original MoE from Jacobs et al. (1991) to a regression setting. More specifically, we consider the following GLoME model, defined by (3.2.1), which is motivated by an inverse regression framework where the role of input and response variables should be exchanged such that $\mathbf{Y}$ becomes the covariates and $\mathbf{X}$ plays the role of a multivariate response. Then its corresponding conditional density is defined as follows:

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) \phi_D\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k\right), \tag{3.2.1}$$

$$g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) = \frac{\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K} \pi_j \phi_L\left(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j\right)}. \tag{3.2.2}$$

Here, $g_k(\cdot; \boldsymbol{\omega})$ and $\phi_D\left(\cdot; \boldsymbol{\upsilon}_{k,d}(\cdot), \boldsymbol{\Sigma}_k\right)$, $k \in [K]$, $K \in \mathbb{N}^\star$, $d \in \mathbb{N}^\star$, are called Gaussian gating functions and Gaussian experts, respectively. Furthermore, we decompose the parameters of the model as follows: $\boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times \mathbf{C}_K \times \boldsymbol{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$, $\boldsymbol{c} = (\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d = (\boldsymbol{\upsilon}_{k,d})_{k\in[K]} \in \boldsymbol{\Upsilon}_{K,d}$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in[K]} \in \mathbf{V}_K$. Note that $\boldsymbol{\Pi}_{K-1} = \left\{(\pi_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \pi_k = 1\right\}$ is a $K-1$ dimensional probability simplex, $\mathbf{C}_K$ is a set of K-tuples of mean vectors of size $L \times 1$, $\boldsymbol{V}'_K$ is a sets of K-tuples of elements in $\mathcal{S}_L^{++}$, where $\mathcal{S}_L^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ is a set of K-tuples of mean functions from $\mathbb{R}^L$ to $\mathbb{R}^D$ depending on a degree d (*e.g.,* a degree of polynomials), and $\mathbf{V}_K$ is a set containing K-tuples from $\mathcal{S}_D^{++}$.

In order to establish our oracle inequality, Theorem 3.2.3, we need to assume that $\mathcal{Y}$ is a bounded set in $\mathbb{R}^L$ and make explicit some classical boundedness conditions on the parameter space.

3.2.1.2 Gaussian gating functions

For a matrix $\mathbf{A}$, let $m(\mathbf{A})$ and $M(\mathbf{A})$ be, respectively, the modulus of the smallest and largest eigenvalues of $\mathbf{A}$. We shall restrict our study to bounded Gaussian gating parameter vectors $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in \boldsymbol{\Omega}_K$. Specifically, we assume that there exist deterministic positive constants $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}}, A_{\boldsymbol{\Gamma}}$, such that $\boldsymbol{\omega}$ belongs to $\widetilde{\boldsymbol{\Omega}}_K$, where

$$\widetilde{\boldsymbol{\Omega}}_K = \left\{\boldsymbol{\omega} \in \boldsymbol{\Omega}_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m\left(\boldsymbol{\Gamma}_k\right) \leq M\left(\boldsymbol{\Gamma}_k\right) \leq A_{\boldsymbol{\Gamma}}, a_{\boldsymbol{\pi}} \leq \pi_k\right\}. \tag{3.2.3}$$

We denote the space of gating functions as

$$\mathcal{P}_K = \left\{\mathbf{g} = \left(g_k\left(\cdot;\boldsymbol{\omega}\right)\right)_{k\in[K]} : \forall k \in [K], g_k\left(\mathbf{y};\boldsymbol{\omega}\right) = \frac{\pi_k \phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^K \pi_j \phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}, \boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}_K\right\}.$$

3.2.1.3 Gaussian experts

Following the same structure for the means of Gaussian experts from Montuelle et al. (2014), the set $\boldsymbol{\Upsilon}_{K,d}$ will be chosen as a tensor product of compact sets of moderate dimension (*e.g.,* a set of polynomials of degree smaller than d, whose coefficients are smaller in absolute values than $T_{\boldsymbol{\Upsilon}}$). Then, $\boldsymbol{\Upsilon}_{K,d}$ is defined as a linear combination of a finite set of bounded functions whose coefficients belong to a compact set. This general setting includes polynomial bases when the covariates are bounded, Fourier bases on an interval, as well as suitably renormalized wavelet dictionaries. More specifically, $\boldsymbol{\Upsilon}_{K,d} = \otimes_{k\in[K]} \boldsymbol{\Upsilon}_{k,d} =: \boldsymbol{\Upsilon}_{k,d}^K$, where $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_{b,d}$, $\forall k \in [K]$, and

$$\boldsymbol{\Upsilon}_{b,d} = \left\{\mathbf{y} \mapsto \left(\sum_{i=1}^d \boldsymbol{\alpha}_i^{(j)} \varphi_{\boldsymbol{\Upsilon},i}(\mathbf{y})\right)_{j\in[D]} =: \left(\boldsymbol{\upsilon}_{d,j}(\mathbf{y})\right)_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \tag{3.2.4}$$

Here, $d \in \mathbb{N}^\star, T_{\boldsymbol{\Upsilon}} \in \mathbb{R}^+$, and $(\varphi_{\boldsymbol{\Upsilon},i})_{i\in[d]}$ is a collection of bounded functions on $\mathcal{Y}$. In particular, we focus on the bounded $\mathcal{Y}$ case and assume that $\mathcal{Y} = [0,1]^L$, without loss of generality. In this case, $\varphi_{\boldsymbol{\Upsilon},i}$ can be chosen as monomials with maximum (non-negative) degree d: $\mathbf{y}^{\mathbf{r}} = \prod_{l=1}^L \mathbf{y}_l^{\mathbf{r}_l}$. Recall that a multi-index $\mathbf{r} = (\mathbf{r}_l)_{l\in[L]}, \mathbf{r}_l \in \mathbb{N}^\star \cup \{0\}, \forall l \in [L]$, is an L-tuple of nonnegative integers. We define $|\mathbf{r}| = \sum_{l=1}^L \mathbf{r}_l$ and the number $|\mathbf{r}|$ is called the order or degree of $\mathbf{y}^{\mathbf{r}}$. Then, $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$, where

$$\boldsymbol{\Upsilon}_{p,d} = \left\{\mathbf{y} \mapsto \left(\sum_{|\mathbf{r}|=0}^d \boldsymbol{\alpha}_{\mathbf{r}}^{(j)} \mathbf{y}^{\mathbf{r}}\right)_{j\in[D]} =: \left(\boldsymbol{\upsilon}_{d,j}(\mathbf{y})\right)_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \tag{3.2.5}$$

For the covariances of Gaussian experts, we follow the classical covariance matrix sets described by Celeux & Govaert (1995). In general situation, sets $\mathbf{V}_K$ depend on the noise model chosen. Formally, we consider the set

$$\mathbf{V}_K = \left\{\left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top\right)_{k\in[K]} : \forall k \in [K], B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}\left(\lambda_-, \lambda_+\right)\right\}, \tag{3.2.6}$$

where any covariance matrix $\boldsymbol{\Sigma}_k$ can be decomposed into the form $B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top$, such that $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$ is a positive scalar corresponding to the volume, $\mathbf{P}_k$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_k$ and $\mathbf{A}_k$ the diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$; $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$, $\mathcal{A}\left(\lambda_-, \lambda_+\right)$ is a set of diagonal matrices $\mathbf{A}_k$, such that $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq \left(\mathbf{A}_k\right)_{i,i} \leq \lambda_+$; and $SO(D)$ is the special orthogonal group of dimension D. For example, in the most general case, we can assume that the matrices $\mathbf{V}_K$ are different for all Gaussian experts. Alternatively, they can share the same volume or diagonalization matrix.

Next, a characterization of GLLiM model, an affine instance of GLoME model, is described in Section 3.2.1.4 and is especially useful for high-dimensional regression data.

3.2.1.4 High-dimensional regression via GLLiM models

A GLLiM model, as originally introduced in Deleforge et al. (2015c), is used to capture the nonlinear relationship between the response and the set of covariates from a high-dimensional regression data, typically in the case when $D \gg L$, by the K locally affine mappings:

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbb{I}\left(Z = k\right)\left(\mathbf{A}_k^*\mathbf{X} + \mathbf{b}_k^* + \mathbf{E}_k^*\right). \tag{3.2.7}$$

Here, $\mathbb{I}$ is an indicator function and Z is a latent variable capturing a cluster relationship, such that $Z = k$ if $\mathbf{Y}$ originates from cluster $k \in [K]$. Cluster specific affine transformations are defined by matrices $\mathbf{A}_k^* \in \mathbb{R}^{L\times D}$ and vectors $\mathbf{b}_k^* \in \mathbb{R}^L$. Furthermore, $\mathbf{E}_k^*$ are an error terms capturing both the reconstruction error due to the local affine approximations and the observation noise in $\mathbb{R}^L$.

Following the common assumption that $\mathbf{E}_k^*$ is a zero-mean Gaussian variable with covariance matrix $\mathbf{\Sigma}_k^* \in \mathbb{R}^{L\times L}$, it holds that

$$p\left(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k; \boldsymbol{\psi}_K^*\right) = \phi_L\left(\mathbf{y}; \mathbf{A}_k^*\mathbf{x} + \mathbf{b}_k^*, \mathbf{\Sigma}_k^*\right), \tag{3.2.8}$$

where we denote by $\boldsymbol{\psi}_K^*$ the vector of model parameters and ϕ_L is the probability density function (PDF) of a Gaussian distribution of dimension L. In order to enforce the affine transformations to be local, $\mathbf{X}$ is defined as a mixture of K Gaussian components as follows:

$$p\left(\mathbf{X} = \mathbf{x}|Z = k; \boldsymbol{\psi}_K^*\right) = \phi_D\left(\mathbf{x}; \mathbf{c}_k^*, \mathbf{\Gamma}_k^*\right), p\left(Z = k; \boldsymbol{\psi}_k^*\right) = \boldsymbol{\pi}_k^*, \tag{3.2.9}$$

where $\mathbf{c}_k^* \in \mathbb{R}^D, \mathbf{\Gamma}_k^* \in \mathbb{R}^{D\times D}$, $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_k^*)_{k\in[K]} \in \mathbf{\Pi}_{K-1}^*$, and $\mathbf{\Pi}_{K-1}^*$ is the $K-1$ dimensional probability simplex. Then, according to formulas for conditional multivariate Gaussian variables and the following hierarchical decomposition

$$\begin{aligned} p\left(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}; \boldsymbol{\psi}_K^*\right) &= \sum_{k=1}^{K} p\left(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k; \boldsymbol{\psi}_K^*\right) p\left(\mathbf{X} = x|Z = k; \boldsymbol{\psi}_K^*\right) p\left(Z = k; \boldsymbol{\psi}_K^*\right), \\ &= \sum_{k=1}^{K} \boldsymbol{\pi}_k^* \phi_D\left(\mathbf{x}; \mathbf{c}_k^*, \mathbf{\Gamma}_k^*\right) \phi_L\left(\mathbf{y}; \mathbf{A}_k^*\mathbf{x} + \mathbf{b}_k^*, \mathbf{\Sigma}_k^*\right), \end{aligned}$$

we obtain the following *forward conditional density* (Deleforge et al., 2015c):

$$p\left(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}; \boldsymbol{\psi}_K^*\right) = \sum_{k=1}^{K} \frac{\boldsymbol{\pi}_k^* \phi_D\left(\mathbf{x}; \mathbf{c}_k^*, \mathbf{\Gamma}_k^*\right)}{\sum_{j=1}^{K} \boldsymbol{\pi}_j^* \phi_D\left(\mathbf{x}; \mathbf{c}_j^*, \mathbf{\Gamma}_j^*\right)} \phi_L\left(\mathbf{y}; \mathbf{A}_k^*\mathbf{x} + \mathbf{b}_k^*, \mathbf{\Sigma}_k^*\right), \tag{3.2.10}$$

where $\boldsymbol{\psi}_K^* = (\boldsymbol{\pi}^*, \boldsymbol{\theta}_K^*) \in \Pi_{K-1} \times \mathbf{\Theta}_K^* =: \mathbf{\Psi}_K^*$. Here, $\boldsymbol{\theta}_K^* = (\mathbf{c}_k^*, \mathbf{\Gamma}_k^*, \mathbf{A}_k^*, \mathbf{b}_k^*, \mathbf{\Sigma}_k^*)_{k\in[K]}$ and

$$\mathbf{\Theta}_K^* = \left(\mathbb{R}^D \times \mathcal{S}_D^{++}(\mathbb{R}) \times \mathbb{R}^{L\times D} \times \mathbb{R}^L \times \mathcal{S}_L^{++}(\mathbb{R})\right)^K.$$

Without assuming anything on the structure of parameters, the dimension of the model (denoted by $\dim(\cdot)$), is defined as the total number of parameters that has to be estimated, as follows:

$$\dim\left(\mathbf{\Psi}_K^*\right) = K\left(1 + D(L+1) + \frac{D(D+1)}{2} + \frac{L(L+1)}{2} + L\right) - 1.$$

It is worth mentioning that $\dim\left(\mathbf{\Psi}_K\right)$ is very large compared to the sample size (see, e.g., Deleforge et al., 2015c, Devijver et al., 2017, Perthame et al., 2018 for more details in their real data sets) whenever $D \gg n$ and $D \gg L$. Furthermore, it is more realistic to make assumption on the residual covariance matrices $\mathbf{\Sigma}_k^*$ of error vectors $\mathbf{E}_k^*$ rather than on $\mathbf{\Gamma}_k^*$ (cf. Deleforge et al., 2015c, Section 3). This justifies the use of the inverse regression trick from Deleforge et al. (2015c), which leads a drastic reduction in the number of parameters to be estimated.

More specifically, in (3.2.10), the roles of input and response variables should be exchanged such that $\mathbf{Y}$ becomes the covariates and $\mathbf{X}$ plays the role of the multivariate response. Therefore, its corresponding *inverse conditional density* is defined as a Gaussian locally-linear mapping (GLLiM) model, based on the previous hierarchical Gaussian mixture model, as follows:

$$p\left(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y},Z=k;\boldsymbol{\psi}_K\right)=\phi_D\left(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k\right), \tag{3.2.11}$$

$$p\left(\mathbf{Y}=\mathbf{y}|Z=k;\boldsymbol{\psi}_K\right)=\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right), p\left(Z=k;\boldsymbol{\psi}_k\right)=\pi_k, \tag{3.2.12}$$

$$p\left(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}_K\right)=\sum_{k=1}^{K}\frac{\pi_k\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K}\pi_j\phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}\phi_D\left(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k\right), \tag{3.2.13}$$

where $\boldsymbol{\Sigma}_k$ is a $D\times D$ covariance structure (usually diagonal, chosen to reduce the number of parameters) automatically learnt from data and $\boldsymbol{\psi}_K$ is the set of parameters, denoted by $\boldsymbol{\psi}_K=(\boldsymbol{\pi},\boldsymbol{\theta}_K)\in\boldsymbol{\Pi}_{K-1}\times\boldsymbol{\Theta}_K=:\boldsymbol{\Psi}_K$. An intriguing feature of the GLLiM model is described in Lemma 3.2.1, which is proved in Section 3.2.5.1.

Lemma 3.2.1. *The parameter $\boldsymbol{\psi}_K^*$ in the forward conditional PDF, defined in (3.2.10), can then be deduced from $\boldsymbol{\psi}_K$ in (3.2.13) via the following one-to-one correspondence:*

$$\boldsymbol{\theta}_K=\begin{pmatrix}\mathbf{c}_k\\ \boldsymbol{\Gamma}_k\\ \mathbf{A}_k\\ \mathbf{b}_k\\ \boldsymbol{\Sigma}_k\end{pmatrix}_{k\in[K]}\mapsto\begin{pmatrix}\mathbf{c}_k^*\\ \boldsymbol{\Gamma}_k^*\\ \mathbf{A}_k^*\\ \mathbf{b}_k^*\\ \boldsymbol{\Sigma}_k^*\end{pmatrix}_{k\in[K]}=\begin{pmatrix}\mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k\\ \boldsymbol{\Sigma}_k+\mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top\\ \boldsymbol{\Sigma}_k^*\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\\ \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1}\mathbf{c}_k-\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k)\\ \left(\boldsymbol{\Gamma}_k^{-1}+\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\end{pmatrix}_{k\in[K]}\in\boldsymbol{\Theta}_K^*, \tag{3.2.14}$$

with the note that $\boldsymbol{\pi}^*\equiv\boldsymbol{\pi}$.

3.2.1.5 Collection of GLoME models

In this paper, we choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}}=[d_{\max}]$ and $\mathcal{K}=[K_{\max}]$, respectively, where $d_{\max}\in\mathbb{N}^\star$ and $K_{\max}\in\mathbb{N}^\star$ may depend on the sample size n. We wish to estimate the unknown inverse conditional density s_0 by conditional densities belonging to the following collection of inverse models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M}=\{(K,d):K\in\mathcal{K},d\in\mathcal{D}_{\boldsymbol{\Upsilon}}\}$,

$$S_{\mathbf{m}}=\left\{(\mathbf{x},\mathbf{y})\mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y})=:s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}):\boldsymbol{\psi}_{K,d}=(\boldsymbol{\omega},\boldsymbol{\upsilon}_d,\boldsymbol{\Sigma})\in\widetilde{\boldsymbol{\Omega}}_K\times\boldsymbol{\Upsilon}_{K,d}\times\mathbf{V}_K=:\widetilde{\boldsymbol{\Psi}}_{K,d}\right\}, \tag{3.2.15}$$

where $\widetilde{\boldsymbol{\Omega}}_K$, $\boldsymbol{\Upsilon}_{K,d}$ and $\mathbf{V}_K$ are define previously in (3.2.3), (3.2.5) (or more general (3.2.4)) and (3.2.6), respectively.

Remark 3.2.2. It is worth mentioning that we can also define the collection of the forward models in the same framework as in (3.2.15). More precisely, the unknown forward conditional density s_0^* is estimated via the following collection of forward model $\mathcal{S}^*=(S_{\mathbf{m}}^*)_{\mathbf{m}\in\mathcal{M}}$, with $\mathcal{M}=\mathcal{K}\times\mathcal{D}_{\boldsymbol{\Upsilon}}$, and

$$S_{\mathbf{m}}^*=\left\{(\mathbf{x},\mathbf{y})\mapsto s_{\boldsymbol{\psi}_K^*}(\mathbf{y}|\mathbf{x})=:s_{\mathbf{m}}^*(\mathbf{y}|\mathbf{x}):\boldsymbol{\psi}_K^*=(\boldsymbol{\omega}^*,\boldsymbol{\upsilon}_d^*,\boldsymbol{\Sigma}^*)\in\widetilde{\boldsymbol{\Omega}}_K^*\times\boldsymbol{\Upsilon}_{K,d}^*\times\mathbf{V}_K^*=:\widetilde{\boldsymbol{\Psi}}_{K,d}^*\right\}, \tag{3.2.16}$$

where $\widetilde{\boldsymbol{\Omega}}_K^*$, $\boldsymbol{\Upsilon}_{K,d}^*$ and $\mathbf{V}_K^*$ are define similar to (3.2.3), (3.2.5) (or more general (3.2.4)) and (3.2.6), respectively.

Note that for sake of simplicity of notation via avoiding the utilization of "$*$" on the parameters, we focus on the collection of inverse models, $\mathcal{S}$, which is defined in (3.2.15), as we are motivated by the inverse conditional densities (3.2.13) of the GLLiM models. However, our finite-sample oracle inequality, Theorem 3.2.3, holds for any collection of GLoME models satisfying the required regularity conditions. In particular, Theorem 3.2.3 can be applied to the forward model $\mathcal{S}^*=(S_{\mathbf{m}}^*)_{\mathbf{m}\in\mathcal{M}}$, established in (3.2.16), if we consider $\mathbf{y}$ and $\mathbf{x}$ as realizations of predictors and response variables, respectively.

3.2.2 Weak oracle inequality

3.2.2.1 Penalized maximum likelihood estimator and losses

In the context of PMLE, given the collection of conditional densities $S_{\mathbf{m}}$, we aim to estimate s_0 by the η-minimizer $\widehat{s}_{\mathbf{m}}$ of the negative log-likelihood (NLL):

$$\sum_{i=1}^{n} -\ln\left(s_{\widehat{s}_{\mathbf{m}}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \sum_{i=1}^{n} -\ln\left(s_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \eta, \tag{3.2.17}$$

where the error term η is necessary when the infimum may not be unique or even not be reached.

As always, using the NLL of the estimate in each model as a criterion is not sufficient. It is an underestimation of the risk of the estimate and this leads to choosing models that are too complex. By adding a suitable penalty pen($\mathbf{m}$), one hopes to compensate between the *variance* term, $\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\mathbf{m}}) - \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right)$, and the *bias*, $\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right)$, where $\mathrm{KL}^{\otimes \mathrm{n}}$ is defined later. For a given choice of pen($\mathbf{m}$), the *selected model* $S_{\widehat{\mathbf{m}}}$ is chosen as the one whose index is an η'-almost minimizer of the sum of the NLL and this penalty:

$$\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \mathrm{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m}\in\mathcal{M}} \left(\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \mathrm{pen}(\mathbf{m}) \right) + \eta'. \tag{3.2.18}$$

Note that $\widehat{s}_{\widehat{\mathbf{m}}}$ is then called the η'-penalized likelihood estimator and depends on both the error terms η and η'. From hereon in, the term *selected model or best data-driven model* or estimate is used to indicate that it satisfies the definition in (3.2.18).

In the maximum likelihood approach, the Kullback–Leibler divergence is the most natural loss function, which is defined for two densities s and t by

$$\mathrm{KL}(s,t) = \begin{cases} \int_{\mathbb{R}^D} \ln\left(\frac{s(\mathbf{y})}{t(\mathbf{y})}\right) s(\mathbf{y}) d\mathbf{y} & \text{if } sd\mathbf{y} \text{ is absolutely continuous w.r.t. } td\mathbf{y}, \\ +\infty & \text{otherwise.} \end{cases}$$

However, to take into account the structure of conditional densities and the random covariates $\mathbf{Y}_{[n]}$, we consider the *tensorized Kullback–Leibler divergence* $\mathrm{KL}^{\otimes \mathrm{n}}$, defined as:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right], \tag{3.2.19}$$

if sdy is absolutely continuous w.r.t. tdy, and $+\infty$ otherwise. Note that if the predictors are fixed, this divergence is the classical fixed design type divergence in which there is no expectation. We refer to our result as a *weak oracle inequality*, because its statement is based on a smaller divergence, when compared to $\mathrm{KL}^{\otimes \mathrm{n}}$, namely the *tensorized Jensen-Kullback–Leibler divergence*:

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), (1-\rho)\, s\left(\cdot|\mathbf{Y}_i\right) + \rho t\left(\cdot|\mathbf{Y}_i\right)\right)\right],$$

with $\rho \in (0,1)$. We note that $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ was first used in Cohen & Le Pennec (2011). However, a version of this divergence appears explicitly with $\rho = \frac{1}{2}$ in Massart (2007), and it is also found implicitly in Birgé et al. (1998). This loss is always bounded by $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ but behaves like $\mathrm{KL}^{\otimes \mathrm{n}}$, when t is close to s. The main tools in the proof of such a weak oracle inequality are deviation inequalities for sums of random variables and their suprema. These tools require a boundedness assumption on the controlled functions which is not satisfied by $-\ln\frac{s_{\mathbf{m}}}{s_0}$, and thus also not satisfied by $\mathrm{KL}^{\otimes \mathrm{n}}$. Therefore, we consider instead the use of $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$. In particular, in general, it holds that $C_\rho\, \mathrm{d}^{2\otimes \mathrm{n}} \leq \mathrm{JKL}_{\rho}^{\otimes \mathrm{n}} \leq \mathrm{KL}^{\otimes \mathrm{n}}$, where $C_\rho = \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho}, 1\right)\left(\ln\left(1+\frac{\rho}{1-\rho}\right) - \rho\right)$ (see Cohen & Le Pennec 2011, Prop. 1) and $\mathrm{d}^{2\otimes \mathrm{n}}$ is a tensorized extension of the squared Hellinger distance $\mathrm{d}^{2\otimes \mathrm{n}}$, defined by

$$\mathrm{d}^{2\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{d}^2\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right].$$

Moreover, if we assume that, for any $\mathbf{m} \in \mathcal{M}$ and any $s_{\mathbf{m}} \in S_{\mathbf{m}}, s_0 d\lambda \ll s_{\mathbf{m}} d\lambda$, then (see Montuelle et al., 2014, Cohen & Le Pennec, 2011)

$$\frac{C_\rho}{2 + \ln \left\| s_0 / s_{\mathbf{m}} \right\|_\infty} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}) \leq \mathrm{JKL}_\rho^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}). \tag{3.2.20}$$

3.2.2.2 Main result

The following result provides a lower bound on the penalty function, $\mathrm{pen}(\mathbf{m})$, which guarantees that the PMLE selects a model that performs almost as well as the best model.

Theorem 3.2.3 (Weak oracle inequality)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, arising from an unknown conditional density* s_0*. Given a collection of GLoME models,* $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m} \in \mathcal{M}}$*, there is a constant* C *such that for any* $\rho \in (0,1)$*, for any* $\mathbf{m} \in \mathcal{M}$*,* $z_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m} \in \mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$ *and any* $C_1 > 1$*, there is a constant* κ_0 *depending only on* ρ *and* C_1*, such that if for every index* $\mathbf{m} \in \mathcal{M}$*,* $\mathrm{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n) \dim\left(S_{\mathbf{m}}\right) + z_{\mathbf{m}} \right]$ *with* $\kappa > \kappa_0$*, then the* η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined in* (3.2.17) *and* (3.2.18)*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}} \left[\mathrm{JKL}_\rho^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \right] \leq C_1 \inf_{\mathbf{m} \in \mathcal{M}} \left(\inf_{s_{\mathbf{m}} \in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{\mathrm{pen}(\mathbf{m})}{n} \right) + \frac{\kappa_0 C_1 \Xi}{n} + \frac{\eta + \eta'}{n}. \tag{3.2.21}$$

Remark 3.2.4. As per the SGaME models from Montuelle et al. (2014), we also have to deal with three potential issues: the differences of divergence on the left ($\mathrm{JKL}_\rho^{\otimes \mathrm{n}}$) and the right ($\mathrm{KL}^{\otimes \mathrm{n}}$) hand side, $C_1 > 1$, and the relationship between $\frac{\mathrm{pen}(\mathbf{m})}{n}$ and the variance.

The first issue is important as in general we have $\mathrm{JKL}_\rho^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}) \leq \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}})$. However, (3.2.20) ensures that the two divergences are *equivalent* under regularity conditions. Namely, when

$$\sup_{\mathbf{m} \in \mathcal{M}} \sup_{s_{\mathbf{m}} \in S_{\mathbf{m}}} \left\| s_0 / s_{\mathbf{m}} \right\|_\infty < \infty.$$

Such a strong assumption is satisfied as long as $\mathcal{X}$ is compact, s_0 is compactly supported, and the regression functions are uniformly bounded, and under the condition that there is a uniform lower bound on the eigenvalues of the covariance matrices.

For a fixed collection $\mathcal{M}$ and $s_0 \notin \mathcal{S}$, the error bound converge to $C_1 \inf_{s_{\mathbf{m}} \in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right)$ as $n \to \infty$, which may be large. To reinforce the power of Theorem 3.2.3, it would be very interesting to prove the consistency result for $\widehat{s}_{\widehat{m}}$ where $\mathcal{M} = \mathcal{M}_n$ grows with the number of observations, using results from approximation theory in the sense of $\mathrm{KL}^{\otimes \mathrm{n}}$. Moreover, we believe that it would be nontrivial to establish a similar adaptive conditional density as in Maugis-Rabusseau & Michel (2013, Theorem 2.9). Therefore, we leave such interesting problems for future research. Nevertheless, as we consider GLoME models, some recent results from Nguyen et al. (2019, 2021a) imply that if we take a sufficiently large number of mixture components, we can approximate a broad class of densities, and thus the term on the right hand side is small for K sufficiently large. This improves the error bound even when s_0 does not belong to $S_{\mathbf{m}}$ for any $\mathbf{m} \in \mathcal{M}$. We aim to provide an oracle inequality with $C_1 = 1$ in future work, which is similar with Rigollet (2012), Dalalyan & Sebbar (2018) .

For the last issue, we claim that $\frac{\mathrm{pen}(\mathbf{m})}{n}$ is approximately proportional to the asymptotic variance in the parametric case: $\frac{\dim(S_{\mathbf{m}})}{n}$. We shall consider the condition (3.2.22) for GLoME models. As shown in the proof of Theorem 3.2.3, in fact we can replace the assumption on $\mathrm{pen}(\mathbf{m})$ by a milder one. More precisely, given a constant $\mathfrak{C}$, which we specify later, there is a constant κ_0 depending only on ρ and C_1, such that with $\kappa > \kappa_0$, for every index $\mathbf{m} \in \mathcal{M}$, we may set

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left(\dim(S_{\mathbf{m}}) \left(2 \left(\sqrt{\mathfrak{C}} + \sqrt{\pi} \right)^2 + \left(\ln \frac{n}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi} \right)^2 \dim\left(S_{\mathbf{m}}\right)} \right)_+ \right) + z_{\mathbf{m}} \right). \tag{3.2.22}$$

Furthermore, based on the Appendix B.4 from Cohen & Le Pennec (2011), we can make explicit the dependence of the constant κ_0, with respect to ρ and C_1 as follows. For any $\rho \in (0,1)$ and $C_1 > 1$, define $\epsilon_{\text{pen}} = 1 - \frac{1}{C_1}$. Then κ_0 is determined by

$$\kappa_0 = \frac{\kappa_0' \left(\kappa_1' + \kappa_2'\right)^2 \left(\sqrt{1 + \frac{72 C_\rho \epsilon_{\text{pen}}}{\rho \kappa_0' \left(\kappa_1' + \kappa_2'\right)^2}} + 1\right)}{2 C_\rho \epsilon_{\text{pen}}} + \frac{18}{\rho},$$

where ϵ_d is a given positive constant and

$$\kappa_0' = \frac{2\left(2 + \epsilon_d\right)}{1 + \epsilon_d}, \kappa_1' = \frac{1}{\sqrt{\rho\left(1-\rho\right)}}\left(3\kappa_3'\sqrt{2} + 12 + 16\sqrt{\frac{1-\rho}{\rho}}\right), \kappa_3' \leq 27, \kappa_2' = \frac{1}{\sqrt{\rho\left(1-\rho\right)}}\left(42 + \frac{3}{4\sqrt{\kappa_0'}}\right).$$

For example, if $\rho = \frac{1}{2}$, $C_1 = 2$, $\epsilon_d = 1$, $\kappa_3' = 27$, then $\epsilon_{\text{pen}} = 1 - \frac{1}{C_1} = \frac{1}{2}$, $\kappa_0' = 3$, $\kappa_1' = 2(81\sqrt{2} + 28) = 56 + 162\sqrt{2}$, $\kappa_2' = 2(42 + \frac{\sqrt{3}}{4}) = 84 + \frac{\sqrt{3}}{2}$, and

$$\begin{aligned}
C_\rho &= \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho}, 1\right)\left(\ln\left(1 + \frac{\rho}{1-\rho}\right) - \rho\right) \\
&= 2\left(\ln 2 - \frac{1}{2}\right) = 2\ln 2 - 1, \\
\kappa_0 &= \frac{\kappa_0' \left(\kappa_1' + \kappa_2'\right)^2 \left(\sqrt{1 + \frac{72(2\ln 2 - 1)}{\kappa_0' \left(\kappa_1' + \kappa_2'\right)^2}} + 1\right)}{2\ln 2 - 1} + 36 \\
&= \frac{3\left(140 + 162\sqrt{2} + \frac{\sqrt{3}}{2}\right)^2 \left(\sqrt{1 + \frac{72(2\ln 2 - 1)}{3\left(140 + 162\sqrt{2} + \frac{\sqrt{3}}{2}\right)^2}} + 1\right)}{2\ln 2 - 1} + 36 \\
&\approx 2126069.
\end{aligned}$$

According to the previous example, we can see that the theoretical penalty is lower bound by κ_0, which can be too large in practice. This result is not surprising since according to Cohen & Le Pennec (2011, Appendix B.4, page 40, line 7), if we choose ϵ_d small enough then κ_0 scales proportionally to

$$\frac{1}{C_\rho \rho(1-\rho)\epsilon_{\text{pen}}} = \frac{\rho}{(1-\rho)^2\left(\ln\left(1 + \frac{\rho}{1-\rho}\right) - \rho\right)}$$

and thus explodes to $+\infty$ when ρ goes to 1 and C_1 goes to 1. Therefore, it is important to study a natural question whether the constant κ_0 appearing in the penalty can be estimated from the data without loosing a theoretical guaranty on the performance? No definite answer exists so far for MoE regression models, however at least our numerical experiment in Section 3.2.3 shows that the slope heuristic proposed by Birgé & Massart (2007) may lead to a good practical solution. In particular, we seek to mathematically and fully justify the slope heuristic in MoE regression models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively Birgé & Massart (2007), Arlot & Massart (2009), Arlot & Bach (2009), Arlot (2019). We summerize this interesting and important problem in Open Problem 5.4.5.

Remark 3.2.5. The main drawback of the previous weak oracle inequality is using different divergences and requiring some strong assumptions for the inequality to be considered a proper oracle inequality. To illustrate the strictness of the compactness assumption for s_0, we only need to consider s_0 as a univariate Gaussian PDF, which obviously does not satisfy such a hypothesis. This motivates us to investigate more an l_1-oracle inequality of GLoME with the LASSO estimator, which is an extension of Nguyen et al. (2020c, Theorem 3.1) for SGaME and is considered as an l_1-ball model selection procedure. Such an l_1-oracle inequality can be considered as a complementary to the Theorem 3.2.3. In particular, the most intriguing property of the l_1-oracle inequality is that it requires only

the boundedness assumption of the parameters of the model, which is also required in Theorem 3.2.3, as well as in Stadler et al. (2010), Meynet (2013), Devijver (2015a), Nguyen et al. (2020c). Note that such a mild assumption is quite common when working with MLE (cf. Baudry 2009, Maugis & Michel 2011b), to tackle the problem of the unboundedness of the likelihood at the boundary of the parameter space McLachlan & Peel 2000, Redner & Walker 1984, and to prevent it from diverging. Nevertheless, by using the smaller divergence: $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ (or more strict assumptions on s_0 and $s_{\mathbf{m}}$, with the same divergence $\mathrm{KL}^{\otimes \mathrm{n}}$ as ours), Theorem 3.2.3 obtain a faster rate of convergence of order $1/n$, while in the l_1-oracle inequality, we can only obtain a rate of convergence of order $1/\sqrt{n}$. It is important to emphasize that we can relatively compare the rate of convergences of PLMEs and Lasso estimators if the error terms $1/n$ of the oracle inequalities resulting from the penalization and the $1/\sqrt{n}$ of the l_1-oracle inequality are both dominant concerning all other constants. Therefore, in the case where there is no guarantee of a compact support of s_0 or uniformly bounded regression functions, the l_1-oracle inequality gives a theoretical foundation for the l_1-ball model selection procedure, with the order of convergence of $1/\sqrt{n}$, with only the boundedness assumption on the parameter space.

3.2.3 Numerical experiments

Note that https://github.com/Trung-TinNGUYEN/NamsGLoME-Simulation contains our numerical experiments in Section 3.2.3, which are written in the **R** programming language (R Core Team, 2020).

3.2.3.1 The procedure

We illustrate our theoretical results in settings similar to those considered by Chamroukhi et al. (2010) and Montuelle et al. (2014), regarding simulated as well as real data sets. We first observe n random samples $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$ from an forward conditional density s_0^*, and look for the best data-driven estimate among $s_{\mathbf{m}}^* \in S_{\mathbf{m}}^*, \mathbf{m} \in \mathcal{M}$, defined in (3.2.16). We considered the simple case where the mean experts are linear functions, which leads to GLoME and supervised GLLiM are identical models. Our aim is to estimate the best data-driven number of components K, as well as the model parameters. As described in more detail in Deleforge et al. (2015c), we use a GLLiM-EM algorithm to estimate the model parameters for each K, and select the optimal model using the penalized approach that was described earlier. More precisely, in the following numerical experiments, the GLoME model is learned using functons from a package xLLiM, available on CRAN. It targets to solve the inverse regression problem, defined in (3.2.13), and obtain the inverse maximum likelihood estimators (MLE) $\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right)_{i\in[N]}$, $\mathbf{m} \in \mathcal{M}$, then via (3.2.14), we obtain the forward MLE $\left(\widehat{s}_{\mathbf{m}}^*\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)_{i\in[N]}$, $\mathbf{m} \in \mathcal{M}$.

According to the general procedure for model selection, we first compute the forward MLE for each model $\mathbf{m} \in \mathcal{M}$, where $\mathcal{M} = \mathcal{K}$. Then, we select the model that satisfies the definition (3.2.18) with $\mathrm{pen}(\mathbf{m}) = \kappa \dim(S_{\mathbf{m}}^*)$, where κ is a positive hyperparameter. In particularly, we need a data-driven method to choose κ, even though our Theorem 3.2.3 and Remark 3.2.2 guarantee that there exists a κ large enough for which the estimate has the desired properties. According to the AIC or the BIC, we can select $\kappa = 1$ or $\kappa = \frac{\ln n}{2}$. An important limitation of these criteria, however, is that they are based on asymptotic theory. To overcome this difficulty, the slope heuristic was proposed by Birgé & Massart (2007) (see, also Baudry et al. 2012). Furthermore, our Theorem 3.2.3 provides some theoretical justifications for the shapes of penalty functions when utilizing the slope heuristic approach in a finite sample setting. Thus, we shall concentrate our attention on the slope heuristic for choosing the number of mixture components in our numerical experiments.

3.2.3.2 Simulated data sets

Note that our main objective here is to investigate how well the empirical tensorized Kullback–Leibler divergence between the true model (s_0^*) and the selected model $\widehat{s}_{\widehat{\mathbf{m}}}^*$ follows the finite-sample oracle inequality of Theorem 3.2.3, as well as the rate of convergence of the error term. Therefore, we focus on 1-dimensional data sets, that is, with $L = D = 1$. Beyond the statistical estimation and model selection objectives considered here, the dimensionality reduction capability of GLLiM in high-dimensional regression data, typically $D \gg L$, can be found in (Deleforge et al., 2015c, Section 6).

We construct simulated data sets following two scenarios: a *well-specified* (WS) case in which the true forward conditional density belongs to the class of proposed models:

$$s_0^*(y|x) = \frac{\phi(x; 0.2, 0.1)}{\phi(x; 0.2, 0.1) + \phi(x; 0.8, 0.15)} \phi(y; -5x + 2, 0.09) + \frac{\phi(x; 0.8, 0.15)}{\phi(x; 0.2, 0.1) + \phi(x; 0.8, 0.15)} \phi(y; 0.1x, 0.09),$$

and a *misspecified* (MS) case, whereupon such an assumption is not true:

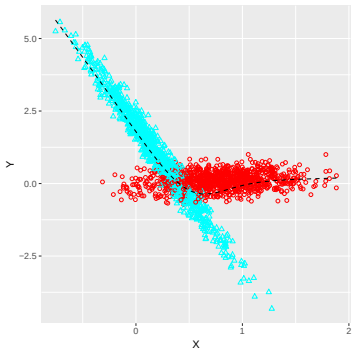
$$s_0^*(y|x) = \frac{\phi(x; 0.2, 0.1)}{\phi(x; 0.2, 0.1) + \phi(x; 0.8, 0.15)} \phi\left(y; x^2 - 6x + 1, 0.09\right) + \frac{\phi(x; 0.8, 0.15)}{\phi(x; 0.2, 0.1) + \phi(x; 0.8, 0.15)} \phi\left(y; -0.4x^2, 0.09\right).$$

Figures 3.1a and 3.1e show some typical realizations of 2000 data points arising from the WS and MS scenarios. Note that by using GLoME, our estimator performs well in the WS setting (Figures 3.1b to 3.1d). In the MS case, we expect our algorithm to automatically balance the model bias and its variance (Figures 3.1f to 3.1h), which leads to the choice of a complex model, with 4 mixture components. This observation will be elaborated upon in the subsequent experiments.

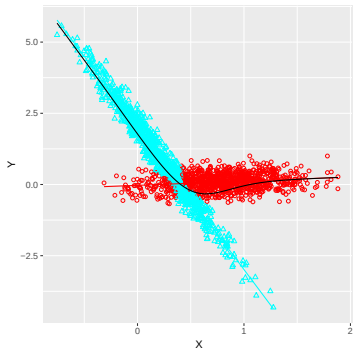
Firstly, by using the capushe (CAlibrating Penalities Using Slope HEuristics) package in **R** (Arlot et al., 2016, Baudry et al., 2012), we can select the penalty coefficient, along with the number of mixture components K. This heuristic comprises two possible criteria: the slope criterion and the jump criterion. The first criterion consists of computing the *asymptotic* slope of the log-likelihood (Figure 3.2), drawn according to the model dimension, and then penalizing the log-likelihood by twice the slope times the model dimension. Regarding the second criterion, one aims to represent the dimension of the selected model according to κ (Figure 3.3), and find $\widehat{\kappa}$, such that if $\kappa < \widehat{\kappa}$, then the dimension of the selected model is large, and of reasonable size, otherwise. The slope heuristic prescribes then the use of $\kappa = 2\widehat{\kappa}$. In our simulated data sets, Figure 3.4 shows that the jump criterion appears to work better. The slope criterion sometimes chooses very highly complex models in the WS case, with the problem exacerbated in the MS case.

Next, a close inspection shows that the bias-variance trade-off differs between the two examples. We run our experiment over 100 trials with $K \in \mathcal{K} = [20]$, using both the jump and slope criteria. The first remark is that the best data-driven choice of $K = 2$ appears to be selected with very high probability, even for large $n = 10000$ in the WS case. This can be observed in Figures 3.4a, 3.4b, 3.4e and 3.4f. In the MS case, the best data-driven choice for K should balance between the model approximation error term and the variance one, which is observed in Figures 3.4c, 3.4d, 3.4g and 3.4h. Here, the larger the number of samples n, the larger the value of K that is selected as optimal.

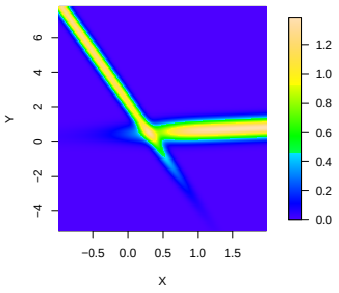
From hereon in, we focus on the jump instead of slope criterion, due to its stability regarding the choice of K. We wish to measure the performances of our chosen GLoME models in term of tensorized Kullback–Leibler divergence, $\mathrm{KL}^{\otimes \mathrm{n}}$, which can not be calculated exactly in the case of Gaussian mixtures. Therefore, we evaluate the divergence using a Monte Carlo simulation, since we know the true density. We should note that the variability of this randomized approximation has been verified to be negligible in practice, which is also supported in the numerical experiments by Montuelle et al. (2014). More precisely, we compute the Monte Carlo approximation for the tensorized Kullback–Leibler divergence as follows. First, note that the Monte Carlo approximation for tensorized Kullback–Leibler divergence between the true model (s_0^*) and the selected model $\widehat{s}^*_{\widehat{\mathbf{m}}}$



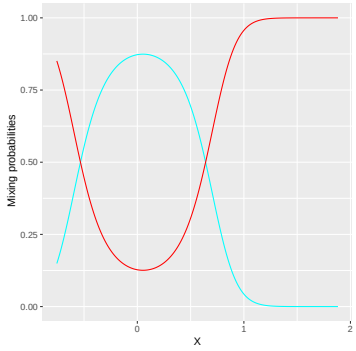
(a) Typical realization of example WS



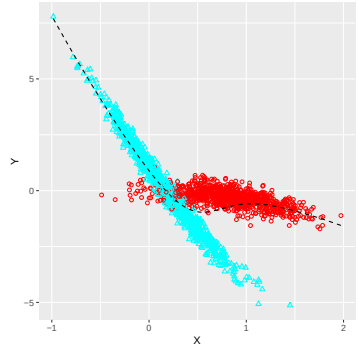
(b) Clustering by GLoME



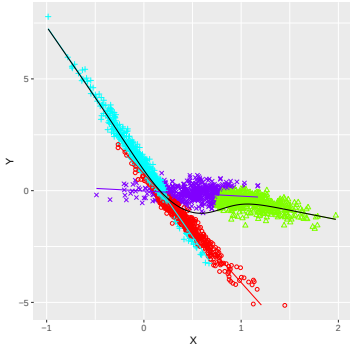
(c) 2D view of the resulting conditional density with the 2 regression components



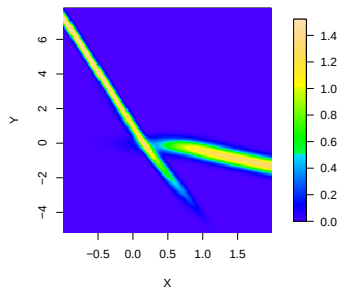
(d) Gating network probabilities



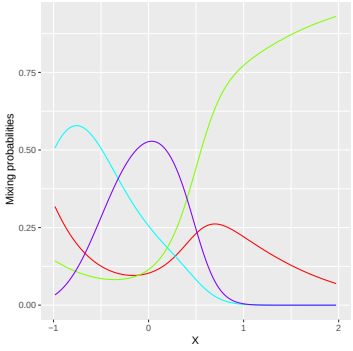
(e) Typical realization of example MS



(f) Clustering by GLoME

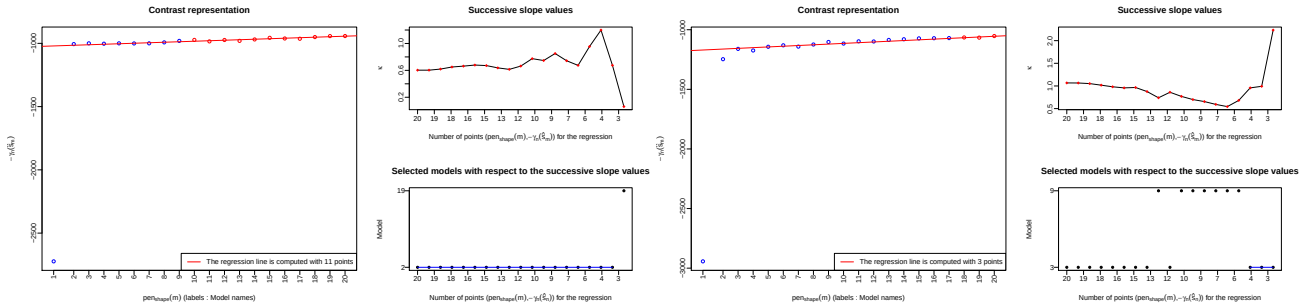


(g) 2D view of the resulting conditional density with the 4 regression components



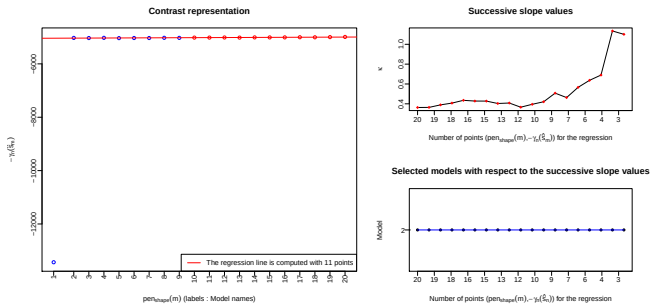
(h) Gating network probabilities

Figure 3.1: Clustering deduced from the estimated conditional density of GLoME by a MAP principle with 2000 data points of example WS and MS. The dash and solid black curves present the true and estimated mean functions.

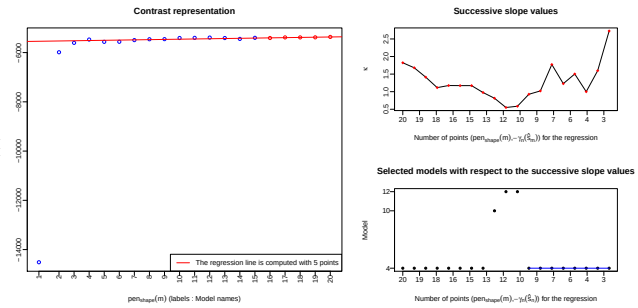


(a) Example WS with 2000 data points

(b) Example MS with 2000 data points



(c) Example WS with 10000 data points



(d) Example MS with 10000 data points

Figure 3.2: Plot of the selected model dimension using the slope criterion.

can be approximated as

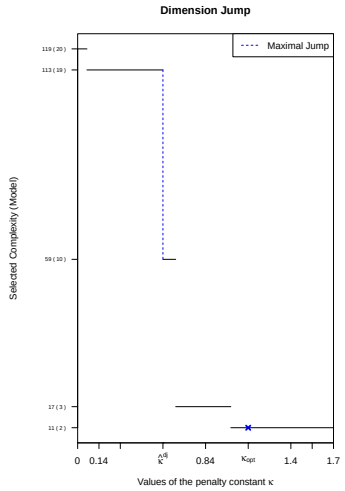
$$\begin{aligned}\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0^*, \widehat{s}^*_{\widehat{\mathbf{m}}}\right) &= \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s_0^*\left(\cdot|\mathbf{x}_i\right), \widehat{s}^*_{\widehat{\mathbf{m}}}\left(\cdot|\mathbf{x}_i\right)\right)\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s_0^*\left(\cdot|\mathbf{x}_i\right), \widehat{s}^*_{\widehat{\mathbf{m}}}\left(\cdot|\mathbf{x}_i\right)\right) \text{(using 1 data point to approximate } \mathbb{E}_{\mathbf{X}}\left[\cdot\right]\text{)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_y}\sum_{j=1}^{n_y}\ln\left(\frac{s_0^*\left(\mathbf{y}_{i,j}|\mathbf{x}_i\right)}{\widehat{s}^*_{\widehat{\mathbf{m}}}\left(\mathbf{y}_{i,j}|\mathbf{x}_i\right)}\right),\end{aligned}$$

where the data $\mathbf{x}_i, i \in [n]$, and $\left(\mathbf{y}_{i,j}\right)_{j\in[n_y]}$ are drawn from $s_0^*\left(\cdot|\mathbf{x}_i\right)$. Then, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0^*, \widehat{s}^*_{\widehat{\mathbf{m}}}\right)\right]$ is approximated again by averaging over $N_t = 100$ Monte Carlo trials. Therefore, the simulated data used for approximation can be written as $\left(\mathbf{x}_i, \mathbf{y}_{i,j}\right)_t$ with $i \in [n], j \in [n_y], t \in [N_t]$.

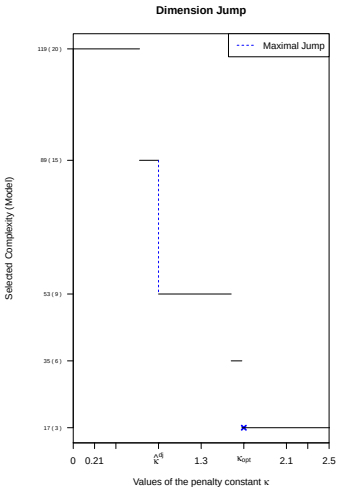
Based on the approximation, Figure 3.5 shows the box plots and the mean of the tensorized Kullback–Leibler divergence over 100 trials, based on the jump criterion. Our box-plots confirm that the mean tensorized Kullback–Leibler divergence between $\widehat{s}^*_K$ and s_0^*, over $K \in \{1, \ldots, 20\}$ number of mixture components, is always larger than the mean of tensorized Kullback–Leibler divergence between the penalized estimator $\widehat{s}^*_{\widehat{K}}$ and s_0^*, which is consistent with Theorem 3.2.3. In particularly, if the true model belongs to our nested collection, the mean tensorized Kullback–Leibler divergence seems to behave like $\frac{\dim(S^*_{\mathbf{m}})}{2n}$ (shown by a dotted line), which can be explained by the AIC heuristic. More precisely, we firstly assume that

$$S^*_{\mathbf{m}} = \left\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto s^*_{\mathbf{m}} := s_{\boldsymbol{\psi}^*_{\mathbf{m}}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\psi}^*_{\mathbf{m}} \in \boldsymbol{\Psi}^*_{\mathbf{m}} \subset \mathbb{R}^{\dim(S^*_{\mathbf{m}})}\right\}$$

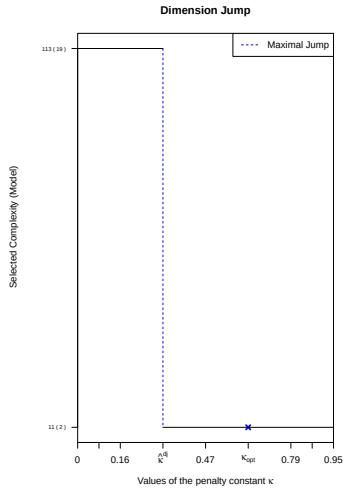
is identifiable and make some strong regularity assumptions on $\boldsymbol{\psi}^*_{\mathbf{m}} \mapsto s^*_{\boldsymbol{\psi}_m}$. Further, we assume the



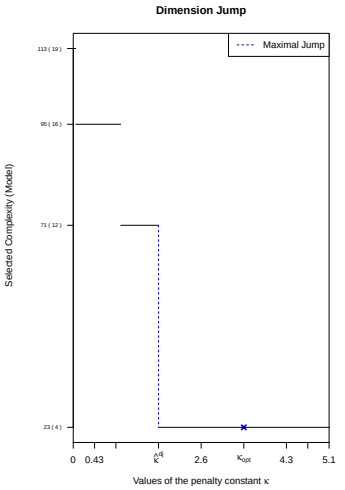
(a) Example WS with 2000 data points



(b) Example MS with 2000 data points

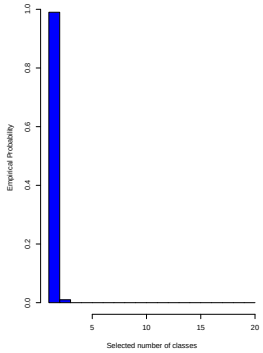


(c) Example WS with 10000 data points

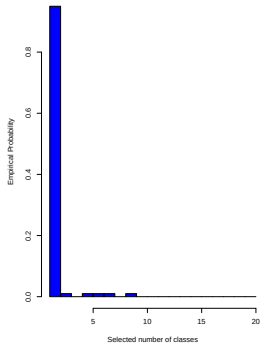


(d) Example MS with 10000 data points

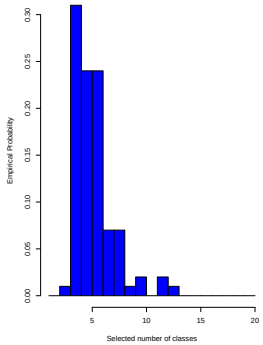
Figure 3.3: Plot of the selected model dimension using the jump criterion.



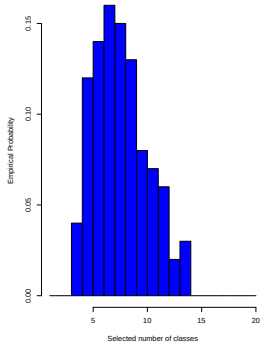
(a) 2000 WS data points using jump criterion



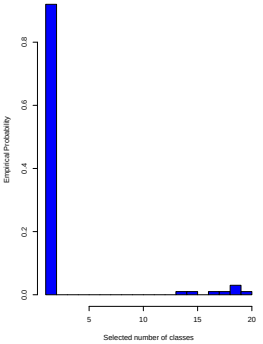
(b) 10000 WS data points using jump criterion



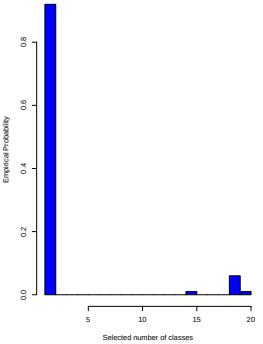
(c) 2000 MS data points using jump criterion



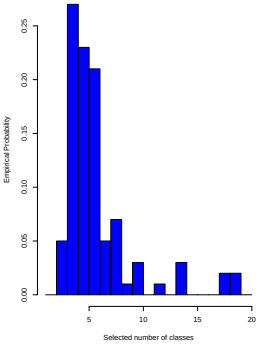
(d) 10000 MS data points using jump criterion



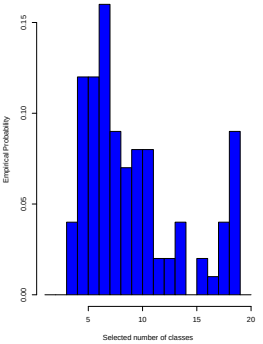
(e) 2000 WS data points using slope criterion



(f) 10000 WS data points using slope criterion

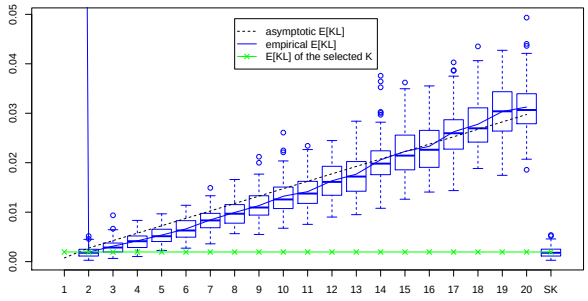


(g) 2000 MS data points using slope criterion

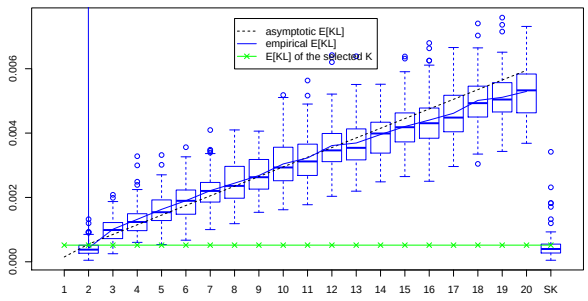


(h) 10000 MS data points using slope criterion.

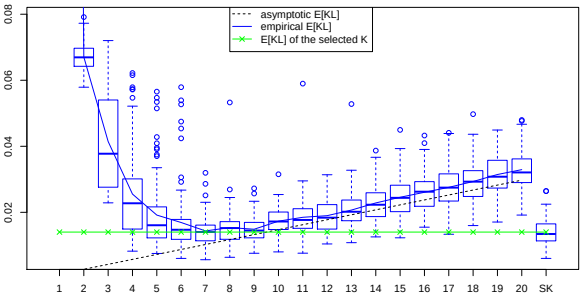
Figure 3.4: Comparison histograms of selected K between WS and MS cases using jump and slope criteria over 100 trials.



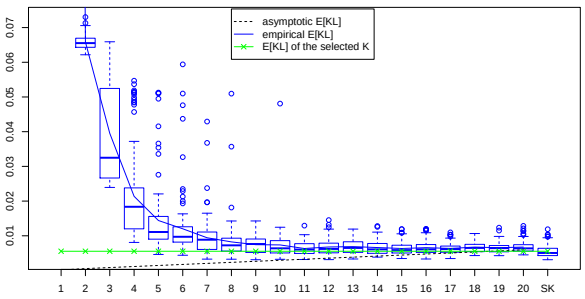
(a) Example WS with 2000 data points



(b) Example WS with 10000 data points



(c) Example MS with 2000 data points



(d) Example MS with 10000 data points

Figure 3.5: Box-plot of the tensorized Kullback–Leibler divergence according to the number of mixture components using the jump criterion over 100 trials. The tensorized Kullback–Leibler divergence of the penalized estimator $\widehat{s}_{\widehat{K}}$ is shown in the right-most box-plot of each graph.

existence of $\dim\left(S^*_{\mathbf{m}}\right) \times \dim\left(S^*_{\mathbf{m}}\right)$ matrices $\mathbf{A}\left(\boldsymbol{\psi}^*_{\mathbf{m}}\right)$ and $\boldsymbol{B}\left(\boldsymbol{\psi}^*_{\mathbf{m}}\right)$, which are defined as follows:

$$\left[\mathbf{A}\left(\boldsymbol{\psi}^*_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{-\partial^2 \ln s_{\boldsymbol{\psi}^*_{\mathbf{m}}}}{\partial \boldsymbol{\psi}^*_{\mathbf{m},k}\partial \boldsymbol{\psi}^*_{\mathbf{m},l}}\left(\mathbf{y}|\mathbf{x}_i\right) s^*_0\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y}\right],$$

$$\left[\boldsymbol{B}\left(\boldsymbol{\psi}^*_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \ln s_{\boldsymbol{\psi}^*_{\mathbf{m}}}}{\partial \boldsymbol{\psi}^*_{\mathbf{m},k}}\left(\mathbf{y}|\mathbf{x}_i\right) \frac{\partial \ln s_{\boldsymbol{\psi}^*_{\mathbf{m}}}}{\partial \boldsymbol{\psi}^*_{\mathbf{m},l}}\left(\mathbf{y}|\mathbf{x}_i\right) s^*_0\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y}\right].$$

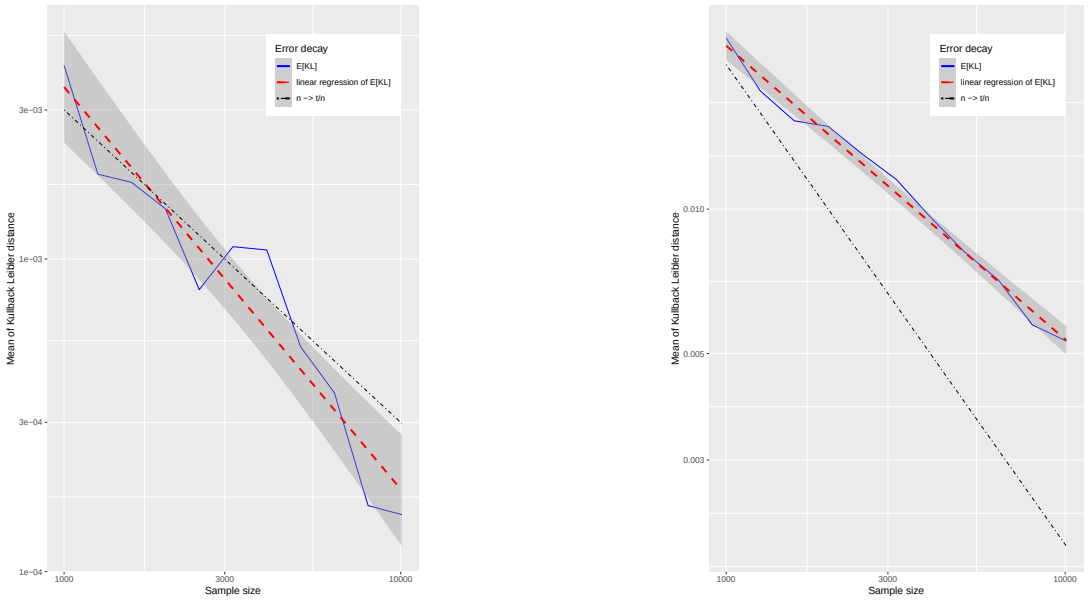
Then, the results from White (1982) and Cohen & Le Pennec (2011) imply that $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s^*_0, \widehat{s}^*_{\mathbf{m}}\right)\right]$ is asymptotically equivalent to

$$\mathrm{KL}^{\otimes \mathrm{n}}\left(s^*_0, s_{\boldsymbol{\psi}^{**}_{\mathbf{m}}}\right) + \frac{1}{2n} tr\left(\boldsymbol{B}\left(\boldsymbol{\psi}^{**}_{\mathbf{m}}\right)\mathbf{A}\left(\boldsymbol{\psi}^{**}_{\mathbf{m}}\right)^{-1}\right),$$

where we defined $\boldsymbol{\psi}^{**}_{\mathbf{m}} = \arg\min_{s_{\boldsymbol{\psi}^*_{\mathbf{m}}} \in S^*_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s^*_0, s_{\boldsymbol{\psi}^*_m}\right)$.

In particularly, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s^*_0, \widehat{s}^*_{\mathbf{m}}\right)\right]$ becomes asymptotically equivalent to $\frac{1}{2n}\dim\left(S^*_{\mathbf{m}}\right)$, whenever s^*_0 belongs to the model collection $S^*_{\mathbf{m}}$. Furthermore, even though there is no theoretical guarantee, the slope of the mean error in the misspecified case seems also to grow at the same rate as $\frac{\dim(S^*_{\mathbf{m}})}{2n}$, for large enough number of mixture components ($K \geq 6$ in the WS case and $K \geq 9$ in the MS case).

Figure 3.6 shows that the error decays when the sample size n grows, when using the penalty based on the jump criterion. The first remark is that we observed the error decay is of order t/n, as predicted in by the theory, where t is some constant, as expected in the well-specified case. The rate of convergence for the misspecified case seems to be slower.



(a) WS Example: The slope of the free regression line is ≈ -1.287 and $t = 3$.

(b) MS Example: The slope of the free regression line is ≈ -0.6120 and $t = 20$.

Figure 3.6: Tensorized Kullback–Leibler divergence between the true and selected densities based on the jump criterion, represented in a log-log scale, using 30 trials. A free least-square regression with standard error and a regression with slope -1 were added to stress the two different behavior for each graph.

3.2.3.3 Ethanol data set

We now consider the use of GLoME models for performing clustering and regression tasks on a real data set. Following the numerical experiments from Young (2014) and Montuelle et al. (2014), we demonstrate our model on the ethanol data set of Brinkman (1981). The data comprises of 88 observations, which represent the relationship between the engine's concentration of nitrogen oxide (NO) emissions and the equivalence ratio (ER), a measure of the air-ethanol mix, used as a spark-ignition engine fuel in a single-cylinder automobile test (Figures 3.8a and 3.8e). Our goal is then to estimate the parameters of a GLoME model, as well as the number of mixture components.

More precisely, we first use the EM algorithm from the xLLiM package to compute the forward PMLE of (3.2.1), for each $K \in [12]$, on the Ethanol data set. Then, based on the slope heuristic (Figure 3.7), we select the best data-driven model. Given the estimators of the model chosen, we obtain the estimated conditional density and clustering by applying the maximum a posteriori probability (MAP) rule (Figures 3.8 and 3.9).

Because we only have 88 data points and roughly 6 parameters per class, the EM algorithm is strongly dependent on the random initialization of the parameters. One solution is that we can modify slightly that procedure in order to guarantee that at least 10 points are assigned to each class so that the estimated parameters are more stable (cf. Montuelle et al. 2014). In this work, we wish to investigate how well our proposed PMLE performs for detecting the best data-driven number of mixture components for the GLoME model. Thus, we run our experiment over 100 trials with different initializations for the EM algorithm. Histograms of selected values of K are presented in Figures 3.7a to 3.7d. Notice that it is quite unlikely that the true conditional PDF of Ethanol data set belongs to our hypothesised collection of GLoME models. In fact, this phenomenon has been observed in the MS case, Figure 3.4, on the simulated data set. We think this is due to the simplistic affine models used in our experiments. Furthermore, it seems that the jump criterion outperformed the slope criterion in the stability of order selection for GLoME models, as previously observed.

Based on the highest empirical probabilities in all situations, our procedure selects $K = 4$ components, which is consistent with the results from Montuelle et al. (2014). It is worth noting that if we consider the regression of NO with respect to ER, our proposed PMLE of GLoME performs very well

for both the clustering and regression tasks (Figure 3.9). Here, instead of considering the variable NO as the covariate, we use it as the response variable. Then, the resulting clustering, the estimated mean function (black curve) and mixing probabilities are more easily interpretable. This is very similar to the results obtained in Montuelle et al. (2014).

3.2.4 Proofs of the oracle inequality

To work with conditional density estimation in the GLoME regression models, in Section 3.2.4.1, we need to present a general theorem for model selection, Theorem 3.2.6, a generalization of Theorem 7.11 from Massart (2007), from Cohen & Le Pennec (2011, Theorem 2) and Montuelle et al. (2014, Theorem 2). Then, we explain how we can use Theorem 3.2.6 to obtain the oracle inequality Theorem 3.2.3 in Section 3.2.4.2. To this end, our model collection has to satisfy some regularity assumptions, which is proved in Section 3.2.5.2 and Section 3.2.5.5. The main difficulty to proving our weak oracle inequality lies in bounding the bracketing entropy of the Gaussian gating functions of the GLoME model. To overcome this difficulty, we propose a reparameterization trick to bound the metric entropy of the Gaussian gating parameters space.

3.2.4.1 General model selection for conditional density

Before stating a general model selection for conditional density, we have to present some regularity assumptions.

First, we need an information theory type assumption to control the complexity of our collection. We assume the existence of a Kraft-type inequality for the collection (Massart, 2007, Barron et al., 2008).

Assumption 3.2.1 (K). *There is a family* $(z_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *of non-negative numbers and a real number* Ξ *such that*

$$\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < +\infty.$$

For technical reasons, a seperability assumption always satisfied in the setting of this paper, is also required. It is a mild condition, which is classical in empirical process theory (Van Der Vaart & Wellner, 1996, van de Geer, 2000). This assumption allows us to work with a countable subset.

Assumption 3.2.2 (Sep). *For every model* $S_{\mathbf{m}}$ *in the collection* $\mathcal{S}$*, there exists some countable subset* $S'_{\mathbf{m}}$ *of* $S_{\mathbf{m}}$ *and a set* $\mathcal{X}'_{\mathbf{m}}$ *with* $\iota\left(\mathcal{X}\setminus\mathcal{X}'_{\mathbf{m}}\right)=0$*, where* ι *denotes Lebesgue measure, such that for every* $t\in S_{\mathbf{m}}$*, there exists some sequence* $(t_k)_{k\geq 1}$ *of elements of* $S'_{\mathbf{m}}$*, such that for every* $\mathbf{y}\in\mathcal{Y}$ *and every* $\mathbf{x}\in\mathcal{X}'_{\mathbf{m}}, \ln\left(t_k\left(\mathbf{x}|\mathbf{y}\right)\right)\xrightarrow{k\to+\infty}\ln\left(t\left(\mathbf{x}|\mathbf{y}\right)\right)$.

Next, recall that the bracketing entropy of a set S with respect to any distance d, denoted by $\mathcal{H}_{[\cdot],d}((\delta,S))$, is defined as the logarithm of the minimal number $\mathcal{N}_{[\cdot],d}\left(\delta,S\right)$ of brackets $[t^-,t^+]$ covering S, such that $d(t^-,t^+)\leq\delta$. That is,

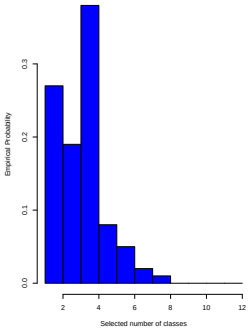
$$N_{[\cdot],d}\left(\delta,S\right):=\min\left\{n\in\mathbb{N}^\star:\exists t_1^-,t_1^+,\ldots,t_n^-,t_n^+\text{ s.t }d(t_k^-,t_k^+)\leq\delta, S\subset\bigcup_{k=1}^{n}\left[t_k^-,t_k^+\right]\right\},\tag{3.2.23}$$

where the bracket $s\in\left[t_k^-,t_k^+\right]$ is defined by $t_k^-(\mathbf{x},\mathbf{y})\leq s(\mathbf{x},\mathbf{y})\leq t_k^+(\mathbf{x},\mathbf{y}),\ \forall(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$.

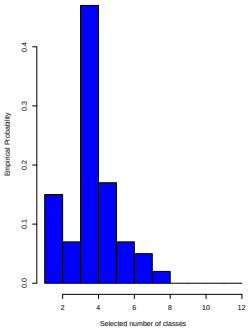
We also need the following important assumption on Dudley-type integral of these bracketing entropies, which is utilized often in empirical process theory (Van Der Vaart & Wellner, 1996, Kosorok, 2007, van de Geer, 2000).

Assumption 3.2.3 (H). *For every model* $S_{\mathbf{m}}$ *in the collection* $\mathcal{S}$*, there is a non-decreasing function* $\phi_{\mathbf{m}}$ *such that* $\delta\mapsto\frac{1}{\delta}\phi_{\mathbf{m}}(\delta)$ *is non-increasing on* $(0,\infty)$ *and for every* $\delta\in\mathbb{R}^+$*,*

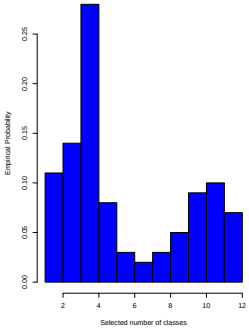
$$\int_0^\delta\sqrt{\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes\mathrm{n}}}\left(\delta,S_{\mathbf{m}}\left(\widetilde{s},\delta\right)\right)}d\delta\leq\phi_{\mathbf{m}}(\delta),$$



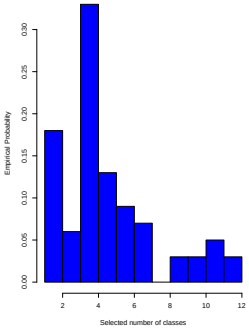
(a) Based on NO and jump criterion



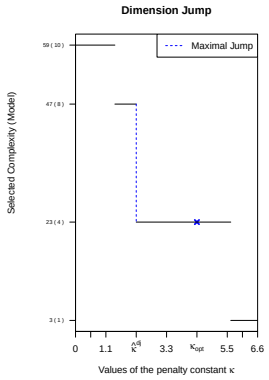
(b) Based on ER and jump criterion



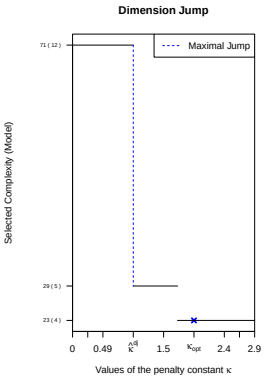
(c) Based on NO and slope criterion



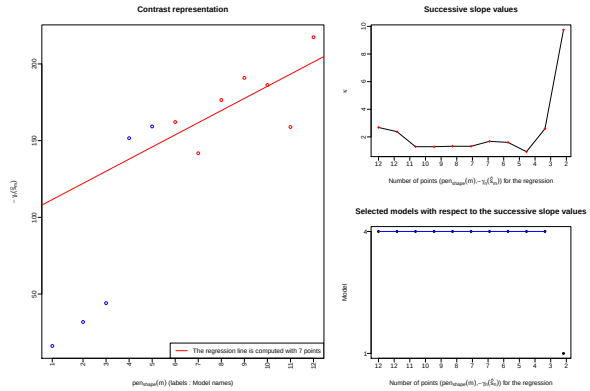
(d) Based on ER and slope criterion



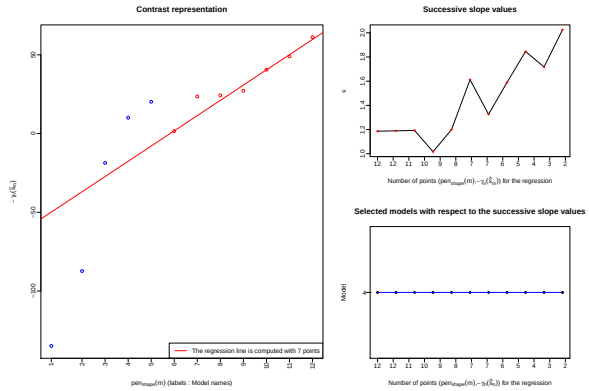
(e) Jump criterion based on NO



(f) Jump criterion based on ER

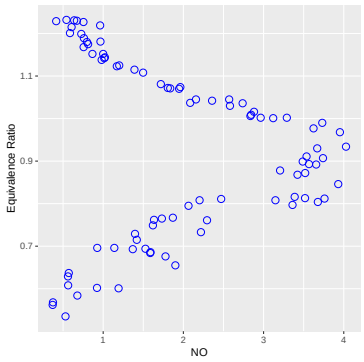


(g) Slope criterion based on NO

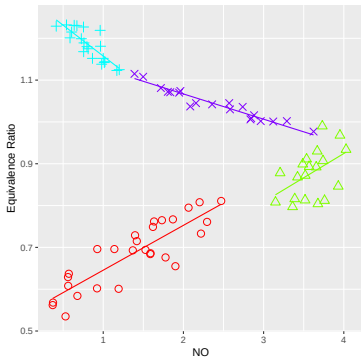


(h) Slope criterion based on ER

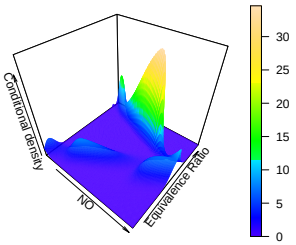
Figure 3.7: Histogram of selected K of GLoME on Ethanol data set based on NO and ER using slope heuristic. We plot the jump and slope criteria corresponding to the models chosen with highest empirical probabilities.



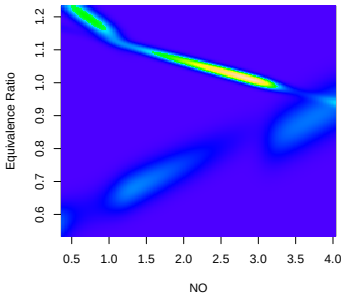
(a) Raw Ethanol data set based on NO.



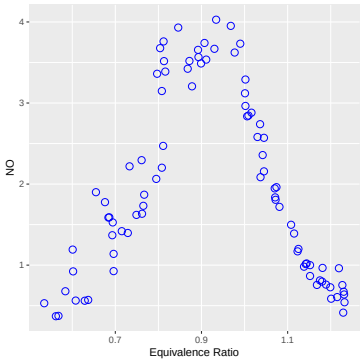
(b) Clustering by GLoME based on NO.



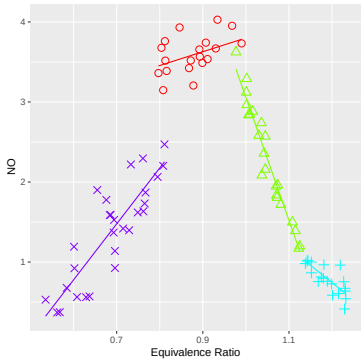
(c) 3D view of the resulting conditional density based on NO with the 4 clusters.



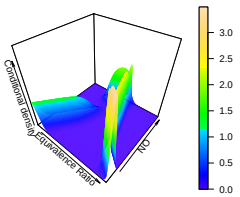
(d) 2D view of the same conditional density on NO.



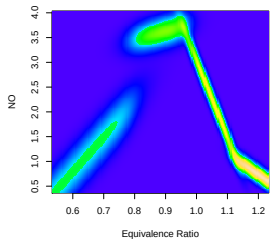
(e) Raw Ethanol data set based on ER.



(f) Clustering by GLoME based on ER.

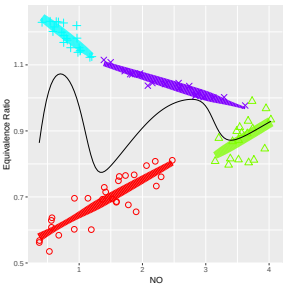


(g) 3D view of the estimated conditional density with the 4 clusters.

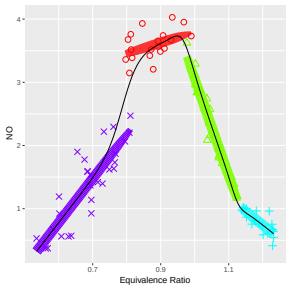


(h) 2D view of the same conditional density.

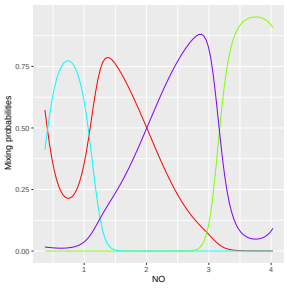
Figure 3.8: Estimated conditional density with 4 components based upon on the covariate NO or ER from the Ethanol data set.



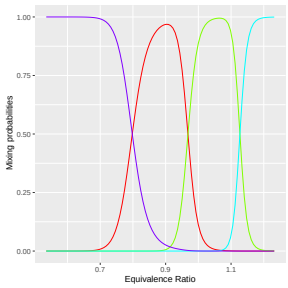
(a) Clustering based on NO



(b) Clustering based on ER



(c) Gating network probabilities based on NO



(d) Gating network probabilities based on ER

Figure 3.9: Clustering of Ethanol data set. The black curves present the estimated mean functions. The size of the component mean functions corresponds to the posterior mixture proportions.

where $S_{\mathbf{m}}(\widetilde{s},\delta) = \{s_{\mathbf{m}} \in S_{\mathbf{m}} : \mathrm{d}^{\otimes \mathrm{n}}(\widetilde{s}, s_{\mathbf{m}}) \leq \delta\}$. *The model complexity* $\mathcal{D}_{\mathbf{m}}$ *of* $S_{\mathbf{m}}$ *is then defined as* $n\delta_{\mathbf{m}}^2$, *where* $\delta_{\mathbf{m}}$ *is the unique root of* $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$.

Observe that the model complexity does not depend on the bracketing entropies of the global models $S_{\mathbf{m}}$, but rather on those of smaller localized sets $S_{\mathbf{m}}(\widetilde{s},\delta)$. Now we are able to state an important weak oracle inequality from Cohen & Le Pennec (2011).

Theorem 3.2.6 (Theorem 2 from Cohen & Le Pennec 2011)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$, *arising from an unknown conditional density* s_0. *Let* $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *be an at most countable conditional density model collection. Assume that Assumption 3.2.1 (K), Assumption 3.2.2 (Sep), and Assumption 3.2.3 (H) hold for every model* $S_{\mathbf{m}} \in \mathcal{S}$. *Then, for any* $\rho \in (0,1)$ *and any* $C_1 > 1$, *there is a constant* κ_0 *depending only on* ρ *and* C_1, *such that for every index* $\mathbf{m} \in \mathcal{M}$,

$$\mathrm{pen}(\mathbf{m}) \geq \kappa\left(n\delta_{\mathbf{m}}^2 + z_{\mathbf{m}}\right)$$

with $\kappa > \kappa_0$ *and* $\delta_{\mathbf{m}}$ *is the unique root of* $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$, *such that the* η'*-penalized likelihood estimator* $\widehat{s}_{\widehat{\mathbf{m}}}$ *satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right) + \frac{\kappa_0 C_1 \Xi}{n} + \frac{\eta+\eta'}{n}. \tag{3.2.24}$$

3.2.4.2 Proof of the Theorem 3.2.3

Sketch of the proof of the Theorem 3.2.3 To prove Theorem 3.2.3, we need to apply Theorem 3.2.6. Then, our model collection has to satisfy Assumption 3.2.1 (K), Assumption 3.2.2 (Sep), and Assumption 3.2.3 (H). Since our model is defined by $\mathcal{M} = \mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} = [K_{\max}] \times [d_{\max}]$, the Assumption 3.2.1 (K) is always satisfied. It is interesting to find the optimal family $(z_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ satisfying Assumption 3.2.1 (K), but that is beyond the scope of this paper. The Assumption 3.2.2 (Sep) is true when we consider Gaussian densities. Therefore, our model has only to satisfy the remaining Assumption 3.2.3 (H). Here, we only present the main steps to prove the Assumption 3.2.3 (H). All the technical details are deferred to Section 3.2.5.2 and Section 3.2.5.5. It is worth noting that a similar procedure has been proposed by Montuelle et al. (2014) in the context of SGaME models.

Firstly, we require the following distance over conditional densities:

$$\sup_{\mathbf{y}} d_{\mathbf{x}}(s,t) = \sup_{\mathbf{y}\in\mathcal{Y}} \left(\int_{\mathcal{X}} \left(\sqrt{s(\mathbf{x}|\mathbf{y})} - \sqrt{t(\mathbf{x}|\mathbf{y})} \right)^2 d\mathbf{x} \right)^{1/2}.$$

This leads straightforwardly to $\mathrm{d}^{2\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d^2_{\mathbf{x}}(s,t)$. Then, we also define

$$\sup_{\mathbf{y}} d_k\left(g,g'\right) = \sup_{\mathbf{y}\in\mathcal{Y}} \left(\sum_{k=1}^{K} \left(\sqrt{g_k(\mathbf{y})} - \sqrt{g'_k(\mathbf{y})} \right)^2 \right)^{1/2},$$

for any gating functions g and g'. To this end, given any densities s and t over $\mathcal{X}$, the following distance, depending on $\mathbf{y}$, is constructed as follows:

$$\sup_{\mathbf{y}} \max_{k} d_{\mathbf{x}}(s,t) = \sup_{\mathbf{y}\in\mathcal{Y}} \max_{k\in[K]} d_{\mathbf{x}}\left(s_k(\cdot,\mathbf{y}), t_k(\cdot,\mathbf{y})\right) = \sup_{\mathbf{y}\in\mathcal{Y}} \max_{k\in[K]} \left(\int_{\mathcal{X}} \left(\sqrt{s_k(\mathbf{x},\mathbf{y})} - \sqrt{t_k(\mathbf{x},\mathbf{y})} \right)^2 d\mathbf{x} \right)^{1/2}.$$

Note that definition of complexity of model $S_{\mathbf{m}}$ in Assumption 3.2.3 (H) is related to an classical entropy dimension with respect to a Hellinger type divergence $\mathrm{d}^{\otimes \mathrm{n}}$, due to Proposition 3.2.7.

Proposition 3.2.7 (Proposition 2 from Cohen & Le Pennec 2011)**.** *For any $\delta \in (0,\sqrt{2}]$, such that $\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right)$, the function*

$$\phi_{\mathbf{m}}(\delta) = \delta\sqrt{\dim(S_{\mathbf{m}})}\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min(\delta,1)}\right)} \right)$$

satisfies Assumption 3.2.3 (H). Furthermore, the unique solution $\delta_{\mathbf{m}}$ of $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$, satisfies

$$n\delta^2_{\mathbf{m}} \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} \right)^2 + \left(\ln \frac{n}{\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} \right)^2 \dim(S_{\mathbf{m}})} \right)_+ \right).$$

Therefore, Proposition 3.2.7 implies that Assumption 3.2.3 (H) is proved via Lemma 3.2.8.

Lemma 3.2.8. *For any $\delta \in (0,\sqrt{2}]$, the collection of GLoME models, $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, satisfies*

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right) \right).$$

Lemma 3.2.8 is then obtained by decomposing the entropy terms between the Gaussian gating functions and the Gaussian experts. For the Gaussian gating parameters, the technique for handling the logistic weights of Montuelle et al. (2014) is not directly applicable to the GLoME setting. Therefore, we propose the following *reparameterization trick* of the Gaussian gating space $\mathcal{P}_K$:

$$\begin{aligned} \mathbf{W}_K &= \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \left(\ln\left(\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right)\right)_{k\in[K]} =: \left(\mathbf{w}_k(\mathbf{y};\boldsymbol{\omega})\right)_{k\in[K]} = \mathbf{w}(\mathbf{y};\boldsymbol{\omega}) : \boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}_K \right\}, \\ \mathcal{P}_K &= \left\{ \mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^{K} e^{\mathbf{w}_l(\mathbf{y})}} \right)_{k\in[K]} =: \left(g_{\mathbf{w},k}(\mathbf{y})\right)_{k\in[K]}, \mathbf{w} \in \mathbf{W}_K \right\}. \end{aligned} \tag{3.2.25}$$

We also require the definition of metric entropy of the set $\mathbf{W}_K$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_K)$, and of the set $\mathbf{\Upsilon}_{K,d}$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{\Upsilon}_{K,d})$, which measure the logarithm of the minimal number of balls of radius at most δ, according to a distance $d_{\|\sup\|_\infty}$, needed to cover $\mathbf{W}_K$ and $\mathbf{\Upsilon}_{K,d}$, respectively, where

$$d_{\|\sup\|_\infty}\left((s_k)_{k\in[K]}, (t_k)_{k\in[K]} \right) = \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left\| s_k(\mathbf{y}) - t_k(\mathbf{y}) \right\|_2, \tag{3.2.26}$$

for any K-tuples of functions $(s_k)_{k\in[K]}$ and $(t_k)_{k\in[K]}$. Here, $s_k, t_k : \mathcal{Y} \ni \mathbf{y} \mapsto s_k(\mathbf{y}), t_k(\mathbf{y}) \in \mathbb{R}^L, \forall k \in [K]$, and given $\mathbf{y} \in \mathcal{X}, k \in [K]$, $\|s_k(\mathbf{y}) - t_k(\mathbf{y})\|_2$ is the Euclidean distance in $\mathbb{R}^L$.

Since $\sum_{k=1}^{K} g_{\mathbf{w},k}(\mathbf{y}) = 1, \forall \mathbf{y} \in \mathcal{Y}, \forall \mathbf{w} \in \mathbf{W}_K$, Lemma 3.2.8 is proved due to the following Lemma 3.2.9 by using the fact that $\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}(\delta, S_{\mathbf{m}})$, which is obtained by definition of bracketing entropy and $\mathrm{d}^{\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)$.

Lemma 3.2.9 (Lemma 5 from Montuelle et al. 2014)**.** *Let*

$$\mathbf{W}_K = \left\{\mathcal{Y} \ni \mathbf{y} \mapsto \left(\ln\left(\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right)\right)_{k\in[K]} =: \left(\mathbf{w}_k(\mathbf{y}; \boldsymbol{\omega})\right)_{k\in[K]} = \mathbf{w}\left(\mathbf{y}; \boldsymbol{\omega}\right) : \boldsymbol{\omega} \in \widetilde{\mathbf{\Omega}}_K\right\},$$

$$\mathcal{P}_K = \left\{\mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^K e^{\mathbf{w}_l(\mathbf{y})}}\right)_{k\in[K]} =: \left(g_{\mathbf{w},k}\left(\mathbf{y}\right)\right)_{k\in[K]}, \mathbf{w} \in \mathbf{W}_K\right\}, \text{ and}$$

$$\mathcal{G}_{K,d} = \left\{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, \mathbf{y}) \mapsto \left(\phi_D\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \mathbf{\Sigma}_k\right)\right)_{k\in[K]} : \boldsymbol{\upsilon}_d \in \mathbf{\Upsilon}_{K,d}, \mathbf{\Sigma} \in \mathbf{V}_K\right\}.$$

For all $\delta \in (0, \sqrt{2}]$ *and* $\mathbf{m} \in \mathcal{M}$,

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} \max_k d_{\mathbf{x}}}\left(\frac{\delta}{5}, \mathcal{G}_{K,d}\right).$$

By making use of the Lemma 3.2.9, the remaining task is to control the bracketing entropy of the Gaussian gating functions and experts separately via Lemmas 3.2.10 and 3.2.11, which are proved in Sections 3.2.5.2 and 3.2.5.5, respectively.

Lemma 3.2.10. *For all* $\delta \in (0, \sqrt{2}]$*, there exists a constant* $C_{\mathbf{W}_K}$ *such that*

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) \leq \mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K}}, \mathbf{W}_K\right) \leq \dim\left(\mathbf{W}_K\right)\left(C_{\mathbf{W}_K} + \ln\left(\frac{20\sqrt{K}}{3\sqrt{3}\delta}\right)\right).$$

Lemma 3.2.11. *For all* $\delta \in (0, \sqrt{2}]$*, there exists a constant* $C_{\mathcal{G}_{K,d}}$ *such that*

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} \max_k d_{\mathbf{x}}}\left(\frac{\delta}{5}, \mathcal{G}_{K,d}\right) \leq \dim\left(\mathcal{G}_{K,d}\right)\left(C_{\mathcal{G}_{K,d}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{3.2.27}$$

To this end, Lemma 3.2.9 allows us to conclude that given $\mathfrak{C} = C_{\mathbf{W}_K} + \ln\left(\frac{5K_{\max}\sqrt{K_{\max}}}{a_W}\right) + C_{\mathcal{G}_{K,d}}$,

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} \max_k d_{\mathbf{x}}}\left(\frac{\delta}{5}, \mathcal{G}_{K,d}\right) \leq \dim(S_{\mathbf{m}})\left(\mathfrak{C} + \ln\left(\frac{1}{\delta}\right)\right).$$

Then, Proposition 3.2.7 leads to

$$n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 \dim\left(S_{\mathbf{m}}\right)}\right)_+\right).$$

Finally, Theorem 3.2.6 implies that for any given collection of GLoME models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, the oracle inequality of Theorem 3.2.3 is satisfied when

$$\mathrm{pen}(\mathbf{m}) \geq \kappa\left(\dim(S_{\mathbf{m}})\left(2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 \dim\left(S_{\mathbf{m}}\right)}\right)_+\right) + z_{\mathbf{m}}\right).$$

3.2.5 Appendix: Lemma proofs

3.2.5.1 Proof of Lemma 3.2.1

In Section 3.2.5.1, we aim to provide the proof of one-to-one correspondence defines the link between the inverse and forward conditional distributions not only for the special case of Gaussian distribution in (3.2.14) but also for elliptical distributions (cf. Cambanis et al., 1981, Fang et al., 1990). It is worth mentioning that the multivariate normal distribution, multivariate t-distribution and multivariate Laplace distribution are some instances of elliptical distributions (cf. Frahm, 2004, Chapter 1, Hult & Lindskog 2002). In fact, a statement similar to the following has been proved in the linear regression setting in (Devijver & Perthame, 2020, Section 2.2). We include a proof for mixture of regression models for completeness, which can be considered as an extension to the aforementioned result.

Elliptically symmetric distributions

Note that we will provide the proof of Lemma 3.2.1 by using some general results regarding elliptical distributions.

Definition 3.2.12. Let $\mathbf{X}$ be a D-dimensional random vector. Then $\mathbf{X}$ is said to be *elliptically distributed* (or simply *elliptical*) if and only if there exist a vector $\boldsymbol{\mu} \in \mathbb{R}^D$, a positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ and a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ such that the characteristic function $\mathbf{t} \mapsto \varphi_{\mathbf{X}-\boldsymbol{\mu}}(\mathbf{t})$ of $\mathbf{X}-\boldsymbol{\mu}$ corresponds to $\mathbf{t} \mapsto \phi\left(\mathbf{t}^\top \boldsymbol{\Sigma}\mathbf{t}\right), \mathbf{t} \in \mathbb{R}^D$. We write $\mathbf{X} \sim \mathcal{E}_D\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi\right)$ to denote that $\mathbf{X}$ is elliptical.

The function ϕ is referred to as the *characteristic generator* of $\boldsymbol{X}$. When $D = 1$ the class of elliptical distributions coincides with the class of univariate symmetric distributions. Thanks to Proposition 1 from Frahm (2004), it holds that every affinely transformed elliptical random vector is elliptically distributed. Moreover, the following stochastic representation theorem, Theorem 3.2.13, shows that the converse is true if the transformation matrix has full rank.

Theorem 3.2.13 (Theorem 1 from Cambanis et al., 1981)**.** *Let $\mathbf{X}$ be a D-dimensional random vector. Then $\mathbf{X} \sim \mathcal{E}_D\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi\right)$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = k$ if and only if*

$$\mathbf{X} = \boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)},$$

where $\boldsymbol{U}^{(k)}$ is a k-dimensional random vector uniformly distributed on the unit hypersphere with $k-1$ dimensions $\mathcal{S}^{k-1} = \left\{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_2 = 1\right\}$, $\mathcal{R}$ is a non-negative random variable with distribution function F related to ϕ being stochastically independent of $\boldsymbol{U}^{(k)}$, $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ is a rank factorization of $\boldsymbol{\Sigma}$ where $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times k}$ with $\operatorname{rank}(\boldsymbol{\Lambda}) = k$.

Note that via the transformation matrix $\boldsymbol{\Lambda}$, the spherical random vector $\boldsymbol{U}^{(k)}$ produces elliptically contoured density surfaces, whereas the generating random variable $\mathcal{R}$ determines the distribution's shape, in particular the heaviness of the distribution's tails. Further, $\boldsymbol{\mu}$ determines the location of the random vector $\mathbf{X}$. The matrix $\boldsymbol{\Sigma}$ is called the *dispersion matrix* or *scatter matrix* of $\mathbf{X}$. Therefore, it holds that every elliptical distribution belongs to a location-scale-family (Kelker, 1970) defined by an underlying spherical standard distribution.

Example 3.2.14 (Multivariate normal distribution)**.** Let $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times k}$ such that $\boldsymbol{\Sigma} := \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top \in \mathbb{R}^{D\times D}$ is positive definite. The random vector $\boldsymbol{X} \sim \phi_D\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is elliptically distributed since $\mathbf{X}$ is representable as $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{\chi^2_k}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}$. The underlying spherical standard distribution is the standard normal distribution. Further, since $s \mapsto \exp(-s/2)$ is the characteristic generator for the class of normal distributions, the characteristic function of $\boldsymbol{X} - \boldsymbol{\mu}$ corresponds to $\mathbf{t} \mapsto \phi_{\boldsymbol{X}-\boldsymbol{\mu}}(\mathbf{t}) = \exp\left(-\mathbf{t}^\top \boldsymbol{\Sigma}\mathbf{t}\right), \mathbf{t} \in \mathbb{R}^d$.

We next describe some important results on the conditional distributions of elliptical random vectors (cf. Cambanis et al., 1981, Corollary 5, Frahm, 2004, Chapter 1).

Theorem 3.2.15. *Let $\mathbf{X} \sim \mathcal{E}_D\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi\right)$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = k$. It holds that:*

(a) $\mathbb{E}\left(X\right) = \boldsymbol{\mu}$.

(b) $var\left(\mathbf{X}\right) = \frac{\mathbb{E}\left(\mathcal{R}^2\right)}{k}\boldsymbol{\Sigma} = -2\phi'(0)\boldsymbol{\Sigma}$, *if ϕ is differentiable at* 0.

(c) The sum of independent elliptically distributed random vector with the same dispersion matrix $\boldsymbol{\Sigma}$ is elliptically too. Furthermore, the sum of two dependent elliptical random vectors with the same dispersion matrix, which are dependent only through their radial parts $\mathcal{R}$, is also elliptical (Hult & Lindskog, 2002, Thereorem 4.1). More precisely, let $\mathcal{R}$ and $\widetilde{\mathcal{R}}$ be nonnegative random variables and let $\mathbf{X} := \boldsymbol{\mu} + \mathcal{R}\boldsymbol{Z} \sim \mathcal{E}_D\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi\right)$ and $\widetilde{\mathbf{X}} := \widetilde{\boldsymbol{\mu}} + \widetilde{\mathcal{R}}\widetilde{\boldsymbol{Z}} \sim \mathcal{E}_D\left(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}, \widetilde{\phi}\right)$, where $\left(\mathcal{R}, \widetilde{\mathcal{R}}\right)$, Z, and $\widetilde{Z}$ are independent. Then $\mathbf{X} + \widetilde{\mathbf{X}} \sim \mathcal{E}_D\left(\boldsymbol{\mu} + \widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}, \phi^*\right)$.

(d) Affine transformation: every affinely transformed and particularly every linearly combined elliptical random vector is elliptical, too. More formally, for any $\mathbf{b} \in \mathbb{R}^L$, $\mathbf{A} \in \mathbb{R}^{L\times D}$, *and* $\mathbf{Y} = \mathbf{b} + \mathbf{A}\mathbf{X}$ *where* $\mathbf{X} = \boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}$ *with* $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times k}$, *it follows that* $\mathbf{Y} \sim \mathcal{E}_L\left(\mathbf{b} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\mathbf{A}^\top, \phi\right) = \mathcal{E}_L\left(\mathbf{b} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top, \phi\right)$ *since*

$$\mathbf{Y} = \mathbf{b} + \mathbf{A}\left(\boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}\right) = (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}) + \mathcal{R}\mathbf{A}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}. \tag{3.2.28}$$

(e) Marginal distribution: let $\mathcal{P}_m \in \{0,1\}^{m\times D}$ $(m \leq D)$ *be a* permutation and deletion *matrix,* i.e., $\mathcal{P}_m$ *has only binary entries of* 0*'s and* 1*'s and* $\mathcal{P}_m\mathcal{P}_m^\top = \mathrm{I}_m$. *So the transformation* $\mathcal{P}_m\mathbf{X} =: \mathbf{Y}$ *permutes and deletes certain components of* $\mathbf{X}$ *such that* $\mathbf{Y}$ *is a* k*-dimensional random vector containing the remaining components of* $\mathbf{X}$ *and having a (multivariate) marginal distribution with respect to the joint distribution of* $\mathbf{X}$. *Then by* (3.2.28), *we obtain* $\mathbf{Y} \sim \mathcal{E}_m\left(\mathcal{P}_m\boldsymbol{\mu}, \mathcal{P}_m\boldsymbol{\Sigma}\mathcal{P}_m^\top, \phi\right)$ *since*

$$\mathbf{Y} = \mathcal{P}_m\left(\boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}\right) = \mathcal{P}_m\boldsymbol{\mu} + \mathcal{R}\mathcal{P}_m\boldsymbol{\Lambda}\boldsymbol{U}^{(k)}. \tag{3.2.29}$$

(f) Conditional distribution: let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, *where* $\mathbf{X}_1$ *is a* k*-dimensional sub-vector of* $\mathbf{X}$, *and let* $\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} \in \mathbb{R}^{D\times D}$. *Provided the conditional random vector* $\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1$ *exists, it is also elliptically distributed:* $\mathbf{X}_2|(\mathbf{X}_1 = \mathbf{x}_1) \sim \mathcal{E}_{D-k}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \phi^*)$. *Moreover, it can be presented stochastically by*

$$\mathbf{X}_2|\left(\mathbf{X}_1 = \mathbf{x}_1\right) = \boldsymbol{\mu}^* + \mathcal{R}^*\boldsymbol{U}^{(D-k)}\Gamma^*$$

and $\boldsymbol{U}^{(D-k)}$ *is uniformly distributed on* $\mathcal{S}^{(D-k-1)}$, *and*

$$\begin{aligned}\mathcal{R}^* &= \mathcal{R}\sqrt{1-\beta}|\left(\mathcal{R}\sqrt{\beta}\boldsymbol{U}^{(k)} = \boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right)\right),\\ \boldsymbol{\mu}^* &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right),\\ \boldsymbol{\Sigma}^* &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.\end{aligned}$$

where $\beta \sim \mathrm{Beta}\left(k/2, (D-k)/2\right)$ *and* $\mathcal{R}, \beta, \boldsymbol{U}^{(k)}$ *and* $\boldsymbol{U}^{(D-k)}$ *are mutually independent, and* $\boldsymbol{\Sigma}^* = \left(\Gamma^*\right)^\top\Gamma^*$.

Relation between forward and inverse regression

Proposition 1 from Deleforge et al. (2015c), a multivariate extension of Ingrassia et al. (2012), leads to a link between GLLiM, defined in (3.2.13) models, and a Gaussian mixture model on the joint variable $[\mathbf{X}; \mathbf{Y}]$. This result motivates us to establish the general proof for the relationship between forward and inverse mixture of elliptical regression models. More precisely, we consider the following generative model, conditionally on the cluster label:

$$[\mathbf{X}; \mathbf{Y}]\,|\,(Z = k) \sim \mathcal{E}_{L+D}\left(\mathbf{m}_k, \mathbf{V}_k, \phi\right), \tag{3.2.30}$$

where $\mathcal{E}_{L+D}$ denotes an elliptical distribution of dimension $D + L$, and are $\mathbf{m}_k$ and $\mathbf{V}_k$ its location and scale parameters, respectively.

When applying the inverse regression strategy in the context of mixture of elliptical locally-linear mapping, the key point is to account for (3.2.31):

$$\mathbf{X} = \sum_{k=1}^{K}\mathbb{I}\left(Z = k\right)\left(\mathbf{A}_k\mathbf{Y} + \mathbf{b}_k + \mathbf{E}_k\right), \tag{3.2.31}$$

where $\mathbf{A}_k^* \in \mathbb{R}^{D\times L}$ and vector $\mathbf{b}_k^* \in \mathbb{R}^D$, and $\mathbf{E}_k$ is an error term capturing both the reconstruction error due to the local affine approximation and the observation noise in $\mathbb{R}^D$, into the parameterization of $\mathbf{m}_k$ and $\mathbf{V}_k$. Given $Z = k$, it follows from (3.2.30) that $\mathbf{Y}$ is also elliptical distribution by using

Theorem 3.2.15 (e) and $\mathbf{Y}$ can be assumed to have a location $\mathbf{c}_k \in \mathbb{R}^L$ and a scale matrix $\mathbf{\Gamma}_k \in \mathbb{R}^{L\times L}$. We assume further that the error term $\mathbf{E}_k \sim \mathcal{E}\left(\mathbf{0}, \phi_{e_k}, \mathbf{\Sigma}_k\right)$ is an unobserved centered elliptical random noise with residual covariance matrix $\mathbf{\Sigma}_k$ of type ϕ_{e_k}, and is independent of $\mathbf{Y}$. Then, using (3.2.31) and Theorem 3.2.15, we have

$$\begin{aligned}
\mathbb{E}\left(\mathbf{X}|\left(Z=k\right)\right) &= \mathbb{E}\left(\mathbf{A}_k\mathbf{Y}+\mathbf{b}_k+\mathbf{E}_k\right) = \mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k,\\
\operatorname{var}\left(\mathbf{X}|\left(Z=k\right)\right) &= \operatorname{var}\left(\mathbf{A}_k\mathbf{Y}+\mathbf{b}_k+\mathbf{E}_k\right) = \operatorname{var}\left(\mathbf{A}_k\mathbf{Y}\right)+\operatorname{var}\left(\mathbf{E}_k\right) = \mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^\top+\mathbf{\Sigma}_k,\\
\operatorname{cov}\left(\mathbf{X},\mathbf{Y}|\left(Z=k\right)\right) &= \operatorname{cov}\left(\mathbf{A}_k\mathbf{Y}+\mathbf{b}_k+\mathbf{E}_k,\mathbf{Y}\right) = \mathbf{A}_k\mathbf{\Gamma}_k, \operatorname{cov}\left(\mathbf{Y},\mathbf{X}\right) = \operatorname{cov}\left(\mathbf{Y},\mathbf{A}_k\mathbf{Y}\right) = \mathbf{\Gamma}_k^\top\mathbf{A}_k^\top,\\
\mathbf{m}_k &= \begin{bmatrix} \mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k \\ \mathbf{c}_k \end{bmatrix},\\
\mathbf{V}_k &= \begin{bmatrix} \mathbf{\Sigma}_k+\mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^\top & \mathbf{A}_k\mathbf{\Gamma}_k \\ \left(\mathbf{A}_k\mathbf{\Gamma}_k\right)^\top & \mathbf{\Gamma}_k \end{bmatrix}.
\end{aligned} \tag{3.2.32}$$

Note that in the forward and inverse regression problems of elliptical locally-linear mapping (containing the Gaussian case (3.2.8), (3.2.9), (3.2.11) and (3.2.12)), by using Theorem 3.2.15, the joint distribution defined in (3.2.30) allows us to consider a mixture of linear regression problem, characterized by the following marginal and conditional distributions:

$$\mathbf{X}|\left(Z=k\right) \sim \mathcal{E}_D\left(\mathbf{c}_k^*, \mathbf{\Gamma}_k^*, \phi\right), \tag{3.2.33}$$
$$\mathbf{Y}|\left(\mathbf{X}, Z=k\right) = \mathbf{A}_k^*\mathbf{X}+\mathbf{b}_k^*+\mathbf{E}_k^*, \tag{3.2.34}$$
$$\mathbf{Y}|\left(Z=k\right) \sim \mathcal{E}_D\left(\mathbf{c}_k, \mathbf{\Gamma}_k, \phi\right), \tag{3.2.35}$$
$$\mathbf{X}|\left(\mathbf{Y}, Z=k\right) = \mathbf{A}_k\mathbf{Y}+\mathbf{b}_k+\mathbf{E}_k, \tag{3.2.36}$$

where $\mathbf{E}_k \sim \mathcal{E}\left(\mathbf{0}, \phi_{e_k}, \mathbf{\Sigma}_k\right)$ and $\mathbf{E}_k^* \sim \mathcal{E}\left(\mathbf{0}, \phi_{e_k^*}, \mathbf{\Sigma}_k^*\right)$.

Then, we claim that the joint distribution, defined in (3.2.30) and (3.2.32), leads to the marginal and the conditional distributions of Equations (3.2.33) to (3.2.36) and to a mapping between their mean and variance parameters. Indeed, by using conditioning properties of elliptical distributions, see more in Theorem 3.2.15, implies the following marginal for $\mathbf{X}$ and conditional distribution for $\mathbf{Y}$ given $\mathbf{X}$ as follows:

$$\mathbf{X}|\left(Z=k\right) \sim \mathcal{E}_D\left(\mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k, \mathbf{\Sigma}_k+\mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^\top, \phi\right), \tag{3.2.37}$$
$$\mathbf{Y}|\left(\mathbf{X}, Z=k\right) \sim \mathcal{E}_L\left(\boldsymbol{m}_k^{yx}, \mathbf{\Sigma}_k^{yx}, \widetilde{\phi}\right), \tag{3.2.38}$$

where the explicit expression of the characteristic function $\widetilde{\phi}$ can be found in Cambanis et al. (1981, Corollary 5), and

$$\begin{aligned}
\boldsymbol{m}_k^{yx} &= \mathbf{c}_k+\mathbf{\Gamma}_k^\top\mathbf{A}_k^\top\left(\mathbf{\Sigma}_k+\mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^\top\right)^{-1}\left(\mathbf{X}-\mathbf{A}_k\mathbf{c}_k-\mathbf{b}_k\right),\\
\mathbf{\Sigma}_k^{yx} &= \mathbf{\Gamma}_k-\mathbf{\Gamma}_k^\top\mathbf{A}_k^\top\left(\mathbf{\Sigma}_k+\mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^\top\right)^{-1}\mathbf{A}_k\mathbf{\Gamma}_k = \left(\mathbf{\Gamma}_k^{-1}+\mathbf{A}_k^\top\mathbf{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1},
\end{aligned}$$

with the fact that the last equality is the Woodbury matrix identity. Note that the locations and scale matrices of the conditional distribution do not depend upon the third parameter of the joint distribution, and consequently, we do not describe the explicit expression for $\widetilde{\phi}$. We then utilize again the Woodbury matrix identity and the symmetric property of $\mathbf{\Gamma}$ to identify (3.2.33) and (3.2.34) with

(3.2.37) and (3.2.38), respectively, which implies the following important connections:

$$
\begin{aligned}
\mathbf{c}_k^* &= \mathbf{A}_k\mathbf{c}_k + \mathbf{b}_k, \boldsymbol{\Gamma}_k^* = \boldsymbol{\Sigma}_k + \mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top, \boldsymbol{\Sigma}_k^* = \left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}, \\
\mathbf{A}_k^* &= \boldsymbol{\Gamma}_k^\top\mathbf{A}_k^\top\left(\boldsymbol{\Sigma}_k + \mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top\right)^{-1} \\
&= \boldsymbol{\Gamma}_k^\top\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Gamma}_k^\top\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} \\
&= \left[\boldsymbol{\Gamma}_k\left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right) - \boldsymbol{\Gamma}_k\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right]\left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} \\
&= \left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Sigma}_k^*\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}, \\
\mathbf{b}_k^* &= \mathbf{c}_k + \boldsymbol{\Gamma}_k^\top\mathbf{A}_k^\top\left(\boldsymbol{\Sigma}_k + \mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top\right)^{-1}(-\mathbf{A}_k\mathbf{c}_k - \mathbf{b}_k) \\
&= \mathbf{c}_k + \left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}(-\mathbf{A}_k\mathbf{c}_k - \mathbf{b}_k) \\
&= \left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\left[\left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)\mathbf{c}_k + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}(-\mathbf{A}_k\mathbf{c}_k - \mathbf{b}_k)\right] \\
&= \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1}\mathbf{c}_k - \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k).
\end{aligned}
$$

Therefore, the desired results then are obtained by using the fact that the multivariate normal distribution not only has the property of having Gaussian marginal and conditional distributions (Bishop, 2006, Sections 2.3.1 and 2.3.2) but also belongs to elliptical distributions (detailed in Example 3.2.14). Furthermore, it should be stressed that several versions of multivariate t-distributions (*e.g.*, Section 5.5, page 94 of Kotz & Nadarajah (2004), Ding (2016)) have the previous property. This leads to the inverse regression model based on the multivariate t-distributions (Perthame et al., 2018). It will be interesting to find other sub-classes of elliptically contoured distributions that have the closedness property on marginal and conditional distribution so that the previous inverse regression trick can be applied.

3.2.5.2 Proof of Lemma 3.2.10

Note that the first inequality of Lemma 3.2.10 comes from Montuelle et al. (2014, Lemma 4) and describes relationship between the bracketing entropy of $\mathcal{P}_K$ and the entropy of $\mathbf{W}_K$. Therefore, Lemma 3.2.10 is obtained by proving that there exists a constant $C_{\mathbf{W}_K}$ such that $\forall\delta \in (0, 2]$,

$$
\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_K) \leq \dim(\mathbf{W}_K)\left(C_{\mathbf{W}_K} + \ln\left(\frac{1}{\delta}\right)\right), \tag{3.2.39}
$$

where $\dim(\mathbf{W}_K) = K - 1 + KL + K\frac{L(L+1)}{2}$.

In order to establish the proof for (3.2.39), we have to construct firstly the $\delta_{\boldsymbol{\pi}}$-covering $\boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$ of $\boldsymbol{\Pi}_{K-1}$ via Lemma 3.2.16, which is proved in Section 3.2.5.3.

Lemma 3.2.16 (Covering number of probability simplex with maximum norm)**.** *Given any $\delta_{\boldsymbol{\pi}} > 0$, any $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$, we can choose $\widehat{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$, an $\delta_{\boldsymbol{\pi}}$-covering of $\boldsymbol{\Pi}_{K-1}$, so that $\max_{k\in[K]}|\pi_k - \widehat{\boldsymbol{\pi}}_k| \leq \delta_{\boldsymbol{\pi}}$. Furthermore, it holds that*

$$
\mathcal{N}(\delta_{\boldsymbol{\pi}}, \boldsymbol{\Pi}_{K-1}, \|\cdot\|_\infty) \leq \frac{K(2\pi e)^{K/2}}{\delta_{\boldsymbol{\pi}}^{K-1}}. \tag{3.2.40}
$$

Then, by definition of the covering number, (3.2.39) is obtained immediately via Lemma 3.2.17, which controls the covering number of $\mathbf{W}_K$ and is proved in Section 3.2.5.4.

Lemma 3.2.17. *Given a bounded set $\mathcal{Y}$ in $\mathbb{R}^L$ such that $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^L : \|\mathbf{y}\|_\infty \leq C_{\mathcal{Y}}\}$, it holds that $\mathbf{W}_K$ has a covering number satisfied $\mathcal{N}\left(\delta, \mathbf{W}_K, \mathrm{d}_{\|\sup\|_\infty}\right) \leq C\delta^{-\dim(\mathbf{W}_K)}$, for some constant C.*

Indeed, Lemma 3.2.17 implies the desired result by noting that

$$\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_K) = \ln \mathcal{N}\left(\delta, \mathbf{W}_K, d_{\|\sup\|_\infty}\right) \leq \ln\left[\frac{C}{\delta^{\dim(\mathbf{W}_K)}}\right]$$
$$= \dim(\mathbf{W}_K)\left[\frac{1}{\dim(\mathbf{W}_K)}\ln C + \ln\left(\frac{1}{\delta}\right)\right] = \dim(\mathbf{W}_K)\left(C_{\mathbf{W}_K} + \ln\left(\frac{1}{\delta}\right)\right).$$

3.2.5.3 Proof of Lemma 3.2.16

Note that Genovese & Wasserman (2000, Lemma 2) provide a result for controlling a $\delta_{\boldsymbol{\pi}}$-Hellinger bracketing of $\boldsymbol{\Pi}_{K-1}$. However, such result can not be applied for our Lemma 3.2.16 since they use $\delta_{\boldsymbol{\pi}}$-Hellinger bracketing entropy while we use $\delta_{\boldsymbol{\pi}}$-covering number for the probability complex with maximum norm.

Then $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$ if and only if $\boldsymbol{\xi} \in Q^+ \cap U$, where U is the surface of the unit sphere and Q^+ is the positive quadrant of $\mathbb{R}^K$. Next, we divide the unit cube in $\mathbb{R}^K$ into disjoint cubes with sides parallel to the axes and sides of length $\delta_{\boldsymbol{\pi}}/\sqrt{K}$. Let $(\mathbf{C}_j)_{j\in[N]}$ is the subset of these cubes that have non-empty intersection with $Q^+ \cap U$. For any $j \in [N]$, let $\boldsymbol{\nu}_j = (\boldsymbol{\nu}_{j,k})_{k\in[K]}$ be the center of the cube $\mathbf{C}_j$ and $\boldsymbol{\nu}_j^2 = \left(\boldsymbol{\nu}_{j,k}^2\right)_{k\in[K]}$.

Then $\{\boldsymbol{\nu}_j\}_{j\in[N]}$ is a $\delta_{\boldsymbol{\pi}}/\left(2\sqrt{K}\right)$-covering of $Q^+\cap U$, since we have for any $\boldsymbol{\xi} = (\boldsymbol{\xi}_k)_{k\in[K]} \in Q^+\cap U$, there exists $j_0 \in [N]$ such that $\boldsymbol{\xi} \in \mathbf{C}_{j_0}$, and

$$\|\boldsymbol{\xi} - \boldsymbol{\nu}_{j_0}\|_\infty = \max_{k\in[K]} |\boldsymbol{\xi}_k - \boldsymbol{\nu}_{j_0,k}| \leq \frac{\delta_{\boldsymbol{\pi}}}{2\sqrt{K}}. \tag{3.2.41}$$

Therefore, it follows that $\boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}} := \left\{\boldsymbol{\nu}_j^2\right\}_{j\in[N]}$ is a $\delta_{\boldsymbol{\pi}}$-covering of $\boldsymbol{\Pi}_{K-1}$, since for any $\boldsymbol{\pi} = (\pi_k)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$, (3.2.41) leads to the existence of $j_0 \in [N]$, such that

$$\left\|\boldsymbol{\pi} - \boldsymbol{\nu}_{j_0}^2\right\|_\infty = \max_{k\in[K]} \left|\boldsymbol{\xi}_k^2 - \boldsymbol{\nu}_{j_0,k}^2\right| = \max_{k\in[K]} \left\{\left|\boldsymbol{\xi}_k - \boldsymbol{\nu}_{j_0,k}\right| \left|\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}\right|\right\} \leq \frac{\delta_{\boldsymbol{\pi}}}{2\sqrt{K}} \max_{k\in[K]} \left|\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}\right| \leq \frac{\delta_{\boldsymbol{\pi}}}{\sqrt{K}} \leq \delta_{\boldsymbol{\pi}},$$

where we used the fact that $\max_{k\in[K]} |\boldsymbol{\xi}_k + \boldsymbol{\nu}_{j_0,k}| \leq 2$. Now, it remains to count the number of cubes N. Let $\mathcal{T}_a = \{\boldsymbol{z} \in Q^+ : \|\boldsymbol{z}\|_2 \leq a\}$ and let $\mathcal{C} = \bigcup_{j\in[N]} \mathbf{C}_j$. Note that $\mathcal{C} \subset \mathcal{T}_{1+\delta_{\boldsymbol{\pi}}} - \mathcal{T}_{1-\delta_{\boldsymbol{\pi}}} \equiv \mathcal{T}$, and so

$$\text{Volume}(\mathcal{T}) \geq \text{Volume}(\mathcal{C}) = N\left(\frac{\delta_{\boldsymbol{\pi}}}{\sqrt{K}}\right)^K.$$

Note that here we use the notation π for the Archimedes' constant, which differs from $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$ for the mixing proportion of the GLoME model. Then, we define $\mathcal{V}_K(a) = a^K \pi^{K/2}$ as the volume of a sphere of radius a. Since $z! \geq z^z e^{-z}$ and $(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K = K\int_{1-\delta_{\boldsymbol{\pi}}}^{1+\delta_{\boldsymbol{\pi}}} z^{K-1}dz \leq 2\delta_{\boldsymbol{\pi}} K (1+\delta_{\boldsymbol{\pi}})^{K-1}$, it follows that

$$\mathcal{N}(\delta_{\boldsymbol{\pi}}, \boldsymbol{\Pi}_{K-1}, \|\cdot\|_\infty) \leq N \leq \frac{\text{Volume}(\mathcal{C})}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K} = \frac{1}{2^K}\frac{\mathcal{V}_K(1+\delta_{\boldsymbol{\pi}}) - \mathcal{V}_K(1-\delta_{\boldsymbol{\pi}})}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K}$$
$$= \frac{1}{2^K}\frac{\left[(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K\right]}{\left(\delta_{\boldsymbol{\pi}}/\sqrt{K}\right)^K}\frac{\pi^{K/2}}{(K/2)!} \leq \left(\frac{\pi e}{2}\right)^{K/2}\frac{\left[(1+\delta_{\boldsymbol{\pi}})^K - (1-\delta_{\boldsymbol{\pi}})^K\right]}{\delta_{\boldsymbol{\pi}}^K}$$
$$\leq \frac{K(2\pi e)^{K/2}}{\delta_{\boldsymbol{\pi}}^{K-1}}.$$

3.2.5.4 Proof of Lemma 3.2.17

In order to find an upper bound for a covering number of $\mathbf{W}_K$, we wish to construct a finite δ-covering $\mathbf{W}_{K,\boldsymbol{\omega}}$ of $\mathbf{W}_K$, with respect to the distance $\mathrm{d}_{\|\sup\|_\infty}$. That is, given any $\delta > 0, \mathbf{w}(\cdot;\boldsymbol{\omega}) \in \mathbf{W}_K$, we aim to prove that there exists $\mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}}) \in \mathbf{W}_{K,\boldsymbol{\omega}}$ such that

$$\mathrm{d}_{\|\sup\|_\infty}(\mathbf{w}(\cdot;\boldsymbol{\omega}), \mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}})) = \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} |\mathbf{w}_k(\mathbf{y};\boldsymbol{\omega}) - \mathbf{w}_k(\mathbf{y};\widehat{\boldsymbol{\omega}})| \leq \delta. \tag{3.2.42}$$

In order to accomplish such task, given any positive constants $\delta_{\boldsymbol{c}}, \delta_{\boldsymbol{\Gamma}}, \delta_{\boldsymbol{\pi}}$, and any $k \in [K]$, let us define

$$\begin{aligned}
\mathcal{F} &= \{\mathcal{Y} \ni \mathbf{y} \mapsto \ln(\phi_L(\mathbf{y};\boldsymbol{c},\boldsymbol{\Gamma})) : \|\boldsymbol{c}\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}) \leq M(\boldsymbol{\Gamma}) \leq A_{\boldsymbol{\Gamma}}\},\\
\mathcal{F}_{\mathbf{c}_k} &= \Big\{\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) : \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F},\\
&\qquad \mathbf{c}_{k,j} \in \{-C_{\mathcal{Y}} + l\delta_{\boldsymbol{c}}/L : l = 0,\ldots,\lceil 2C_{\mathcal{Y}}L/\delta_{\boldsymbol{c}}\rceil\}, j \in [L]\Big\},
\end{aligned} \tag{3.2.43}$$

$$\begin{aligned}
\mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k} = \Bigg\{ &\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) : \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k},\\
&[\mathrm{vec}(\boldsymbol{\Gamma}_k)]_{i,j} = \gamma_{i,j}\frac{\delta_{\boldsymbol{\Gamma}}}{L^2}; \gamma_{i,j} = \gamma_{j,i} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{L^2A_{\boldsymbol{\Gamma}}}{\delta_{\boldsymbol{\Gamma}}}\right\rfloor, \left\lfloor\frac{L^2A_{\boldsymbol{\Gamma}}}{\delta_{\boldsymbol{\Gamma}}}\right\rfloor\right], i \in [L], j \in [L]\Bigg\},
\end{aligned} \tag{3.2.44}$$

$$\mathbf{W}_{K,\boldsymbol{\omega}} = \{\mathbf{w}(\cdot;\boldsymbol{\omega}) : \mathbf{w}(\cdot;\boldsymbol{\omega}) \in \mathbf{W}_K, \forall k \in [K], \ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}, \boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}\}. \tag{3.2.45}$$

Here, $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are ceiling and floor functions, respectively, and $\mathrm{vec}(\cdot)$ is an operator that stacks matrix columns into a column vector. In particular, we denote $\boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$ as a $\delta_{\boldsymbol{\pi}}$-covering of $\boldsymbol{\Pi}_{K-1}$, which is defined in Lemma 3.2.16. By the previous definition, it holds that $\forall k \in [K]$, $\mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k} \subset \mathcal{F}_{\mathbf{c}_k} \subset \mathcal{F}$, and $\mathbf{W}_{K,\boldsymbol{\omega}} \subset \mathbf{W}_K$.

Next, we claim that $\mathbf{W}_{K,\boldsymbol{\omega}}$ is a finite δ-covering of $\mathbf{W}_K$ with respect to the distance $\mathrm{d}_{\|\sup\|_\infty}$. To do this, for any $\mathbf{w}(\cdot;\boldsymbol{\omega}) = (\ln(\pi_k\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)))_{k\in[K]} \in \mathbf{W}_K$, $\ln(\phi_L(\cdot;\mathbf{c}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}, \boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$, and for any $k \in [K]$, by (3.2.43), we first choose a function $\ln(\phi_L(\cdot;\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k}$ so that

$$\|\widehat{\boldsymbol{c}}_k - \mathbf{c}_k\|_1 = \sum_{j=1}^{L} |\widehat{\boldsymbol{c}}_{k,j} - \mathbf{c}_{k,j}| \leq L\frac{\delta_{\boldsymbol{c}}}{L} = \delta_{\boldsymbol{c}}.$$

Furthermore, by (3.2.44), we can obtain a result to construct the covariance matrix lattice. That is, any $\ln(\phi_L(\cdot;\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) \in \mathcal{F}_{\mathbf{c}_k}$ can be approximated by $\ln\left(\phi_L\left(\cdot;\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) \in \mathcal{F}_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}$ such that

$$\left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) - \mathrm{vec}(\boldsymbol{\Gamma}_k)\right\|_1 \equiv \left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right\|_1 = \sum_{i=1}^{L}\sum_{j=1}^{L}\left|\left[\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right]_{i,j}\right| \leq \frac{L^2\delta_{\boldsymbol{\Gamma}}}{L^2} = \delta_{\boldsymbol{\Gamma}}.$$

Note that since for any $k \in [K]$, $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))$ is differentiable, it is also continuous w.r.t. $\mathbf{y}$ and its parameters $\mathbf{c}_k$ and $\boldsymbol{\Gamma}_k$. Thus, for every fixed $\mathbf{y} \in \mathcal{Y}$, for every $\widehat{\boldsymbol{c}}_k, \mathbf{c}_k \in \mathcal{X}$ with $\widehat{\boldsymbol{c}}_k \leq \mathbf{c}_k$, and for every $\widehat{\boldsymbol{\Gamma}}_k$, $\boldsymbol{\Gamma}_k$, where $\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) \leq \mathrm{vec}(\boldsymbol{\Gamma}_k)$, we can apply the mean value theorem (see Duistermaat & Kolk 2004, Lemma 2.5.1) to $\ln(\phi_L(\mathbf{y};\cdot,\boldsymbol{\Gamma}_k))$ and $\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\cdot))$ on the intervals $[\widehat{\boldsymbol{c}}_k,\mathbf{c}_k]$ and $\left[\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right), \mathrm{vec}(\boldsymbol{\Gamma}_k)\right]$ for some $z_{\mathbf{c}_k} \in (\widehat{\boldsymbol{c}}_k,\mathbf{c}_k)$ and $z_{\boldsymbol{\Gamma}_k} \in \left(\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right), \mathrm{vec}(\boldsymbol{\Gamma}_k)\right)$, respectively, to get

$$\begin{aligned}
\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)) &= (\widehat{\boldsymbol{c}}_k - \mathbf{c}_k)^\top \nabla_{\mathbf{c}_k}\ln(\phi_L(\mathbf{y};z_{\mathbf{c}_k},\boldsymbol{\Gamma}_k)),\\
\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) &= \left(\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k\right) - \mathrm{vec}(\boldsymbol{\Gamma}_k)\right)^\top \nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)}\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,z_{\boldsymbol{\Gamma}_k})).
\end{aligned}$$

Moreover, $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \nabla_{\mathbf{c}_k}\ln(\phi_L(\mathbf{y};z_{\mathbf{c}_k},\boldsymbol{\Gamma}_k))$ and $(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k)) \mapsto \nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)}\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,z_{\boldsymbol{\Gamma}_k}))$ are continuous functions on the compact set $\mathcal{U} := \mathcal{Y}\times\mathcal{Y}\times[a_{\boldsymbol{\Gamma}},A_{\boldsymbol{\Gamma}}]^{L^2}$ leads to they attain minimum

and maximum values (see Duistermaat & Kolk 2004, Theorem 1.8.8). That is, we can set

$$\mathbf{0} < (C_{\boldsymbol{c}})_{1,\ldots,L}^{\top} := \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k))\in\mathcal{U}} |\nabla_{\mathbf{c}_k} \ln|\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)|| < \infty,$$
$$\mathbf{0} < (C_{\boldsymbol{\Gamma}})_{1,\ldots,L^2}^{\top} := \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\boldsymbol{\Gamma}_k))\in\mathcal{U}} \left|\nabla_{\mathrm{vec}(\boldsymbol{\Gamma}_k)} \ln|\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)(x)|\right| < \infty.$$

Therefore, by the Cauchy–Schwarz inequality, we have

$$\sup_{\mathbf{y}\in\mathcal{Y}} |\ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k)) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))| \leq |\widehat{\boldsymbol{c}}_k - \mathbf{c}_k|^{\top} (C_{\boldsymbol{c}})_{1,\ldots,L}^{\top} = C_{\boldsymbol{c}} \|\widehat{\boldsymbol{c}}_k - \mathbf{c}_k\|_1 \leq C_{\boldsymbol{c}}\delta_{\boldsymbol{\pi}},$$
$$\sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\widehat{\boldsymbol{c}}_k,\boldsymbol{\Gamma}_k))\right| \leq C_{\boldsymbol{\Gamma}} \left\|\mathrm{vec}\left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right)\right\|_1 \leq C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}},$$

and by using the triangle inequality, it follows that

$$\max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right) - \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k))\right| \leq C_{\boldsymbol{c}}\delta_{\boldsymbol{c}} + C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}}. \tag{3.2.46}$$

Moreover, for every $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}$, Lemma 3.2.16 implies that we can choose $\widehat{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_{K-1,\boldsymbol{\omega}}$ so that $\max_{k\in[K]} |\pi_k - \widehat{\boldsymbol{\pi}}_k| \leq \delta_{\boldsymbol{\pi}}$. Notice that $[a_{\boldsymbol{\pi}},\infty) \ni t \mapsto \ln(t), a_{\boldsymbol{\pi}} > 0$ is a Lipschitz continuous function on $[a_{\boldsymbol{\pi}},\infty]$. Indeed, by the mean value theorem, it holds that there exists $c \in (t_1,t_2)$, such that

$$|\ln(t_1) - \ln(t_2)| = \ln'(c)\,|t_1 - t_2| \leq \frac{1}{a_{\boldsymbol{\pi}}}\,|t_1 - t_2|, \text{ for all } t_1,t_2 \in [a_{\boldsymbol{\pi}},\infty). \tag{3.2.47}$$

Therefore, (3.2.42) can be obtained by the following evaluation

$$\begin{aligned}
\mathrm{d}_{\|\sup\|_\infty}(\mathbf{w}(\cdot;\boldsymbol{\omega}),\mathbf{w}(\cdot;\widehat{\boldsymbol{\omega}})) &= \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln(\pi_k\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)) - \ln\left(\widehat{\boldsymbol{\pi}}_k\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&= \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln(\pi_k) - \ln(\widehat{\boldsymbol{\pi}}_k) + \ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)) - \ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&\leq \max_{k\in[K]} |\ln(\pi_k) - \ln(\widehat{\boldsymbol{\pi}}_k)| + \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left|\ln(\phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)) - \ln\left(\phi_L\left(\mathbf{y};\widehat{\boldsymbol{c}}_k,\widehat{\boldsymbol{\Gamma}}_k\right)\right)\right| \\
&\leq \frac{1}{a_{\boldsymbol{\pi}}} \max_{k\in[K]} |\pi_k - \widehat{\boldsymbol{\pi}}_k| + C_{\boldsymbol{c}}\delta_{\boldsymbol{c}} + C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}} \text{ (using (3.2.47) and (3.2.46))} \\
&\leq \frac{\delta_{\boldsymbol{\pi}}}{a_{\boldsymbol{\pi}}} + C_{\boldsymbol{c}}\delta_{\boldsymbol{c}} + C_{\boldsymbol{\Gamma}}\delta_{\boldsymbol{\Gamma}} \text{ (using Lemma 3.2.16)} \leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta,
\end{aligned}$$

where we choose $\delta_{\boldsymbol{\pi}} = \frac{\delta a_{\boldsymbol{\pi}}}{3}, \delta_{\boldsymbol{c}} = \frac{\delta}{3C_\mu}, \delta_{\boldsymbol{\Gamma}} = \frac{\delta}{3C_{\boldsymbol{\Gamma}}}$. Finally, we get the covering number

$$\begin{aligned}
\mathcal{N}\left(\delta,\mathbf{W}_K,\mathrm{d}_{\|\sup\|_\infty}\right) &\leq \mathrm{card}(\mathbf{W}_{K,\boldsymbol{\omega}}) = \left[\frac{2C_{\mathcal{Y}}L}{\delta_{\boldsymbol{c}}}\right]^{KL} \left[\frac{2A_{\boldsymbol{\Gamma}}L^2}{\delta_{\boldsymbol{\Gamma}}}\right]^{\frac{L(L+1)}{2}K} \mathcal{N}(\delta_{\boldsymbol{\pi}},\boldsymbol{\Pi}_{K-1},\|\cdot\|_\infty) \\
&= \frac{C}{\delta^{KL+K\frac{L(L+1)}{2}+K-1}} = \frac{C}{\delta^{\dim(\mathbf{W}_K)}},
\end{aligned}$$

where

$$C = (6C_{\boldsymbol{c}}C_{\mathcal{Y}}L)^{KL} \left(6C_{\boldsymbol{\Gamma}}A_{\boldsymbol{\Gamma}}L^2\right)^{\frac{L(L+1)}{2}K} \left(\frac{3}{a_{\boldsymbol{\pi}}}\right)^{K-1} K(2\pi e)^{K/2}.$$

3.2.5.5 Proof of Lemma 3.2.11

Note that Lemmas 1 and 2 from Montuelle et al. (2014) imply that there exists a constant $C_{\boldsymbol{\Upsilon}_{K,d}} = \ln\left(\sqrt{2} + \sqrt{D}dT_{\boldsymbol{\Upsilon}}\right)$ (when $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{b,d}^K$) or $C_{\boldsymbol{\Upsilon}_{K,d}} = \ln\left(\sqrt{2} + \sqrt{D}\binom{d+L}{L}T_{\boldsymbol{\Upsilon}}\right)$ (when $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$) such that, $\forall\delta \in (0,\sqrt{2})$,

$$\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta,\boldsymbol{\Upsilon}_{K,d}) \leq \dim(\boldsymbol{\Upsilon}_{K,d})\left(C_{\boldsymbol{\Upsilon}_{K,d}} + \ln\frac{1}{\delta}\right). \tag{3.2.48}$$

Next, we rely on Proposition 3.2.18 for constructing of Gaussian brackets to for the Gaussian experts.

Proposition 3.2.18 (Proposition 2 from Montuelle et al. 2014). *Let* $\kappa \geq \frac{17}{29}$ *and* $\gamma_\kappa = \frac{25(\kappa-\frac{1}{2})}{49\left(1+\frac{2\kappa}{5}\right)}$. *For any* $0 < \delta \leq \sqrt{2}$, *any* $D \geq 1$, *and any* $\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac{1}{2}}}\frac{\delta}{D}$, *let* $(\boldsymbol{\upsilon}_d, B, \mathbf{A}, \mathbf{P}) \in \boldsymbol{\Upsilon}_{K,d} \times [B_-, B_+] \times \mathcal{A}(\lambda_-, \lambda_+) \times SO(D)$ *and* $\left(\widetilde{\boldsymbol{\upsilon}}_d, \widetilde{B}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{P}}\right) \in \boldsymbol{\Upsilon}_{K,d} \times [B_-, B_+] \times \mathcal{A}(\lambda_-, +\infty) \times SO(D)$, *and define* $\boldsymbol{\Sigma} = B\mathbf{P}\mathbf{A}\mathbf{P}^\top$, $\widetilde{\boldsymbol{\Sigma}} = \widetilde{B}\widetilde{\mathbf{P}}\widetilde{\mathbf{A}}\widetilde{\mathbf{P}}^\top$,

$$t^-(\mathbf{x},\mathbf{y}) = (1+\kappa\delta_{\boldsymbol{\Sigma}})^{-D}\,\phi_D\left(\mathbf{x};\widetilde{\boldsymbol{\upsilon}}_d(\mathbf{y}),(1+\delta_{\boldsymbol{\Sigma}})^{-1}\widetilde{\boldsymbol{\Sigma}}\right) \text{ and } t^+(\mathbf{x},\mathbf{y}) = (1+\kappa\delta_{\boldsymbol{\Sigma}})^{D}\,\phi_D\left(\mathbf{x};\widetilde{\boldsymbol{\upsilon}}_d(\mathbf{y}),(1+\delta_{\boldsymbol{\Sigma}})\widetilde{\boldsymbol{\Sigma}}\right).$$

If

$$\begin{cases} \forall \mathbf{y}\in\mathcal{Y}, \left\|\boldsymbol{\upsilon}_d(\mathbf{y})-\widetilde{\boldsymbol{\upsilon}}_d(\mathbf{y})\right\|^2 \leq D\gamma_\kappa\mathcal{L}_-\lambda_-\frac{\lambda_-}{\lambda_+}\delta_{\boldsymbol{\Sigma}}^2 \\ (1+\frac{2}{25}\delta_{\boldsymbol{\Sigma}})^{-1}\widetilde{B} \leq B \leq \widetilde{B} \\ \forall i\in[D], \left|\mathbf{A}_{i,i}^{-1}-\widetilde{\mathbf{A}}_{i,i}^{-1}\right| \leq \frac{1}{10}\frac{\delta_{\boldsymbol{\Sigma}}}{\lambda_+} \\ \forall \mathbf{x}\in\mathbb{R}^D, \left\|\mathbf{P}\mathbf{x}-\widetilde{\mathbf{P}}\mathbf{x}\right\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_{\boldsymbol{\Sigma}}\left\|\mathbf{x}\right\| \end{cases}$$

then $[t^-, t^+]$ *is a* $\frac{\delta}{5}$ *Hellinger bracket such that* $t^-(\mathbf{x},\mathbf{y}) \leq \phi_D(\mathbf{x};\boldsymbol{\upsilon}_d(\mathbf{y}),\boldsymbol{\Sigma}) \leq t^+(\mathbf{x},\mathbf{y})$.

Then, following the same argument as in Montuelle et al. (2014, Appendix B.2.3), Proposition 3.2.18 allows us to construct nets over the spaces of the means, the volumes, the eigenvector matrices, the normalized eigenvalue matrices and then control the bracketing entropy of $\mathcal{G}_{K,d}$. More precisely, three different contexts are considered for the mean, volume and matrix parameters. They can be all known ($\star = 0$), unknown but common to all classes ($\star = c$), unknown and possibly different for every class ($\star = K$). For example, $[\boldsymbol{\upsilon}_K, B_0, \mathbf{P}_0, \mathbf{A}_0]$ denotes a model in which only mean vectors are assumed to be free. Then, we obtain

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}}\max_k d_{\mathbf{x}}}\left(\frac{\delta}{5},\mathcal{G}_{K,d}\right) \leq \dim\left(\mathcal{G}_{K,d}\right)\left(C_{\mathcal{G}_{K,d}}+\ln\left(\frac{1}{\delta}\right)\right), \tag{3.2.49}$$

where $\dim\left(\mathcal{G}_{K,d}\right) = Z_{\boldsymbol{\upsilon},\star} + Z_{B,\star} + \frac{D(D-1)}{2}Z_{\mathbf{P},\star} + (D-1)Z_{\mathbf{A},\star}$. Here, $Z_{\boldsymbol{\upsilon},K} = \dim\left(\boldsymbol{\Upsilon}_{K,d}\right), Z_{\boldsymbol{\upsilon},c} = \dim\left(\boldsymbol{\Upsilon}_1\right), Z_{\boldsymbol{\upsilon},0} = 0$, $Z_{B,0} = Z_{\mathbf{P},0} = Z_{\mathbf{A},0} = 0$, $Z_{B,c} = Z_{\mathbf{P},c} = Z_{\mathbf{A},c} = 1$, $Z_{B,K} = Z_{\mathbf{P},K} = Z_{\mathbf{A},K} = K$, and given a universal constant c_U,

$$\begin{aligned} C_{\mathcal{G}_{K,d}} = {} & \ln\left(5D\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac{1}{2}}\right) + C_{\boldsymbol{\Upsilon}_{K,d}} + \frac{1}{2}\ln\left(\frac{\lambda_+}{D\gamma_k B_-\lambda_-^2}\right) \\ & + \ln\left(\frac{4+129\ln\left(\frac{B_+}{B_-}\right)}{10}\right) + \frac{D(D-1)}{2}\ln\left(c_U\right) + \ln\left(\frac{10\lambda_+}{\lambda_-}\right) + \ln\left(\frac{4}{5}+\frac{52\lambda_+}{5\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right). \end{aligned}$$

3.3 A non-asymptotic model selection in the block-diagonal localized mixture of experts regression model

It is interesting to point out that block-diagonal covariance for Gaussian locally-linear mapping (BLLiM) model in Devijver et al. (2017) is an affine instance of a BLoME model, where linear combination of bounded functions (*e.g.,* polynomials) are considered instead of affine mean functions for the Gaussian experts. The BLLiM framework aims to model a sample of high-dimensional regression data issued from a heterogeneous population with hidden graph-structured interaction between covariates. In particular, the BLLiM model is considered as a good candidate for performing a model-based clustering and predicting the response in situations affected by the curse of dimensionality phenomenon, where the number of parameters could be larger than the sample size. Indeed, to deal with high-dimensional regression problems, the BLLiM model, initially proposed by Li (1991), is based on an inverse regression strategy, which inverts the role of the high-dimensional predictor and the multivariate response. Therefore, the number of parameters to estimate is drastically reduced. More precisely,

BLLiM utilizes the Gaussian locally-linear mapping (GLLiM), described in Deleforge et al. (2015a,c), and Perthame et al. (2018), in conjunction with a block-diagonal structure hypothesis on the residual covariance matrices to make a trade-off between complexity and sparsity.

This prediction model is fully parametric and highly interpretable. For instance, it might be useful for the analysis of transcriptomic data in molecular biology to classify observations or predict phenotypic states, as for example disease versus non disease or tumor versus normal (Golub et al., 1999, Nguyen & Rocke, 2002, Lê Cao et al., 2008). Indeed, if predictor variables are gene expression data measured by microarrays or by the RNA-seq technologies and the response is a phenotypic variables, situations affected by the BLLiM not only provides clusters of individuals based on the relation between gene expression data and the phenotype but also implies a gene regulatory network specific for each cluster of individuals (see Devijver et al. (2017) for more details).

It is worth noting that two hyperparameters must be estimated to construct a BLLiM model: the number of mixtures components (or clusters) and the block structure of large covariance matrices specific of each cluster (the size and the number of blocks). Data driven choices of hyperparameters of learning algorithms belong to the model selection class of problems, which has attracted much attention in statistics and machine learning over the last 50 years (Akaike, 1974, Mallows, 1973, Anderson & Burnham, 2002, Massart, 2007). This is a particular instance of the estimator (or model) selection problem: given a family of estimators, how do we choose, using data, one among them whose risk is as small as possible? Note that penalization is one of the main strategies proposed for model selection. It suggests to choose the estimator minimizing the sum of its empirical risk and some penalty terms corresponding to how well the model fits the data, while avoiding overfitting.

In general, model selection can be performed using the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Akaike, 1974, Schwarz et al., 1978). Nevertheless, these approaches are asymptotic, which implies that there are no finite sample guarantees for choosing between different levels of complexity. Their use in small sample settings is thus ad hoc. To overcome such difficulties, Birgé & Massart (2007) proposed a novel approach, called slope heuristics, supported by a non-asymptotic oracle inequality. This method leads to an optimal data-driven choice of multiplicative constants for penalties. Practical issues and recent surveys for the slope heuristic can be found in Baudry et al. (2012), and Arlot (2019).

It should be stressed that a general model selection result, originally established by Massart (2007, Theorem 7.11), guarantees a penalized criterion leads to a good model selection and the penalty being only known up to multiplicative constants and proportional to the dimensions of models. In particular, such multiplicative constants can be calibrated by the slope heuristic approach in a finite sample setting. Then, in the spirit of the concentration inequality-based methods developed in Massart (2007), Massart & Meynet (2011), and Cohen & Le Pennec (2011), a huge number of finite-sample oracle results have been proposed in several statistical frameworks including high dimensional Gaussian graphical models (Devijver & Gallopin, 2018), Gaussian mixture model selection (Maugis & Michel, 2011b,a), finite mixture regression models (Meynet, 2013, Devijver, 2015a,b, 2017b,a), softmax-gated mixture of experts (SGaME) (Montuelle et al., 2014, Nguyen et al., 2020c), and Gaussian-gated localized MoE (GLoME) models (Nguyen et al., 2021c). However, to the best of our knowledge, we are the first to provide a finite-sample oracle inequality: Theorem 3.3.2, for the BLoME regression model. In particular, our proof strategy makes use of recent novel approaches comprising a model selection theorem for maximum likelihood estimator (MLE) among a random subcollection (Devijver, 2015b), a non-asymptotic model selection result for detecting a good block-diagonal structure in high-dimensional graphical models (Devijver & Gallopin, 2018) and a reparameterization trick to bound the metric entropy of the Gaussian gating parameter space in GLoME models (Nguyen et al., 2021c), see also Section 3.2 for more details.

The main contribution of Section 3.3 is an important theoretical result: a finite-sample oracle inequality that provides a non-asymptotic bound on the risk, and a lower bound on the penalty function that ensures such non asymptotic theoretical control on the estimator under the Kullback–Leibler loss. It also provides a theoretical justification for the penalty shape when using the slope heuristic for the BLoME as well as BLLiM models.

The rest of Section 3.3 is organized as follows. In Section 3.3.1, we discuss the model construction

and framework for BLoME and BLLiM models. Then, we present the main results of Section 3.3, an oracle inequality satisfied by the penalized maximum likelihood of BLoME, in Section 3.3.2. Section 3.3.3 is devoted to the proof of these main results based on a general model selection theorem. Proofs of lemmas are provided in Section 3.3.4.

3.3.1 Notation and framework

3.3.1.1 BLoME models

In order to accommodate a potential hidden graph-structured interaction and make a trade-off between complexity and sparsity, we consider an extension of the GLoME model from Nguyen et al. (2021c), which generalized the SGaME and GLLiM models (Jacobs et al., 1991, Xu et al., 1995, Deleforge et al., 2015c). More specifically, we consider the following BLoME model, defined by (3.3.1), which is motivated by an inverse regression framework, where the role of response variables and high-dimensional predictors are exchanged such that the response $\mathbf{Y}$ becomes the covariate and the predictor $\mathbf{X}$ plays the role of a multivariate response.

Then the BLoME model is defined by the following conditional density:

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y};\boldsymbol{\omega}\right) \phi_D\left(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right), \tag{3.3.1}$$

with

$$g_k\left(\mathbf{y};\boldsymbol{\omega}\right) = \frac{\pi_k \phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K} \pi_j \phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}. \tag{3.3.2}$$

Here, $g_k(\cdot;\boldsymbol{\omega})$ and $\phi_D\left(\cdot;\boldsymbol{\upsilon}_{k,d}(\cdot),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right), k \in [K]$, $K \in \mathbb{N}^\star$, $d \in \mathbb{N}^\star$, are called Gaussian gating functions and Gaussian experts, respectively. Furthermore, we decompose the parameters of the model as follows: $\boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}\left(\mathbf{B}\right)) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\left(\mathbf{B}\right) =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times \mathbf{C}_K \times \boldsymbol{V}_K') =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$, $\boldsymbol{c} = (\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d = (\boldsymbol{\upsilon}_{k,d})_{k\in[K]} \in \boldsymbol{\Upsilon}_{K,d}$, and $\boldsymbol{\Sigma}\left(\mathbf{B}\right) = \left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)_{k\in[K]} \in \mathbf{V}_K\left(\mathbf{B}\right)$. Note that $\boldsymbol{\Pi}_{K-1} = \left\{(\pi_k)_{k\in[K]} \in \left(\mathbb{R}^+\right)^K, \sum_{k=1}^K \pi_k = 1\right\}$ is a $K-1$ dimensional probability simplex, $\mathbf{C}_K$ is a set of K-tuples of mean vectors of size $L \times 1$, $\boldsymbol{V}_K'$ is a sets of K-tuples of elements in $\mathcal{S}_L^{++}$, where $\mathcal{S}_L^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ is a set of K-tuples of mean functions from $\mathbb{R}^L$ to $\mathbb{R}^D$ depending on a degree d (*e.g.,* a degree of polynomials), and $\mathbf{V}_K\left(\mathbf{B}\right)$ is a set containing K-tuples from $\mathcal{S}_D^{++}$ with the following block-diagonal structures defined in (3.3.3) (Devijver et al., 2017, Devijver & Gallopin, 2018).

More precisely, for $k \in [K]$, we decompose $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)$ into G_k blocks, $G_k \in \mathbb{N}^\star$, and we denote by $d_k^{[g]}$ the set of variables into the gth group, for $g \in [G_k]$, and by $\operatorname{card}\left(d_k^{[g]}\right)$ the number of variables in the corresponding set. Then, we denote by $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$ a block structure for the cluster k, and $\mathbf{B} = \left(\mathbf{B}_k\right)_{k\in[K]}$ the covariate indexes into each group for each cluster. Hence, up to a permutation, we can construct the following block-diagonal covariance matrices: $\mathbf{V}_K\left(\mathbf{B}\right) = \left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right)_{k\in[K]}$, for every $k \in [K]$,

$$\mathbf{V}_k\left(\mathbf{B}_k\right) = \left\{ \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathcal{S}_D^{++} \;\middle|\; \begin{array}{l} \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) = \mathbf{P}_k \begin{pmatrix} \boldsymbol{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \boldsymbol{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\operatorname{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}, \tag{3.3.3}$$

where $\mathbf{P}_k$ corresponds to the permutation leading to a block-diagonal matrix in cluster k. It is worth mentioning that outside the blocks, all coefficients of the matrix are zeros and we also authorize reordering of the blocks: *e.g.,* $\{(1,3);(2,4)\}$ is identical to $\{(2,4);(1,3)\}$, and the permutation inside blocks: *e.g.,* the partition of 4 variables into blocks $\{(1,3);(2,4)\}$ is the same as the partition $\{(3,1);(4,2)\}$.

Remark 3.3.1. The block-diagonal structures for covariance matrices $(\boldsymbol{\Sigma}_k(\mathbf{B}_k))_{k\in[K]}$, defined in (3.3.3), are not only used for a trade-off between complexity and sparsity but also motivated by some real applications, where we want to perform prediction on data sets with heterogeneous observations and hidden graph-structured interactions between covariates. For instance, for gene expression data set in which conditionally on the phenotypic response, genes interact with few other genes only, *i.e.,* there are small modules of correlated genes (see Devijver et al., 2017, Devijver & Gallopin, 2018 for more details).

In order to establish our oracle inequality, Theorem 3.3.2, we need to assume that $\mathcal{Y}$ is a bounded set in $\mathbb{R}^L$ and make explicit some classical boundedness conditions on the parameter space, see Sections 3.2.1.2 and 3.2.1.3 for more details.

In particular, for the block-diagonal covariances of Gaussian experts, we assume that there exist some positive constants λ_m and λ_M such that, for every $k \in [K]$,

$$0 < \lambda_m \leq m\left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)\right) \leq M\left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)\right) \leq \lambda_M. \tag{3.3.4}$$

Note that this is a quite general assumption and is also used in the block-diagonal covariance selection for Gaussian graphical models of Devijver & Gallopin (2018).

Next, a characterization of BLLiM model, an affine instance of BLoME model, is described in Section 3.3.1.2 and is especially useful for high-dimensional regression data. Note that the BLLiM model relies on an inverse regression trick from a GLLiM model (Deleforge et al., 2015c) and the block-diagonal structure hypothesis on the residual covariance matrices (Devijver & Gallopin, 2018).

3.3.1.2 High-dimensional regression via BLLiM models

A BLLiM model, as originally introduced in Devijver et al. (2017), is used to capture the nonlinear relationship between the response and the set of covariates, imposed by a potential hidden graph-structured interaction, from a high-dimensional regression data, typically in the case when $D \gg L$, by the following K locally affine mappings:

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbb{I}(Z=k)\left(\mathbf{A}_k^*\mathbf{X} + \mathbf{b}_k^* + \mathbf{E}_k^*\right). \tag{3.3.5}$$

Note that BLLiM model follows the same framework as GLLiM model, defined in Section 3.2.1.4, except for the fact that BLLiM model imposes the block-diagonal structures on $(\boldsymbol{\Sigma}_k)_{k\in[K]}$, established in (3.3.3), to make a trade-off between complexity and sparsity.

3.3.1.3 Collection of BLoME models

In this paper, we choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, where $d_{\max} \in \mathbb{N}^\star$ and $K_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. Moreover, $\mathbf{B}$ is selected among a list of candidate structures $(\mathcal{B}_k)_{k\in[K]} \equiv (\mathcal{B})_{k\in[K]}$, where $\mathcal{B}$ denotes the set of all possible partitions of the covariables indexed by $[D]$, for each cluster of individuals. We wish to estimate the unknown conditional density s_0 by conditional densities belonging to the following collection of models: $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \left\{(K, d, \mathbf{B}) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}, \mathbf{B} \in (\mathcal{B})_{k\in[K]}\right\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}(\mathbf{B})) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K(\mathbf{B}) =: \widetilde{\boldsymbol{\Psi}}_{K,d}(\mathbf{B})\right\}. \tag{3.3.6}$$

In theory, we would like to consider the whole collection of model $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$. However, the cardinality of $\mathcal{B}$ is large; its size is a Bell number. Even for a moderate number of variables D, it is not possible to explore the set $\mathcal{B}$, exhaustively. We restrict our attention to a random subcollection $\mathcal{B}^R$ of moderate size. For example, we can consider the BLLiM procedure from Devijver et al. (2017, Section 2.2).

In Section 3.3.2, we state our main contribution: a finite-sample oracle type inequality, which ensures that if we have penalized the log-likelihood in an approximate approach, we are able to select a model, which is as good as the oracle.

3.3.2 Main result on oracle inequality

Note that in Section 3.3, the constructed collection of models with block-diagonal structures for each cluster of individuals is designed, for example, by the BLLiM procedure from Devijver et al. (2017), where each collection of partition is sorted by sparsity level. Nevertheless, our finite-sample oracle inequality still holds for any random subcollection of $\mathcal{M}$, which is constructed by some suitable tools in the framework of BLoME regression models.

Theorem 3.3.2 (Oracle inequality)**.** *Let $(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}$ be the observations coming from an unknown conditional density s_0. For each $\mathbf{m} = (K, d, \mathbf{B}) \in (\mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} \times \mathcal{B}) \equiv \mathcal{M}$, let $S_{\mathbf{m}}$ be define by (3.3.6). Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$, such that*

$$\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \bar{s}_{\mathbf{m}}) \leq \inf_{t \in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + \frac{\epsilon_{KL}}{n}, \text{ and } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0.$$

Next, we construct some random subcollection $(S_{\mathbf{m}})_{\mathbf{m} \in \widetilde{\mathcal{M}}}$ of $(S_{\mathbf{m}})_{\mathbf{m} \in \mathcal{M}}$ by letting $\widetilde{\mathcal{M}} \equiv (\mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} \times \mathcal{B}^R) \subset \mathcal{M}$ such that $\mathcal{B}^R$ is a random subcollection $\mathcal{B}$, of moderate size, as described in Section 3.3.1.3. Consider the collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m} \in \widetilde{\mathcal{M}}}$ of η-log likelihood minimizers satisfying (3.2.17) for all $\mathbf{m} \in \widetilde{\mathcal{M}}$. Then, there is a constant C such that for any $\rho \in (0, 1)$, and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, for every index, $\mathbf{m} \in \mathcal{M}$, $\xi_{\mathbf{m}} \in \mathbb{R}^+$, $\Xi = \sum_{\mathbf{m} \in \mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$ and

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n) \dim(S_{\mathbf{m}}) + (1 \vee \tau) \xi_{\mathbf{m}} \right],$$

with $\kappa > \kappa_0$, the η'-penalized likelihood estimator $\widehat{s}_{\widehat{\mathbf{m}}}$, defined as in (3.2.18) on the subset $\widetilde{\mathcal{M}} \subset \mathcal{M}$, satisfies

$$\mathbb{E}_{\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}} \left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) \right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}} \left[\inf_{\mathbf{m} \in \widetilde{\mathcal{M}}} \left(\inf_{t \in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + 2 \frac{\mathrm{pen}(\mathbf{m})}{n} \right) \right] + C_2 (1 \vee \tau) \frac{\Xi^2}{n} + \frac{\eta' + \eta}{n}.$$

Remark 3.3.3. In Theorem 3.3.2, the finite-sample oracle inequality compares performances of our estimator with the best model in the collection. However, Theorem 3.3.2 allows us to approximate well a rich class of conditional densities if we take enough degree of polynomials of Gaussian expert means and/or enough clusters in the context of mixture of Gaussian experts (Jiang & Tanner, 1999a, Mendes & Jiang, 2012, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2021a). This leads to the term on the right hand side being small, for $\mathcal{D}_{\boldsymbol{\Upsilon}}$ and $\mathcal{K}$ well-chosen.

Furthermore, in the context of MoE regression models, our non-asymptotic oracle inequality, Theorem 3.3.2, can be considered as a complementary result to a classical asymptotic theory (Khalili, 2010, Theorems 1,2, and 3), to a finite-sample oracle inequality on the whole collection of models (Montuelle et al., 2014, Nguyen et al., 2021c) and to an l_1-oracle inequality focusing on the Lasso estimation properties rather than the model selection procedure (Nguyen et al., 2020c).

In particular, aside from important theoretical issues regarding the tightness of the bounds, the way to integrate a priori information and the minimax analysis of our proposed PMLE, we hope that our finite-sample oracle inequalities and corresponding interesting numerical experiments help to partially answer the two following important questions raised in the area of MoE regression models: (1) What number of mixture components K should be chosen, given the sample size n, and (2) Whether it is better to use a few complex experts or combine many simple experts, given the total number of parameters. Note that, such problems are considered in the work of Mendes & Jiang (2012, Proposition 1), where the authors provided some qualitative insights and only suggested a practical method for choosing K and d involving a complexity penalty or cross-validation. Furthermore, their model is only for a non-regularized maximum-likelihood estimation, and thus is not suitable in the high-dimensional setting.

3.3.3 Proof of the oracle inequality

Sketch of the proof To work with conditional density estimation in the BLoME regression models, in Section 3.3.3.1, we need to present a general theorem for model selection: Theorem 3.3.4. It is worth mentioning that, because the model collection constructed by the BLLiM procedure is random, we have to use a model selection theorem for MLE among a random subcollection (cf. Devijver, 2015b, Theorem 5.1 and Devijver & Gallopin, 2018, Theorem 7.3), which is an extension of a whole collection of conditional densities from Cohen & Le Pennec (2011, Theorem 2), and of Massart (2007, Theorem 7.11), working only for density estimation. Then, we explain how we use Theorem 3.3.4 to get the oracle inequality: Theorem 3.3.2 in Section 3.3.3.2. To this end, our model collection has to satisfy some regularity assumptions, which are proved in Section 3.3.4. The main difficulty in proving our oracle inequality lies in bounding the bracketing entropy of the Gaussian gating functions of the BLoME model and Gaussian experts with block-diagonal covariance matrices. To overcome the former issue, we follow a reparameterization trick of the Gaussian gating parameters space (Nguyen et al., 2021c). For the second one, we utilize the recent novel result on block-diagonal covariance matrices in Devijver & Gallopin (2018).

3.3.3.1 Model selection theorem for MLE among a random subcollection

Before stating the general theorem, we begin by discussing our assumptions. We work here in a more general context, with $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$, and $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ defining a model collection indexed by $\mathcal{M}$. Follow the same framework in Section 3.2.4.1, we aim to make use of Assumption 3.2.1 (K), Assumption 3.2.2 (Sep), and Assumption 3.2.3 (H).

We can now state the main result of (Devijver, 2015b, Theorem 5.1) for the model selection theorem for MLE among a random subcollection.

Theorem 3.3.4 (Theorem 5.1 from Devijver (2015b))**.** *Let $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$ be observations coming from an unknown conditional density s_0. Let the model collection $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ be an at most countable collection of conditional density sets. Assume that Assumption 3.2.1 (K), Assumption 3.2.2 (Sep), and Assumption 3.2.3 (H) hold for every $\mathbf{m} \in \mathcal{M}$. Let $\epsilon_{KL} > 0$, and $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$, such that*

$$\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \bar{s}_{\mathbf{m}}\right) \leq \inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, t\right) + \frac{\epsilon_{KL}}{n};$$

and let $\tau > 0$, such that

$$\bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0. \tag{3.3.7}$$

Introduce $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$, a random subcollection of $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$. Consider the collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of η-log likelihood minimizer satisfying (3.2.17) for all $\mathbf{m} \in \widetilde{\mathcal{M}}$. Then, for any $\rho \in (0,1)$, and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1, such that, for every index $\mathbf{m} \in \mathcal{M}$,

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left(\mathcal{D}_{\mathbf{m}} + (1 \vee \tau)\xi_{\mathbf{m}}\right),$$

with $\kappa > \kappa_0$, and where the model complexity $\mathcal{D}_{\mathbf{m}}$ is defined in Assumption 3.2.3, the η'-penalized likelihood estimator $\widehat{s}_{\widehat{\mathbf{m}}}$, defined as in (3.2.18) on the subset $\widetilde{\mathcal{M}} \subset \mathcal{M}$, satisfies

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, t\right) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta' + \eta}{n}.$$

In the next section, we apply Theorem 3.3.4 to prove Theorem 3.3.2. Consequently, the penalty can be chosen roughly proportional to the intrinsic dimension of the model, and thus of the order of the variance.

3.3.3.2 Proof of Theorem 3.3.2

It should be stressed that all we need is to verify that Assumption 3.2.3 (H), Assumption 3.2.2 (Sep) and Assumption 3.2.1 (K) hold for every $\mathbf{m} \in \mathcal{M}$. According to the result from Devijver (2015b, Section 5.3), Assumption 3.2.2 (Sep) holds when we consider Gaussian densities and the assumption defined by (3.3.7) is true if we assume further that the true conditional density s_0 is bounded and compactly supported. Furthermore, since we restricted d and K to $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, it is true that there exists a family $(\xi_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ and $\Xi > 0$ such that, Assumption 3.2.1 (K) is satisfied. Therefore, the main steps of the proof for the remaining Assumption 3.2.3 (H) are presented in this Section 3.3.3.2. All technical results are deferred to Section 3.3.4.

Note that the definition of complexity of model $S_{\mathbf{m}}$ in Assumption 3.2.3 (H) is related to a classical entropy dimension of a compact set w.r.t. a Hellinger type divergence $\mathrm{d}^{\otimes \mathrm{n}}$, thanks to the following Proposition 3.3.5, which is established in (Cohen & Le Pennec, 2011, Proposition 2).

Proposition 3.3.5 (Proposition 2 from Cohen & Le Pennec (2011)). *If, for any $\delta \in (0, \sqrt{2}]$, $\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right)$, then the function*

$$\phi_{\mathbf{m}}(\delta) = \delta\sqrt{\dim(S_{\mathbf{m}})}\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min(\delta, 1)}\right)}\right)$$

satisfies Assumption 3.2.3 (H). Furthermore, the unique solution $\delta_{\mathbf{m}}$ of $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$ satisfies

$$n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 \dim(S_{\mathbf{m}})}\right)_+\right).$$

Then, Assumption 3.2.3 (H) is proved via Proposition 3.3.5 using the fact that

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \tag{3.3.8}$$

where $C_{\mathbf{m}}$ is a constant depending on the model. Before proving the previous statement (3.3.8), we need to define the following distance over conditional densities:

$$\sup_{\mathbf{y}} d_{\mathbf{x}}(s, t) = \sup_{\mathbf{y}\in\mathcal{Y}}\left(\int_{\mathcal{X}}\left(\sqrt{s(\mathbf{x}|\mathbf{y})} - \sqrt{t(\mathbf{x}|\mathbf{y})}\right)^2 d\mathbf{x}\right)^{1/2}.$$

This leads straightforwardly to $\mathrm{d}^{2\otimes \mathrm{n}}(s, t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}^2(s, t)$. Then, we also define

$$\sup_{\mathbf{y}} d_k(g, g') = \sup_{\mathbf{y}\in\mathcal{Y}}\left(\sum_{k=1}^{K}\left(\sqrt{g_k(\mathbf{y})} - \sqrt{g'_k(\mathbf{y})}\right)^2\right)^{1/2},$$

for any gating functions g and g'. To this end, given any densities s and t over $\mathcal{X}$, the following distances, depending on $\mathbf{y}$, is constructed as follows:

$$\sup_{\mathbf{y}}\max_{k} d_{\mathbf{x}}(s, t) = \sup_{\mathbf{y}\in\mathcal{Y}}\max_{k\in[K]} d_{\mathbf{x}}(s_k(\cdot, \mathbf{y}), t_k(\cdot, \mathbf{y})) = \sup_{\mathbf{y}\in\mathcal{Y}}\max_{k\in[K]}\left(\int_{\mathcal{X}}\left(\sqrt{s_k(\mathbf{x}, \mathbf{y})} - \sqrt{t_k(\mathbf{x}, \mathbf{y})}\right)^2 d\mathbf{x}\right)^{1/2}.$$

Then (3.3.8) can be established by first decomposing the entropy term between the Gaussian gating functions and the Gaussian experts. Indeed, there are two possible ways to decompose the bracketing entropy of $S_{\mathbf{m}}$ based on the reparameterization trick (Nguyen et al., 2021c), for $\mathcal{P}_K$ via $\mathcal{W}_k$ and Gaussian experts $\mathcal{G}_{K,d,\mathbf{B}}$.

$$\mathcal{W}_K = \left\{\mathcal{Y} \ni \mathbf{y} \mapsto (\ln(\pi_k \phi(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k)))_{k\in[K]} =: (\mathbf{w}_k(\mathbf{y}; \boldsymbol{\omega}))_{k\in[K]} = \mathbf{w}(\mathbf{y}; \boldsymbol{\omega}) : \boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}_K\right\},$$

$$\mathcal{P}_K = \left\{\mathcal{Y} \ni \mathbf{y} \mapsto \left(\frac{e^{\mathbf{w}_k(\mathbf{y})}}{\sum_{l=1}^{K} e^{\mathbf{w}_l(\mathbf{y})}}\right)_{k\in[K]} =: (g_k(\mathbf{y}; \mathbf{w}))_{k\in[K]}, \mathbf{w} \in \mathcal{W}_K\right\}, \text{ and}$$

$$\mathcal{G}_{K,d,\mathbf{B}} = \left\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x}, \mathbf{y}) \mapsto (\phi(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)))_{k\in[K]} : \boldsymbol{\upsilon}_d \in \boldsymbol{\Upsilon}_{K,d}, \boldsymbol{\Sigma}(\mathbf{B}) \in \mathbf{V}_K(\mathbf{B})\right\}.$$

For the first approach, we can use Lemma 3.3.6 (Montuelle et al., 2014, Lemma 5):

Lemma 3.3.6. *For all* $\delta \in (0, \sqrt{2}]$ *and* $\mathbf{m} \in \mathcal{M}$,

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{5}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} \max_k d_{\mathbf{x}}}\left(\frac{\delta}{5}, \mathcal{G}_{K,d,\mathbf{B}}\right).$$

As mentioning in Appendix B.2.1 from Montuelle et al. (2014), Lemma 3.3.6 boils down to assuming that $\mathbf{Y}$ is bounded. To weaken this assumption, we are going to use the smaller distance: $\mathrm{d}^{\otimes \mathrm{n}}$, for the bracketing entropy of $S_{\mathbf{m}}$ although bounding such bracketing entropies for $\mathcal{W}_K$ and $\mathcal{G}_{K,\mathcal{B}}$ becomes much more challenging. Consequently, this leads to the second approach via Lemma 3.3.7 (Montuelle et al., 2014, Lemma 6).

Lemma 3.3.7. *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],d_{\mathcal{P}_K}}\left(\frac{\delta}{2}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right),$$

where

$$d^2_{\mathcal{P}_K}\left(g^+, g^-\right) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} d_k^2\left(g^+\left(\mathbf{Y}_i\right), g^-\left(\mathbf{Y}_i\right)\right)\right] = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\sqrt{g_k^+\left(\mathbf{Y}_i\right)} - \sqrt{g_k^-\left(\mathbf{Y}_i\right)}\right)^2\right],$$

$$\begin{aligned} d^2_{\mathcal{G}_{K,d,\mathbf{B}}}\left(\phi^+, \phi^-\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} d_{\mathbf{x}}^2\left(\phi_k^+\left(\cdot, \mathbf{Y}_i\right), \phi_k^-\left(\cdot, \mathbf{Y}_i\right)\right)\right] \\ &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_{\mathcal{X}}\left(\sqrt{\phi_k^+\left(\mathbf{x}, \mathbf{Y}_i\right)} - \sqrt{\phi_k^+\left(\mathbf{x}, \mathbf{Y}_i\right)}\right)^2 d\mathbf{x}\right]. \end{aligned}$$

Next, we make use of Lemma 3.3.8, which is proved in Section 3.3.4.1, to provide an upper bound on the bracketing entropy of $S_{\mathbf{m}}\left(\mathcal{P}_K\right)$ on distances $\mathrm{d}^{\otimes \mathrm{n}}\left(d_{\mathcal{P}_K}\right)$, respectively.

Lemma 3.3.8. *It holds that*

$$\mathrm{d}^{\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t), \text{ and } \mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}(\delta, S_{\mathbf{m}}), \tag{3.3.9}$$

$$d_{\mathcal{P}_K}\left(g^+, g^-\right) \leq \sup_{\mathbf{y}} d_k(g^+, g^-), \text{ and } \mathcal{H}_{[\cdot],d_{\mathcal{P}_K}}\left(\frac{\delta}{2}, \mathcal{P}_K\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right). \tag{3.3.10}$$

Lemmas 3.3.7 and 3.3.8 imply that

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right).$$

We next define the metric entropy of the set $\mathcal{W}_K$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathcal{W}_K)$, which measures the logarithm of the minimal number of balls of radius at most δ, according to a distance $d_{\|\sup\|_\infty}$, needed to cover $\mathcal{W}_K$, where

$$d_{\|\sup\|_\infty}\left(\left(\mathbf{s}_k\right)_{k\in[K]}, \left(\mathbf{t}_k\right)_{k\in[K]}\right) = \max_{k\in[K]} \sup_{\mathbf{y}\in\mathcal{Y}} \left\|\mathbf{s}_k(\mathbf{y}) - \mathbf{t}_k(\mathbf{y})\right\|_2, \tag{3.3.11}$$

for any K-tuples of functions $\left(\mathbf{s}_k\right)_{k\in[K]}$ and $\left(\mathbf{t}_k\right)_{k\in[K]}$. Here, $\mathbf{s}_k, \mathbf{t}_k : \mathcal{Y} \ni \mathbf{y} \mapsto \mathbf{s}_k(\mathbf{y}), \mathbf{t}_k(\mathbf{y}) \in \mathbb{R}^L, \forall k \in [K]$, and given $\mathbf{y} \in \mathcal{Y}, k \in [K]$, $\left\|\mathbf{s}_k(\mathbf{y}) - \mathbf{t}_k(\mathbf{y})\right\|_2$ is the Euclidean distance in $\mathbb{R}^L$.

Based on this metric, one can first relate the bracketing entropy of $\mathcal{P}_K$ to $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathcal{W}_K)$, and then obtain the upper bound for its entropy via Lemma 3.3.9. It is worth mentioning that for the Gaussian gating parameters, the technique for handling the logistic weights of Montuelle et al. (2014) is not directly applicable to the BLoME setting. Therefore, by using the previous reparameterization trick, Nguyen et al. (2021c, Lemmas 5.4 and 5.8) allow for the control of the metric entropy of the parameters of Gaussian gating functions.

Lemma 3.3.9 (Lemmas 5.5 from Nguyen et al. (2021c)). *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right) \leq \mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K}}, \mathcal{W}_K\right) \leq \dim\left(\mathcal{W}_K\right)\left(C_{\mathcal{W}} + \ln\left(\frac{8\sqrt{K}}{3\sqrt{3}\delta}\right)\right),$$

where $C_{\mathcal{Y}} := \sup_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y}\|_\infty < \infty$ *whenever* $\mathcal{Y}$ *is bounded,* $\mathcal{U} := \mathcal{Y} \times \mathcal{Y} \times [a_{\mathbf{\Gamma}}, A_{\mathbf{\Gamma}}]^{L^2}$,

$$C_{\mathcal{W}} := \frac{1}{\dim\left(\mathcal{W}_K\right)} \ln C_0, C_0 := \left(6C_{\boldsymbol{c}} C_{\mathcal{Y}} L\right)^{KL} \left(6C_{\mathbf{\Gamma}} A_{\mathbf{\Gamma}} L^2\right)^{\frac{L(L+1)}{2}K} \left(\frac{3}{a_{\boldsymbol{\pi}}}\right)^{K-1} K\left(2\pi e\right)^{K/2},$$

$$0 < \left(C_{\boldsymbol{c}}\right)^{\top}_{1,\ldots,L} := \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\mathbf{\Gamma}_k))\in\mathcal{U}} |\nabla_{\mathbf{c}_k} \ln|\phi_L(\mathbf{y};\mathbf{c}_k,\mathbf{\Gamma}_k)|| < \infty,$$

$$0 < \left(C_{\mathbf{\Sigma}}\right)^{\top}_{1,\ldots,L^2} := \max_{k\in[K]} \sup_{(\mathbf{y},\mathbf{c}_k,\mathrm{vec}(\mathbf{\Gamma}_k))\in\mathcal{U}} \left|\nabla_{\mathrm{vec}(\mathbf{\Gamma}_k)} \ln|\phi_L(\mathbf{y};\mathbf{c}_k,\mathbf{\Gamma}_k)|\right| < \infty,$$

and $\mathrm{vec}(\cdot)$ *denotes the vectorization operator that stacks the columns of a matrix into a vector.*

Lemma 3.3.10 allows us to construct the Gaussian brackets to handle the metric entropy for Gaussian experts, which is established in Section 3.3.4.2.

Lemma 3.3.10.

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \leq \dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{3.3.12}$$

Finally, (3.3.8) is proved via Lemmas 3.3.9 and 3.3.10. Indeed, with the fact that $\dim(S_{\mathbf{m}}) = \dim(\mathcal{W}_K) + \dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)$, it follows

$$\begin{aligned}
&\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \\
&\leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_k}\left(\frac{\delta}{2}, \mathcal{P}_K\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \\
&\leq \dim\left(\mathcal{W}_K\right)\left(C_{\mathcal{W}} + \ln\left(\frac{8\sqrt{K}}{3\sqrt{3}\delta}\right)\right) + \dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right) \\
&= \dim\left(S_{\mathbf{m}}\right)\left[\frac{\dim\left(\mathcal{W}_K\right)}{\dim\left(S_{\mathbf{m}}\right)}\left(C_{\mathcal{W}} + \ln\left(\frac{8\sqrt{K}}{3\sqrt{3}}\right) + \ln\left(\frac{1}{\delta}\right)\right) + \frac{\dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)}{\dim\left(S_{\mathbf{m}}\right)}\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right)\right] \\
&= \dim(S_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ where} \\
C_{\mathbf{m}} &= \frac{\dim(\mathcal{W}_K)}{\dim(S_{\mathbf{m}})}\left(C_{\mathcal{W}} + \ln\left(\frac{8\sqrt{K}}{3\sqrt{3}}\right)\right) + \frac{\dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right) C_{\mathcal{G}_{K,d,\mathbf{B}}}}{\dim\left(S_{\mathbf{m}}\right)} \\
&\leq C_{\mathcal{W}} + \ln\left(\frac{8\sqrt{K_{\max}}}{3\sqrt{3}}\right) + C_{\mathcal{G}_{K,d,\mathbf{B}}} := \mathfrak{C}.
\end{aligned}$$

It is interesting that the constant $\mathfrak{C}$ does not depend on the dimension $\dim\left(S_{\mathbf{m}}\right)$ of the model, thanks to the hypothesis that $C_{\mathcal{W}}$ is common for every model $S_{\mathbf{m}}$ in the collection. Therefore, Proposition 3.3.5 implies that, give $C = 2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2$, the model complexity $\mathcal{D}_{\mathbf{m}}$ satisfies

$$\mathcal{D}_{\mathbf{m}} \equiv n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 \dim\left(S_{\mathbf{m}}\right)}\right)_+\right) \leq \dim(S_{\mathbf{m}})\left(C + \ln n\right).$$

To this end, Theorem 3.3.4 implies that when a collection of BLoME models $\left(S_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ with the penalty functions satisfies $\mathrm{pen}(\mathbf{m}) \geq \kappa\left[\dim(S_{\mathbf{m}})\left(C + \ln n\right) + (1 \vee \tau)\xi_{\mathbf{m}}\right]$ with $\kappa > \kappa_0$, the oracle inequality in Theorem 3.3.2 holds.

3.3.4 Appendix: Lemma proofs

3.3.4.1 Proof of Lemma 3.3.8

We first aim to prove that $\mathrm{d}^{2\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d^2_{\mathbf{x}}(s,t)$. Indeed, by definition, it follows that

$$\begin{aligned}\mathrm{d}^{2\otimes \mathrm{n}}(s,t) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} d^2_{\mathbf{x}}\left(s\left(\cdot|\mathbf{y}\right), t\left(\cdot|\mathbf{y}\right)\right) s_{\mathbf{x},0}(\mathbf{y})d\mathbf{y} \leq \sup_{\mathbf{y}} d^2_{\mathbf{x}}\ (s,t)\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} s_{\mathbf{x},0}(\mathbf{y})d\mathbf{y} = \sup_{\mathbf{y}} d^2_{\mathbf{x}}\ (s,t),\end{aligned}$$

where $s_{\mathbf{x},0}$ denotes that marginal PDF of s_0, w.r.t. $\mathbf{x}$. Consequently, it holds that $\mathrm{d}^{\otimes \mathrm{n}}(s,t) = \sqrt{\mathrm{d}^{2\otimes \mathrm{n}}(s,t)} \leq \sqrt{\sup_{\mathbf{y}} d^2_{\mathbf{x}}(s,t)} = \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)$. To prove that

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right),$$

it is sufficient to check that

$$\mathcal{N}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right) \leq \mathcal{N}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}}\left(\delta, S_{\mathbf{m}}\right).$$

By using the definition of bracketing entropy in (3.2.23) and $\mathrm{d}^{\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)$, given

$$\begin{aligned}A &= \left\{n \in \mathbb{N}^{\star} : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t } \sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)\left(t_k^-, t_k^+\right) \leq \delta, S_{\mathbf{m}} \subset \bigcup_{k=1}^{n}\left[t_k^-, t_k^+\right]\right\}, \\ B &= \left\{n \in \mathbb{N}^{\star} : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t } \mathrm{d}^{\otimes \mathrm{n}}\left(t_k^-, t_k^+\right) \leq \delta, S_{\mathbf{m}} \subset \bigcup_{k=1}^{n}\left[t_k^-, t_k^+\right]\right\},\end{aligned}$$

it leads to that $A \subset B$ and then (3.3.9) follows, since

$$\mathcal{N}_{[\cdot],\sup_{\mathbf{y}} d_{\mathbf{x}}(s,t)}\left(\delta, S_{\mathbf{m}}\right) = \min A \geq \min B = \mathcal{N}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, S_{\mathbf{m}}\right).$$

With the similar argument as in the proof of (3.3.9), it holds that $d_{\mathcal{P}_K}\left(g^+, g^-\right) \leq \sup_{\mathbf{y}} d_k(g^+, g^-)$ and (3.3.10) is proved.

3.3.4.2 Proof of Lemma 3.3.10

It is worth mentioning that without any structures on covariance matrices of Gaussian experts from the collection $\mathcal{M}$, Lemma 3.3.10 can be proved using Proposition 2 from Montuelle et al. (2014) and Montuelle et al. (2014, Appendix B.2.3), for constructing of Gaussian brackets to deal with the Gaussian experts. However, dealing with block-diagonal covariance matrices with random subcollection is much more challenging. We have to establish more constructive bracketing entropies in the spirits of Maugis & Michel (2011b), Devijver (2015b), Devijver & Gallopin (2018).

Given any $k \in [K]$, by defining

$$\mathcal{G}_{d,\mathbf{B}_k} = \left\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) =: \phi_k : \boldsymbol{\upsilon}_{k,d} \in \boldsymbol{\Upsilon}_{k,d}, \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k(\mathbf{B}_k)\right\}, \tag{3.3.13}$$

it follows that $\mathcal{G}_{K,d,\mathbf{B}} = \prod_{k=1}^{K}\mathcal{G}_{d,\mathbf{B}_k}$, where $\prod$ stands for the cartesian product. By using Lemma 3.3.11, which is proved in Section 3.3.4.3, it follows that

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2}, \mathcal{G}_{K,d,\mathbf{B}}\right) \leq \sum_{k=1}^{K}\mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}}, \mathcal{G}_{d,\mathbf{B}_k}\right). \tag{3.3.14}$$

Lemma 3.3.11. *Given* $\mathcal{G}_{K,d,\mathbf{B}} = \prod_{k=1}^{K} \mathcal{G}_{d,\mathbf{B}_k}$, *where* $\mathcal{G}_{d,\mathbf{B}_k}$ *is defined in* (3.3.13), *it holds that*

$$\mathcal{N}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2},\mathcal{G}_{K,d,\mathbf{B}}\right) \leq \prod_{k=1}^{K} \mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}},\mathcal{G}_{d,\mathbf{B}_k}\right),$$

where for any $\phi^+, \phi^- \in \mathcal{G}_{K,d,\mathbf{B}}$ *and any* $\phi_k^+, \phi_k^- \in \mathcal{G}_{d,\mathbf{B}_k}, k \in [K]$,

$$d^2_{\mathcal{G}_{K,d,\mathbf{B}}}\left(\phi^+,\phi^-\right) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathrm{d}^2\left(\phi_k^+\left(\cdot,\mathbf{Y}_i\right),\phi_k^-\left(\cdot,\mathbf{Y}_i\right)\right)\right],$$

$$d^2_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^+,\phi_k^-\right) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{d}^2\left(\phi_k^+\left(\cdot,\mathbf{Y}_i\right),\phi_k^-\left(\cdot,\mathbf{Y}_i\right)\right)\right].$$

Lemma 3.3.10 is proved via (3.3.14) and Lemma 3.3.12, which is proved in Section 3.3.4.4.

Lemma 3.3.12. *By defining* $\mathcal{G}_{d,\mathbf{B}_k}$ *as in* (3.3.13), *for all* $\delta \in (0,\sqrt{2}]$, *it holds that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2},\mathcal{G}_{d,\mathbf{B}_k}\right) \leq \dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ where} \tag{3.3.15}$$

$$D_{\mathbf{B}_k} = \sum_{g=1}^{G_k} \frac{\mathrm{card}\left(d_k^{[g]}\right)\left(\mathrm{card}\left(d_k^{[g]}\right)-1\right)}{2},$$

$$C_{\mathcal{G}_{d,\mathbf{B}_k}} = \frac{D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M D^2(D-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)}{\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)}.$$

Indeed, (3.3.14) and (3.3.15) lead to

$$\begin{aligned}
\mathcal{H}_{[\cdot],d_{\mathcal{G}_{K,d,\mathbf{B}}}}\left(\frac{\delta}{2},\mathcal{G}_{K,d,\mathbf{B}}\right) &\leq \sum_{k=1}^{K} \mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2\sqrt{K}},\mathcal{G}_{d,\mathbf{B}_k}\right) \\
&\leq \sum_{k=1}^{K} \dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\sqrt{K}\right) + \ln\left(\frac{1}{\delta}\right)\right) \\
&\leq \dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right)\left(C_{\mathcal{G}_{K,d,\mathbf{B}}} + \ln\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

Here, $C_{\mathcal{G}_{K,d,\mathbf{B}}} = \sum_{k=1}^{K} C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\sqrt{K}\right)$ and note that $\dim\left(\mathcal{G}_{K,d,\mathbf{B}}\right) = \sum_{k=1}^{K}\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)$, $\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right) = D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)$, $\dim\left(\boldsymbol{\Upsilon}_{k,d}\right) = Dd_{\boldsymbol{\Upsilon}_{k,d}}, C_{\boldsymbol{\Upsilon}_{k,d}} = \sqrt{D}d_{\boldsymbol{\Upsilon}_{k,d}}T_{\boldsymbol{\Upsilon}_{k,d}}$ (in cases where linear combination of bounded functions are used for means, *i.e.*, $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_b$) or $\dim\left(\boldsymbol{\Upsilon}_{k,d}\right) = D\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}$, $C_{\boldsymbol{\Upsilon}_{k,d}} = \sqrt{D}\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}T_{\boldsymbol{\Upsilon}_{k,d}}$ (in cases where we use polynomial means, *i.e.*, $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_p$).

3.3.4.3 Proof of Lemma 3.3.11

By the definition of the bracketing entropy in (3.2.23), for each $k \in [K]$, let $\left\{\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ be a minimal covering of δ_k brackets for $d_{\mathcal{G}_{d,\mathbf{B}_k}}$ of $\mathcal{G}_{d,\mathbf{B}_k}$, with cardinality $\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}$. This leads to

$$\forall l \in \left[\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}\right], d_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l,-},\phi_k^{l,+}\right) \leq \delta_k.$$

Therefore, we claim that the set $\left\{\prod_{k=1}^{K}\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ is a covering of $\frac{\delta}{2}$-bracket for $d_{\mathcal{G}_{K,d,\mathbf{B}}}$ of $\mathcal{G}_{K,d,\mathbf{B}}$ with cardinality $\prod_{k=1}^{K}\mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}(\delta_k,\mathcal{G}_{d,\mathbf{B}_k})$. Indeed, let any $\phi = (\phi_k)_{k\in[K]} \in \mathcal{G}_{K,d,\mathbf{B}}$. Consequently, for each $k \in [K], \phi_k \in \mathcal{G}_{d,\mathbf{B}_k}$, there exists $l(k) \in \left[\mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}\right]$, such that

$$\phi_k^{l(k),-} \leq \phi_k \leq \phi_k^{l(k),+}, d^2_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l(k),+},\phi_k^{l(k),-}\right) \leq \left(\delta_k\right)^2.$$

Then, it follows that $\phi \in [\phi^-, \phi^+] \in \left\{\prod_{k=1}^K \left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$, with $\phi^- = \left(\phi_k^{l(k),-}\right)_{k\in[K]}, \phi^+ = \left(\phi_k^{l(k),+}\right)_{k\in[K]}$, which implies that $\left\{\prod_{k=1}^K \left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{d,\mathbf{B}_k}}}$ is a bracket covering of $\mathcal{G}_{K,d,\mathbf{B}}$.

Now, we want to verify that the size of this bracket is $\delta/2$ by choosing $\delta_k = \frac{\delta}{2\sqrt{K}}, \forall k \in [K]$. It follows that

$$\begin{aligned} d^2_{\mathcal{G}_{K,d,\mathbf{B}}}\left(\phi^-, \phi^+\right) &= \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \mathrm{d}^2\left(\phi_k^{l(k),-}\left(\cdot, \mathbf{Y}_i\right), \phi_k^{l(k),+}\left(\cdot, \mathbf{Y}_i\right)\right)\right] \\ &= \sum_{k=1}^K \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \mathrm{d}^2\left(\phi_k^{l(k),-}\left(\cdot, \mathbf{Y}_i\right), \phi_k^{l(k),+}\left(\cdot, \mathbf{Y}_i\right)\right)\right] \\ &= \sum_{k=1}^K d^2_{\mathcal{G}_{d,\mathbf{B}_k}}\left(\phi_k^{l(k),-}, \phi_k^{l(k),+}\right) \leq K\left(\frac{\delta}{2\sqrt{K}}\right)^2 = \left(\frac{\delta}{2}\right)^2. \end{aligned}$$

To this end, by definition of a minimal $\frac{\delta}{2}$-bracket covering number for $\mathcal{G}_{K,d,\mathbf{B}}$, Lemma 3.3.11 is proved.

3.3.4.4 Proof of Lemma 3.3.12

To provide the upper bound of the bracketing entropy in (3.3.15), our technique is adapted from the work of Genovese & Wasserman (2000) for unidimensional Gaussian mixture families, which is then generalized to multidimensional case by Maugis & Michel (2011b). Furthermore, we make use of the results from Devijver & Gallopin (2018) to deal with block-diagonal covariance matrices, $\mathbf{V}_k\left(\mathbf{B}_k\right), k \in [K]$, and from Montuelle et al. (2014) to handle the means of Gaussian experts $\boldsymbol{\Upsilon}_{k,d}, k \in [K]$. The main idea is to define firstly a net over the parameter spaces of Gaussian experts, $\boldsymbol{\Upsilon}_{k,d} \times \mathbf{V}_k\left(\mathbf{B}_k\right), k \in [K]$, and to construct a bracket covering of $\mathcal{G}_{d,\mathbf{B}_k}$ according to the tensorized Hellinger distance. Note that $\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right) = \dim\left(\boldsymbol{\Upsilon}_{k,d}\right) + \dim\left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right)$.

Step 1: Construction of a net for the block-diagonal covariance matrices. Firstly, for $k \in [K]$, we denote by $\mathrm{Adj}\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)$ the adjacency matrix associated to the covariance matrix $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)$. Note that this matrix of size D^2 can be defined by a vector of concatenated upper triangular vectors. We are going to make use of the result from Devijver & Gallopin (2018) to handle the block-diagonal covariance matrices $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)$, via its corresponding adjacency matrix. To do this, we need to construct a discrete space for $\{0,1\}^{D(D-1)/2}$, which is a one-to-one correspondence (bijection) with

$$\mathcal{A}_{\mathbf{B}_k} = \left\{\mathbf{A}_{\mathbf{B}_k} \in \mathcal{S}_D\left(\{0,1\}\right) : \exists \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k\left(\mathbf{B}_k\right) \text{ s.t } \mathrm{Adj}\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) = \mathbf{A}_{\mathbf{B}_k}\right\},$$

where $\mathcal{S}_D\left(\{0,1\}\right)$ is the set of symmetric matrices of size D taking values on $\{0,1\}$.

Then, we want to deduce a discretization of the set of covariance matrices. Let h denotes Hamming distance on $\{0,1\}^{D(D-1)/2}$ defined by

$$d(z, z') = \sum_{i=1}^n \mathbb{I}\left\{z \neq z'\right\}, \text{ for all } z, z' \in \{0,1\}^{D(D-1)/2}.$$

Let $\{0,1\}^{D(D-1)/2}_{\mathbf{B}_k}$ be the subset of $\{0,1\}^{D(D-1)/2}$ of vectors for which the corresponding graph has structure $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$. Corollary 1 and Proposition 2 from Supplementary Material A of Devijver & Gallopin (2018) imply that there exists some subset $\mathcal{R}$ of $\{0,1\}^{D(D-1)/2}$, as well as its equivalent $\mathcal{A}^{\mathrm{disc}}_{\mathbf{B}_k}$ for adjacency matrices such that, given $\epsilon > 0$, and

$$\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon) = \left\{\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathcal{S}_D^{++}(\mathbb{R}) : \mathrm{Adj}\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \in \mathcal{A}^{\mathrm{disc}}_{\mathbf{B}_k}, \left[\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right]_{i,j} = \sigma_{i,j}\epsilon, \sigma_{i,j} \in \left[\frac{-\lambda_M}{\epsilon}, \frac{\lambda_M}{\epsilon}\right] \bigcap \mathbb{Z}\right\},$$

it holds that

$$\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2 \leq \frac{D_{\mathbf{B}_k}}{2}\wedge\epsilon^2, \forall\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right),\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right)\in\left(\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)\right)^2 s.t.\ \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\neq\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right),$$

$$\operatorname{card}\left(\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)\right)\leq\left(\left\lfloor\frac{2\lambda_M}{\epsilon}\right\rfloor\frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}}, \tag{3.3.16}$$

$$D_{\mathbf{B}_k}=\dim\left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right)=\sum_{g=1}^{G_k}\frac{\operatorname{card}\left(d_k^{[g]}\right)\left(\operatorname{card}\left(d_k^{[g]}\right)-1\right)}{2}. \tag{3.3.17}$$

By choosing $\epsilon^2\leq\frac{D_{\mathbf{B}_k}}{2}$, given $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\in\mathbf{V}_k\left(\mathbf{B}_k\right)$, then there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\in\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)$, such that

$$\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2\leq\epsilon^2. \tag{3.3.18}$$

Step 2: Construction of a net for the mean functions. Based on $\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)$, we can construct the following bracket covering of $\mathcal{G}_{d,\mathbf{B}_k}$ by defining the nets for the means of Gaussian experts. The proof of Lemma 1, page 1693, from Montuelle et al. (2014) implies that

$$\mathcal{N}_{[\cdot],\sup_{\mathbf{y}}\|\cdot\|_2}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}},\boldsymbol{\Upsilon}_{k,d}\right)\leq\left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta_{\boldsymbol{\Upsilon}_{k,d}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)}.$$

Here $\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)=Dd_{\boldsymbol{\Upsilon}_{k,d}}$, and $C_{\boldsymbol{\Upsilon}_{k,d}}=\sqrt{D}d_{\boldsymbol{\Upsilon}_{k,d}}T_{\boldsymbol{\Upsilon}_{k,d}}$ in the general case or $\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)=D\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}$, and $C_{\boldsymbol{\Upsilon}_{k,d}}=\sqrt{D}\binom{d_{\boldsymbol{\Upsilon}_{k,d}}+L}{L}T_{\boldsymbol{\Upsilon}_{k,d}}$ in the special case of polynomial means. Then, by the definition of bracketing entropy in (3.2.23), for any minimal $\delta_{\boldsymbol{\Upsilon}_{k,d}}$-bracketing covering of the means from Gaussian experts, denoted by $G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$, it is true that

$$\operatorname{card}\left(G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)\right)\leq\left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta_{\boldsymbol{\Upsilon}_{k,d}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)}. \tag{3.3.19}$$

Therefore, given $\alpha>0$, which is specified later, we claim that the set

$$\left\{[l,u]\;\middle|\;\begin{array}{l} l(\mathbf{x},\mathbf{y})=(1+2\alpha)^{-D}\,\phi\left(\mathbf{x};\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right),\\ u(\mathbf{x},\mathbf{y})=(1+2\alpha)^{D}\,\phi\left(\mathbf{x};\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right),\\ \widetilde{\boldsymbol{\upsilon}}_{k,d}\in G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right),\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\in\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)\end{array}\right\},$$

is a $\delta_{\boldsymbol{\Upsilon}_{k,d}}$-brackets set over $\mathcal{G}_{d,\mathbf{B}_k}$. Indeed, let $\mathcal{X}\times\mathcal{Y}\ni(\mathbf{x},\mathbf{y})\mapsto f(\mathbf{x},\mathbf{y})=\phi\left(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)$ be a function of $\mathcal{G}_{d,\mathbf{B}_k}$, where $\boldsymbol{\upsilon}_{k,d}\in\boldsymbol{\Upsilon}_{k,d}$ and $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\in\mathbf{V}_k\left(\mathbf{B}_k\right)$. According to (3.3.18), there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\in\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)$, such that

$$\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2\leq\epsilon^2.$$

By definition of $G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$, there exists $\widetilde{\boldsymbol{\upsilon}}_{k,d}\in G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$, such that

$$\sup_{\mathbf{y}\in\mathcal{Y}}\|\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y})-\boldsymbol{\upsilon}_{k,d}(\mathbf{y})\|_2^2\leq\delta^2_{\boldsymbol{\Upsilon}_{k,d}}. \tag{3.3.20}$$

Step 3: Upper bound of the number of the bracketing entropy. Next, we wish to make use of Lemma 3.3.13 to evaluate the ratio of two Gaussian densities.

Lemma 3.3.13 (Proposition C.1 from Maugis & Michel (2011b)). *Let $\phi\left(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)$ and $\phi\left(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)$ be two Gaussian densities. If $\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1$ is a positive definite matrix then for all $\mathbf{x}\in\mathbb{R}^D$,*

$$\frac{\phi\left(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)}{\phi\left(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)}\leq\sqrt{\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}}\exp\left[\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^\top\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right)^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\right].$$

The following Lemma 3.3.14 allows us to fulfill the assumptions of Lemma 3.3.13.

Lemma 3.3.14 (Similar to Lemma B.8 from Maugis & Michel (2011b)). *Assume that* $0 < \epsilon < \lambda_m^2/9$, *and set* $\alpha = 3\sqrt{\epsilon}/\lambda_m$. *Then, for every* $k \in [K]$, $(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)$ *and* $\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)$ *are both positive definite matrices. Moreover, for all* $\mathbf{x} \in \mathbb{R}^D$,

$$\mathbf{x}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{x} \geq \epsilon\,\|\mathbf{x}\|_2^2, \quad \mathbf{x}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{x} \geq \epsilon\,\|\mathbf{x}\|_2^2.$$

Proof of Lemma 3.3.14. For all $\mathbf{x} \neq \mathbf{0}$, since $\sup_{\lambda\in \mathrm{vp}\left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)-\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right)}|\lambda| = \left\|\boldsymbol{\Sigma}_k(\mathbf{B}_k) - \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right\|_2 \leq \epsilon$, where vp denotes the spectrum of matrix, $-\epsilon \geq -\lambda_m/3$, and $\alpha = 3\epsilon/\lambda_m$, it follow that

$$\begin{aligned}
\mathbf{x}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{x} &= (1+\alpha)\,\mathbf{x}^\top\left[\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{x} + \alpha\mathbf{x}^\top\boldsymbol{\Sigma}_k(\mathbf{B}_k)\,\mathbf{x} \\
&\geq -(1+\alpha)\left\|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right\|_2\|\mathbf{x}\|_2^2 + \alpha\lambda_m\,\|\mathbf{x}\|_2^2 \\
&\geq (\alpha\lambda_m - (1+\alpha)\,\epsilon)\,\|\mathbf{x}\|_2^2 = (\alpha\lambda_m - \alpha\epsilon - \epsilon)\,\|\mathbf{x}\|_2^2 \\
&\geq \left(\frac{2}{3}\alpha\lambda_m - \epsilon\right)\|\mathbf{x}\|_2^2 = \epsilon\,\|\mathbf{x}\|_2^2 > 0, \text{ and} \\
\mathbf{x}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{x} &= (1+\alpha)^{-1}\,\mathbf{x}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{x} + \left(1-(1+\alpha)^{-1}\right)\mathbf{x}^\top\boldsymbol{\Sigma}_k(\mathbf{B}_k)\,\mathbf{x} \\
&\geq \left(\frac{\alpha\lambda_m - \epsilon}{1+\alpha}\right)\|\mathbf{x}\|_2^2 = \frac{2\epsilon}{1+\alpha}\,\|\mathbf{x}\|_2^2 \geq \epsilon\,\|\mathbf{x}\|_2^2 > 0\,(\text{ since } 0 < \alpha < 1).
\end{aligned}$$

□

By Lemma 3.3.13 and the same argument as in the proof of Lemma B.9 from Maugis & Michel (2011b), given $0 < \epsilon < \lambda_m/3$, where ϵ is chosen later, and $\alpha = 3\epsilon/\lambda_m$, we obtain

$$\max\left\{\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}, \frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right\} \leq (1+2\alpha)^{-\frac{D}{2}}\exp\left(\frac{\|\boldsymbol{\upsilon}_{k,d}(\mathbf{y}) - \widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y})\|_2^2}{2\epsilon}\right). \tag{3.3.21}$$

Because $\ln(\cdot)$ is a non-decreasing function, $\ln(1+2\alpha) \geq \alpha, \forall\alpha \in [0,1]$. Combined with (3.3.20) where $\delta^2_{\boldsymbol{\Upsilon}_{k,d}} = D\alpha\epsilon$, we conclude that

$$\max\left\{\ln\left(\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right), \ln\left(\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right)\right\} \leq -\frac{D}{2}\ln(1+2\alpha) + \frac{\delta^2_{\boldsymbol{\Upsilon}_{k,d}}}{2\epsilon} \leq -\frac{D}{2}\alpha + \frac{\delta^2_{\boldsymbol{\Upsilon}_{k,d}}}{2\epsilon} = 0.$$

This means that $l(\mathbf{x},\mathbf{y}) \leq f(\mathbf{x},\mathbf{y}) \leq u(\mathbf{x},\mathbf{y}), \forall(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$. Hence, it remains to bound the size of bracket $[l,u]$ w.r.t. $d_{\mathcal{G}_{d,\mathbf{B}_k}}$. To this end, we aim to verify that $d^2_{\mathcal{G}_{d,\mathbf{B}_k}}(l,u) \leq \frac{\delta}{2}$. To do that, we make use of the following Lemma 3.3.15.

Lemma 3.3.15 (Proposition C.3 from Maugis & Michel (2011b)). *Let* $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ *and* $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ *be two Gaussian densities with full rank covariance. It holds that*

$$\begin{aligned}
&d^2\left(\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1), \phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right) \\
&= 2\left\{1 - 2^{D/2}\,|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2|^{-1/4}\left|\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right|^{-1/2}\exp\left[-\frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]\right\}.
\end{aligned}$$

Therefore, using the fact that $\cosh(t) = \frac{e^{-t}+e^{t}}{2}$, Lemma 3.3.15 leads to, for all $\mathbf{y} \in \mathcal{Y}$:

$$\begin{aligned}
\mathrm{d}^2(l(\cdot,\mathbf{y}),u(\cdot,\mathbf{y})) &= \int_{\mathcal{X}} \left[l(\mathbf{x},\mathbf{y}) + u(\mathbf{x},\mathbf{y}) - 2\sqrt{l(\mathbf{x},\mathbf{y})u(\mathbf{x},\mathbf{y})}\right] d\mathbf{x} \\
&= (1+2\alpha)^{-D} + (1+2\alpha)^{D} - 2 \\
&+ \mathrm{d}^2\left(\phi\left(\cdot;\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right),\phi\left(\cdot;\widetilde{\boldsymbol{\upsilon}}_{k,d}(\mathbf{y}),(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right)\right) \\
&= 2\cosh\left[D\ln\left(1+2\alpha\right)\right] - 2 \\
&\quad + 2\left[1 - 2^{D/2}\left[(1+\alpha)^{-1} + (1+\alpha)\right]^{-D/2}\left|\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right|^{-1/2}\left|\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right|^{1/2}\right] \\
&= 2\cosh\left[D\ln\left(1+2\alpha\right)\right] - 2 + 2 - 2\left[\cosh\left(\ln\left(1+\alpha\right)\right)\right]^{-D/2} \\
&= 2g\left(D\ln\left(1+2\alpha\right)\right) + 2h\left(\ln\left(1+\alpha\right)\right),
\end{aligned}$$

where $g(t) = \cosh(t) - 1 = \frac{e^{-t}+e^{t}}{2} - 1$, and $h(t) = 1 - \cosh(t)^{-D/2}$. The upper bounds of terms g and h separately imply that, for all $\mathbf{y} \in \mathcal{Y}$,

$$\mathrm{d}^2(l(\cdot,\mathbf{y}),u(\cdot,\mathbf{y})) \leq 2\left(2\cosh\left(\frac{1}{\sqrt{6}}\right)\alpha^2D^2 + \frac{1}{4}\alpha^2D^2\right) \leq 6\alpha^2D^2 = \frac{\delta^2}{4},$$

where we choose $\alpha = \frac{3\epsilon}{\lambda_m}, \epsilon = \frac{\delta\lambda_m}{6\sqrt{6D}}, \forall\delta \in (0,1], D \in \mathbb{N}^\star, \lambda_m > 0$, which appears in (3.3.21) and satisfies $\alpha = \frac{\delta}{2\sqrt{6D}}$ and $0 < \epsilon < \frac{\lambda_m}{3}$. Indeed, studying functions g and h yields

$$\begin{aligned}
g'(t) &= \sinh(t), g''(t) = \cosh(t) \leq \cosh(c), \forall t \in [0,c], c \in \mathbb{R}_+, \\
h'(t) &= \frac{D}{2}\cosh(t)^{-D/2-1}\sinh(t), \\
h''(t) &= \frac{D}{2}\left(-\frac{D}{2} - 1\right)\cosh(t)^{-D/2-2}\sinh^2(t) + \frac{D}{2}\cosh(t)^{-D/2} \\
&= \frac{D}{2}\left(1 - \left(\frac{D}{2}+1\right)\left(\frac{\sinh(t)}{\cosh(t)}\right)^2\right)\cosh(t)^{-D/2} \leq \frac{D}{2},
\end{aligned}$$

where we used the fact that $\cosh(t) \geq 1$. Then, since $g(0) = 0, g'(0) = 0, h(0) = 0, h'(0) = 0$, by applying Taylor's Theorem, it is true that

$$\begin{aligned}
g(t) &= g(t) - g(0) - g'(0)t = R_{0,1}(t) \leq \cosh(c)\frac{t^2}{2}, \forall t \in [0,c], \\
h(t) &= h(t) - h(0) - h'(0)t = R_{0,1}(t) \leq \frac{D}{2}\frac{t^2}{2} \leq \frac{D^2}{2}\frac{t^2}{2}, \forall t \geq 0.
\end{aligned}$$

We wish to find an upper bound for $t = D\ln(1+2\alpha)$, $D \in \mathbb{N}^\star$, $\alpha = \frac{\delta}{2\sqrt{6D}}$, $\delta \in (0,1]$. Since ln is an increasing function, then we have

$$t = D\ln\left(1 + \frac{\delta}{\sqrt{6}D}\right) \leq D\ln\left(1 + \frac{1}{\sqrt{6}D}\right) \leq D\frac{1}{\sqrt{6}D} = \frac{1}{\sqrt{6}}, \forall\delta \in (0,1],$$

since $\ln\left(1 + \frac{1}{\sqrt{6}D}\right) \leq \frac{1}{\sqrt{6}D}$, $\forall D \in \mathbb{N}^\star$. Then, since $\ln(1+2\alpha) \leq 2\alpha, \forall\alpha \geq 0$,

$$\begin{aligned}
g\left(D\ln\left(1+2\alpha\right)\right) &\leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{\left(D\ln\left(1+2\alpha\right)\right)^2}{2} \leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{D^2}{2}4\alpha^2, \\
h\left(\ln\left(1+\alpha\right)\right) &\leq \frac{D^2}{2}\frac{\left(\ln\left(1+\alpha\right)\right)^2}{2} \leq \frac{D^2\alpha^2}{4}.
\end{aligned}$$

Note that the set of $\delta/2$-brackets $[l,u]$ over $\mathcal{G}_{d,\mathbf{B}_k}$ is totally defined by the parameter spaces $\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)$ and $G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)$. This leads to an upper bound of the $\delta/2$-bracketing entropy of $\mathcal{G}_{d,\mathbf{B}_k}$ evaluated

from an upper bound of the two set cardinalities. Hence, given any $\delta > 0$, by choosing $\epsilon = \frac{\delta\lambda_m}{6\sqrt{6}D}$, $\alpha = \frac{3\epsilon}{\lambda_m} = \frac{\delta}{2\sqrt{6}D}$, and $\delta^2_{\boldsymbol{\Upsilon}_{k,d}} = D\alpha\epsilon = D\frac{\delta}{2\sqrt{6}D}\frac{\delta\lambda_m}{6\sqrt{6}D} = \frac{\delta^2\lambda_m}{72D}$, it holds that

$$\begin{aligned}
\mathcal{N}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2}, \mathcal{G}_{d,\mathbf{B}_k}\right) &\leq \text{card}\left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right) \times \text{card}\left(G_{\boldsymbol{\Upsilon}_{k,d}}\left(\delta_{\boldsymbol{\Upsilon}_{k,d}}\right)\right) \\
&\leq \left(\left\lfloor \frac{2\lambda_M}{\epsilon} \right\rfloor \frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta_{\boldsymbol{\Upsilon}_{k,d}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \quad \text{(using (3.3.17) and (3.3.19))} \\
&\leq \left(\frac{2\lambda_M 6\sqrt{6}D}{\delta\lambda_m}\frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\delta\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \\
&= \left(\frac{6\sqrt{6}\lambda_M D^2\left(D-1\right)}{\lambda_m D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)} \left(\frac{1}{\delta}\right)^{D_{\mathbf{B}_k}+\dim\left(\boldsymbol{\Upsilon}_{k,d}\right)}.
\end{aligned}$$

Finally, by definition of bracketing entropy in (3.2.23), we obtain

$$\begin{aligned}
\mathcal{H}_{[\cdot],d_{\mathcal{G}_{d,\mathbf{B}_k}}}\left(\frac{\delta}{2}, \mathcal{G}_{d,\mathbf{B}_k}\right) &\leq D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M D^2\left(D-1\right)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right) \\
&\quad + \left(D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\right)\ln\left(\frac{1}{\delta}\right) = \dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)\left(C_{\mathcal{G}_{d,\mathbf{B}_k}} + \ln\left(\frac{1}{\delta}\right)\right),
\end{aligned}$$

where $\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right) = D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)$ and

$$C_{\mathcal{G}_{d,\mathbf{B}_k}} = \frac{D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M D^2(D-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{k,d}\right)\ln\left(\frac{6\sqrt{2D}\exp\left(C_{\boldsymbol{\Upsilon}_{k,d}}\right)}{\sqrt{\lambda_m}}\right)}{\dim\left(\mathcal{G}_{d,\mathbf{B}_k}\right)}.$$

Chapter 4

Joint rank and variable selection in the softmax-gated block-diagonal mixture of experts regression model

Chapter 4 is based on the following works:

(C7) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *An l_1-oracle inequality for the lasso in mixture of experts regression models.* arXiv preprint arXiv:2009.10622. 2020. Link: https://arxiv.org/pdf/2009.10622.pdf (Nguyen et al., 2020c).

(C8) *Joint rank and variable selection by a non-asymptotic model selection in mixture of polynomial experts models.* Ongoing work.

Contents

Recall that in Sections 1.2.8 and 1.2.9, we highlighted the main oracle inequalities without detailed proofs regarding an l_1-oracle inequality satisfied by the Lasso estimator in SGaME models and a weak oracle type inequality satisfied by PMLEs is constructed for PSGaBloME models, which is particularly useful for nonlinear regression models for high-dimensional heterogeneous data. In particular, our oracle inequalities show that the performance in Jensen–Kullback–Leibler type loss of our penalized maximum likelihood estimators are roughly comparable to that of oracle models if we take large enough the constants in front of the penalties, whose forms are only known up to multiplicative constants and proportional to the dimensions of models. Such theoretical justifications of the penalty shapes

motivate us to make use of the slope heuristic criterion to select several hyperparameters, including the number of mixture components, the amount of sparsity (the coefficients and ranks sparsity levels), the degree of polynomial mean functions, and the potential hidden block-diagonal structures of the covariance matrices of the multivariate predictor or response variable. In Chapter 4, we aim to present such non-asymptotic oracle inequalities in as much detail as possible.

4.1 Introduction

We recall that MoE models are used to estimate the conditional distribution of a random variable $\mathbf{Y} \in \mathbb{R}^q$, given certain features from n observations $\{\mathbf{x}_i\}_{i\in[n]} = \{(x_{i1}, \ldots, x_{ip})\}_{i\in[n]} \in (\mathbb{R}^p)^n$, where $q, p, n \in \mathbb{N}^\star$, $[n] := \{1, \ldots, n\}$, $n \in \mathbb{N}^\star$ denotes the positive integer numbers, and $\mathbb{R}^p$ means the p-dimensional real number. The use of MoE models in the high-dimensional regression setting, when the number of explanatory variables can be much larger than the sample size, remains a challenge, particularly from a theoretical point of view, where there is still a lack of results in the literature regarding both statistical estimation and model selection. In such settings, we are required to reduce the dimension of the problem by seeking the most relevant relationships, to avoid numerical identifiability problems.

In Chapter 4, we focus on the use of an l_1-penalized maximum likelihood estimator (PMLE), as originally proposed as the Lasso by Tibshirani (1996), which tends to produce sparse solutions and can be viewed as a convex surrogate for the non-convex l_0-penalization problem. These methods have attractive computational and theoretical properties (cf. Fan & Li, 2001). First introduced in Tibshirani (1996) for the linear regression model, the Lasso estimator has since been extended to many statistical problems, including for high-dimensional regression of non-homogeneous data by using finite mixture regression models as considered by Khalili & Chen (2007), Stadler et al. (2010), and Lloyd-Jones et al. (2018). In Stadler et al. (2010), they consider scalar response, *i.e.*, $q = 1$, and it is assumed that, for $i \in [n]$, the observations y_i, conditionally on $\mathbf{X}_i = \mathbf{x}_i$, come from a conditional density $s_{\psi_0(\cdot|\mathbf{x}_i)}$, which is a finite mixture of $K \in \mathbb{N}^\star$ Gaussian conditional densities with mixing proportions $(\pi_{0,1}, \ldots, \pi_{0,K})$, where

$$Y_i|\mathbf{X}_i = \mathbf{x}_i \sim s_{\boldsymbol{\psi}_0}(y_i|\mathbf{x}_i) = \sum_{k=1}^{K} \pi_{0,k}\phi(y_i; \boldsymbol{\beta}_{0,k}^\top \mathbf{x}_i, \sigma_{0,k}^2). \tag{4.1.1}$$

Here

$$\phi(\cdot; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\cdot - \mu)^2}{2\sigma^2}\right)$$

is the univariate Gaussian probability density function (PDF), with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$, and $\boldsymbol{\psi}_0 = (\pi_{0,k}, \boldsymbol{\beta}_{0,k}, \sigma_{0,k})_{k\in[K]}$ is the vector of model parameters.

Then, considering a model $\mathcal{S}$, defined by the form (4.1.1). To estimate the true generative model $s_{\boldsymbol{\psi}_0}$, Stadler et al. (2010) proposed a Lasso-regularization based estimator, which consists of a minimiser of the penalized negative conditional log-likelihood that is defined by

$$\begin{aligned}
\widehat{s}^{\text{Lasso}}(\lambda) &= \underset{s_{\boldsymbol{\psi}} \in \mathcal{S}}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\psi}}\left(y_i|\mathbf{x}_i\right)\right) + \text{pen}_\lambda(\boldsymbol{\psi}) \right\}, \\
\text{pen}_\lambda(\boldsymbol{\psi}) &= \lambda \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} \left|\sigma_k^{-1}\beta_{kj}\right|, \lambda > 0, \boldsymbol{\psi} = (\pi_k, \boldsymbol{\beta}_k, \sigma_k)_{k\in[K]}.
\end{aligned} \tag{4.1.2}$$

For this estimator, the authors provided an l_0-oracle inequality, satisfied by $\widehat{s}^{\text{Lasso}}(\lambda)$, conditional on the restricted eigenvalue condition and margin condition, which leads to link the Kullback-Leibler loss function to the l_2-norm of the parameters.

Another direction of study regarding $\widehat{s}^{\text{Lasso}}(\lambda)$ is to look at its l_1-regularization properties; see, for example, Massart & Meynet (2011), Meynet (2013), and Devijver (2015a). As indicated by Devijver (2015a), contrary to results for the l_0 penalty, some results for the l_1 penalty are valid with no

assumptions, neither on the Gram matrix nor on the margin. However, such results can be achieved only at a rate of convergence of $1/n$, rather than at order $1/\sqrt{n}$.

In the framework of finite mixtures of Gaussian regression models, Meynet (2013) considered the case for a univariate response, and Devijver (2015a) extended these results to the case of a multivariate responses, *i.e.,* the Gaussian conditional PDF in (4.1.1) is replaced by a multivariate Gaussian PDF of the form $\phi(\cdot;\mu,\Sigma)$ with mean vector μ and a covariance matrix Σ. In particular, Devijver (2015a) considered an extension of the Lasso-estimator (4.1.2), with a regularization term defined by $\text{pen}_\lambda(\boldsymbol{\psi}) = \lambda \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q}\left|[\boldsymbol{\beta}_k]_{z,j}\right|$.

In Section 4.2, we shall extend such result for the finite mixture of Gaussian regressions models, which is considered as a special case of the MoE models, where only the mixture components depend on the features, to the more general mixture of Gaussian experts regression models with softmax gating functions, as defined in (4.2.1). Since each mixing proportion is modeled by a softmax function of the covariates, the dependence on each feature appears both in the experts PDFs and in the mixing proportion functions (gating functions), which allows us to capture more complex non-linear relationships between the response and predictors arising from different subpopulations, compared to the finite mixture of Gaussian regression models. This is demonstrated via numerical experiments in several articles such as Nguyen & Chamroukhi (2018), Chamroukhi & Huynh (2018), and Chamroukhi & Huynh (2019).

In the context of studying the statistical properties of the penalized maximum likelihood approach for MoE models with softmax gating functions, we may consider the prior works of Khalili (2010) and Montuelle et al. (2014). In Khalili (2010), for feature selection, two extra penalty terms are applied to the l_2-penalized conditional log-likelihood function. Their penalized conditional log-likelihood estimator is given by

$$\widehat{s}^{\text{PL}}(\boldsymbol{\lambda}) = \underset{s_{\boldsymbol{\psi}}\in\mathcal{S}}{\arg\min}\left\{-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_{\boldsymbol{\psi}}\left(y_i|\mathbf{x}_i\right)\right) + \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi})\right\}, \tag{4.1.3}$$

$$s_{\boldsymbol{\psi}}(y|\mathbf{x}) = \sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\phi\left(y;\beta_{k0}+\boldsymbol{\beta}_k^\top\mathbf{x},\sigma_k^2\right), \boldsymbol{\psi} = (\boldsymbol{\gamma}_k,\boldsymbol{\beta}_k,\sigma_k)_{k\in[K]}, \tag{4.1.4}$$

$$\text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K}\lambda_k^{[1]}\sum_{j=1}^{p}|\gamma_{kj}| + \sum_{k=1}^{K}\lambda_k^{[2]}\sum_{j=1}^{p}|\beta_{kj}| + \frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\|\boldsymbol{\gamma}_k\|_2^2, \tag{4.1.5}$$

where $\boldsymbol{\lambda} = \left(\lambda_1^{[1]},\ldots,\lambda_K^{[1]},\lambda_1^{[2]},\ldots,\lambda_K^{[2]},\frac{\lambda^{[3]}}{2}\right)$ is a vector of non-negative regularization parameters, $\mathcal{S}$ contains all functions of form (4.1.4), $\|\cdot\|_2^2$ is the Euclidean norm in $\mathbb{R}^p$, and

$$g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) = \frac{\exp\left(\gamma_{k0}+\boldsymbol{\gamma}_k^\top\mathbf{x}\right)}{\sum_{l=1}^{K}\exp\left(\gamma_{l0}+\boldsymbol{\gamma}_l^\top\mathbf{x}\right)}$$

is a softmax gating function. Note that the first two terms from (4.1.5) are the normal Lasso functions (l_1 penalty function), while the l_2 penalty function for the gating network is added to excessively wildly large estimates of the regression coefficients corresponding to the mixing proportions. This behavior can be observed in logistic/multinomial regression when the number of potential features is large and highly correlated (see *e.g.,* Park & Hastie, 2008 and Bunea et al., 2008). However, this also affects the sparsity of the regularization model, which is confirmed via the numerical experiments of Chamroukhi & Huynh (2018) and Chamroukhi & Huynh (2019).

By extending the theoretical developments for mixture of linear regression models in Khalili & Chen (2007), standard asymptotic theorems for MoE models are established in Khalili (2010). More precisely, under several strict regularity conditions on the true joint density function $s_{\boldsymbol{\psi}_0}(y,\mathbf{x})$ and the choice of tuning parameter $\boldsymbol{\lambda}$, the estimator of the true parameter vector $\widehat{\boldsymbol{\psi}}_n^{\text{PL}}(\boldsymbol{\lambda})$, defined via $\widehat{s}^{\text{PL}}(\boldsymbol{\lambda})$ from (4.1.3) but using the Scad penalty function from Fan & Li (2001), instead of Lasso, is proved to be both consistent in feature selection and maintains root-n consistency. Differing from Scad, for Lasso, the estimator $\widehat{\boldsymbol{\psi}}_n^{\text{PL}}(\boldsymbol{\lambda})$ cannot achieve both properties, simultaneously. In other words, Lasso is

consistent in feature selection but introduces bias to the estimators of the true nonzero coefficients. Therefore, our non-asymptotic result in Theorem 4.2.2 can be considered as a complement to such asymptotics for MoE regression models with softmax gating functions. To obtain our oracle inequality, Theorem 4.2.2, we shall restrict our study to the Lasso estimator without the l_2-norm. While studying the oracle inequality within the context of the $(l_1 + l_2)$-norm may also be interesting. It has been demonstrated, in Huynh & Chamroukhi (2019), that the regularized maximum-likelihood estimation of MoE models for generalized linear models, better encourages sparsity under the l_1-norm, compared to when using the $(l_1 + l_2)$-norm, which may affect sparsity. However, in Section 4.2, we shall not consider such approaches.

To the best of our knowledge, we are the first to study the l_1-regularization properties of the MoE regression models. The main contribution of Section 4.2 is a theoretical result: an oracle inequality, that provides the lower bound on the regularization parameter of the Lasso penalty that ensures such non-asymptotic theoretical control on the Kullback-Leibler loss of the Lasso estimator for SGaME models. More precisely, we assume that we have n observations, $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right) \in ([0,1]^p \times \mathbb{R}^q)$, where $\mathbf{x}_{[n]}$ are fixed values, coming from the unknown conditional mixture of Gaussian experts regression models $s_0 := s_{\boldsymbol{\psi}_0} \in \mathcal{S}$, cf. (4.2.6), and a Lasso estimator $\widehat{s}^{\text{Lasso}}(\lambda)$, defined as in (4.2.8), where $\lambda \geq 0$ is a regularization parameter to be tuned. Then, if $\lambda \geq \kappa \frac{c_1 \ln n}{\sqrt{n}}$ for some absolute constant $\kappa \geq 148$, the estimator $\widehat{s}^{\text{Lasso}}(\lambda)$ satisfies the following l_1-oracle inequality:

$$\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}^{\text{Lasso}}(\lambda)\right)\right] \leq \left(1+\kappa^{-1}\right) \inf_{s_{\boldsymbol{\psi}} \in \mathcal{S}} \left(\text{KL}_n\left(s_0, s_{\boldsymbol{\psi}}\right) + \lambda \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right) + \lambda + \sqrt{\frac{K}{n}} c_2 \ln n,$$

where we suppress the dependence of positive constants c_1, c_2 on p, q, K, and boundedness constants defining our conditional class $\mathcal{S}$, for the sake of ease of interpretation. For the explicit dependences, we refer the reader to Theorem 4.2.2.

Next, in Section 4.3, we focus on joint rank and variable selection by a non-asymptotic model selection via weak oracle inequalities. Note that, in Section 4.3, we proposed a Lasso+l2-Rank procedure for PSGaBloME regression models to deal with high-dimensional data. This allows us to study the model selection procedure instead of investigating the l_1-regularization properties for the Lasso estimator. Furthermore, the results from Section 4.3 extend the LinBoSGaME from Montuelle et al. (2014, Theorem 1) for handling high-dimensional data via selecting relevant variables and low-rank regression matrices. Note that, these weak oracle inequalities also extend the results of some simple MoE models from Devijver (2015b, 2017a), Devijver & Gallopin (2018) to one of the most general MoE models, namely LinBoSGaBloME models. A detailed comparison between our work and their results can be found in Section 1.2.11.

Note that our results in Sections 4.2 and 4.3 are non-asymptotic; *i.e.,* the number of observations n is fixed, while the number of predictors p and the dimension of the response q can grow, with respect to n, and can be much larger than n. Good discussions regarding non-asymptotic statistics are provided in Massart (2007) and Wainwright (2019).

The goal of Chapter 4 is to provide lower bounds of penalty functions that guarantee an l_1-oracle inequality or a weak oracle inequality for softmax-gated MoE models, which is particularly useful for nonlinear regression models for high-dimensional heterogeneous data. As such, the remainder of Chapter 4 progresses as follows. In Sections 4.2.1 and 4.3.1.1, we discuss constructions and frameworks of SGaME and PSGaBloME regression models, respectively. Then, we establish an l_1-oracle inequality satisfied by the Lasso estimator in SGaME models in Section 4.2.2. Furthermore, a weak oracle type inequality satisfied by PMLEs is constructed for PSGaBloME regression models in Section 4.3.2. Next, Sections 4.2.3 and 4.3.3 are devoted to the corresponding proofs of these main results. The proof of technical lemmas and additional technical results can be found in Sections 4.2.4, 4.2.5 and 4.3.4, respectively.

4.2 An l_1-oracle inequality for the Lasso estimator in the softmax-gated mixture of experts regression models

4.2.1 Notation and framework

4.2.1.1 SGaME models

We consider the statistical framework in which we model a sample of high-dimensional regression data generated from a heterogeneous population via the mixtures of Gaussian experts regression models with Gaussian gating functions, named softmax-gated MoE (SGaME) regression models. We observe n independent couples $((\mathbf{x}_i, \mathbf{y}_i))_{i\in[n]} \in (\mathcal{X} \times \mathbb{R}^q)^n \subset (\mathbb{R}^p \times \mathbb{R}^q)^n \, (p, q, n \in \mathbb{N}^\star)$, where typically $p \gg n$, $\mathbf{x}_i$ is fixed and $\mathbf{y}_i$ is a realization of the random variable $\mathbf{Y}_i$, for all $i \in [n]$. We assume that $\mathcal{X}$ is a compact set of $\mathbb{R}^p$. We also assume that the response variable $\mathbf{Y}_i$ depends on the set of explanatory variables (covariates) through a regression-type model. The conditional probability density function (PDF) of the model is approximated by SGaME models.

More precisely, we assume that, conditionally to the $\{\mathbf{x}_i\}_{i\in[n]}$, $\{\mathbf{Y}_i\}_{i\in[n]}$ are independent and identically distributed with conditional density $s_0\left(\cdot|\mathbf{x}_i\right)$, which is approximated by a MoE model. Our goal is to estimate this conditional density function s_0 from the observations.

For any $K \in \mathbb{N}^\star$, the K-component MoE model can be defined as

$$\mathrm{MoE}\left(\mathbf{y}|\mathbf{x};\theta\right) = \sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) f_k\left(\mathbf{y}|\mathbf{x};\boldsymbol{\eta}\right),$$

where $g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) > 0$ and $\sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) = 1$, and $f_k\left(\mathbf{y}|\mathbf{x};\boldsymbol{\eta}\right)$ is a conditional PDF (cf. Nguyen & Chamroukhi, 2018). In our proposal, we consider the MoE model of Jordan & Jacobs (1994), which extended the original MoE from Jacobs et al. (1991), for a regression model. More precisely, we utilize the following mixtures of Gaussian experts regression models with softmax gating functions:

$$s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) \phi\left(\mathbf{y};\boldsymbol{\upsilon}_k(\mathbf{x}),\boldsymbol{\Sigma}_k\right), \tag{4.2.1}$$

to estimate s_0, where given any $k \in [K]$, $\phi\left(\cdot;\boldsymbol{\upsilon}_k,\boldsymbol{\Sigma}_k\right)$ is the multivariate Gaussian density with mean $\boldsymbol{\upsilon}_k$, which is a function of $\mathbf{x}$ that specifies the mean of the kth component, and with covariance matrix $\boldsymbol{\Sigma}_k$. Here, $(\boldsymbol{\upsilon},\boldsymbol{\Sigma}) := \left((\boldsymbol{\upsilon}_1,\ldots,\boldsymbol{\upsilon}_K),(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K)\right) \in (\boldsymbol{\Upsilon} \times \boldsymbol{V})$, where $\boldsymbol{\Upsilon}$ is a set of K-tuples of mean functions from $\mathcal{X}$ to $\mathbb{R}^q$ and $\boldsymbol{V}$ is a sets of K-tuples of symmetric positive definite matrices on $\mathbb{R}^q$, and the softmax gating function $g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)$ is defined as in (4.2.2):

$$g_k\left(\mathbf{x};\boldsymbol{\gamma}\right) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)}, w_k(\mathbf{x}) = \gamma_{k0} + \boldsymbol{\gamma}_k^\top\mathbf{x}, \boldsymbol{\gamma} = \left(\gamma_{k0},\boldsymbol{\gamma}_k^\top\right)_{k\in[K]} \in \boldsymbol{\Gamma} = \mathbb{R}^{(p+1)K}. \tag{4.2.2}$$

We shall define the parameter vector $\boldsymbol{\psi}$ in the sequel.

Fixed predictors and number of components with linear mean functions

Inspired by the framework in Meynet (2013) and Devijver (2015a), the explanatory variables $\mathbf{x}_i$ and the number of components $K \in \mathbb{N}^\star$ are both fixed. We assume that the observed $\mathbf{x}_i, i \in [n]$, are finite. Without loss of generality, we choose to rescale $\mathbf{x}$, so that $\|\mathbf{x}\|_\infty \leq 1$. Therefore, we can assume that the explanatory variables $\mathbf{x}_i \in \mathcal{X} = [0,1]^p$, for all $i \in [n]$. Note that such a restriction is also used in Devijver (2015a). Under only the assumption of bounded parameters, we provide a lower bound on the Lasso regularization parameter λ, which guarantees an oracle inequality. Note that in this non-random explanatory variables setting, we focus on the Lasso for its l_1-regularization properties rather than as a model selection procedure, as in the case of random explanatory variables and unknown K, as in Montuelle et al. (2014).

For simplicity, we consider the case where the means of Gaussian experts are linear functions of the explanatory variables; *i.e.,*

$$\boldsymbol{\Upsilon} = \left\{\boldsymbol{\upsilon} : \mathcal{X} \mapsto \boldsymbol{\upsilon}_{\boldsymbol{\beta}}(\mathbf{x}) := \left(\boldsymbol{\beta}_{k0} + \boldsymbol{\beta}_k\mathbf{x}\right)_{k\in[K]} \in \left(\mathbb{R}^q\right)^K \,\middle|\, \boldsymbol{\beta} = \left(\boldsymbol{\beta}_{k0},\boldsymbol{\beta}_k\right)_{k\in[K]} \in \mathcal{B} = \left(\mathbb{R}^{q\times(p+1)}\right)^K\right\},$$

where $\boldsymbol{\beta}_{k0}$ and $\boldsymbol{\beta}_k$ are respectively the $q \times 1$ vector of bias and the $q \times p$ regression coefficients matrix for the kth expert.

In summary, we wish to estimate s_0 via conditional densities belonging to the class:

$$\{(\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) \,|\, \boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \boldsymbol{\Psi}\}, \tag{4.2.3}$$

where $\boldsymbol{\Psi} = \boldsymbol{\Gamma} \times \boldsymbol{\Xi}$, and $\boldsymbol{\Xi} = \mathcal{B} \times \boldsymbol{V}$.

From hereon in, for a vector $\mathbf{x} \in \mathbb{R}^p$, we assume that $\mathbf{x} = (x_1, \ldots, x_p)$ is in the column form. Similarly, the parameter of the entire model, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, is also a column vector, where we consider any matrix as a vector produced using vec$(\cdot)$: the vectorization operator that stacks the columns of a matrix into a vector.

Boundedness assumption on the softmax gating and Gaussian parameters

We shall restrict our study to estimate s_0 by conditional PDFs belonging to the model class $\mathcal{S}$, which has boundedness assumptions on the softmax gating and Gaussian expert parameters. Specifically, we assume that there exists deterministic constants $A_{\boldsymbol{\gamma}}, A_{\boldsymbol{\beta}}, a_{\boldsymbol{\Sigma}}, A_{\boldsymbol{\Sigma}} > 0$, such that $\boldsymbol{\psi} \in \widetilde{\boldsymbol{\Psi}}$, where

$$\begin{aligned}
\widetilde{\boldsymbol{\Gamma}} &= \left\{ \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \mid \forall k \in [K], \sup_{\mathbf{x} \in \mathcal{X}} \left(|\gamma_{k0}| + \left| \boldsymbol{\gamma}_k^\top \mathbf{x} \right| \right) \leq A_{\boldsymbol{\gamma}} \right\}, \\
\widetilde{\boldsymbol{\Xi}} &= \left\{ \xi \in \boldsymbol{\Xi} \mid \forall k \in [K], \max_{z \in \{1, \ldots, q\}} \sup_{\mathbf{x} \in \mathcal{X}} \left(|[\boldsymbol{\beta}_{k0}]_z| + |[\boldsymbol{\beta}_k \mathbf{x}]_z| \right) \leq A_{\boldsymbol{\beta}}, a_{\boldsymbol{\Sigma}} \leq m\left(\boldsymbol{\Sigma}_k^{-1}\right) \leq M\left(\boldsymbol{\Sigma}_k^{-1}\right) \leq A_{\boldsymbol{\Sigma}} \right\}, \\
\widetilde{\boldsymbol{\Psi}} &= \widetilde{\boldsymbol{\Gamma}} \times \widetilde{\boldsymbol{\Xi}}.
\end{aligned} \tag{4.2.4}$$

Since

$$a_G := \frac{\exp\left(-A_{\boldsymbol{\gamma}}\right)}{\sum_{l=1}^K \exp\left(A_{\boldsymbol{\gamma}}\right)} \leq \sup_{\mathbf{x} \in \mathcal{X}, \boldsymbol{\gamma} \in \widetilde{\boldsymbol{\Gamma}}} \frac{\exp\left(\gamma_{k0} + \boldsymbol{\gamma}_k^\top \mathbf{x}\right)}{\sum_{l=1}^K \exp\left(\gamma_{l0} + \boldsymbol{\gamma}_l^\top \mathbf{x}\right)} \leq \frac{\exp\left(A_{\boldsymbol{\gamma}}\right)}{\sum_{l=1}^K \exp\left(-A_{\boldsymbol{\gamma}}\right)} =: A_G,$$

there exists deterministic positive constants a_G, A_G, such that

$$a_G \leq \sup_{\mathbf{x} \in \mathcal{X}, \boldsymbol{\gamma} \in \widetilde{\boldsymbol{\Gamma}}} g_k\left(\mathbf{x}; \boldsymbol{\gamma}\right) \leq A_G. \tag{4.2.5}$$

We wish to use the model class $\mathcal{S}$ of conditional PDFs to estimate s_0, where

$$\mathcal{S} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) \,\middle|\, \boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Psi}} \right\}. \tag{4.2.6}$$

To simplify the proofs, we shall assume that the true density s_0 belongs to $\mathcal{S}$. That is to say, there exists $\boldsymbol{\psi}_0 = (\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \in \widetilde{\boldsymbol{\Psi}}$, such that $s_0 = s_{\boldsymbol{\psi}_0}$.

4.2.1.2 Losses and the penalized maximum likelihood estimator

In maximum likelihood estimation, we consider the Kullback-Leibler information as the loss function, which is defined for densities s and t by

$$\mathrm{KL}(s, t) = \begin{cases} \int_{\mathbb{R}^q} \ln\left(\frac{s(\mathbf{y})}{t(\mathbf{y})}\right) s(\mathbf{y}) d\mathbf{y} & \text{if } sdy \text{ is absolutely continuous with respect to } tdy, \\ +\infty & \text{otherwise.} \end{cases}$$

Since we are working with conditional PDFs and not with classical densities, we define the following adapted Kullback-Leibler information that takes into account the structure of conditional PDFs. For fixed explanatory variables $(\mathbf{x}_i)_{1 \leq i \leq n}$, we consider the average loss function

$$\mathrm{KL}_n(s, t) = \frac{1}{n} \sum_{i=1}^n \mathrm{KL}\left(s\left(\cdot|\mathbf{x}_i\right), t\left(\cdot|\mathbf{x}_i\right)\right) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^q} \ln\left(\frac{s\left(\mathbf{y}|\mathbf{x}_i\right)}{t\left(\mathbf{y}|\mathbf{x}_i\right)}\right) s\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y}. \tag{4.2.7}$$

The maximum likelihood estimation approach suggests to estimate s_0 by the conditional PDF $s_{\boldsymbol{\psi}}$ that maximizes the likelihood, conditioned on $(\mathbf{x}_i)_{1\leq i\leq n}$, defined as

$$\ln\left(\prod_{i=1}^{n} s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) = \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right).$$

Or equivalently, that minimizes the empirical contrast:

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right).$$

However, since we want to handle high-dimensional data, we have to regularize the maximum likelihood estimator (MLE) in order to obtain reasonable estimates. Here, we shall consider l_1-regularization and the associated so-called Lasso estimator, which is the l_1-norm penalized MLE and is defined as in (4.1.2) but with a modified penalty function as follows:

$$\widehat{s}^{\text{Lasso}}(\lambda) := \underset{s_{\boldsymbol{\psi}}\in\mathcal{S}}{\arg\min}\left\{-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}_{\lambda}(\boldsymbol{\psi})\right\}, \tag{4.2.8}$$

where $\lambda \geq 0$ is a regularization parameter to be tuned, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ and

$$\text{pen}_{\lambda}(\boldsymbol{\psi}) = \lambda\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 := \lambda\left(\left\|\boldsymbol{\psi}^{[1]}\right\|_1 + \left\|\boldsymbol{\psi}^{[2]}\right\|_1\right), \tag{4.2.9}$$

$$\left\|\boldsymbol{\psi}^{[1]}\right\|_1 = \|\boldsymbol{\gamma}\|_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}|\gamma_{kj}|, \tag{4.2.10}$$

$$\left\|\boldsymbol{\psi}^{[2]}\right\|_1 = \|\text{vec}(\boldsymbol{\beta})\|_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q}\left|[\boldsymbol{\beta}_k]_{z,j}\right|. \tag{4.2.11}$$

From now on, we denote $\|\boldsymbol{\beta}\|_p$ ($p \in \{1, 2, \infty\}$) by the induced p-norm of a matrix; see Definition 4.2.18, which differs from $\|\text{vec}(\boldsymbol{\beta})\|_p$.

Note that $\text{pen}_{\lambda}(\boldsymbol{\psi})$ is a Lasso regularization term encouraging sparsity for both the gating and expert parameters. Recall that this penalty is also studied in Khalili (2010), Chamroukhi & Huynh (2018), and Chamroukhi & Huynh (2019), in which the authors studied the univariate case: $\mathbf{Y} \in \mathbb{R}$. Notice that, without considering the l_2-norm, the penalty function considered in (4.1.5) belongs to our framework and the l_1-oracle inequality from Theorem 4.2.2 can be obtained for it. Indeed, by considering $\lambda = \min\left\{\lambda_1^{[1]}, \ldots, \lambda_K^{[1]}, \lambda_1^{[2]}, \ldots, \lambda_K^{[2]}, \frac{\lambda^{[3]}}{2}\right\}$, the condition for a regularization parameter's lower bound, (4.2.13) from Theorem 4.2.2, can also be applied to model (4.1.3), which leads to an l_1-oracle inequality.

4.2.2 An l_1-oracle inequality for the Lasso estimator

In this section, we state Theorem 4.2.2, which is proved in Section 4.2.3.3. This result provides an l_1-oracle inequality for the Lasso estimator for mixtures of Gaussian experts regression models with softmax gating functions. It is the primary contribution of this article and is motivated by the problem studied in Meynet (2013) and Devijver (2015a).

Firstly, we aim to prove that the negative of differential entropy (see its definition, *e.g.,* from Mansuripur (1987, Chapter 9)) of the true unknown conditional density $s_0 \in S$, defined in (4.2.1), is finite, see more in Lemma 4.2.1, which is proved in Section 4.2.4.3.

Lemma 4.2.1 (Differential entropy of SGaME regression model with boundedness assumptions on parameter spaces)**.** *There exist a nonnegative constant* $H_{s_0} = \max\{0, \ln C_{s_0}\}$, $C_{s_0} = (2\pi)^{-q/2}\left(2A_{\Sigma}^{-1}\right)^{-q/2}$, *such that*

$$\max\left\{0, \sup_{x\in\mathcal{X}}\int_{\mathbb{R}^q} \ln\left(s_0\left(y|x\right)\right) s_0\left(y|x\right) dy\right\} \leq H_{s_0} < \infty. \tag{4.2.12}$$

Theorem 4.2.2 (l_1-oracle inequality from Nguyen et al., 2020c, Theorem 3.1)**.** *We observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right) \in ([0,1]^p \times \mathbb{R}^q)$*, coming from the unknown conditional mixture of Gaussian experts regression models* $s_0 := s_{\boldsymbol{\psi}_0} \in \mathcal{S}$*, cf.* (4.2.6)*. We define the Lasso estimator* $\widehat{s}^{\text{Lasso}}(\lambda)$*, by* (4.2.8)*, where* $\lambda \geq 0$ *is a regularization parameter to be tuned. Then, if*

$$\lambda \geq \kappa \frac{K B'_n}{\sqrt{n}} \left(q \ln n \sqrt{\ln(2p+1)} + 1\right), \tag{4.2.13}$$

$$B'_n = \max\left(A_{\boldsymbol{\Sigma}}, 1 + K A_G\right)\left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right), \tag{4.2.14}$$

for some absolute constants $\kappa \geq 148$*, the estimator* $\widehat{s}^{\text{Lasso}}(\lambda)$ *satisfies the following* l_1*-oracle inequality:*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n\left(s_0, \widehat{s}^{\text{Lasso}}(\lambda)\right)\right] \leq\ & \left(1 + \kappa^{-1}\right) \inf_{s_{\boldsymbol{\psi}} \in \mathcal{S}} \left(\mathrm{KL}_n\left(s_0, s_{\boldsymbol{\psi}}\right) + \lambda \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right) + \lambda \\
& + \sqrt{\frac{K}{n}} \left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}} + H_{s_0}\right) \sqrt{2qA_{\boldsymbol{\gamma}}} \\
& + 302q\sqrt{\frac{K}{n}} \max\left(A_{\boldsymbol{\Sigma}}, 1 + K A_G\right)\left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right) \\
& \times K \left(1 + \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right).
\end{aligned} \tag{4.2.15}$$

Next, we state the following Theorem 4.2.3, which is an l_1-ball MoE regression model selection theorem for l_1-penalized maximum conditional likelihood estimation in the Gaussian mixture framework. Note that Theorem 4.2.2 is an immediate consequence of Theorem 4.2.3.

Theorem 4.2.3. *Assume that we observe* $\left(\left(\mathbf{x}_i, \mathbf{y}_i\right)\right)_{i \in [n]}$ *with unknown conditional Gaussian mixture PDF* s_0*. For all* $m \in \mathbb{N}^\star$*, consider the* l_1*-ball*

$$S_m = \left\{s_{\boldsymbol{\psi}} \in \mathcal{S}, \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 \leq m\right\}, \tag{4.2.16}$$

and let $\widehat{s}_m$ *be a* η_m*-*ln*-likelihood minimizer in* S_m *for some* $\eta_m \geq 0$*:*

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) \leq \inf_{s_m \in S_m} \left(-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)\right) + \eta_m. \tag{4.2.17}$$

Assume that, for all $m \in \mathbb{N}^\star$*, the penalty function satisfies* $\mathrm{pen}(m) = \lambda m$*, where* λ *is defined later. Then, we define the penalized likelihood estimator* $\widehat{s}_{\widehat{m}}$*, where* $\widehat{m}$ *is defined via the satisfaction of the inequality*

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_{\widehat{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \mathrm{pen}(\widehat{m}) \leq \inf_{m \in \mathbb{N}^\star} \left(-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \mathrm{pen}(m)\right) + \eta, \tag{4.2.18}$$

for some $\eta \geq 0$*. Then, if*

$$\lambda \geq \kappa \frac{K B'_n}{\sqrt{n}} \left(q \ln n \sqrt{\ln(2p+1)} + 1\right), \tag{4.2.19}$$

$$B'_n = \max\left(A_{\boldsymbol{\Sigma}}, 1 + K A_G\right)\left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right), \tag{4.2.20}$$

for some absolute constants $\kappa \geq 148$*, then*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\right] \leq\ & \left(1 + \kappa^{-1}\right) \inf_{m \in \mathbb{N}^\star} \left(\inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \eta_m\right) + \eta \\
& + \sqrt{\frac{K}{n}} \left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}} + H_{s_0}\right) \sqrt{2qA_{\boldsymbol{\gamma}}} \\
& + 302q\sqrt{\frac{K}{n}} \max\left(A_{\boldsymbol{\Sigma}}, 1 + K A_G\right)\left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right) \\
& \times K \left(1 + \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right).
\end{aligned} \tag{4.2.21}$$

Theorem 4.2.2 provides information about the performance of the Lasso as an l_1 regularization estimator for mixtures of Gaussian experts regression models. If the regularization parameter λ is properly chosen, the Lasso estimator, which is the solution of the l_1-penalized empirical risk minimization problem, behaves in a comparable manner to the deterministic Lasso (or so-called *oracle*), which is the solution of the l_1-penalized true risk minimization problem, up to an error term of order λ. Note that the best model defined via

$$\inf_{s_{\boldsymbol{\psi}} \in \mathcal{S}} \left(\mathrm{KL}_n \left(s_0, s_{\boldsymbol{\psi}} \right) + \lambda \left\| \boldsymbol{\psi}^{[1,2]} \right\|_1 \right) \tag{4.2.22}$$

is the one with the smallest l_1-penalized true risk. However, since we do not have access to the true density s_0, we can not select that best model, which we call the *oracle*. The oracle is by definition the model belonging to the collection that minimizes the l_1-penalized risk (4.2.22), which is generally assumed unknown.

To the best of our knowledge, Theorem 2.8 from Maugis-Rabusseau & Michel (2013) is the only result in the literature which studies the minimax estimator for Gaussian mixture models. We believe that the establishment of such an extension to MoE regression models is not trivial. We hope to overcome such a challenge in the future.

4.2.3 Proof of the oracle inequality

Sketch of the proof. Motivated by the idea from Meynet (2013) and Devijver (2015a), we study the Lasso as the solution of a penalized maximum likelihood model selection procedure over countable collections of models in an l_1-ball. Therefore, the main Theorem 4.2.2 is an immediate consequence of Theorem 4.2.3, which is an l_1-ball MoE regression model selection theorem for l_1-penalized maximum conditional likelihood estimation in the Gaussian mixture framework. The proof of Theorem 4.2.3 can be deduced from Proposition 4.2.4 and Proposition 4.2.5, which address the cases for large and small values of $\mathbf{Y}$. Proposition 4.2.4 constitutes our main technical contribution. Its proof follows the arguments developed in the proof of a more general model selection theorem for maximum likelihood estimators: Massart (2007, Theorem 7.11). More precisely, the proof of Proposition 4.2.4 is in the spirit of Vapnik's method of structural risk minimization, which is established initially in Vapnik (1982) and briefly summarized in Section 8.2 in Massart (2007). In particularly, to obtain an upper bound of the empirical process in expectation, we shall use concentration inequalities combined with symmetrization arguments.

4.2.3.1 Main propositions used in this proof

Theorem 4.2.3 can be deduced from the two following propositions, which address the cases for large and small values of $\mathbf{Y}$.

Proposition 4.2.4. *Assume that we observe* $((\mathbf{x}_i, \mathbf{y}_i))_{i \in [n]}$, *with unknown conditional PDF* s_0. *Let* $M_n > 0$ *and consider the event*

$$\mathcal{T} = \left\{ \max_{i=1,\dots,n} \left\| \mathbf{Y}_i \right\|_\infty = \max_{i=1,\dots,n} \max_{z \in \{1,\dots,q\}} \left| [\mathbf{Y}_i]_z \right| \leq M_n \right\}.$$

For all $m \in \mathbb{N}^\star$, *consider the* l_1*-ball*

$$S_m = \left\{ s_{\boldsymbol{\psi}} \in \mathcal{S}, \left\| \boldsymbol{\psi}^{[1,2]} \right\|_1 \leq m \right\}$$

and let $\widehat{s}_m$ *be a* η_m*-*ln*-likelihood minimizer in* S_m, *for some* $\eta_m \geq 0$*:*

$$-\frac{1}{n} \sum_{i=1}^{n} \ln \left(\widehat{s}_m \left(\mathbf{y}_i | \mathbf{x}_i \right) \right) \leq \inf_{s_m \in S_m} \left(-\frac{1}{n} \sum_{i=1}^{n} \ln \left(s_m \left(\mathbf{y}_i | \mathbf{x}_i \right) \right) \right) + \eta_m.$$

Assume that for all $m \in \mathbb{N}^\star$, *the penalty function satisfies* $\text{pen}(m) = \lambda m$, *where* λ *is defined later. Then, we define the penalized likelihood estimator* $\widehat{s}_{\widehat{m}}$ *with* $\widehat{m}$ *defined via the inequality*

$$-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_{\widehat{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^\star}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(m)\right)+\eta, \tag{4.2.23}$$

for some $\eta \geq 0$. *Then, if*

$$\lambda \geq \kappa\frac{KB_n}{\sqrt{n}}\left(q\ln n\sqrt{\ln(2p+1)}+1\right),$$
$$B_n = \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+q\sqrt{q}\left(M_n+A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right),$$

for some absolute constants $\kappa \geq 148$, *then*

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}}\right] \leq \left(1+\kappa^{-1}\right)\inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\text{KL}_n\left(s_0,s_m\right)+\text{pen}(m)+\eta_m\right)\\ +\frac{302K^{3/2}qB_n}{\sqrt{n}}\left(1+\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right)+\eta.\end{aligned} \tag{4.2.24}$$

Proposition 4.2.5. *Consider* $s_0, \mathcal{T}$, *and* $\widehat{s}_m$ *as defined in Proposition 4.2.4. Denote by* $\mathcal{T}^C$ *the complement of* $\mathcal{T}$,i.e.,

$$\mathcal{T}^C = \left\{\max_{i=1,\ldots,n}\left\|\mathbf{Y}_i\right\|_\infty = \max_{i=1,\ldots,n}\max_{z\in\{1,\ldots,q\}}\left|\left[\mathbf{Y}_i\right]_z\right| > M_n\right\}.$$

Then,

$$\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}^C}\right] \leq \left(\frac{e^{q/2-1}\pi^{q/2}}{A_\Sigma^{q/2}}+H_{s_0}\right)\sqrt{2Knqa_\gamma}e^{-\frac{M_n^2-2M_nA_\beta}{4A_\Sigma}}.$$

Theorem 4.2.3, and Propositions 4.2.4 and 4.2.5 are proved in Sections 4.2.3.4 to 4.2.3.6, respectively.

4.2.3.2 Additional notation

We first introduce some definitions and notations that we shall use in the proofs. For any measurable function $f : \mathbb{R} \to \mathbb{R}$, consider its empirical norm

$$\left\|f\right\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}f^2\left(\mathbf{y}_i|\mathbf{x}_i\right)},$$

and its conditional expectation

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[f\right] := \mathbb{E}\left[f\left(\mathbf{Y}|\mathbf{X}\right)|\mathbf{X}=\mathbf{x}\right] = \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{x})s_0(\mathbf{y}|\mathbf{x})d\mathbf{y},$$

as well as its empirical process

$$P_n(f) := \frac{1}{n}\sum_{i=1}^{n}f\left(\mathbf{Y}_i|\mathbf{x}_i\right), \tag{4.2.25}$$

with expectation

$$P(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f\left(\mathbf{Y}_i|\mathbf{x}_i\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^q}f\left(\mathbf{y}|\mathbf{x}_i\right)s_0\left(\mathbf{y}|\mathbf{x}_i\right)d\mathbf{y} \tag{4.2.26}$$

and the recentered process

$$\nu_n(f) := P_n(f) - P(f) = \frac{1}{n}\sum_{i=1}^{n}\left[f\left(\mathbf{y}_i|\mathbf{x}_i\right) - \int_{\mathbb{R}^q} f\left(\mathbf{y}|\mathbf{x}_i\right) s_0\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y}\right]. \tag{4.2.27}$$

For all $m \in \mathbb{N}^\star$, recall that we consider the model

$$S_m = \left\{s_{\boldsymbol{\psi}} \in \mathcal{S}, \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 \leq m\right\},$$

and define

$$F_m = \left\{f_m = -\ln\left(\frac{s_m}{s_0}\right) = \ln(s_0) - \ln(s_m), s_m \in S_m\right\}. \tag{4.2.28}$$

By using the basic properties of the infimum: for every $\epsilon > 0$, there exists $x_\epsilon \in A$, such that $x_\epsilon < \inf A + \epsilon$. Then let $\delta_{\mathrm{KL}} > 0$ for all $m \in \mathbb{N}^\star$, and let $\eta_m \geq 0$. It holds that there exist two functions $\widehat{s}_m$ and $\overline{s}_m$ in S_m, such that

$$P_n\left(-\ln \widehat{s}_m\right) \leq \inf_{s_m \in S_m} P_n\left(-\ln s_m\right) + \eta_m, \text{and} \tag{4.2.29}$$

$$\mathrm{KL}_n\left(s_0, \overline{s}_m\right) \leq \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \delta_{\mathrm{KL}}. \tag{4.2.30}$$

Define

$$\widehat{f}_m := -\ln\left(\frac{\widehat{s}_m}{s_0}\right), \text{and } \overline{f}_m := -\ln\left(\frac{\overline{s}_m}{s_0}\right). \tag{4.2.31}$$

Let $\eta \geq 0$ and fix $m \in \mathbb{N}^\star$. Further, define

$$\widehat{\mathcal{M}}(m) = \left\{m' \in \mathbb{N}^\star | P_n\left(-\ln \widehat{s}_{m'}\right) + \mathrm{pen}(m') \leq P_n\left(-\ln \widehat{s}_m\right) + \mathrm{pen}(m) + \eta\right\}. \tag{4.2.32}$$

4.2.3.3 Proof of Theorem 4.2.2

Let $\lambda > 0$ and define $\widehat{m}$ to be the smallest integer such that $\widehat{s}^{\mathrm{Lasso}}(\lambda)$ belongs to $S_{\widehat{m}}$, *i.e.*, $\widehat{m} := \left\lceil \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 \right\rceil \leq \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 + 1$. Then using the definition of $\widehat{m}$, (4.2.8), (4.2.16), and $\mathcal{S} = \bigcup_{m\in\mathbb{N}^\star} S_m$, we get

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}^{\mathrm{Lasso}}(\lambda)\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda\widehat{m} \\
&\leq -\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}^{\mathrm{Lasso}}(\lambda)\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda\left(\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1 + 1\right) \\
&= \inf_{s_{\boldsymbol{\psi}}\in\mathcal{S}}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right) + \lambda \\
&= \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_{\boldsymbol{\psi}}\in S_m}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right)\right) + \lambda \\
&= \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_{\boldsymbol{\psi}}\in\mathcal{S}, \left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\leq m}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right)\right) + \lambda \\
&\leq \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(s_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \lambda m\right)\right) + \lambda,
\end{aligned}$$

which implies

$$-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}^{\mathrm{Lasso}}(\lambda)\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \mathrm{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^\star}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \mathrm{pen}(m)\right) + \eta$$

with $\text{pen}(m) = \lambda m, \eta = \lambda$, and $\widehat{s}_m$ is a η_m-ln-likelihood minimizer in S_m, with $\eta_m \geq 0$ defined by (4.2.17). Thus, $\widehat{s}^{\text{Lasso}}(\lambda)$ satisfies (4.2.18) with $\widehat{s}^{\text{Lasso}}(\lambda) \equiv \widehat{s}_{\widehat{m}}$, *i.e.*,

$$-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_{\widehat{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^\star}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(m)\right)+\eta. \tag{4.2.33}$$

Then, Theorem 4.2.3 implies that if

$$\begin{aligned}\lambda &\geq \kappa\frac{KB'_n}{\sqrt{n}}\left(q\ln n\sqrt{\ln(2p+1)}+1\right),\\ B'_n &= \max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2+4A_{\boldsymbol{\Sigma}}\ln n\right)\right),\end{aligned}$$

for some absolute constants $\kappa \geq 148$, it holds that

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}^{\text{Lasso}}(\lambda)\right)\right] \leq\ & \left(1+\kappa^{-1}\right)\inf_{s_{\boldsymbol{\psi}}\in\mathcal{S}}\left(\text{KL}_n\left(s_0, s_{\boldsymbol{\psi}}\right)+\lambda\left\|\boldsymbol{\psi}^{[1,2]}\right\|_1\right)+\lambda\\ &+\sqrt{\frac{K}{n}}\left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\Sigma}^{q/2}}+H_{s_0}\right)\sqrt{2qA_{\boldsymbol{\gamma}}}\\ &+302q\sqrt{\frac{K}{n}}\max\left(A_{\boldsymbol{\Sigma}}, 1+KA_G\right)\left(1+2q\sqrt{q}A_{\boldsymbol{\Sigma}}\left(5A_{\boldsymbol{\beta}}^2+4A_{\boldsymbol{\Sigma}}\ln n\right)\right)\\ &\times K\left(1+\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right),\end{aligned}$$

as required.

4.2.3.4 Proof of Theorem 4.2.3

Let $M_n > 0$ and $\kappa \geq 148$. Assume that, for all $m \in \mathbb{N}^\star$, the penalty function satisfies $\text{pen}(m) = \lambda m$, with

$$\lambda \geq \kappa\frac{KB_n}{\sqrt{n}}\left(q\ln n\sqrt{\ln(2p+1)}+1\right). \tag{4.2.34}$$

We derive, from Propositions 4.2.4 and 4.2.5, that any penalized likelihood estimator $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$, satisfying

$$-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_{\widehat{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^\star}\left(-\frac{1}{n}\sum_{i=1}^{n}\ln\left(\widehat{s}_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)+\text{pen}(m)\right)+\eta,$$

for some $\eta \geq 0$, yields

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\right] =& \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}}\right]+\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\text{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}^c}\right]\\ \leq\ & \left(1+\kappa^{-1}\right)\inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\text{KL}_n\left(s_0, s_m\right)+\text{pen}(m)+\eta_m\right)\\ &+\frac{302K^{3/2}qB_n}{\sqrt{n}}\left(1+\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right)+\eta\\ &+\left(\frac{e^{q/2-1}\pi^{q/2}}{A_{\Sigma}^{q/2}}+H_{s_0}\right)\sqrt{2KnqA_{\boldsymbol{\gamma}}}e^{-\frac{M_n^2-2M_nA_\beta}{4A_\Sigma}}.\end{aligned} \tag{4.2.35}$$

To obtain inequality (4.2.21), it only remains to optimize the inequality (4.2.35), with respect M_n. Since the two terms depending on M_n, in (4.2.35), have opposite monotonicity with respect to M_n, we are looking for a value of M_n such that these two terms are of the same order with respect to

n. Consider the positive solution $M_n = A_{\boldsymbol{\beta}} + \sqrt{A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n}$ of the equation $\frac{X\left(X-2A_{\boldsymbol{\beta}}\right)}{4A_{\boldsymbol{\Sigma}}} - \ln n = 0$. Then, on the one hand,

$$e^{-\frac{M_n^2 - 2M_n A_{\boldsymbol{\beta}}}{4A_{\boldsymbol{\Sigma}}}} \sqrt{n} = e^{-\ln n} \sqrt{n} = \frac{1}{\sqrt{n}}.$$

On the other hand, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\begin{aligned} B_n &= \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right) \left(1 + q\sqrt{q}\left(M_n + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right) \\ &= \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right) \left(1 + q\sqrt{q} A_{\boldsymbol{\Sigma}} \left(2A_{\boldsymbol{\beta}} + \sqrt{A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n}\right)^2\right) \\ &\leq \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right) \left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}} \left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right), \end{aligned}$$

hence (4.2.35) implies (4.2.21). Indeed, it holds that

$$\begin{aligned} \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\right] \leq\ & \left(1 + \kappa^{-1}\right) \inf_{m \in \mathbb{N}^\star} \left(\inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \eta_m \right) + \eta \\ &+ \sqrt{\frac{K}{n}} \frac{e^{\frac{q}{2}-1} \pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}} \sqrt{2qA_{\boldsymbol{\gamma}}} \\ &+ 302q\sqrt{\frac{K}{n}} \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right) \left(1 + 2q\sqrt{q} A_{\boldsymbol{\Sigma}} \left(5A_{\boldsymbol{\beta}}^2 + 4A_{\boldsymbol{\Sigma}} \ln n\right)\right) \\ &\times K \left(1 + \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right). \end{aligned} \tag{4.2.36}$$

4.2.3.5 Proof of Proposition 4.2.4

For every $m' \in \widehat{\mathcal{M}}(m)$, from (4.2.32), (4.2.31), and (4.2.29), we obtain

$$\begin{aligned} P_n\left(\widehat{f}_{m'}\right) + \mathrm{pen}(m') &= P_n\left(\ln(s_0) - \ln\left(\widehat{s}_{m'}\right)\right) + \mathrm{pen}(m') \text{ (using (4.2.31))} \\ &\leq P_n\left(\ln(s_0) - \ln\left(\widehat{s}_m\right)\right) + \mathrm{pen}(m) + \eta \text{ (using (4.2.32))} \\ &\leq P_n\left(\ln(s_0) - \ln\left(\overline{s}_m\right)\right) + \eta_m + \mathrm{pen}(m) + \eta \text{ (using (4.2.29))} \\ &= P_n\left(\overline{f}_m\right) + \mathrm{pen}(m) + \eta_m + \eta \text{ (using (4.2.31))}, \end{aligned}$$

which implies that

$$P\left(\widehat{f}_{m'}\right) + \mathrm{pen}(m') \leq P\left(\overline{f}_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta + \eta_m.$$

Taking into account (4.2.7) and (4.2.25), we obtain

$$\begin{aligned} \mathrm{KL}_n(s_0, \widehat{s}_{m'}) &= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^q} \ln\left(\frac{s_0\left(\mathbf{y}|\mathbf{x}_i\right)}{\widehat{s}_{m'}\left(\mathbf{y}|\mathbf{x}_i\right)}\right) s_0\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y} \\ &= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^q} \widehat{f}_{m'}\left(\mathbf{y}|\mathbf{x}_i\right) s_0\left(\mathbf{y}|\mathbf{x}_i\right) d\mathbf{y} \text{ (using (4.2.31))} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\widehat{f}_{m'}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right] = P\left(\widehat{f}_{m'}\right) \text{ (using (4.2.25))}. \end{aligned}$$

Similarly, we also obtain $\mathrm{KL}_n(s_0, \overline{s}_m) = P\left(\overline{f}_m\right)$. Hence (4.2.30) implies that

$$\begin{aligned} &\mathrm{KL}_n(s_0, \widehat{s}_{m'}) + \mathrm{pen}(m') \\ &\leq \mathrm{KL}_n(s_0, \overline{s}_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta + \eta_m \\ &\leq \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta_m + \delta_{\mathrm{KL}} + \eta. \end{aligned} \tag{4.2.37}$$

All that remains is to control the deviation of $-\nu_n\left(\widehat{f}_{m'}\right) = \nu_n\left(-\widehat{f}_{m'}\right)$. To handle the randomness of $\widehat{f}_{m'}$, we shall control the deviation of $\sup_{f_{m'}\in F_{m'}} \nu_n\left(-f_{m'}\right)$, since $\widehat{f}_{m'} \in F_{m'}$. Such control is provided by Lemma 4.2.6.

Control of deviation

Lemma 4.2.6. *Let $M_n > 0$. Consider the event*

$$\mathcal{T} = \left\{ \max_{i=1,\ldots,n} \|\mathbf{Y}_i\|_\infty = \max_{i=1,\ldots,n} \max_{z\in\{1,\ldots,q\}} |[\mathbf{Y}_i]_z| \leq M_n \right\},$$

and set

$$B_n = \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right)\left(1 + q\sqrt{q}\left(M_n + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right), \text{ and} \tag{4.2.38}$$

$$\Delta_{m'} = m'\sqrt{\ln(2p+1)}\ln n + 2\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right). \tag{4.2.39}$$

Then, on the event $\mathcal{T}$, for all $m' \in \mathbb{N}^\star$, and for all $t > 0$, with probability greater than $1 - e^{-t}$,

$$\sup_{f_{m'}\in F_{m'}} |\nu_n\left(-f_{m'}\right)|\,\mathbb{1}_{\mathcal{T}} \leq \frac{4KB_n}{\sqrt{n}}\left[37q\Delta_{m'} + \sqrt{2}\left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_\Sigma}\right)\sqrt{t}\right]. \tag{4.2.40}$$

Proof. The proof appears in Section 4.2.4.1. □

From (4.2.37) and (4.2.40), we derive that on the event $\mathcal{T}$, which means that we multiply the considered term by the indicator function $\mathbb{1}_{\mathcal{T}}$, for all $m \in \mathbb{N}^\star$, and $t > 0$, with probability larger than $1 - e^{-t}$,

$$\begin{aligned}
\mathrm{KL}_n(s_0, \widehat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m\in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta_m + \delta_{\mathrm{KL}} + \eta. \\
&\leq \inf_{s_m\in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
&\quad + \frac{4KB_n}{\sqrt{n}}\left[37q\Delta_{m'} + \sqrt{2}\left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\sqrt{t}\right] \\
&\leq \inf_{s_m\in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
&\quad + \frac{4KB_n}{\sqrt{n}}\left[37q\Delta_{m'} + \frac{1}{2}\left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + t\right], \text{ if } m' \in \widehat{\mathcal{M}}(m).
\end{aligned} \tag{4.2.41}$$

Here we get the last inequality using the fact that

$$2ab \leq a^2 + b^2 \text{ for } b = \sqrt{t}, \text{ and } a = \left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)/\sqrt{2}.$$

It remains to sum up the tail bounds (4.2.41) over all possible values of $m \in \mathbb{N}^\star$ and $m' \in \widehat{\mathcal{M}}(m)$. To get an inequality valid on a set of high probability, we need to adequately choose the value of the parameter t, depending on $m \in \mathbb{N}^\star$ and $m' \in \widehat{\mathcal{M}}(m)$. Let $z > 0$, for all $m \in \mathbb{N}^\star$ and $m' \in \widehat{\mathcal{M}}(m)$, and apply (4.2.41) to obtain $t = z + m + m'$. Then, on the event $\mathcal{T}$, for all $m, m' \in \mathbb{N}^\star$, with probability larger than $1 - e^{-(z+m+m')}$,

$$\begin{aligned}
\mathrm{KL}_n(s_0, \widehat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m\in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
&\quad + \frac{4KB_n}{\sqrt{n}}\left[37q\Delta_{m'} + \frac{1}{2}\left(A_\gamma + qA_\beta + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + \left(z + m + m'\right)\right],
\end{aligned}$$

if $m' \in \widehat{\mathcal{M}}(m)$. (4.2.42)

Here, (4.2.42) is equivalent to

$$\begin{aligned}\mathrm{KL}_n(s_0,\widehat{s}_{m'})-\nu_n\left(\overline{f}_m\right)\leq & \inf_{s_m\in S_m}\mathrm{KL}_n\left(s_0,s_m\right)+\left[\mathrm{pen}(m)+\frac{4KB_n}{\sqrt{n}}m\right]+\eta_m+\delta_{\mathrm{KL}}+\eta\\ &+\left[\frac{4KB_n}{\sqrt{n}}\left(37q\Delta_{m'}+m'\right)-\mathrm{pen}(m')\right]\\ &+\frac{4KB_n}{\sqrt{n}}\left[\frac{1}{2}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2+z\right].\end{aligned}\tag{4.2.43}$$

Note that with probability larger than $1-e^{-z}$, (4.2.42) holds simultaneously for all $m\in\mathbb{N}^\star$ and $m'\in\widehat{\mathcal{M}}(m)$. Indeed, by defining the event

$$\cap_{(m,m')\in\mathbb{N}^\star\times\widehat{\mathcal{M}}(m)}\Omega_{m,m'}=\left\{w:w\in\Omega\text{ such that the event in (4.2.42) holds}\right\},$$

it holds that, on the event $\mathcal{T}$,

$$\begin{aligned}\mathbb{P}\left(\cap_{(m,m')\in\mathbb{N}^\star\times\widehat{\mathcal{M}}(m)}\Omega_{m,m'}\right)&\geq\mathbb{P}\left(\cap_{(m,m')\in\mathbb{N}^\star\times\mathbb{N}^\star}\Omega_{m,m'}\right)\\&=1-\mathbb{P}\left(\cup_{(m,m')\in\mathbb{N}^\star\times\mathbb{N}^\star}\Omega^C_{m,m'}\right)\\&\geq 1-\sum_{(m,m')\in\mathbb{N}^\star\times\mathbb{N}^\star}\mathbb{P}\left(\Omega^C_{m,m'}\right)\\&\geq 1-\sum_{(m,m')\in\mathbb{N}^\star\times\mathbb{N}^\star}e^{-(z+m+m')}\\&=1-e^{-z}\left(\sum_{m\in\mathbb{N}^\star}e^{-m}\right)^2\\&\geq 1-e^{-z},\end{aligned}$$

where we get the last inequality by using the the geometric series

$$\sum_{m=1}^{\infty}\left(e^{-1}\right)^m=\sum_{m=0}^{\infty}\left(e^{-1}\right)^m-1=\frac{1}{1-e^{-1}}-1=\frac{e}{e-1}-1=\frac{1}{e-1}<1.$$

Taking into account (4.2.39), we get

$$\begin{aligned}\mathrm{KL}_n(s_0,\widehat{s}_{m'})-\nu_n\left(\overline{f}_m\right)\leq & \inf_{s_m\in S_m}\mathrm{KL}_n\left(s_0,s_m\right)+\left[\mathrm{pen}(m)+\frac{4KB_n}{\sqrt{n}}m\right]+\eta_m+\delta_{\mathrm{KL}}+\eta\\ &+\left[\frac{4KB_n}{\sqrt{n}}\left(37q\ln n\sqrt{\ln(2p+1)}+1\right)m'-\mathrm{pen}(m')\right]\\ &+\frac{4KB_n}{\sqrt{n}}\left[\frac{1}{2}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2+74q\sqrt{K}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)+z\right].\end{aligned}\tag{4.2.44}$$

Now, let $\kappa\geq 1$ and assume that $\mathrm{pen}(m)=\lambda m$, for all $m\in\mathbb{N}^\star$ with

$$\lambda\geq\kappa\frac{4KB_n}{\sqrt{n}}\left(37q\ln n\sqrt{\ln(2p+1)}+1\right).\tag{4.2.45}$$

Then, (4.2.44) implies

$$\begin{aligned}
\mathrm{KL}_n(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \leq & \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left[\lambda m + \frac{4KB_n}{\sqrt{n}} m\right] + \eta_m + \delta_{\mathrm{KL}} + \eta \\
& + \left[\underbrace{\frac{4KB_n}{\sqrt{n}}\left(37q \ln n \sqrt{\ln(2p+1)} + 1\right)}_{\leq \lambda\kappa^{-1}} m' - \lambda m'\right] \\
& + \frac{4KB_n}{\sqrt{n}}\left[\frac{1}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + 74q\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) + z\right] \\
\leq & \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left[\mathrm{pen}(m) + \underbrace{\frac{4KB_n}{\sqrt{n}} m}_{\leq \kappa^{-1}\mathrm{pen}(m)}\right] + \eta_m + \delta_{\mathrm{KL}} + \eta \\
& + \underbrace{\left[\lambda\kappa^{-1} m' - \lambda m'\right]}_{\leq 0} \\
& + \frac{4KB_n}{\sqrt{n}}\left[\frac{1}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + 74q\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) + z\right] \\
\leq & \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left(1 + \kappa^{-1}\right)\mathrm{pen}(m) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
& + \frac{4KB_n}{\sqrt{n}}\left[\frac{1}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + 74q\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) + z\right].
\end{aligned}$$

Next, using the inequality $2ab \leq \boldsymbol{\beta}^{-1}a^2 + \boldsymbol{\beta}^{-1}b^2$ for $a = \sqrt{K}$, $b = K\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)$, and $\boldsymbol{\beta} = \sqrt{K}$, and the fact that $K \leq K^{3/2}$, for all $K \in \mathbb{N}^\star$, it follows that

$$\begin{aligned}
& \mathrm{KL}_n(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \\
\leq & \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left(1 + \kappa^{-1}\right)\mathrm{pen}(m) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
& + \frac{4B_n}{\sqrt{n}}\left[\frac{qK^{3/2}}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + \underbrace{74q\sqrt{K}K\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)}_{37q \times 2ab} + Kz\right] \\
\leq & \inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left(1 + \kappa^{-1}\right)\mathrm{pen}(m) + \eta_m + \delta_{\mathrm{KL}} + \eta \\
& + \frac{4B_n}{\sqrt{n}}\left[37qK^{1/2} + \frac{75qK^{3/2}}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + Kz\right]. \qquad (4.2.46)
\end{aligned}$$

By (4.2.23) and (4.2.32), $\widehat{m}$ belongs to $\widehat{\mathcal{M}}(m)$, for all $m \in \mathbb{N}^\star$, so we deduce from (4.2.46) that on the event $\mathcal{T}$, for all $z > 0$, with probability greater than $1 - e^{-z}$,

$$\begin{aligned}
\mathrm{KL}_n(s_0, \widehat{s}_{\widehat{m}}) - \nu_n\left(\overline{f}_m\right) \leq & \inf_{m \in \mathbb{N}^\star}\left(\inf_{s_m \in S_m} \mathrm{KL}_n\left(s_0, s_m\right) + \left(1 + \kappa^{-1}\right)\mathrm{pen}(m) + \eta_m\right) + \eta + \delta_{\mathrm{KL}} \\
& + \frac{4B_n}{\sqrt{n}}\left[37qK^{1/2} + \frac{75qK^{3/2}}{2}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2 + Kz\right]. \qquad (4.2.47)
\end{aligned}$$

By integrating (4.2.47) over $z > 0$, using the fact that for any non-negative random variable Z and any $a > 0, \mathbb{E}\left[Z\right] = a\int_{z \geq 0} \mathbb{P}(Z > az)dz$. Then, note that $\mathbb{E}\left[\nu_n\left(\overline{f}_m\right)\right] = 0$, and that $\delta_{\mathrm{KL}} > 0$ can be

chosen arbitrary small, we obtain that

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n(s_0,\widehat{s}_{\widehat{m}})\mathbb{1}_{\mathcal{T}}\right] \leq & \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\mathrm{KL}_n\left(s_0,s_m\right)+\left(1+\kappa^{-1}\right)\mathrm{pen}(m)+\eta_m\right)+\eta \\
& +\frac{4B_n}{\sqrt{n}}\left[37qK^{1/2}+\frac{75qK^{3/2}}{2}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2+K\right] \\
\leq & \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\mathrm{KL}_n\left(s_0,s_m\right)+\left(1+\kappa^{-1}\right)\mathrm{pen}(m)+\eta_m\right)+\eta \\
& +\frac{4B_n}{\sqrt{n}}\left[37qK^{3/2}+\frac{75qK^{3/2}}{2}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2+qK^{3/2}\right] \\
\leq & \inf_{m\in\mathbb{N}^\star}\left(\inf_{s_m\in S_m}\mathrm{KL}_n\left(s_0,s_m\right)+\left(1+\kappa^{-1}\right)\mathrm{pen}(m)+\eta_m\right)+\eta \\
& +\frac{302K^{3/2}qB_n}{\sqrt{n}}\left(1+\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)^2\right). \qquad (4.2.48)
\end{aligned}$$

4.2.3.6 Proof of Proposition 4.2.5

By the Cauchy-Schwarz inequality,

$$\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n\left(s_0,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n^2\left(s_0,\widehat{s}_{\widehat{m}}\right)\right]}\sqrt{\mathbb{P}\left(\mathcal{T}^C\right)}. \qquad (4.2.49)$$

We seek to bound the two terms of the right-hand side of (4.2.49).

For the first term, let us bound $\mathrm{KL}\left(s_0\left(\cdot|\mathbf{x}\right),s_{\boldsymbol{\psi}}\left(\cdot|\mathbf{x}\right)\right)$, for all $s_{\boldsymbol{\psi}}\in\mathcal{S}$ and $\mathbf{x}\in\mathcal{X}$. Let $s_{\boldsymbol{\psi}}\in\mathcal{S}$ and $\mathbf{x}\in\mathcal{X}$. Then, we obtain,

$$\begin{aligned}
\mathrm{KL}\left(s_0\left(\cdot|x\right),s_\psi\left(\cdot|x\right)\right) = & \int_{\mathbb{R}^q}\ln\left(\frac{s_0(y|x)}{s_\psi(y|x)}\right)s_0(y|x)dy \\
= & \int_{\mathbb{R}^q}\ln\left(s_0(y|x)\right)s_0(y|x)dy-\int_{\mathbb{R}^q}\ln\left(s_\psi(y|x)\right)s_0(y|x)dy \\
\leq & -\int_{\mathbb{R}^q}\ln\left(s_\psi(y|x)\right)s_0(y|x)dy+H_{s_0},\forall x\in\mathcal{X}\left(\text{using (4.2.12)}\right). \qquad (4.2.50)
\end{aligned}$$

Thus, for all $\mathbf{y} \in \mathbb{R}^q$,

$$
\begin{aligned}
&\ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right) s_0(\mathbf{y}|\mathbf{x}) \\
&= \ln\left[\sum_{k=1}^{K} \frac{g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left(-\frac{\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)}{2}\right)\right] \\
&\quad \times \sum_{k=1}^{K} \frac{g_{0,k}\left(\mathbf{x};\boldsymbol{\gamma}\right)}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_{0,k})^{1/2}} \exp\left(-\frac{\left(\mathbf{y}-\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)\right)^\top \boldsymbol{\Sigma}_{0,k}^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)\right)}{2}\right) \\
&\geq \ln\left[\sum_{k=1}^{K} \frac{a_G \det(\boldsymbol{\Sigma}_k^{-1})^{1/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{y}+\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\mathbf{x}\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)\right)\right] \\
&\quad \times \sum_{k=1}^{K} \frac{a_G \det(\boldsymbol{\Sigma}_{0,k}^{-1})^{1/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\boldsymbol{\Sigma}_{0,k}^{-1}\mathbf{y}+\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)^\top\boldsymbol{\Sigma}_{0,k}^{-1}\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)\right)\right) \\
&\Big(\text{using (4.2.5) and } -(\boldsymbol{a}-\boldsymbol{b})^\top\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{b})/2 \geq -(\boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{a}+\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{b}),\ e.g.,\ \boldsymbol{a}=\mathbf{y}, \boldsymbol{b}=\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x},\ \boldsymbol{A}=\boldsymbol{\Sigma}_k\Big) \\
&\geq \ln\left[\sum_{k=1}^{K} \frac{a_G a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{y}+\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\mathbf{x}\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)\right)\right] \\
&\quad \times \sum_{k=1}^{K} \frac{a_G a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\boldsymbol{\Sigma}_{0,k}^{-1}\mathbf{y}+\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)^\top\boldsymbol{\Sigma}_{0,k}^{-1}\left(\boldsymbol{\beta}_{0,k0}+\boldsymbol{\beta}_{0,k}\mathbf{x}\right)\right)\right) \text{ (using (4.2.4))} \\
&\geq \ln\left[K\frac{a_G a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\mathbf{y}+qA_{\boldsymbol{\beta}}^2\right)A_{\boldsymbol{\Sigma}}\right)\right] \times K\frac{a_G a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}} \exp\left(-\left(\mathbf{y}^\top\mathbf{y}+qA_{\boldsymbol{\beta}}^2\right)A_{\boldsymbol{\Sigma}}\right) \text{ (using (4.2.4))},
\end{aligned}
\tag{4.2.51}
$$

where, in the last inequality, we use the fact that for all $u \in \mathbb{R}^q$. By using the eigenvalue decomposition of $\boldsymbol{\Sigma}_1 = \boldsymbol{P}^\top\boldsymbol{D}\boldsymbol{P}$,

$$
\left|\boldsymbol{u}^\top\boldsymbol{\Sigma}_1\boldsymbol{u}\right| = \left|\boldsymbol{u}^\top\boldsymbol{P}^\top\boldsymbol{D}\boldsymbol{P}\boldsymbol{u}\right| \leq \|\boldsymbol{P}\boldsymbol{u}\|_2 \leq M(\boldsymbol{D})\|\boldsymbol{P}\boldsymbol{u}\|_2^2 \leq A_{\boldsymbol{\Sigma}}\|\boldsymbol{u}\|_2^2 \leq A_{\boldsymbol{\Sigma}} q \|\boldsymbol{u}\|_\infty^2,
$$

where in the last inequality, we used the fact that (4.2.82). Therefore, setting $\boldsymbol{u} = \sqrt{2A_{\boldsymbol{\Sigma}}}\mathbf{y}$ and $h(t) = t\ln t$, for all $t \in \mathbb{R}$, and noticing that $h(t) \geq h\left(e^{-1}\right) = -e^{-1}$, for all $t \in \mathbb{R}$, and from (4.2.50) and (4.2.51), we get that

$$
\begin{aligned}
&\mathrm{KL}\left(s_0\left(\cdot|\mathbf{x}\right), s_{\boldsymbol{\psi}}\left(\cdot|\mathbf{x}\right)\right) \\
&\leq -\int_{\mathbb{R}^q}\left[\ln\left[K\frac{a_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}}\exp\left(-\left(\mathbf{y}^\top\mathbf{y}+qA_{\boldsymbol{\beta}}^2\right)A_{\boldsymbol{\Sigma}}\right)\right] K\frac{a_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}}\exp\left(-\left(\mathbf{y}^\top\mathbf{y}+qA_{\boldsymbol{\beta}}^2\right)A_{\boldsymbol{\Sigma}}\right)\right) d\mathbf{y} \\
&= -\frac{Ka_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}e^{-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}}}{(2A_{\boldsymbol{\Sigma}})^{q/2}}\int_{\mathbb{R}^q}\left[\ln\left(K\frac{a_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}}\right)-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}-\frac{\boldsymbol{u}^\top\boldsymbol{u}}{2}\right]\frac{e^{-\frac{\boldsymbol{u}^\top\boldsymbol{u}}{2}}}{(2\pi)^{q/2}}d\boldsymbol{u} \\
&= -\frac{Ka_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}e^{-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}}}{(2A_{\boldsymbol{\Sigma}})^{q/2}}\mathbb{E}_U\left[\left[\ln\left(K\frac{a_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}}\right)-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}-\frac{\boldsymbol{U}^\top\boldsymbol{U}}{2}\right]\right] \text{ (with } \boldsymbol{U}\sim\mathcal{N}_q(0,\mathrm{I}_q)) \\
&= -\frac{Ka_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}e^{-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}}}{(2A_{\boldsymbol{\Sigma}})^{q/2}}\left[\ln\left(K\frac{a_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}}{(2\pi)^{q/2}}\right)-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}-\frac{q}{2}\right] \\
&= -\frac{Ka_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}e^{-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}-\frac{q}{2}}}{(2\pi)^{q/2}\left(A_{\boldsymbol{\Sigma}}\right)^{q/2}}e^{q/2}\pi^{q/2}\ln\left(\frac{Ka_{\boldsymbol{\gamma}}a_{\boldsymbol{\Sigma}}^{q/2}e^{-qA_{\boldsymbol{\beta}}^2A_{\boldsymbol{\Sigma}}-\frac{q}{2}}}{(2\pi)^{q/2}}\right) \\
&\leq \frac{e^{q/2-1}\pi^{q/2}}{A_{\boldsymbol{\Sigma}}^{q/2}},
\end{aligned}
\tag{4.2.52}
$$

where we used the fact that $t\ln(t) \geq -e^{-1}$, for all $t \in \mathbb{R}$. Then, for all $s_\psi \in S$,

$$\mathrm{KL}_n\left(s_0, s_\psi\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s_0\left(\cdot|x_i\right), s_\psi\left(.|x_i\right)\right) \leq \frac{e^{q/2-1}\pi^{q/2}}{A_\Sigma^{q/2}} + H_{s_0},$$

and note that $\widehat{s}_{\widehat{m}} \in S$, thus

$$\sqrt{\mathbb{E}\left[\mathrm{KL}_n^2\left(s_0, \widehat{s}_{\widehat{m}}\right)\right]} \leq \frac{e^{q/2-1}\pi^{q/2}}{A_\Sigma^{q/2}} + H_{s_0}. \tag{4.2.53}$$

We now provide an upper bound for $\mathbb{P}\left(\mathcal{T}^C\right)$:

$$\mathbb{P}\left(\mathcal{T}^C\right) \leq \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sum_{i=1}^{n}\mathbb{P}\left(\left\|\mathbf{Y}_i\right\|_\infty > M_n\right)\right]. \tag{4.2.54}$$

For all $i \in [n]$,

$$\mathbf{Y}_i|\mathbf{x}_i \sim \sum_{k=1}^{K} g_k\left(\mathbf{x}_i;\boldsymbol{\gamma}\right)\mathcal{N}_q\left(\boldsymbol{\beta}_{k0} + \boldsymbol{\beta}_k\mathbf{x}_i, \boldsymbol{\Sigma}_k\right),$$

so we see from (4.2.54) that we need to provide an upper bound on $\mathbb{P}\left(|\mathbf{Y}_x| > M_n\right)$, with

$$\mathbf{Y}_x \sim \sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\mathcal{N}_q\left(\boldsymbol{\beta}_{k0} + \boldsymbol{\beta}_k\mathbf{x}, \boldsymbol{\Sigma}_k\right), \mathbf{x} \in \mathcal{X}.$$

First, using Chernoff's inequality for a centered Gaussian variable (see Lemma 4.2.23), and the fact

that $\boldsymbol{\psi}$ belongs to the bounded space $\widetilde{\boldsymbol{\Psi}}$ (defined by (4.2.4)), and that $\sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) = 1$, we get

$$
\begin{aligned}
&\mathbb{P}\left(\|\mathbf{Y}_x\|_\infty > M_n\right) \\
&= \int_{\{\|\mathbf{y}\|_\infty > M_n\}} \sum_{k=1}^{K} \frac{g_k(\mathbf{x};\boldsymbol{\gamma})}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left(-\frac{(\mathbf{y}-(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}))^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}-(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}))}{2}\right) d\mathbf{y} \\
&= \sum_{k=1}^{K} \frac{g_k(\mathbf{x};\boldsymbol{\gamma})}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_k)^{1/2}} \int_{\{\|\mathbf{y}\|_\infty > M_n\}} \exp\left(-\frac{(\mathbf{y}-(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}))^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}-(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}))}{2}\right) d\mathbf{y} \\
&= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma})\,\mathbb{P}\left(\|\mathbf{Y}_{\mathbf{x},k}\|_\infty > M_n\right) \leq \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} \mathbb{P}\left(\left|[\mathbf{Y}_{\mathbf{x},k}]_z\right| > M_n\right) \\
&= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} \left(\mathbb{P}\left([\mathbf{Y}_{\mathbf{x},k}]_z < -M_n\right) + \mathbb{P}\left([\mathbf{Y}_{\mathbf{x},k}]_z > M_n\right)\right) \\
&= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} \left(\mathbb{P}\left(\boldsymbol{U} > \frac{M_n - [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right) + \mathbb{P}\left(\boldsymbol{U} < \frac{-M_n - [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right)\right) \\
&= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} \left(\mathbb{P}\left(\boldsymbol{U} > \frac{M_n - [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right) + \mathbb{P}\left(\boldsymbol{U} > \frac{M_n + [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right)\right) \\
&\leq \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} \left[e^{-\frac{1}{2}\left(\frac{M_n - [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right)^2} + e^{-\frac{1}{2}\left(\frac{M_n + [\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right)^2}\right] \text{(using Lemma 4.2.23, (4.2.93))} \\
&\leq 2\sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} e^{-\frac{1}{2}\left(\frac{M_n - \left|[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z\right|}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}}\right)^2} \\
&= 2\sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} e^{-\frac{1}{2}\frac{M_n^2 - 2M_n\left|[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z\right| + |[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]|_z^2}{[\boldsymbol{\Sigma}_k]_{z,z}}} \\
&\leq 2\sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\gamma}) \sum_{z=1}^{q} e^{-\frac{1}{2}\frac{M_n^2 - 2M_n\left|[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z\right| + |[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]|_z^2}{[\boldsymbol{\Sigma}_k]_{z,z}}} \leq 2KA_{\boldsymbol{\gamma}} q e^{-\frac{M_n^2 - 2M_n A_{\boldsymbol{\beta}}}{2A_{\boldsymbol{\Sigma}}}},
\end{aligned}
\tag{4.2.55}
$$

where

$$
\mathbf{Y}_{\mathbf{x},k} \sim \mathcal{N}_q\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}, \boldsymbol{\Sigma}_k\right),
$$

$$
\mathbf{Y}_{\mathbf{x},k} \sim \mathcal{N}\left([\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]_z, [\boldsymbol{\Sigma}_k]_{z,z}\right), \text{ and}
$$

$$
\boldsymbol{U} = \frac{[\mathbf{Y}_{\mathbf{x},k}]_z - [\boldsymbol{\beta}\mathbf{x}]_z}{[\boldsymbol{\Sigma}_k]_{z,z}^{1/2}} \sim \mathcal{N}(0,1),
$$

and using the facts that $e^{-\frac{1}{2}\frac{|[\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}]|_z^2}{A_{\boldsymbol{\Sigma}}}} \leq 1$ and $\max_{1\leq z\leq q}\left|[\boldsymbol{\Sigma}_k]_{z,z}\right| \leq \|\boldsymbol{\Sigma}_k\|_2 = M(\boldsymbol{\Sigma}_k) = m\left(\boldsymbol{\Sigma}_k^{-1}\right) \leq A_{\boldsymbol{\Sigma}}$. We derive from (4.2.54) and (4.2.55) that

$$
\mathbb{P}\left(\mathcal{T}^c\right) \leq 2KnqA_{\boldsymbol{\gamma}} e^{-\frac{M_n^2 - 2M_n A_{\boldsymbol{\beta}}}{2A_{\boldsymbol{\Sigma}}}},
\tag{4.2.56}
$$

and finally from (4.2.49), (4.2.53), and (4.2.56), we obtain

$$
\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\mathrm{KL}_n\left(s_0, \widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\mathcal{T}^C}\right] \leq \left(\frac{e^{q/2-1}\pi^{q/2}}{A_\Sigma^{q/2}} + H_{s_0}\right)\sqrt{2KnqA_\gamma} e^{-\frac{M_n^2 - 2M_n A_\beta}{4A_\Sigma}}.
\tag{4.2.57}
$$

4.2.4 Proofs of technical lemmas

4.2.4.1 Proof of Lemma 4.2.6

First, we give some tools to prove Lemma 4.2.6. Recall that

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f^2\left(\mathbf{y}_i|\mathbf{x}_i\right)},$$

for any measurable function f.

Let $m \in \mathbb{N}^\star$, we have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left| \frac{1}{n}\sum_{i=1}^{n} \left(f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right) - \mathbb{E}_{\mathbf{Y}_{[n]}}\left[f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right)\right] \right) \right|. \tag{4.2.58}$$

To control the deviation of (4.2.58), we shall use concentration and symmetrization arguments. We shall first use the following concentration inequality, which can be found in Boucheron et al. (2013).

Lemma 4.2.7 (See Boucheron et al., 2013)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Assume that there exists R_n, a non-random constant, such that $\sup_{f\in\mathcal{F}} \|f\|_n \leq R_n$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \left[f(\mathbf{Z}_i) - \mathbb{E}\left[f\left(\mathbf{Z}_i\right)\right]\right] \right| > \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \left[f(\mathbf{Z}_i) - \mathbb{E}\left[f\left(\mathbf{Z}_i\right)\right]\right] \right| \right] + 2\sqrt{2}R_n\sqrt{\frac{t}{n}} \right) \leq e^{-t}. \tag{4.2.59}$$

Then, we propose to bound $\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n}\left[f(\mathbf{Z}_i) - \mathbb{E}\left[f\left(\mathbf{Z}_i\right)\right]\right]\right|\right]$ due to the following symmetrization argument. The proof of this result can be found in Van Der Vaart & Wellner (1996).

Lemma 4.2.8 (See Lemma 2.3.6 in Van Der Vaart & Wellner, 1996)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Let $(\epsilon_1, \ldots, \epsilon_n)$ be a Rademacher sequence independent of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \left[f(\mathbf{Z}_i) - \mathbb{E}\left[f\left(\mathbf{Z}_i\right)\right]\right] \right| \right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(\mathbf{Z}_i) \right| \right]. \tag{4.2.60}$$

From (4.2.60), the problem is to provide an upper bound on

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(\mathbf{Z}_i) \right| \right].$$

To do so, we shall apply the following lemma, which is adapted from Lemma 6.1 in Massart (2007).

Lemma 4.2.9 (See Lemma 6.1 in Massart, 2007)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Let $(\epsilon_1, \ldots, \epsilon_n)$ be a Rademacher sequence, independent of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. Define R_n, a non-random constant, such that*

$$\sup_{f\in\mathcal{F}} \|f\|_n \leq R_n. \tag{4.2.61}$$

Then, for all $S \in \mathbb{N}^\star$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(\mathbf{Z}_i) \right| \right] \leq R_n \left(\frac{6}{\sqrt{n}} \sum_{s=1}^{S} 2^{-s} \sqrt{\ln\left[1 + M\left(2^{-s}R_n, \mathcal{F}, \|.\|_n\right)\right]} + 2^{-S} \right), \tag{4.2.62}$$

where $M\left(\delta, \mathcal{F}, \|.\|_n\right)$ stands for the δ-packing number (see Definition 4.2.20) of the set of functions $\mathcal{F}$, equipped with the metric induced by the norm $\|\cdot\|_n$.

In our case, from (4.2.58), we apply a conditional version of Lemmas 4.2.7–4.2.9 to $\mathcal{F} = F_m$, $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n) = (\mathbf{Y}_1|\mathbf{x}_1, \ldots, \mathbf{Y}_n|\mathbf{x}_n)$, and $f(\mathbf{Z}_i) = f_m(\mathbf{Y}_i|\mathbf{x}_i)$, so as to control $\sup_{f_m \in F_m} |\nu_n(-f_m)|$. On the one hand, we see from (4.2.61) that we need an upper bound of $\sup_{f_m \in F_m} \|f_m\|_n$. On the other hand, we see from (4.2.62) that we need to bound the entropy of the set of functions F_m, equipped with the metric induced by the norm $\|\cdot\|_n$. Such bounds are provided by the two following lemmas.

Let $M_n > 0$ and consider the event

$$\mathcal{T} = \left\{ \max_{i=1,\ldots,n} \|\mathbf{Y}_i\|_\infty = \max_{i=1,\ldots,n} \max_{z \in \{1,\ldots,q\}} |[\mathbf{Y}_i]_z| \leq M_n \right\},$$

and put $B_n = \max(A_{\boldsymbol{\Sigma}}, 1 + KA_G)\left(1 + q\sqrt{q}\,(M_n + A_{\boldsymbol{\beta}})^2 A_{\boldsymbol{\Sigma}}\right)$.

Lemma 4.2.10. *On the event $\mathcal{T}$, for all $m \in \mathbb{N}^\star$,*

$$\sup_{f_m \in F_m} \|f_m\|_n \mathbb{1}_{\mathcal{T}} \leq 2KB_n \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}} \right) =: R_n. \tag{4.2.63}$$

Proof. See Section 4.2.4.2. □

Lemma 4.2.11. *Let $\delta > 0$ and $m \in \mathbb{N}^\star$. On the event $\mathcal{T}$, we have the following upper bound of the δ-packing number of the set of functions F_m, equipped with the metric induced by the norm $\|\cdot\|_n$:*

$$\begin{aligned} &M(\delta, F_m, \|\cdot\|_n) \\ &\leq (2p+1)^{\frac{72B_n^2 q^2 K^2 m^2}{\delta^2}} \left(1 + \frac{18B_n K q A_{\boldsymbol{\beta}}}{\delta}\right)^K \left(1 + \frac{18B_n K A_{\boldsymbol{\gamma}}}{\delta}\right)^K \left(1 + \frac{18B_n K q\sqrt{q}}{a_{\boldsymbol{\Sigma}}\delta}\right)^K. \end{aligned}$$

Proof. See Section 4.2.4.2. □

Lemma 4.2.12 (Lemma 5.9 from Meynet, 2013)**.** *Let $\delta > 0$ and $(x_{ij})_{i=1,\ldots,n;j=1,\ldots,p} \in \mathbb{R}^{np}$. There exists a family $\mathcal{B}$ of $(2p+1)^{\|\mathbf{x}\|^2_{\max,n}/\delta^2}$ vectors in $\mathbb{R}^p$, such that for all $\boldsymbol{\beta} \in \mathbb{R}^p$, with $\|\boldsymbol{\beta}\|_1 \leq 1$, where $\|\mathbf{x}\|^2_{\max,n} = \frac{1}{n}\sum_{i=1}^n \max_{j\in\{1,\ldots,p\}} x_{ij}^2$, there exists $\boldsymbol{\beta}' \in \mathcal{B}$, such that*

$$\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p (\beta_j - \beta'_j) x_{ij} \right)^2 \leq \delta^2.$$

Proof. See in the proof of Lemma 5.9 Meynet (2013). □

Via the upper bounds provided in Lemmas 4.2.10 and 4.2.11, we can apply Lemma 4.2.9 to get an upper bound on $\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m \in \mathcal{F}_m} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(\mathbf{Y}_i|\mathbf{x}_i)\right|\right]$. We thus obtain the following results.

Lemma 4.2.13. *Let $m \in \mathbb{N}^\star$, consider $(\epsilon_1, \ldots, \epsilon_n)$, a Rademacher sequence independent of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Then, on the event $\mathcal{T}$,*

$$\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m \in \mathcal{F}_m} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(\mathbf{Y}_i|\mathbf{x}_i) \right| \right] \leq \frac{74KB_n q}{\sqrt{n}} \Delta_m, \tag{4.2.64}$$

$$\Delta_m := m\sqrt{\ln(2p+1)}\ln n + 2\sqrt{K}\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}} \right). \tag{4.2.65}$$

Proof. See Section 4.2.4.2. □

Now using (4.2.64) and applying both Lemmas 4.2.7 and 4.2.8 to $\mathcal{F} = F_m, (\mathbf{Z}_1, \ldots, \mathbf{Z}_n) = (\mathbf{Y}_1|\mathbf{x}_1, \ldots, \mathbf{Y}_n|\mathbf{x}_n)$ and $f(\mathbf{Z}_i) = f_m(\mathbf{Y}_i|\mathbf{x}_i)$, we get for all $m \in \mathbb{N}^\star$ and $t > 0$, with probability

greater than $1-e^{-t}$,

$$
\begin{aligned}
&\sup_{f_m\in F_m}\left|\nu_n\left(-f_m\right)\right| \\
&=\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right)-\mathbb{E}_{\mathbf{Y}_{[n]}}\left[f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right)\right]\right)\right| \\
&\leq \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right)-\mathbb{E}_{\mathbf{Y}_{[n]}}\left[f_m\left(\mathbf{Y}_i|\mathbf{x}_i\right)\right]\right)\right|\right]+2\sqrt{2}R_n\sqrt{\frac{t}{n}}\ \text{(Lemma 4.2.7)} \\
&\leq 2\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(\mathbf{Y}_i|\mathbf{x}_i)\right|\right]+2\sqrt{2}R_n\sqrt{\frac{t}{n}}\ \text{(using Lemma 4.2.8)} \\
&\leq \frac{148KB_nq}{\sqrt{n}}\Delta_m+4\sqrt{2}KB_n\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\sqrt{\frac{t}{n}} \\
&\left(\text{using Lemma 4.2.13 and } R_n=2KB_n\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\right) \\
&\leq \frac{4KB_n}{\sqrt{n}}\left[37q\Delta_m+\sqrt{2}\left(A_{\boldsymbol{\gamma}}+qA_{\boldsymbol{\beta}}+\frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\sqrt{t}\right].
\end{aligned}
$$

4.2.4.2 Proofs of Lemmas 4.2.10–4.2.13

The proofs of Lemmas 4.2.10–4.2.11 require an upper bound on the uniform norm of the gradient of $\ln s_{\boldsymbol{\psi}}$, for $s_{\boldsymbol{\psi}}\in\mathcal{S}$. We begin by providing such an upper bound.

Lemma 4.2.14. *Given $s_{\boldsymbol{\psi}}$, as described in (4.2.6), it holds that*

$$
\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}}\left\|\frac{\partial\ln\left(s_{\boldsymbol{\psi}}(\cdot|\mathbf{x})\right)}{\partial\boldsymbol{\psi}}\right\|_{\infty}\leq G(\cdot), \\
&G:\mathbb{R}^q\ni\mathbf{y}\mapsto G(\mathbf{y})=\max\left(A_{\boldsymbol{\Sigma}},1+KA_G\right)\left(1+q\sqrt{q}\left(\|\mathbf{y}\|_{\infty}+A_{\boldsymbol{\beta}}\right)^2A_{\boldsymbol{\Sigma}}\right). \qquad (4.2.66)
\end{aligned}
$$

Proof. Let $s_{\boldsymbol{\psi}}\in\mathcal{S}$, with $\boldsymbol{\psi}=(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\Sigma})$. From now on, we consider any $\mathbf{x}\in\mathcal{X}$, any $\mathbf{y}\in\mathbb{R}^q$, and any $k\in[K]$. We can write

$$
\begin{aligned}
\ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)&=\ln\left(\sum_{k=1}^{K}g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\phi\left(\mathbf{y};\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x},\boldsymbol{\Sigma}_k\right)\right)=\ln\left(\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})\right), \\
g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)&=\frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)},w_k(\mathbf{x})=\boldsymbol{\gamma}_{k0}+\boldsymbol{\gamma}_k^{\top}\mathbf{x}, \\
\phi\left(\mathbf{y};\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x},\boldsymbol{\Sigma}_k\right)&=\frac{1}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_k)^{1/2}}\exp\left(-\frac{\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)}{2}\right), \\
f_k(\mathbf{x},\mathbf{y})&=g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\phi\left(\mathbf{y};\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x},\boldsymbol{\Sigma}_k\right) \\
&=\frac{g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)}{(2\pi)^{q/2}\det(\boldsymbol{\Sigma}_k)^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{k0}+\boldsymbol{\beta}_k\mathbf{x}\right)\right)\right].
\end{aligned}
$$

By using the chain rule, for all $l\in[K]$,

$$
\begin{aligned}
\frac{\partial\ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\gamma_{l0}}&=\sum_{k=1}^{K}\frac{f_k(\mathbf{x},\mathbf{y})}{g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})}\frac{\partial g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)}{\partial w_l(\mathbf{x})}\underbrace{\frac{\partial w_l(\mathbf{x})}{\partial\gamma_{l0}}}_{=1},\text{and} \\
\frac{\partial\ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\left(\boldsymbol{\gamma}_l^{\top}\mathbf{x}\right)}&=\sum_{k=1}^{K}\frac{f_k(\mathbf{x},\mathbf{y})}{g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})}\frac{\partial g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)}{\partial w_l(\mathbf{x})}\underbrace{\frac{\partial w_l(\mathbf{x})}{\partial\left(\boldsymbol{\gamma}_l^{\top}\mathbf{x}\right)}}_{=1}.
\end{aligned}
$$

Furthermore,

$$\begin{aligned}
\frac{\partial g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)}{\partial w_l(\mathbf{x})} &= \frac{\partial}{\partial w_l(\mathbf{x})}\left(\frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)}\right) \\
&= \frac{\frac{\partial}{\partial w_l(\mathbf{x})}\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)} - \frac{\exp\left(w_k(\mathbf{x})\right)}{\left(\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)\right)^2}\frac{\partial}{\partial w_l(\mathbf{x})}\sum_{i=1}^{K}\exp\left(w_i(\mathbf{x})\right) \\
&\left(\text{using } \frac{\partial}{\partial \mathbf{x}}\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) = \frac{f'(\mathbf{x})g(\mathbf{x}) - g'(\mathbf{x})f(\mathbf{x})}{g^2(\mathbf{x})}\right) \\
&= \frac{\delta_{lk}\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)} - \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)}\frac{\exp\left(w_l(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_l(\mathbf{x})\right)} \\
&= g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\left(\delta_{lk} - g_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right), \text{ where } \delta_{lk} = \begin{cases} 1 & \text{if } l = k, \\ 0 & \text{if } l \neq k. \end{cases}
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\left|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\left(\boldsymbol{\gamma}_l^\top\mathbf{x}\right)}\right| &= \left|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial \gamma_{l0}}\right| \\
&= \left|\sum_{k=1}^{K}\frac{f_k(\mathbf{x},\mathbf{y})}{g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})}g_k\left(\mathbf{x};\boldsymbol{\gamma}\right)\left(\delta_{lk} - g_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right)\right| \\
&= \left|\sum_{k=1}^{K}\frac{f_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})}\left(\delta_{lk} - g_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right)\right| \\
&\leq \left|\sum_{k=1}^{K}\left(\delta_{lk} - g_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right)\right| \left(\text{since}\frac{f_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})} \leq 1\right) \\
&= \left|1 - \sum_{k=1}^{K}g_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right| = \left|1 - Kg_l\left(\mathbf{x};\boldsymbol{\gamma}\right)\right| \\
&\leq 1 + Kg_l\left(\mathbf{x};\boldsymbol{\gamma}\right) \leq 1 + KA_G \left(\text{using } (4.2.5)\right).
\end{aligned}$$

Similarly, by using the fact that $\boldsymbol{\psi}$ belongs to the bounded space $\widetilde{\boldsymbol{\Psi}}$, $f_l(\mathbf{x},\mathbf{y})/\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y}) \leq 1$,

$$\begin{aligned}
&\left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\boldsymbol{\beta}_{l0}}\right\|_\infty = \left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\left(\boldsymbol{\beta}_l\mathbf{x}\right)}\right\|_\infty \\
&= \left\|\frac{f_l(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{K}f_k(\mathbf{x},\mathbf{y})}\frac{\partial}{\partial\left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)}\left[-\frac{1}{2}\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)\right]\right\|_\infty \\
&\leq \left\|\frac{\partial}{\partial\left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)}\left[-\frac{1}{2}\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)\right]\right\|_\infty \\
&= \left\|\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)\right\|_\infty \leq \left\|\boldsymbol{\Sigma}_l^{-1}\right\|_\infty\left\|\left(\mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)\right)\right\|_\infty \left(\text{using } (4.2.83)\right) \\
&\leq \sqrt{q}\left\|\boldsymbol{\Sigma}_l^{-1}\right\|_2\left(\|\mathbf{y}\|_\infty + \|\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\|_\infty\right) \left(\text{using } (4.2.88)\right) \\
&\leq \sqrt{q}M\left(\boldsymbol{\Sigma}_l^{-1}\right)\left(\|\mathbf{y}\|_\infty + \|\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\|_\infty\right) \left(\text{using } (4.2.87)\right) \\
&\leq \sqrt{q}A_{\boldsymbol{\Sigma}}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right) \left(\text{using } (4.2.4)\right).
\end{aligned}$$

Now, we need to calculate the gradient w.r.t. to the covariance matrices of the Gaussian experts.

To do this, we need the following result: given any $l \in [K]$, $\boldsymbol{\upsilon}_l = \boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l \mathbf{x}$, it holds that

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_l} \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right) \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left[(2\pi)^{-p/2} \det(\boldsymbol{\Sigma}_l)^{-1/2} \exp\left(-\frac{\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)}{2}\right)\right] \\
&= \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right)\left[-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left(\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\right) + \det(\boldsymbol{\Sigma}_l)^{1/2}\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left(\det(\boldsymbol{\Sigma}_l)^{-1/2}\right)\right] \\
&= \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right)\left[\frac{1}{2}\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1} - \frac{1}{2}\det(\boldsymbol{\Sigma}_l)^{-1}\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left(\det(\boldsymbol{\Sigma}_l)\right)\right] \\
&= \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right)\left[\frac{1}{2}\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1} - \frac{1}{2}\det(\boldsymbol{\Sigma}_l)^{-1}\det(\boldsymbol{\Sigma}_l)\left(\boldsymbol{\Sigma}_l^{-1}\right)^\top\right] \\
&= \phi\left(\mathbf{x}; \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right) \underbrace{\frac{1}{2}\left[\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1} - \left(\boldsymbol{\Sigma}_l^{-1}\right)^\top\right]}_{T\left(\mathbf{x}, \boldsymbol{\upsilon}_l, \boldsymbol{\Sigma}_l\right)}, \qquad (4.2.67)
\end{aligned}$$

noting that

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left(\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\right) = -\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)\left(\mathbf{x}-\boldsymbol{\upsilon}_l\right)^\top \boldsymbol{\Sigma}_l^{-1} \text{ (using Lemma 4.2.16)}, \qquad (4.2.68)$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left(\det(\boldsymbol{\Sigma}_l)\right) = \det(\boldsymbol{\Sigma}_l)\left(\boldsymbol{\Sigma}_l^{-1}\right)^\top \text{ (using Jacobi formula, Lemma 4.2.17)}. \qquad (4.2.69)$$

For any $l \in [K]$,

$$\begin{aligned}
\left|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial\left([\boldsymbol{\Sigma}_l]_{z_1,z_2}\right)}\right| &\leq \left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial \boldsymbol{\Sigma}_l}\right\|_2 \text{ (using (4.2.87))} \\
&= \left|\frac{f_l(\mathbf{x},\mathbf{y})}{\sum_{k=1}^K f_k(\mathbf{x},\mathbf{y})}\right| \left\|\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left[-\frac{1}{2}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)\right]\right\|_2 \\
&\leq \left\|\frac{\partial}{\partial \boldsymbol{\Sigma}_l}\left[-\frac{1}{2}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)\right]\right\|_2 \\
&= \frac{1}{2}\left\|\boldsymbol{\Sigma}_l^{-1}\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top \boldsymbol{\Sigma}_l^{-1} - \left(\boldsymbol{\Sigma}_l^{-1}\right)^\top\right\|_2 \text{ (using (4.2.67))} \\
&\leq \frac{1}{2}\left[A_{\boldsymbol{\Sigma}} + \sqrt{q}\left\|\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)\left(\mathbf{y}-\left(\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_l\mathbf{x}\right)\right)^\top\right\|_\infty A_{\boldsymbol{\Sigma}}^2\right] \text{ (using (4.2.88))} \\
&\leq \frac{1}{2}\left[A_{\boldsymbol{\Sigma}} + q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}^2\right] \text{ (using (4.2.4))},
\end{aligned}$$

where, in the last inequality given $\boldsymbol{a} = \mathbf{y} - \left(\boldsymbol{\beta}_{l0} + \boldsymbol{\beta}_l\mathbf{x}\right)$, we use the fact that

$$\left\|\boldsymbol{a}\boldsymbol{a}^\top\right\|_\infty = \max_{1\leq i\leq q}\sum_{j=1}^q\left|[\boldsymbol{a}\boldsymbol{a}^\top]_{i,j}\right| = \max_{1\leq i\leq q}\sum_{j=1}^q |a_i a_j| = \max_{1\leq i\leq q}|a_i|\sum_{j=1}^q |a_j| \leq q\|\boldsymbol{a}\|_\infty^2.$$

Thus,

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}}\left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x})\right)}{\partial \boldsymbol{\psi}}\right\|_\infty \\
&\leq \max\left[1 + KA_G, \sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)A_{\boldsymbol{\Sigma}}, \frac{1}{2}\left[A_{\boldsymbol{\Sigma}} + q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}^2\right]\right] \\
&\leq \max\left[1 + KA_G, \max\left(A_{\boldsymbol{\Sigma}}, 1\right)\left(1 + q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right)\right] \\
&\leq \max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right)\left(1 + q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right) =: G(\mathbf{y}),
\end{aligned}$$

where we use the fact that

$$\begin{aligned}\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right) A_{\boldsymbol{\Sigma}} =: \theta \leq 1 + \theta^2 &= 1 + q\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}^2 \\ &\leq \max\left(A_{\boldsymbol{\Sigma}}, 1\right)\left(1 + q\sqrt{q}\left(\|\mathbf{y}\|_\infty + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right).\end{aligned}$$

□

Proof of Lemma 4.2.10

Let $m \in \mathbb{N}^\star$ and $f_m \in F_m$. By (4.2.28), there exists $s_m \in S_m$, such that $f_m = -\ln\left(s_m/s_0\right)$. For all $\mathbf{x} \in \mathcal{X}$, let $\boldsymbol{\psi}(\mathbf{x}) = \left(\gamma_{k0}, \boldsymbol{\gamma}_k^\top \mathbf{x}, \boldsymbol{\beta}_{k0}, \boldsymbol{\beta}_k \mathbf{x}, \boldsymbol{\Sigma}_k\right)_{k\in[K]}$ be the parameters of $s_m\left(\cdot|\mathbf{x}\right)$. In our case, we approximate $f(\boldsymbol{\psi}) = \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)$ around $\boldsymbol{\psi}_0(\mathbf{x}_i)$ by the $n = 0^{th}$ degree Taylor polynomial of $f(\boldsymbol{\psi})$. That is,

$$\begin{aligned}\left|\ln\left(\underbrace{\left(s_m\right)}_{s_{\boldsymbol{\psi}}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) - \ln\left(s_0\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)\right| &=: \left|f(\boldsymbol{\psi}) - f(\boldsymbol{\psi}_0)\right| = \left|R_0(\boldsymbol{\psi})\right| \text{ (defined in Lemma 4.2.24)} \\ &\leq \sup_{\mathbf{x}\in\mathcal{X}}\sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}}\left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}_i|\mathbf{x})\right)}{\partial \boldsymbol{\psi}}\right\|_\infty \left\|\boldsymbol{\psi}\left(\mathbf{x}_i\right) - \boldsymbol{\psi}_0\left(\mathbf{x}_i\right)\right\|_1.\end{aligned}$$

First applying Taylor's inequality and then Lemma 4.2.14 on the event $\mathcal{T}$. For all $i \in [n]$, it holds that

$$\begin{aligned}&\left|f_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right| \mathbb{1}_{\mathcal{T}} = \left|\ln\left(s_m\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) - \ln\left(s_0\left(\mathbf{y}_i|\mathbf{x}_i\right)\right)\right| \mathbb{1}_{\mathcal{T}} \\ &\leq \sup_{\mathbf{x}\in\mathcal{X}}\sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}}\left\|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}_i|\mathbf{x})\right)}{\partial \boldsymbol{\psi}}\right\|_\infty \left\|\boldsymbol{\psi}\left(\mathbf{x}_i\right) - \boldsymbol{\psi}_0\left(\mathbf{x}_i\right)\right\|_1 \mathbb{1}_{\mathcal{T}} \\ &\leq \underbrace{\max\left(A_{\boldsymbol{\Sigma}}, 1 + KA_G\right)\left(1 + q\sqrt{q}\left(M_n + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right)}_{=:B_n}\left\|\boldsymbol{\psi}\left(\mathbf{x}_i\right) - \boldsymbol{\psi}_0\left(\mathbf{x}_i\right)\right\|_1 \text{ (using Lemma 4.2.14)} \\ &\leq B_n \sum_{k=1}^{K}\Bigg(\left|\gamma_{k0} - \gamma_{0,k0}\right| + \left|\boldsymbol{\gamma}_k^\top \mathbf{x}_i - \boldsymbol{\gamma}_{0,k}^\top \mathbf{x}_i\right| \\ &\qquad + \left\|\boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{0,k0}\right\|_1 + \left\|\boldsymbol{\beta}_k \mathbf{x}_i - \boldsymbol{\beta}_{0,k}\mathbf{x}_i\right\|_1 + \left\|\operatorname{vec}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{0,k}\right)\right\|_1\Bigg) \\ &\leq 2B_n \sum_{k=1}^{K}\left(\left|\gamma_{k0}\right| + \left|\boldsymbol{\gamma}_k^\top \mathbf{x}_i\right| + \left\|\boldsymbol{\beta}_{k0}\right\|_1 + \left\|\boldsymbol{\beta}_k \mathbf{x}_i\right\|_1 + q\left\|\boldsymbol{\Sigma}_k\right\|_1\right) \text{ (using (4.2.85))} \\ &\leq 2KB_n\left(A_{\boldsymbol{\gamma}} + q\left\|\boldsymbol{\beta}_{k0}\right\|_\infty + q\left\|\boldsymbol{\beta}_k \mathbf{x}_i\right\|_\infty + q\sqrt{q}\left\|\boldsymbol{\Sigma}_k\right\|_2\right) \text{ (using (4.2.4), (4.2.80), (4.2.81), (4.2.89))} \\ &\leq 2KB_n\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) \text{ (using (4.2.4))}.\end{aligned}$$

Therefore,

$$\sup_{f_m\in F_m}\left\|f_m\right\|_n \mathbb{1}_{\mathcal{T}} \leq 2KB_n\left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) =: R_n.$$

Proof of Lemma 4.2.11

Let $m \in \mathbb{N}^\star$, $f_m^{[1]} \in F_m$, and $\mathbf{x} \in [0,1]^p$. By (4.2.28), there exists $s_m^{[1]} \in S_m$, such that $f_m^{[1]} = -\ln\left(s_m^{[1]}/s_0\right)$. Introduce the notation $s_m^{[2]} \in \mathcal{S}$ and $f_m^{[2]} = -\ln\left(s_m^{[2]}/s_0\right)$. Let

$$\boldsymbol{\psi}^{[1]}(\mathbf{x}) = \left(\gamma_{k0}^{[1]}, \boldsymbol{\gamma}_k^{[1]}\mathbf{x}, \boldsymbol{\beta}_{k0}^{[1]}, \boldsymbol{\beta}_k^{[1]}\mathbf{x}, \boldsymbol{\Sigma}_k^{[1]}\right)_{k\in[K]}, \text{and } \boldsymbol{\psi}^{[2]}(\mathbf{x}) = \left(\gamma_{k0}^{[2]}, \boldsymbol{\gamma}_k^{[2]}\mathbf{x}, \boldsymbol{\beta}_{k0}^{[2]}, \boldsymbol{\beta}_k^{[2]}\mathbf{x}, \boldsymbol{\Sigma}_k^{[2]}\right)_{k\in[K]},$$

be the parameters of the PDFs $s_m^{[1]}(\cdot|\mathbf{x})$ and $s_m^{[2]}(\cdot|\mathbf{x})$, respectively. By applying Taylor's inequality and then Lemma 4.2.14 on the event $\mathcal{T}$, for all $i \in [n]$, it holds that

$$
\begin{aligned}
&\left|f_m^{[1]}(\mathbf{y}_i|\mathbf{x}_i) - f_m^{[2]}(\mathbf{y}_i|\mathbf{x}_i)\right| \mathbb{1}_{\mathcal{T}} = \left|\ln\left(s_m^{[1]}(\mathbf{y}_i|\mathbf{x}_i)\right) - \ln\left(s_m^{[2]}(\mathbf{y}_i|\mathbf{x}_i)\right)\right| \mathbb{1}_{\mathcal{T}} \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \sup_{\boldsymbol{\psi}\in\widetilde{\boldsymbol{\Psi}}} \left|\frac{\partial \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}_i|\mathbf{x})\right)}{\partial \boldsymbol{\psi}}\right| \left\|\boldsymbol{\psi}^{[1]}(\mathbf{x}_i) - \boldsymbol{\psi}^{[1]}(\mathbf{x}_i)\right\|_1 \mathbb{1}_{\mathcal{T}} \text{ (using Taylor's inequality in Lemma 4.2.24)} \\
&\leq \underbrace{\max\left(A_{\boldsymbol{\Sigma}}, C(p,K)\right)\left(1 + q\sqrt{q}\left(M_n + A_{\boldsymbol{\beta}}\right)^2 A_{\boldsymbol{\Sigma}}\right)}_{B_n} \left\|\boldsymbol{\psi}^{[1]}(\mathbf{x}_i) - \boldsymbol{\psi}^{[2]}(\mathbf{x}_i)\right\|_1 \text{ (using Lemma 4.2.14)} \\
&\leq B_n \sum_{k=1}^{K} \Bigg(\left|\gamma_{k0}^{[1]} - \gamma_{k0}^{[2]}\right| + \left|{\boldsymbol{\gamma}_k^{[1]}}^\top \mathbf{x}_i - {\boldsymbol{\gamma}_k^{[2]}}^\top \mathbf{x}_i\right| \\
&\qquad\qquad + \left\|\boldsymbol{\beta}_{k0}^{[1]} - \boldsymbol{\beta}_{k0}^{[2]}\right\|_1 + \left\|\boldsymbol{\beta}_k^{[1]}\mathbf{x}_i - \boldsymbol{\beta}_k^{[2]}\mathbf{x}_i\right\|_1 + \left\|\mathrm{vec}\left(\boldsymbol{\Sigma}_k^{[1]} - \boldsymbol{\Sigma}_k^{[2]}\right)\right\|_1 \Bigg).
\end{aligned}
$$

By the Cauchy-Schwarz inequality, $\left(\sum_{i=1}^m a_i\right)^2 \leq m \sum_{i=1}^m a_i^2$ $(m \in \mathbb{N}^\star)$, we get

$$
\begin{aligned}
&\left|f_m^{[1]}(\mathbf{y}_i|\mathbf{x}_i) - f_m^{[2]}(\mathbf{y}_i|\mathbf{x}_i)\right|^2 \mathbb{1}_{\mathcal{T}} \\
&\leq 3B_n^2 \Bigg[\left(\sum_{k=1}^{K} \left|{\boldsymbol{\gamma}_k^{[1]}}^\top \mathbf{x}_i - {\boldsymbol{\gamma}_k^{[2]}}^\top \mathbf{x}_i\right|\right)^2 + \left(\sum_{k=1}^{K}\sum_{z=1}^{q} \left|\left[\boldsymbol{\beta}_k^{[1]}\mathbf{x}_i\right]_z - \left[\boldsymbol{\beta}_k^{[2]}\mathbf{x}_i\right]_z\right|\right)^2 \\
&\qquad\qquad + \left(\left\|\boldsymbol{\beta}_0^{[1]} - \boldsymbol{\beta}_0^{[2]}\right\|_1 + \left\|\boldsymbol{\gamma}_0^{[1]} - \boldsymbol{\gamma}_0^{[2]}\right\|_1 + \left\|\mathrm{vec}\left(\boldsymbol{\Sigma}^{[1]} - \boldsymbol{\Sigma}^{[2]}\right)\right\|_1\right)^2 \Bigg] \\
&\leq 3B_n^2 \Bigg[K \sum_{k=1}^{K} \left(\sum_{j=1}^{p} {\boldsymbol{\gamma}_{kj}^{[1]}}^\top x_{ij} - \sum_{j=1}^{p} {\boldsymbol{\gamma}_{kj}^{[2]}}^\top x_{ij}\right)^2 \\
&\quad + Kq \sum_{k=1}^{K}\sum_{z=1}^{q} \left(\sum_{j=1}^{p} \left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j} x_{ij} - \sum_{j=1}^{p} \left[\boldsymbol{\beta}_k^{[2]}\right]_{z,j} x_{ij}\right)^2 \\
&\quad + \left(\left\|\boldsymbol{\beta}_0^{[1]} - \boldsymbol{\beta}_0^{[2]}\right\|_1 + \left\|\boldsymbol{\gamma}_0^{[1]} - \boldsymbol{\gamma}_0^{[2]}\right\|_1 + \left\|\mathrm{vec}\left(\boldsymbol{\Sigma}^{[1]} - \boldsymbol{\Sigma}^{[2]}\right)\right\|_1\right)^2 \Bigg],
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|f_m^{[1]} - f_m^{[2]}\right\|_n^2 \mathbb{1}_{\mathcal{T}} &= \frac{1}{n}\sum_{i=1}^{n} \left|f_m^{[1]}(\mathbf{y}_i|\mathbf{x}_i) - f_m^{[2]}(\mathbf{y}_i|\mathbf{x}_i)\right|^2 \mathbb{1}_{\mathcal{T}} \\
&\leq 3B_n^2 \underbrace{K \sum_{k=1}^{K} \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \gamma_{kj}^{[1]} x_{ij} - \sum_{j=1}^{p} \gamma_{kj}^{[2]} x_{ij}\right)^2}_{=:a} \\
&\quad + 3B_n^2 \underbrace{Kq \sum_{k=1}^{K}\sum_{z=1}^{q} \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j} x_{ij} - \sum_{j=1}^{p} \left[\boldsymbol{\beta}_k^{[2]}\right]_{z,j} x_{ij}\right)^2}_{=:b} \\
&\quad + 3B_n^2 \left(\left\|\boldsymbol{\beta}_0^{[1]} - \boldsymbol{\beta}_0^{[2]}\right\|_1 + \left\|\boldsymbol{\gamma}_0^{[1]} - \boldsymbol{\gamma}_0^{[2]}\right\|_1 + \left\|\mathrm{vec}\left(\boldsymbol{\Sigma}^{[1]} - \boldsymbol{\Sigma}^{[2]}\right)\right\|_1\right)^2.
\end{aligned}
$$

So, for all $\delta > 0$, if

$$\begin{aligned} a &\leq \delta^2/\left(36B_n^2\right), \\ b &\leq \delta^2/\left(36B_n^2\right), \\ \left\|\boldsymbol{\beta}_0^{[1]} - \boldsymbol{\beta}_0^{[2]}\right\|_1 &\leq \delta/\left(18B_n\right), \\ \left\|\boldsymbol{\gamma}_0^{[1]} - \boldsymbol{\gamma}_0^{[2]}\right\|_1 &\leq \delta/\left(18B_n\right), \text{and} \\ \left\|\operatorname{vec}\left(\boldsymbol{\Sigma}^{[1]} - \boldsymbol{\Sigma}^{[2]}\right)\right\|_1 &\leq \delta/\left(18B_n\right), \end{aligned}$$

then $\left\|f_m^{[1]} - f_m^{[2]}\right\|_n^2 \mathbb{1}_{\mathcal{T}} \leq \delta^2/4$. To bound a and b, we can write

$$\begin{aligned} a &= Km^2 \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\gamma_{kj}^{[1]}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\gamma_{kj}^{[2]}}{m} x_{ij} \right)^2, \text{and} \\ b &= Kqm^2 \sum_{k=1}^{K} \sum_{z=1}^{q} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\left[\boldsymbol{\beta}_k^{[2]}\right]_{z,j}}{m} x_{ij} \right)^2. \end{aligned}$$

Then, we apply Lemma 4.2.12 to obtain $\frac{\boldsymbol{\gamma}_{k,\cdot}^{[1]}}{m} = \left(\frac{\gamma_{kj}^{[1]}}{m}\right)_{j\in[q]}$ and $\frac{\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,\cdot}}{m} = \left(\frac{\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j}}{m}\right)_{j\in[q]}$, for all $k \in [K], z \in [q]$. Since $s_m^{[1]} \in S_m$, and using (4.2.16), we have $\left\|\boldsymbol{\gamma}_k^{[1]}\right\| \leq m$ and $\left\|\operatorname{vec}\left(\boldsymbol{\beta}_k^{[1]}\right)\right\|_1 \leq m$, which leads to $\sum_{j=1}^{p}\left|\frac{\gamma_{kj}^{[1]}}{m}\right| \leq 1$ and $\sum_{z=1}^{q}\sum_{j=1}^{p}\left|\frac{\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j}}{m}\right| \leq 1$, respectively. Furthermore, given $\mathbf{x} \in \mathcal{X} = [0,1]^p$, we have $\|\mathbf{x}\|_{\max,n}^2 = 1$. Thus, there exist families $\mathcal{A}$ of $(2p+1)^{36B_n^2K^2m^2/\delta^2}$ vectors and $\mathcal{B}$ of $(2p+1)^{16B_n^2q^2K^2m^2/\delta^2}$ vectors of $\mathbb{R}^p$, such that for all $k \in [K]$, $z \in [q]$, $\boldsymbol{\gamma}_{k,\cdot}^{[1]}$, and $\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,\cdot}$, there exist $\boldsymbol{\gamma}_{k,\cdot}^{[1]} \in \mathcal{A}$ and $\left[\boldsymbol{\beta}_k^{[2]}\right]_{z,\cdot} \in \mathcal{B}$, such that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\gamma_{kj}^{[1]}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\gamma_{kj}^{[2]}}{m} x_{ij} \right)^2 &\leq \frac{\delta^2}{36B_n^2K^2m^2}, \text{and} \\ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\left[\boldsymbol{\beta}_k^{[1]}\right]_{z,j}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\left[\boldsymbol{\beta}_k^{[2]}\right]_{z,j}}{m} x_{ij} \right)^2 &\leq \frac{\delta^2}{36B_n^2q^2K^2m^2}, \end{aligned}$$

which leads to $a \leq \delta^2/36B_n^2$ and $b \leq \delta^2/36B_n^2$. Moreover, (4.2.4) leads to

$$\begin{aligned} \left\|\boldsymbol{\beta}_0^{[1]}\right\|_1 &= \sum_{k=1}^{K} \left\|\boldsymbol{\beta}_{0k}^{[1]}\right\|_1 \leq Kq \left\|\boldsymbol{\beta}_{0k}^{[1]}\right\|_\infty \leq KqA_{\boldsymbol{\beta}} \text{ (using (4.2.80))}, \\ \left\|\boldsymbol{\gamma}_0^{[1]}\right\|_1 &= \sum_{k=1}^{K} \left|\boldsymbol{\gamma}_{0k}^{[1]}\right| \leq KA_{\boldsymbol{\gamma}}, \text{and} \\ \left\|\operatorname{vec}\left(\boldsymbol{\Sigma}^{[1]}\right)\right\|_1 &= \sum_{k=1}^{K} \left\|\operatorname{vec}\left(\boldsymbol{\Sigma}_k^{[1]}\right)\right\|_1 \leq \frac{Kq\sqrt{q}}{a_{\boldsymbol{\Sigma}}}. \end{aligned}$$

Therefore, on the event $\mathcal{T}$,

$$
\begin{aligned}
M\left(\delta, F_m, \|\cdot\|_n\right) &\leq N\left(\delta/2, F_m, \|\cdot\|_n\right) \text{ (using Lemma 4.2.22)} \\
&\leq \operatorname{card}(\mathcal{A}) \operatorname{card}(\mathcal{B}) N\left(\frac{\delta}{18B_n}, B_1^K\left(KqA_{\boldsymbol{\beta}}\right), \|\cdot\|_1\right) \\
&\quad N\left(\frac{\delta}{18B_n}, B_1^K\left(KA_{\boldsymbol{\gamma}}\right), \|\cdot\|_1\right) N\left(\frac{\delta}{18B_n}, B_1^K\left(\frac{Kq\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right), \|\cdot\|_1\right) \\
&\leq (2p+1)^{\frac{72B_n^2q^2K^2m^2}{\delta^2}} \left(1+\frac{18B_nKqA_{\boldsymbol{\beta}}}{\delta}\right)^K \left(1+\frac{18B_nKA_{\boldsymbol{\gamma}}}{\delta}\right)^K \left(1+\frac{18B_nKq\sqrt{q}}{a_{\boldsymbol{\Sigma}}\delta}\right)^K.
\end{aligned}
$$

Proof of Lemma 4.2.13

Let $m \in \mathbb{N}^\star$. From Lemma 4.2.10, on the event $\mathcal{T}$,

$$
\sup_{f_m \in F_m} \|f_m\|_n \mathbb{1}_{\mathcal{T}} \leq 2KB_n \left(A_{\boldsymbol{\gamma}} + qA_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) =: R_n. \tag{4.2.70}
$$

From Lemma 4.2.11, on the event $\mathcal{T}$ for all $S \in \mathbb{N}^\star$,

$$
\begin{aligned}
&\sum_{s=1}^{S} 2^{-s} \sqrt{\ln\left[1+M\left(2^{-s}R_n, F_m, \|\cdot\|_n\right)\right]} \\
&\leq \sum_{s=1}^{S} 2^{-s} \sqrt{\ln\left[2M\left(\delta, F_m, \|\cdot\|_n\right)\right]} \text{ with } \delta = 2^{-s}R_n \\
&\leq \sum_{s=1}^{S} 2^{-s} \Bigg[\sqrt{\ln 2} + \frac{6\sqrt{2}B_nqKm}{\delta}\sqrt{\ln(2p+1)} \\
&\quad + \sqrt{K \ln\left[\left(1+\frac{18B_nKqA_{\boldsymbol{\beta}}}{\delta}\right)\left(1+\frac{18B_nKA_{\boldsymbol{\gamma}}}{\delta}\right)\left(1+\frac{18B_nKq\sqrt{q}}{a_{\boldsymbol{\Sigma}}\delta}\right)\right]}\Bigg] \\
&\leq \sum_{s=1}^{S} 2^{-s} \Bigg[\sqrt{\ln 2} + \frac{2^s6\sqrt{2}B_nqKm}{R_n}\sqrt{\ln(2p+1)} \\
&\quad + \sqrt{K \ln\left[\left(1+\frac{2^s18B_nKqA_{\boldsymbol{\beta}}}{R_n}\right)\left(1+\frac{2^s18B_nKA_{\boldsymbol{\gamma}}}{R_n}\right)\left(1+\frac{2^s18B_nKq\sqrt{q}}{a_{\boldsymbol{\Sigma}}R_n}\right)\right]}\Bigg].
\end{aligned} \tag{4.2.71}
$$

Notice from (4.2.70), that $R_n \geq 2KB_n \max\left(A_{\boldsymbol{\gamma}}, qA_{\boldsymbol{\beta}}, \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)$. Moreover, it holds that $1 \leq 2^{s+3}$, and $\sum_{s=1}^{S} 2^{-s} = 1 - 2^{-S} \leq 1, \sum_{s=1}^{S} \left(\sqrt{e}/2\right)^s \leq \sqrt{e}/\left(2-\sqrt{e}\right)$, and since for all $s \in \mathbb{N}^\star, e^s \geq s$, and thus

$2^{-s}\sqrt{s} \leq (\sqrt{e}/2)^s$. Therefore, from (4.2.71):

$$
\begin{aligned}
&\sum_{s=1}^{S} 2^{-s} \sqrt{\ln\left[1 + M\left(2^{-s}R_n, F_m, \|\cdot\|_n\right)\right]} \\
&\leq \sum_{s=1}^{S} 2^{-s} \left[\sqrt{\ln 2} + \frac{2^s 6\sqrt{2} B_n q K m}{R_n} \sqrt{\ln\left(2p+1\right)} + \sqrt{K \ln\left[\left(2^{s+1}3^2\right)\left(2^{s+1}3^2\right)\left(2^{s+1}3^2\right)\right]}\right] \\
&= \sum_{s=1}^{S} 2^{-s} \left[\sqrt{\ln 2} + \frac{2^s 6\sqrt{2} B_n q K m}{R_n} \sqrt{\ln(2p+1)} + \sqrt{K}\sqrt{3\left((s+1)\ln 2 + 2\ln 3\right)}\right] \\
&\leq \frac{6\sqrt{2} B_n K q m}{R_n} \sqrt{\ln(2p+1)} S + \sqrt{K}\sqrt{3\ln 2} \sum_{s=1}^{S} 2^{-s}\sqrt{s} + \sqrt{\ln 2}\left(1 + \sqrt{3K}\right) + \sqrt{6\ln 3 K} \\
&\leq \frac{6\sqrt{2} B_n K q m}{R_n} \sqrt{\ln(2p+1)} S + \sqrt{K}\sqrt{3\ln 2} \sum_{s=1}^{S} \left(\frac{\sqrt{e}}{2}\right)^s + \sqrt{\ln 2}\left(1 + \sqrt{3K}\right) + \sqrt{6\ln 3 K} \\
&\leq \frac{6\sqrt{2} B_n q K m}{R_n} \sqrt{\ln(2p+1)} S + \sqrt{K \ln 2} \underbrace{\left(\frac{\sqrt{3e}}{2-\sqrt{e}} + 1 + \sqrt{3} + \sqrt{\frac{6\ln 3}{\ln 2}}\right)}_{=:C_1}.
\end{aligned} \tag{4.2.72}
$$

Then, from (4.2.62) and (4.2.72), for all $S \in \mathbb{N}^\star$:

$$
\mathbb{E}\left[\sup_{f_m \in \mathcal{F}_m} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_m(\mathbf{Z}_i)\right|\right] \leq R_n \left[\frac{6}{\sqrt{n}}\left(\frac{6\sqrt{2} B_n K m q}{R_n}\sqrt{\ln(2p+1)} S + \sqrt{K\ln 2} C_1\right) + 2^{-S}\right]. \tag{4.2.73}
$$

We choose $S = \ln n / \ln 2$ so that the two terms depending on S in (4.2.73) are of the same order. In particular, for this value of S, $2^{-S} \leq 1/n$, and we deduce from (4.2.73) and (4.2.70) that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{Z}_{[n]}}\left[\sup_{f_m \in \mathcal{F}_m} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_m(\mathbf{Z}_i)\right|\right] \\
&\leq \frac{36\sqrt{2} B_n K m q}{\sqrt{n}} \sqrt{\ln(2p+1)} \frac{\ln n}{\ln 2} + 2K B_n \left(A_{\boldsymbol{\gamma}} + q A_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\left(6\sqrt{\ln 2} C_1 \frac{\sqrt{K}}{\sqrt{n}} + \frac{1}{n}\right) \\
&\leq \frac{B_n K m q}{\sqrt{n}} \sqrt{\ln(2p+1)} \ln n \underbrace{\frac{36\sqrt{2}}{\ln 2}}_{\approx 73.45} + \frac{K\sqrt{K}}{\sqrt{n}} B_n \left(A_{\boldsymbol{\gamma}} + q A_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right) \underbrace{2\left(6\sqrt{\ln 2} C_1 + 1\right)}_{\approx 141.32} \\
&< \frac{74 K B_n}{\sqrt{n}} \left[mq\sqrt{\ln(2p+1)} \ln n + 2\sqrt{K}\left(A_{\boldsymbol{\gamma}} + q A_{\boldsymbol{\beta}} + \frac{q\sqrt{q}}{a_{\boldsymbol{\Sigma}}}\right)\right].
\end{aligned}
$$

4.2.4.3 Proof of Lemma 4.2.1

Since $\ln(z)$ is concave in z, Jensen's inequality implies that $\ln\left(\mathbb{E}_Z\left[Z\right]\right) \geq \mathbb{E}_Z\left[\ln\left(Z\right)\right]$, where Z is a random variable. Thus, for all $x \in \mathcal{X}$, Jensen's inequality and Lemma 4.2.15 lead us the following

upper bound

$$
\begin{aligned}
&\int_{\mathbb{R}^q} \ln\left(s_0\left(y|x\right)\right) s_0\left(y|x\right) dy \\
&= \int_{\mathbb{R}^q} \ln\left[s_0\left(y|x\right)\right] \sum_{k=1}^{K} \left[g_k\left(x;\gamma_0\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right)\right] dy \\
&= \sum_{k=1}^{K} g_k\left(x;\gamma_0\right) \int_{\mathbb{R}^q} \ln\left[s_0\left(y|x\right)\right] \phi\left(y; v_{0k}(x), \Sigma_{0k}\right) dy \\
&\leq \sum_{k=1}^{K} g_k\left(x;\gamma_0\right) \ln\left[\int_{\mathbb{R}^q} s_0\left(y|x\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right) dy\right] \\
&= \sum_{k=1}^{K} g_k\left(x;\gamma_0\right) \ln\left[\int_{\mathbb{R}^q} \sum_{l=1}^{K} g_l\left(x;\gamma_0\right) \phi\left(y; v_{0l}(x), \Sigma_{0l}\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right) dy\right] \\
&= \sum_{k=1}^{K} g_k\left(x;\gamma_0\right) \ln\left[\sum_{l=1}^{K} g_l\left(x;\gamma_0\right) \int_{\mathbb{R}^q} \phi\left(y; v_{0l}(x), \Sigma_{0l}\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right) dy\right] \\
&\leq \sum_{k=1}^{K} g_k\left(x;\gamma_0\right) \ln\left[\sum_{l=1}^{K} g_l\left(x;\gamma_0\right) C_{s_0}\right], C_{s_0} = (2\pi)^{-q/2}\left(2A_{\Sigma}^{-1}\right)^{-q/2}, \text{(using Lemma 4.2.15)} \\
&= \ln C_{s_0} < \infty.
\end{aligned} \tag{4.2.74}
$$

Therefore, we obtain

$$
\max\left\{0, \sup_{x\in\mathcal{X}} \int_{\mathbb{R}^q} \ln\left(s_0\left(y|x\right)\right) s_0\left(y|x\right) dy\right\} \leq \max\left\{0, \ln C_{s_0}\right\} =: H_{s_0} < \infty.
$$

Next, we state the following important Lemma Lemma 4.2.15, which is used in the proof of Lemma 4.2.1.

Lemma 4.2.15. *There exists a positive constant* $C_{s_0} := (2\pi)^{-q/2}\left(2A_{\Sigma}^{-1}\right)^{-q/2}$, $0 < C_{s_0} < \infty$, *such that for all* $k \in [K], l \in [L]$,

$$
\int_{\mathbb{R}^q} \phi\left(y; v_{0l}(x), \Sigma_{0l}\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right) dy < C_{s_0}, \quad \forall x \in \mathcal{X}. \tag{4.2.75}
$$

Proof of Lemma 4.2.15. Firstly, for all $k \in [K], l \in [L]$, given

$$
c_{lk}(x) = C_{lk}\left[\Sigma_{0l}^{-1} v_{0l}(x) + \Sigma_{0k}^{-1} v_{0k}(x)\right], \quad C_{lk} = \left(\Sigma_{0l}^{-1} + \Sigma_{0k}^{-1}\right)^{-1},
$$

Lemma 4.2.25 leads to

$$
\begin{aligned}
&\int_{\mathbb{R}^q} \left[\phi\left(y; v_{0l}(x), \Sigma_{0l}\right) \phi\left(y; v_{0k}(x), \Sigma_{0k}\right)\right] dy \\
&= Z_{lk}^{-1} \underbrace{\int_{\mathbb{R}^q} \phi\left(y; c_{lk}(x), C_{lk}\right) dy}_{=1}, \text{ where} \\
&= (2\pi)^{-q/2} \det\left(\Sigma_{0l} + \Sigma_{0k}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(v_{0l}(x) - v_{0k}(x)\right)^{\top}\left(\Sigma_{0l} + \Sigma_{0k}\right)^{-1}\left(v_{0l}(x) - v_{0k}(x)\right)\right)
\end{aligned} \tag{4.2.76}
$$

Next, since the determinant is the product of the eigenvalues, counted with multiplicity, and Weyl's inequality, see *e.g.,* Lemma 4.2.26, for all $k \in [K], l \in [L]$, we have

$$
\begin{aligned}
\det\left(\Sigma_{0l} + \Sigma_{0k}\right) &\geq \left[m\left(\Sigma_{0l} + \Sigma_{0k}\right)\right]^q \\
&\geq \left[m\left(\Sigma_{0l}\right) + m\left(\Sigma_{0k}\right)\right]^q \text{ (using (4.2.96) from Lemma 4.2.26)} \\
&= \left[M\left(\Sigma_{0l}^{-1}\right)^{-1} + M\left(\Sigma_{0k}^{-1}\right)^{-1}\right]^q \\
&\geq \left(2A_{\Sigma}^{-1}\right)^q \text{ (using boundedness assumptions in (4.2.4))}.
\end{aligned}
$$

Therefore, for all $k \in [K], l \in [L]$, it holds that

$$\det\left(\Sigma_{0l}+\Sigma_{0k}\right)^{-1/2} \leq \left(2A_{\Sigma}^{-1}\right)^{-q/2} \text{(using boundedness assumptions in (4.2.4))}. \tag{4.2.77}$$

Since $\left(\Sigma_{0l}+\Sigma_{0k}\right)^{-1}$ is a positive definite matrix, it holds that

$$\left(v_{0l}(x)-v_{0k}(x)\right)^{\top}\left(\Sigma_{0l}+\Sigma_{0k}\right)^{-1}\left(v_{0l}(x)-v_{0k}(x)\right) \geq 0, \quad \forall x \in \mathcal{X}, l \in [L], k \in [K].$$

Then, since the exponential function is increasing, $\forall x \in \mathcal{X}, l \in [L], k \in [K]$, we have

$$\exp\left(-\frac{1}{2}\left(v_{0l}(x)-v_{0k}(x)\right)^{\top}\left(\Sigma_{0l}+\Sigma_{0k}\right)^{-1}\left(v_{0l}(x)-v_{0k}(x)\right)\right) \leq \exp(0)=1. \tag{4.2.78}$$

Finally, from (4.2.76), (4.2.77) and (4.2.78), we obtain

$$\int_{\mathbb{R}^q}\left[\phi\left(y; v_{0l}(x), \Sigma_{0l}\right)\phi\left(y; v_{0k}(x), \Sigma_{0k}\right)\right]dy \leq (2\pi)^{-q/2}\left(2A_{\Sigma}^{-1}\right)^{-q/2} =: C_{s_0} < \infty.$$

□

4.2.5 Technical results

We denote the vector space of all q-by-q real matrices by $\mathbb{R}^{q\times q}$ ($q \in \mathbb{N}^{\star}$):

$$\boldsymbol{A} \in \mathbb{R}^{q\times q} \Longleftrightarrow \boldsymbol{A}=(A_{i,j})=\begin{bmatrix} A_{1,1} & \cdots & A_{1,q} \\ \vdots & & \vdots \\ A_{q,1} & \cdots & A_{q,q} \end{bmatrix}, A_{i,j} \in \mathbb{R}.$$

If a capital letter is used to denote a matrix (*e.g.*, $\boldsymbol{A}, \boldsymbol{B}$), then the corresponding lower-case letter with subscript i, j refers to the (i, j)th entry (*e.g.*, $A_{i,j}, B_{i,j}$). When required, we also designate the elements of a matrix with the notation $[\boldsymbol{A}]_{i,j}$ or $\boldsymbol{A}(i, j)$. Denote the q-by-q identity and zero matrices by I_q and $\mathbf{0}_q$, respectively.

Lemma 4.2.16 (Derivative of quadratic form, Magnus & Neudecker, 2019)**.** *Assume that $\boldsymbol{X}$ and $\boldsymbol{a}$ are non-singular matrix in $\mathbb{R}^{q\times q}$ and vector in $\mathbb{R}^{q\times 1}$, respectively. Then*

$$\frac{\partial \boldsymbol{a}^{\top}\boldsymbol{X}^{-1}\boldsymbol{a}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-1}\boldsymbol{a}\boldsymbol{a}^{\top}\boldsymbol{X}^{-1}.$$

Lemma 4.2.17 (Jacobi's formula, Theorem 8.1 from Magnus & Neudecker, 2019)**.** *If $\boldsymbol{X}$ is a differentiable map from the real numbers to q-by-q matrices,*

$$\frac{d}{dt}\det\left(\boldsymbol{X}(t)\right) = tr\left(\operatorname{Adj}\left(\boldsymbol{X}(t)\right)\frac{d\boldsymbol{X}(t)}{dt}\right).$$

In particular,

$$\frac{\partial \det(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left(\operatorname{Adj}(\boldsymbol{X})\right)^{\top} = \det(\boldsymbol{X})\left(\boldsymbol{X}^{-1}\right)^{\top}.$$

Definition 4.2.18 (Operator (induced) p-norm)**.** We recall an operator (induced) p-norms of a matrix $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ ($q \in \mathbb{N}^{\star}, p \in \{1, 2, \infty\}$),

$$\|\boldsymbol{A}\|_p = \max_{\mathbf{x}\neq 0}\frac{\|\boldsymbol{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\mathbf{x}\neq 0}\left\|\boldsymbol{A}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_p}\right)\right\|_p = \max_{\|\mathbf{x}\|_p=1}\|\boldsymbol{A}\mathbf{x}\|_p, \tag{4.2.79}$$

where for all $\mathbf{x}=(x_i)_{i\in[q]} \in \mathbb{R}^q$,

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_1 = \sum_{i=1}^{q}|x_i| \leq q\|\mathbf{x}\|_{\infty}, \tag{4.2.80}$$

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{q}|x_i|^2\right)^{\frac{1}{2}} = \left(\mathbf{x}^{\top}\mathbf{x}\right)^{\frac{1}{2}} \leq \|\mathbf{x}\|_1 \leq \sqrt{q}\|\mathbf{x}\|_2, \text{and} \tag{4.2.81}$$

$$\|\mathbf{x}\|_{\infty} = \max_{1\leq i\leq q}|x_i| \leq \|\mathbf{x}\|_2 \leq \sqrt{q}\|\mathbf{x}\|_{\infty}. \tag{4.2.82}$$

Lemma 4.2.19 (Some matrix p-norm properties, Golub & Van Loan, 2012)**.** *By definition, we always have the important property that for every* $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ *and* $\mathbf{x} \in \mathbb{R}^q$,

$$\|\boldsymbol{A}\mathbf{x}\|_p \leq \|\boldsymbol{A}\|_p \|\mathbf{x}\|_p , \tag{4.2.83}$$

and every induced p-norm is submultiplicative, i.e., *for every* $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ *and* $\boldsymbol{B} \in \mathbb{R}^{q\times q}$,

$$\|\boldsymbol{A}\boldsymbol{B}\|_p \leq \|\boldsymbol{A}\|_p \|\boldsymbol{B}\|_p . \tag{4.2.84}$$

In particular, it holds that

$$\|\boldsymbol{A}\|_1 = \max_{1\leq j\leq q} \sum_{i=1}^{q} |A_{ij}| \leq \sum_{j=1}^{q}\sum_{i=1}^{q} |A_{ij}| := \|\mathrm{vec}(\boldsymbol{A})\|_1 \leq q \|\boldsymbol{A}\|_1 , \tag{4.2.85}$$

$$\|\mathrm{vec}(\boldsymbol{A})\|_\infty := \max_{1\leq i\leq q, 1\leq j\leq q} |A_{ij}| \leq \|\boldsymbol{A}\|_\infty = \max_{1\leq j\leq q} \sum_{i=1}^{q} |A_{ij}| \leq q \|\mathrm{vec}(\boldsymbol{A})\|_\infty , \tag{4.2.86}$$

$$\|\mathrm{vec}(\boldsymbol{A})\|_\infty \leq \|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A}) \leq q \|\mathrm{vec}(\boldsymbol{A})\|_\infty , \tag{4.2.87}$$

where $\lambda_{\max}$ *is the largest eigenvalue of a positive definite symmetric matrix* $\boldsymbol{A}$*. The p-norms, when* $p \in \{1, 2, \infty\}$*, satisfy*

$$\frac{1}{\sqrt{q}} \|\boldsymbol{A}\|_\infty \leq \|\boldsymbol{A}\|_2 \leq \sqrt{q} \|\boldsymbol{A}\|_\infty , \tag{4.2.88}$$

$$\frac{1}{\sqrt{q}} \|\boldsymbol{A}\|_1 \leq \|\boldsymbol{A}\|_2 \leq \sqrt{q} \|\boldsymbol{A}\|_1 . \tag{4.2.89}$$

Given $\delta > 0$, we need to define the δ-packing number and δ-covering number.

Definition 4.2.20 (δ-packing number,*e.g.,* Definition 5.4 from Wainwright, 2019)**.** Let $(\mathcal{F}, \|\cdot\|)$ be a normed space and let $\mathcal{G} \subset \mathcal{F}$. With $(g_i)_{i=1,\ldots,m} \in \mathcal{G}$, $\{g_1, \ldots, g_m\}$ is an δ-packing of $\mathcal{G}$ of size $m \in \mathbb{N}^\star$, if $\|g_i - g_j\| > \delta, \forall i \neq j, i, j \in \{1, \ldots, m\}$, or equivalently, $\bigcap_{i=1}^{n} \boldsymbol{B}(g_i, \delta/2) = \emptyset$. Upon defining δ-packing, we can measure the maximal number of disjoint closed balls with radius $\delta/2$ that can be "packed" into $\mathcal{G}$. This number is called the *δ-packing number* and is defined as

$$M(\delta, \mathcal{G}, \|\cdot\|) := \max\{m \in \mathbb{N}^\star : \exists \delta\text{-packing of } \mathcal{G} \text{ of size } m\} . \tag{4.2.90}$$

Definition 4.2.21 (δ-covering number, Definition 5.1 from Wainwright, 2019)**.** Let $(\mathcal{F}, \|\cdot\|)$ be a normed space and let $\mathcal{G} \subset \mathcal{F}$. With $(g_i)_{i=1,\ldots,n} \in \mathcal{G}$, $\{g_1, \ldots, g_n\}$ is an δ-covering of $\mathcal{G}$ of size n if $\mathcal{G} \subset \cup_{i=1}^{n} \boldsymbol{B}(g_i, \delta)$, or equivalently, $\forall g \in \mathcal{G}, \exists i$ such that $\|g - g_i\| \leq \delta$. Upon defining the δ-covering, we can measure the minimal number of closed balls with radius δ, which is necessary to cover $\mathcal{G}$. This number is called the *δ-covering number* and is defined as

$$N(\delta, \mathcal{G}, \|\cdot\|) := \min\{n \in \mathbb{N}^\star : \exists \delta\text{-covering of } \mathcal{G} \text{ of size } n\} . \tag{4.2.91}$$

The covering entropy (metric entropy) is defined as follows $H_{\|\cdot\|}(\delta, \mathcal{G}) = \ln(N(\delta, \mathcal{G}, \|\cdot\|))$.

The relation between the packing number and the covering number is described in the following lemma.

Lemma 4.2.22 (Lemma 5.5 from Wainwright, 2019)**.** *Let* $(\mathcal{F}, \|\cdot\|)$ *be a normed space and let* $\mathcal{G} \subset \mathcal{F}$*. Then*

$$M(2\delta, \mathcal{G}, \|\cdot\|) \leq N(\delta, \mathcal{G}, \|\cdot\|) \leq M(\delta, \mathcal{G}, \|\cdot\|) .$$

Lemma 4.2.23 (Chernoff's inequality, *e.g.,* Chapter 2 in Wainwright, 2019)**.** *Assume that the random variable has a moment generating function in a neighborhood of zero, meaning that there is some*

constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}\left[e^{\lambda(U-\mu)}\right]$ exists for all $\lambda \leq |b|$. In such a case, we may apply Markov's inequality to the random variable $Y = e^{\lambda(U-\mu)}$, thereby obtaining the upper bound

$$\mathbb{P}\left(U - \mu \geq a\right) = \mathbb{P}\left(e^{\lambda(U-\mu)} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left[e^{\lambda(U-\mu)}\right]}{e^{\lambda t}}.$$

Optimizing our choice of λ so as to obtain the tightest result yields the Chernoff bound

$$\ln\left(\mathbb{P}\left(U - \mu \geq a\right)\right) \leq \sup_{\lambda \in [0,b]} \left\{\lambda t - \ln\left(\mathbb{E}\left[e^{\lambda(U-\mu)}\right]\right)\right\}. \tag{4.2.92}$$

In particular, if $U \sim \mathcal{N}(\mu, \sigma)$ is a Gaussian random variable with mean μ and variance σ^2. By a straightforward calculation, we find that U has the moment generating function

$$\mathbb{E}\left[e^{\lambda U}\right] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \text{valid for all } \lambda \in \mathbb{R}.$$

Substituting this expression into the optimization problem defining the optimized Chernoff bound (4.2.92), we obtain

$$\sup_{\lambda \geq 0} \left\{\lambda t - \ln\left(\mathbb{E}\left[e^{\lambda(U-\mu)}\right]\right)\right\} = \sup_{\lambda \geq 0} \left\{\lambda t - \frac{\sigma^2\lambda^2}{2}\right\} = -\frac{t^2}{2\sigma^2},$$

where we have taken derivatives in order to find the optimum of this quadratic function. So, (4.2.92) leads to

$$\mathbb{P}\left(X \geq \mu + t\right) \leq e^{-\frac{t^2}{2\sigma^2}}, \text{ for all } t \geq 0. \tag{4.2.93}$$

Recall that a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p), \alpha_i \in \mathbb{N}^\star, \forall i \in [p]$ is an p-tuple of non-negative integers. Let

$$|\boldsymbol{\alpha}| = \sum_{i=1}^{p} \alpha_i, \quad \boldsymbol{\alpha}! = \prod_{i=1}^{p} \alpha_i!,$$
$$\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^{p} \mathbf{x}_i^{\alpha_i}, \mathbf{x} \in \mathbb{R}^p, \quad \partial^{\boldsymbol{\alpha}} f = \partial_1^{\alpha_1}\partial_2^{\alpha_2}\cdots\partial_p^{\alpha_p} = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial \mathbf{x}_1^{\alpha_1}\partial \mathbf{x}_2^{\alpha_2}\cdots\partial \mathbf{x}_p^{\alpha_p}}.$$

The number $|\boldsymbol{\alpha}|$ is called the *order* or *degree* of $\boldsymbol{\alpha}$. Thus, the order of $\boldsymbol{\alpha}$ is the same as the order of $\mathbf{x}^\alpha$ as a monomial or the order of $\partial^{\boldsymbol{\alpha}}$ as a partial derivative.

Lemma 4.2.24 (Taylor's Theorem in Several Variables from Duistermaat & Kolk, 2004)**.** *Suppose $f : \mathbb{R}^p \mapsto \mathbb{R}$ is in the class C^{k+1}, of continuously differentiable functions, on an open convex set $\mathcal{S}$. If $\boldsymbol{a} \in \mathcal{S}$ and $\boldsymbol{a} + \boldsymbol{h} \in \mathcal{S}$, then*

$$f(\boldsymbol{a} + \boldsymbol{h}) = \sum_{|\boldsymbol{\alpha}| \leq k} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{a})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} + R_{\boldsymbol{a},k}(\boldsymbol{h}),$$

where the remainder is given in Lagrange's form by

$$R_{\boldsymbol{a},k}(\boldsymbol{h}) = \sum_{|\boldsymbol{\alpha}| = k+1} \partial^{\boldsymbol{\alpha}} f(\boldsymbol{a} + ch)\frac{\boldsymbol{h}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \text{ for some } c \in (0,1),$$

or in integral form by

$$R_{\boldsymbol{a},k}(\boldsymbol{h}) = (k+1)\sum_{|\boldsymbol{\alpha}| = k+1} \frac{\boldsymbol{h}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}\int_0^1 (1-t)^k \partial^{\boldsymbol{\alpha}} f(\boldsymbol{a} + th)dt.$$

In particular, we can estimate the remainder term if $|\partial^{\boldsymbol{\alpha}} f(\mathbf{x})| \leq M$ for $\mathbf{x} \in \mathcal{S}$ and $|\boldsymbol{\alpha}| = k+1$, then

$$|R_{\boldsymbol{a},k}(\boldsymbol{h})| \leq \frac{M}{(k+1)!}\|\boldsymbol{h}\|_1^{k+1}, \|\boldsymbol{h}\|_1 = \sum_{i=1}^{p} |h_i|.$$

Recall that the multivariate Gaussian (or Normal) distribution has a joint density given by

$$\phi\left(y;\mu;\Sigma\right)=\left(2\pi\right)^{-q/2}\det\left(\Sigma\right)^{-1/2}\exp\left(-\frac{1}{2}\left(y-\mu\right)^{\top}\Sigma^{-1}\left(y-\mu\right)\right), \tag{4.2.94}$$

where μ is the mean vector (of length q) and Σ is the symmetric, positive definite covariance matrix (of size $q\times q$). Then, we have the following well-known Gaussian identity, see more in Lemma 4.2.25.

Lemma 4.2.25 (Product of two Gaussians, see *e.g.,* Equation (A.7) in Williams & Rasmussen (2006))**.** *The product of two Gaussians gives another (un-normalized) Gaussian*

$$\begin{aligned}\phi\left(y;a,A\right)\phi\left(y;b,B\right)&=Z^{-1}\phi\left(y;c,C\right),\ \textit{where,}\\ c&=C\left(A^{-1}a+B^{-1}b\right)\ \textit{and}\ C=\left(A^{-1}+B^{-1}\right)^{-1},\\ Z^{-1}&=\left(2\pi\right)^{-q/2}\det\left(A+B\right)^{-1/2}\exp\left(-\frac{1}{2}\left(a-b\right)^{\top}\left(A+B\right)^{-1}\left(a-b\right)\right).\end{aligned} \tag{4.2.95}$$

We recall the following inequality of Hermann Weyl, see *e.g.,* Horn & Johnson (2012, Theorem 4.3.1)

Lemma 4.2.26 (Weyl's inequality, *e.g.,* Theorem 4.3.1 from Horn & Johnson (2012))**.** *Let $A,B\in\mathbb{R}^{q\times q}$ be Hermitian and let the respective eigenvalues of A,B, and $A+B$ be $\left\{\lambda_i\left(A\right)\right\}_{i\in[q]}$, $\left\{\lambda_i\left(B\right)\right\}_{i\in[q]}$, and $\left\{\lambda_i\left(A+B\right)\right\}_{i\in[q]}$, each algebraically nondecreasing order as follows:*

$$m\left(A\right)=\lambda_1\left(A\right)\leq\lambda_2\left(A\right)\leq\ldots\leq\lambda_q\left(A\right)=M\left(A\right).$$

Then, for each $i\in[q]$,

$$\begin{aligned}\lambda_i\left(A+B\right)&\leq\lambda_{i+j}\left(A\right)+\lambda_{q-j}\left(B\right),\quad j\in\{0\}\cup[q-i],\\ \lambda_{i-j+1}\left(A\right)+\lambda_j\left(B\right)&\leq\lambda_i\left(A+B\right),\quad j\in[i].\end{aligned}$$

In particular, we have

$$\begin{aligned}M\left(A+B\right)&\leq M\left(A\right)+M\left(B\right),\\ m\left(A+B\right)&\geq m\left(A\right)+m\left(B\right).\end{aligned} \tag{4.2.96}$$

Given vectors $z,z'\in\mathbb{R}^n$ and an index $k\in[n]$, we define a new vector $z^{\backslash k}$ as follows

$$z_j^{\backslash k}:=\begin{cases}z_j & \text{if } j\neq k,\\ z'_j & \text{if } j=k.\end{cases} \tag{4.2.97}$$

4.3 Joint rank and variable selection by a non-asymptotic model selection in the softmax-gated block-diagonal mixture of experts regression model

The goal of Section 4.3 is to provide a treatment regarding penalizations that guarantee an Lasso $+l_2$-Rank-oracle inequality of PSGaBloME. As such, the remainder of the article progresses as follows. In Section 4.3.1, we discuss the construction the framework for PSGaBloME regression models and its collection of models. Section 4.3.5 is aimed to introduce the Lasso $+l_2$-Rank procedure for PSGaBloME regression models to deal with high-dimensional data. In Section 4.3.2, we state one of the main results of this thesis: a finite-sample oracle inequality satisfied by PMLEs in PSGaBloME regression models. Section 4.3.3 is devoted to the proof of these main results based on a general model selection theorem. Some conclusions and proofs of lemmas can be founded in Section 5.4 and Section 4.3.4, respectively.

4.3.1 Notation and framework

4.3.1.1 PSGaBloME models

We will consider the statistical frameworks in which we model a sample of high-dimensional regression data issued from a heterogeneous population via a suitable MoE model with softmax gating functions. We emphasize that the dimension of the input $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ and/or the output $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^q$ variable are/is typically much higher than the sample size n.

In Section 4.3, based on the original MoE models from Jacobs et al. (1991), we aim to establish a MoE model with softmax gating functions as generic as possible such that it can be used to handle with high-dimensional regression datasets and to study oracle inequalities. To do that, we first define $s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x})$ to be a conditional PDF of MoE model as follows:

$$s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} g_{\mathbf{w},k}(\mathbf{x})\, \phi_q\left(\mathbf{y}; \boldsymbol{\upsilon}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), \text{ where,} \tag{4.3.1}$$

$$g_{\mathbf{w},k}(\mathbf{x}) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\mathbf{x})\right)}, \mathbf{w}(\mathbf{x}) = \left(w_k(\mathbf{x})\right)_{k\in[K]}. \tag{4.3.2}$$

Here, $g_{\mathbf{w},k}(\cdot)$ and $\phi_q\left(\cdot; \boldsymbol{\upsilon}_k(\cdot), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), k \in [K]$, are called softmax gating functions (or gating networks) and Gaussian experts, respectively. Note that for every $\mathbf{x} \in \mathcal{X}$, $\left(g_{\mathbf{w},k}(\mathbf{x})\right)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$. Furthermore, we decompose the set of model parameters as follows: $\boldsymbol{\psi}_K = (\mathbf{w}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \mathbf{W}_K \times \boldsymbol{\Upsilon}_K \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_K$, $\mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_K$, $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}_k)_{k\in[K]} \in \boldsymbol{\Upsilon}_K$, and $\boldsymbol{\Sigma}(\mathbf{B}) = \left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)\right)_{k\in[K]} \in \boldsymbol{\Sigma}_k(\mathbf{B}_k)$. It is worth noting that $\mathbf{W}_K$ and $\boldsymbol{\Upsilon}_K$ are sets of K-tuples of functions defined in logistic schemes (weights) and mean functions from $\mathbb{R}^p$ to $\mathbb{R}^+$ and $\mathbb{R}^p$ to $\mathbb{R}^q$, respectively; and $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$ is a set containing K-tuples of $\mathcal{S}_q^{++}$ with the block-diagonal structures defined in (1.1.14), where $\mathcal{S}_q^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^q$. Since we need to bound the model complexity using the dimension of model, we have to restrict our attention to LinBoSGaBloME models, where $\mathbf{W}_K$ and $\boldsymbol{\Upsilon}_K$ are defined as the linear combination of a finite set of bounded functions whose coefficients belong to a compact set. When the dimension of both inputs and outputs are not too large, we do not need to select relevant variables. Then, we can work on the previous LinBoSGaBloME models with general structures for means, weights and multi-block-diagonal covariance matrices. In some situation, we do not need to take into account the trade-off between complexity and sparsity for covariance matrices, in LinBoSGaBloME models, we can consider 1-block-diagonal covariance matrices, which is well studied in Montuelle et al. (2014) and will be referred to be as *linear-combination-of-bounded-functions softmax-gated mixture of experts* (LinBoSGaME) regression models. However, to deal with high-dimensional data and to simplify the interpretation of sparsity, in LinBoSGaBloME model, we propose to utilize polynomials for the weights of the softmax gating functions and the Gaussian expert means, which will be referred to as *polynomial softmax-gated block-diagonal mixture of experts* (PSGaBloME) regression models. In particular, we simply refer to affine instances of LinBoSGaBloME models as *softmax-gated mixture of experts* (SGaME) regression models.

In order to establish our oracle inequality, Theorem 4.3.2, we need to assume that $\mathcal{X}$ is a bounded set in $\mathbb{R}^p$ and make explicit some classical boundedness conditions on the parameter space. We further assume that the covariates $\mathbf{X}$ belong to an hypercube, *e.g.,* $\mathcal{X} = [0,1]^p$, for the simplicity of notation. In particular, just for the interpretation of sparsity, the weights of softmax gating functions and means of Gaussian experts are considered as monomials and polynomial functions of the explanatory variables, respectively.

Linear combination of bounded functions for the weights and the means

We follow the idea from Montuelle et al. (2014) to restrict our attention on a finite set of bounded functions whose coefficients belong to a compact set. It is worth mentioning that such quite general setting includes the polynomial basis when the predictors are bounded, the suitable renormalized wavelet dictionaries as well as the Fourier basis on an interval. More precisely, we first define the following two collections of bounded functions for the weights and means: $\mathcal{X} \ni \mathbf{x} \mapsto \left(\theta_{\mathbf{W},d}(\mathbf{x})\right)_{d\in[d_{\mathbf{W}}]} \in$

$[-1,1]^{d_{\mathbf{W}}}$ and $\mathcal{X} \ni \mathbf{x} \mapsto \left(\theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right)\right)_{d\in[d_{\boldsymbol{\Upsilon}}]} \in [-1,1]^{d_{\boldsymbol{\Upsilon}}}$, where $d_{\mathbf{W}} \in \mathbb{N}^{\star}$ and $d_{\boldsymbol{\Upsilon}} \in \mathbb{N}^{\star}$ indicate its degrees, respectively. Then, by making use of these collections, we are able to define the corresponding desired bounded spaces via tensorial constructions as follows:

$$\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}^{K-1}, \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{d=1}^{d_{\mathbf{W}}} \omega_d \theta_{\mathbf{W},d}\left(\mathbf{x}\right) \in \mathbb{R} : \max_{d\in[d_{\mathbf{W}}]} |\omega_d| \leq T_{\mathbf{W}} \right\},$$

$$\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}} = \boldsymbol{\Upsilon}^{K}, \boldsymbol{\Upsilon} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(\sum_{d=1}^{d_{\boldsymbol{\Upsilon}}} \beta_d^{(z)} \theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right) \right)_{z\in[q]} : \max_{d\in[d_{\boldsymbol{\Upsilon}}],z\in[q]} \left|\beta_d^{(z)}\right| \leq T_{\boldsymbol{\Upsilon}} \right\}. \tag{4.3.3}$$

We do not need to select relevant variables. Then, we can work on the previous LinBoSGaBloME models with general structures for means, weights and multi-block-diagonal covariance matrices or with LinBoSGaME models as in Montuelle et al. (2014). However, in PSGaBloME models, to handle with high-dimensional data and to simplify the interpretation of sparsity, we propose to utilize polynomials for weights and polynomial regression models for the softmax gating functions and the means of Gaussian experts as follows:

$$\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}^{K-1}, \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \max_{\boldsymbol{\alpha}\in\mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}} \right\}, \tag{4.3.4}$$

$$\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(\boldsymbol{\beta}_{k0} + \sum_{d=1}^{d_{\boldsymbol{\Upsilon}}} \boldsymbol{\beta}_{kd} \mathbf{x}^{d} \right)_{k\in[K]} : \max\left\{ \left\|\left\|\boldsymbol{\beta}_{kd}\right\|\right\|_{\infty} : k \in [K], d \in (\{0\} \cup [d_{\boldsymbol{\Upsilon}}]) \right\} \leq T_{\boldsymbol{\Upsilon}} \right\}.$$

Here, note that the multi-index $\boldsymbol{\alpha} = \left(\alpha_t\right)_{t\in[p]}, \alpha_t \in \mathbb{N}^{\star} \bigcup \{0\} =: \mathbb{N}, \forall t \in [p]$, is an p-tuple of nonnegative integers that satisfies $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{p} x_j^{\alpha_j}$ and $|\boldsymbol{\alpha}| = \sum_{t=1}^{p} \alpha_t$. Then, for all $l \in [d_{\mathbf{W}}]$, we define $\mathcal{A} = \bigcup_{l=0}^{d_{\mathbf{W}}} \mathcal{A}_l$, $\mathcal{A}_l = \left\{ \boldsymbol{\alpha} = \left(\alpha_t\right)_{t\in[p]} \in \mathbb{N}^p, |\boldsymbol{\alpha}| = l \right\}$. The number $\boldsymbol{\alpha}$ is called the order or degree of monomials $\mathbf{x}^{\boldsymbol{\alpha}}$. By using the well-known stars and bars methods, *e.g.,* Feller (1957, Chapter 2), the cardinality of the set $\mathcal{A}$, denoted by $\mathrm{card}\left(\mathcal{A}\right)$, equals $\binom{d_{\mathbf{W}}+p}{p}$. Note that, for all $d \in [d_{\boldsymbol{\Upsilon}}]$, we define $\mathbf{x}^d$ as $\left(x_j^d\right)_{j\in[p]}$ for the means, which are often used for polynomial regression models. Moreover, given any matrix $\mathbf{A} \in \mathbb{R}^{q\times p}$, the following notations are used for matrix norms: the *max norm* $\left\|\left\|\mathbf{A}\right\|\right\|_{\infty} = \max_{i\in[q],j\in[p]} \left|\left[\mathbf{A}\right]_{i,j}\right|$, the 2-*norm* $\left\|\mathbf{A}\right\|_2 = \sup_{\left\|\mathbf{x}\right\|_2=1} \left|\mathbf{x}^{\top}\mathbf{A}\mathbf{x}\right| = \sup_{\lambda\in\mathrm{vp}(\mathbf{A})} |\lambda|$, where $\mathrm{vp}(\mathbf{A})$ denotes the spectrum of $\mathbf{A}$, and the *Frobenius norm* $\left\|\mathbf{A}\right\|_F^2 = \sum_{i=1}^{q}\sum_{j=1}^{p} \left|\left[\mathbf{A}\right]_{i,j}\right|^2$. Then, it holds that $\left\|\mathbf{A}\right\|_2 \leq \left\|\mathbf{A}\right\|_F$, $\left\|\mathbf{A}\right\|_2 \leq \sqrt{qp}\left\|\left\|\mathbf{A}\right\|\right\|_{\infty}$, and for any $\mathbf{x} \in \mathbb{R}^p$, $\left\|\mathbf{x}\right\|_2 \leq \sqrt{p}\left\|\mathbf{x}\right\|_{\infty}$, $\left\|\mathbf{A}\mathbf{x}\right\|_2 \leq \left\|\mathbf{A}\right\|_2 \left\|\mathbf{x}\right\|_2$, *e.g.,* Golub & Van Loan (2013, Chapter 2).

Gaussian expert covariance matrices

For the block-diagonal covariances of Gaussian experts, we assume that there exist some positive constants λ_m and λ_M such that, for every $k \in [K]$,

$$0 < \lambda_m \leq m\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq M\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq \lambda_M. \tag{4.3.5}$$

Note that this is a quite general assumption and is also used in the block-diagonal covariance selection for Gaussian graphical models of Devijver & Gallopin (2018).

On Convergence Rates of Mixtures of Polynomial Experts We should refer to Mendes & Jiang (2012) for discussing about the optimal convergence rate on a mixture of experts structure where K experts are mixed, with each expert being related to a polynomial regression model of order $d_{\boldsymbol{\Upsilon}}$.

When using PSGaBloME to study complex data with high number of responses and covariates, the number of parameters could be quickly larger than the sample sizes. Following the same spirit of the frameworks in Devijver (2015b, 2017a,b), the extension from a linear model of Bunea et al. (2012) to finite mixture models, we propose to work with parsimonious models combing two well-know approaches: selecting relevant variables and rank sparse models; Section 4.3.1.2, respectively.

4.3.1.2 High-dimensional regression via variable selection and rank sparse models

Variable selection via selecting relevant variables

The Lasso estimator, originally established by Tibshirani (1996), is a classical choice and has been extended to deal with multiple multivariate regression models for column sparsity using the Group-Lasso estimator (Yuan & Lin, 2006). Note that the Group-Lasso penalty can be used to select a subset of variables for one choice of regularization parameter in the Lasso-Rank procedure, as done, *e.g.,* Devijver (2015b, 2017a,b) or to get a ranking of the variables, as done, e.g., in Bach (2008).

Recall that, for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $\boldsymbol{\beta}_{kd}$ is the matrix of d-th term of regression coefficients, $\boldsymbol{\Sigma}_k$ ($\mathbf{B}_k$) is the covariance matrix in the mixture component k, and the g_k is the mixture proportion k with the $\boldsymbol{\alpha}$-th order term of its monomials is $\omega_{k\boldsymbol{\alpha}}$. Furthermore, given a regressor $\mathbf{x}$, for all $k \in [K]$, for all $d \in [d_{\boldsymbol{\Upsilon}}]$ and for all $z \in [q]$, $\left[\boldsymbol{\beta}_{kd}\mathbf{x}^d\right]_z = \sum_{j=1}^p \left[\boldsymbol{\beta}_{kd}\right]_{z,j} x_j^d$ is the z-th component of the d-th terms of means for the mixture components k. In particular, for all $l \in [d_{\mathbf{W}}]$, $j \in [p]$, we define $\boldsymbol{\omega}_k^{[j,l]} = \left\{\omega_{k\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} = (\alpha_t)_{t\in[p]} \in \mathcal{A}_l, \boldsymbol{\alpha}_j > 0\right\}$.

We have to deal with high-dimensional data where we estimate many coefficients while given a small number of target variables. Therefore, we need to focus on selecting relevant variables via the notion of irrelevant indices in Definition 4.3.1.

Definition 4.3.1 (Relevant variables in PSGaBloME models)**.** A couple $(\mathbf{Y}_z, \mathbf{X}_j)$ and its corresponding indices $(z, j) \in [q] \times [p]$ are said to be *irrelevant* if, for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $l \in [d_{\mathbf{W}}]$, $\left[\boldsymbol{\beta}_{kd}\right]_{z,j} = 0, \boldsymbol{\omega}_k^{[j,l]} = \mathbf{0}$. This means that the variable $\mathbf{X}_j$ does not explain the variable $\mathbf{Y}_z$ for the regression models. A couple and its corresponding indices are relevant if they are not irrelevant. A model is said to be sparse if there are few of relevant variables. We denote by $\mathbf{J}$ the set of indices (z, j) of relevant couples $(\mathbf{Y}_z, \mathbf{X}_j)$. Then, we define the set of relevant variables (columns) as $\mathbf{J}_{\boldsymbol{\omega}} = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$. We denote by $\mathbf{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ and $\mathbf{b}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ the matrix and vector with vectors $\mathbf{0}$ on the columns indexed by the set $\mathbf{J}_{\boldsymbol{\omega}}^C$ and values 0 on the set $\mathbf{J}_{\boldsymbol{\omega}}^C$, respectively. Here, $\mathbf{J}_{\boldsymbol{\omega}}^C$ is the complement of the set $\mathbf{J}_{\boldsymbol{\omega}}$.

Remark that $\mathbf{J} \subset \mathcal{P}([q] \times [p])$ and $\mathbf{J}_{\boldsymbol{\omega}} \subset \mathcal{P}([q])$, where $\mathcal{P}([q] \times [p])$ contains all subsets of $[q] \times [p]$. In our context, we focus on the Group-Lasso estimator to detect relevant variables, where the groups correspond to the columns. Therefore, if for all $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, a matrix $\boldsymbol{\beta}_{kd}$ has $\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})$ relevant columns, there are $q \operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})$ coefficients to be estimated instead of qp per clusters and coefficient matrices. This leads to the number of parameters to be estimated is then drastically reduced when $\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}}) \ll p$. Furthermore, such column sparsity may enhance the interpretation since the responses are described by only few relevant columns. To construct the regularization for coefficients of polynomial functions, we can consider the sparse Group-Lasso estimator from Simon et al. (2013) and Hastie et al. (2015, Chapter 4).

Rank sparse models

This approach is based on rank sparse models, introduced by Anderson et al. (1998). More precisely, if regression matrices have low rank or at least can be well approximated by low-rank matrices, then its corresponding regression models are called rank sparse. In the PSGaBloME model, for every $k \in [K]$, $d \in [L]$, the matrix $\boldsymbol{\beta}_{kd}$ is fully determined by $R_{kd}(p + q - R_{kd})$ coefficients if it has rank R_{kd}. This advantage will be very useful because the total parameters to estimate may be smaller than the sample size nq. It is worth noting that such low-rank estimation generalizes the classical principal component analysis for reducing the dimension of multivariate data and appears in many applications: *e.g.,* Friston et al. (2003, 2019, analysis of fMRI image data), Anderson et al. (1998, analysis of EEG data decoding).

By combining the previous rank and column sparsity, we consider the matrices of regression coefficients $\boldsymbol{\beta}_{kd}$ of rank R_{kd} and a vector of ranks $\mathbf{R} = (R_{kd})_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}]}$ belongs to $[\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{d_{\boldsymbol{\Upsilon}}K}$, where in general, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

We describe in more detail the collection of models with relevant variables and rank sparse models in the sequel.

4.3.1.3 Collection of models

To simplify the notations, L and D stand for $\binom{d_{\mathbf{W}}+\text{card}(\mathbf{J}_{\boldsymbol{\omega}})}{\text{card}(\mathbf{J}_{\boldsymbol{\omega}})}$ and $d_{\boldsymbol{\Upsilon}}$, which are related to the dimensions of $\mathbf{W}_{K,d_{\mathbf{W}}}$ and $\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}}$, respectively. Combining all the previous structures in Sections 4.3.1.1 and 4.3.1.2 , given $\mathbf{m} = (K, L, D, \mathbf{B}, \mathbf{J}, \mathbf{R}) \in \mathbb{N}^\star \times \mathbb{N}^\star \times \mathbb{N}^\star \times (\mathcal{B}_k)_{k\in[K]} \times \mathcal{P}([q] \times [p]) \times [\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{DK}$, some real positive constants $A_{\mathbf{u},\mathbf{v}} > 0$, $A_\sigma > 0$, we obtain the following model:

$$
\begin{aligned}
S_{\mathbf{m}} &= \left\{ (\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}}(\mathbf{y}|\mathbf{x}) =: s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \in \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \right\}, \\
\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \left((\omega_{k\boldsymbol{\alpha}})_{k\in[K],\boldsymbol{\alpha}\in\mathcal{A}}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]} \right)_{k\in[K]}, (\boldsymbol{\Sigma}_k(\mathbf{B}_k))_{k\in[K]} \right) \\
&\in \left(\mathbb{R}^L\right)^{K-1} \times \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}, \\
\boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \Bigg\{ \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]}^{[\mathbf{J}_{\boldsymbol{\omega}}]} \right)_{k\in[K]} \in \left(\mathbb{R}^{q\times 1} \times \left(\mathbb{R}^{q\times p}\right)^D \right)^K : \forall k \in [K], \forall d \in [D], \\
&\qquad \boldsymbol{\beta}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} [\sigma_{kd}]_r [\mathbf{u}_{kd}]_{\bullet,r} \left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}, \text{rank}\left(\boldsymbol{\beta}_{kd}^{R_{kd}}\right) = R_{kd}, \forall r \in [R_{kd}], [\sigma_{kd}]_r < A_\sigma, \\
&\qquad \max_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}],r\in[R_{kd}]} \left\{ \|\boldsymbol{\beta}_{k0}\|_\infty, \left\|[\mathbf{u}_{kd}]_{\bullet,r}\right\|_\infty, \left\|\left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}\right\|_\infty \right\} \le A_{\mathbf{u},\mathbf{v}} \Bigg\}. \qquad (4.3.6)
\end{aligned}
$$

In the above, for $k \in [K], d \in [D]$, $[\sigma_{kd}]_r, r \in [R_{kd}]$, denote the singular values of $\boldsymbol{\beta}_{kd}^{R_{kd}}$, with corresponding orthogonal unit vectors $\left([\mathbf{u}_{kd}]_{\bullet,r}\right)_{r\in[R_{kd}]}$ and $\left(\left[\mathbf{v}_{kd}^\top\right]_{r,\bullet}\right)_{r\in[R_{kd}]}$ (Strang, 2019, I.8). The dimension of $S_{\mathbf{m}}$ is

$$
\dim(S_{\mathbf{m}}) = (K-1)L + qK + \sum_{k=1}^{K}\sum_{d=1}^{D} R_{kd}(\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) + q - R_{kd}) + \sum_{k=1}^{K}\sum_{g=1}^{G_k} \frac{\text{card}\left(d_k^{[g]}\right)\left(\text{card}\left(d_k^{[g]}\right)+1\right)}{2}.
$$

Remark that the collection of models in (4.3.6) is generally large and therefore not tractable in practice. This motivates us to restrict the numbers of components K, the orders of monomial weights L and polynomial means D among finite sets $\mathcal{K} = [K_{\max}]$, $\mathcal{L} = [L_{\max}]$ and $\mathcal{D} = [D_{\max}]$, respectively, where $K_{\max} \in \mathbb{N}^\star$, $L_{\max} \in \mathbb{N}^\star$ and $D_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. Furthermore, we focus on a (potentially random) subcollection $\mathcal{J}$ of $\mathcal{P}([q] \times [p])$, the controlled size being required in high-dimension case. Moreover, the number of possible vectors of ranks considered is reduced by working on a subset (potentially random) $\mathcal{R}_{(K,\mathbf{J},D)}$ of $[\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{DK}$.

In particular, recall that $\mathbf{B}$ is selected among a list of candidate structures $(\mathcal{B}_k)_{k\in[K]} \equiv (\mathcal{B})_{k\in[K]}$, where $\mathcal{B}$ denotes the set of all possible partitions of the covariables indexed by $[p]$ for each cluster of individuals. It is worth mentioning that the size of $\mathcal{B}$ (Bell number) is very large even for a moderate number of variables p. This prevents us to consider an exhaustive exploration of the set $\mathcal{B}$. Motivated by the recent novel work from Devijver & Gallopin (2018), for each cluster $k \in [K]$, we restrict our attention to the sub-collection $\mathcal{B}_{k,\boldsymbol{\Lambda}} = (\mathcal{B}_{k,\lambda})_{\lambda\in\boldsymbol{\Lambda}}$ of $\mathcal{B}_k$. Here $\mathcal{B}_{k,\boldsymbol{\Lambda}}$ is the partition of the variables corresponding to the block-diagonal structure of the adjacency matrix $\mathbf{E}_{k,\lambda} = \left[\mathbb{I}\left\{\left|[\mathbf{S}_k]_{z,z'}\right| > \lambda\right\}\right]_{z\in[q],z'\in[q]}$, which is based on the thresholded absolute value of the sample covariance matrix $\mathbf{S}_k$ in each cluster $k \in [K]$. It is important to point out that the class of block-diagonal structures detected by the graphical lasso algorithm when the regularization parameter varies is identical to the block-diagonal structures $\mathcal{B}_{k,\lambda}$ detected by the thresholding of the sample covariance for each cluster $k \in [K]$ (Mazumder & Hastie, 2012).

Finally, given $S_{\mathbf{m}}$ defined as in (4.3.6), our full model collection and random subcollection are defined, respectively, as follows:

$$
\mathcal{S} = \{S_{\mathbf{m}} : \mathbf{m} \in \mathcal{M}\}, \mathcal{M} = \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times (\mathcal{B}_k)_{k\in[K]} \times \mathcal{P}([q] \times [p]) \times [\text{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{D_{\max}K}, \qquad (4.3.7)
$$

$$
\widetilde{\mathcal{S}} = \left\{S_{\mathbf{m}} : \mathbf{m} \in \widetilde{\mathcal{M}}\right\}, \widetilde{\mathcal{M}} = \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times (\mathcal{B}_{k,\boldsymbol{\Lambda}})_{k\in[K]} \times \mathcal{J} \times \mathcal{R}_{(K,\mathbf{J},D_{\max})}. \qquad (4.3.8)
$$

4.3.2 Oracle inequality

In Section 4.3.2, we state our main contribution, a finite-sample oracle type inequality, which ensures that if we have penalized the log-likelihood in an approximate approach, we are able to select a model which is as good as the oracle. Note that in Section 4.3, the block-diagonal structures, the relevant variables and rank sparse models are designed, for example, by the Lasso $+l_2$-Rank procedure in Section 4.3.5. Nevertheless, our finite-sample oracle inequality in Theorem 4.3.2 still holds for any random subcollection of $\mathcal{M}$ which is constructed by some suitable tools in the framework of PSGaBloME regression models.

Theorem 4.3.2 (Oracle inequality)**.** *Let $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$ be the observations arising from the unknown conditional density s_0. For each $\mathbf{m} \equiv (K, L, D, \mathbf{B}, \mathbf{J}, \mathbf{R}) \in \mathcal{M}$, let $S_{\mathbf{m}}$ be define by (4.3.7). Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$ such that*

$$\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \bar{s}_{\mathbf{m}}) \leq \inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + \frac{\epsilon_{KL}}{n}, \text{ and } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0.$$

Furthermore, we construct a random subcollection $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ as in (4.3.8) and consider the collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of η-log likelihood minimizers defined in (3.2.17). Then, there is a constant C such that for any $\rho \in (0,1)$, and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, for every index $\mathbf{m} \in \mathcal{M}$, $\xi_{\mathbf{m}} \in \mathbb{R}^+$, $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$,

$$\mathrm{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n)\dim(S_{\mathbf{m}}) + (1 \vee \tau)\xi_{\mathbf{m}}\right], \kappa > \kappa_0,$$

the η'-penalized likelihood estimator $\widehat{s}_{\widehat{\mathbf{m}}}$, defined in (3.2.18) on the subset $\widetilde{\mathcal{M}}$ instead of $\mathcal{M}$, satisfies

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta' + \eta}{n}.$$

Remark 4.3.3. This finite-sample oracle inequality, Theorem 4.3.2, compares performances of our estimator with the best model in the collection. However, Theorem 4.3.2 allows us to approximate well a rich class of conditional densities if we take enough degree of monomials and/or polynomials of weights and Gaussian expert means, respectively, or enough clusters in the context of mixture of Gaussian experts (Jiang & Tanner, 1999a, Mendes & Jiang, 2012, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2021a). This leads to the term on the right hand side being small, for $\mathcal{L}, \mathcal{D}$, and $\mathcal{K}$ well-chosen.

Especially, aside from important theoretical issues regarding the tightness of the bounds, the way to integrate a priori information and the minimax analysis of our proposed PMLE, we hope that our finite-sample oracle inequalities and corresponding interesting numerical experiments help to partially answer the two following important questions raised in the area of MoE regression models: (1) What number of mixture components K should be chosen, given the sample size n, and (2) Whether it is better to use a few complex experts or combine many simple experts, given the total number of parameters. Note that, such problems are considered in the work of Mendes & Jiang (2012, Proposition 1), where the authors provided some qualitative insights and only suggested a practical method for choosing K and d involving a complexity penalty or cross-validation. Furthermore, their model is only for a non-regularized maximum-likelihood estimation, and thus is not suitable in the high-dimensional setting.

Furthermore, in the context of MoE regression models, our non-asymptotic oracle inequality, Theorem 4.3.2, can be considered as a complementary result to a classical asymptotic theory (Khalili, 2010, Theorems 1,2, and 3), to a finite-sample oracle inequality on the whole collection of models in low-dimensional setting (Montuelle et al., 2014, Nguyen et al., 2021c). Furthermore, Theorem 4.3.2 also complements alternative structures of MoE using Gaussian gating functions instead of softmax functions, *e.g.,* GLoME and BLoMPE models in Nguyen et al. (2021c,b), a practical point of view of regularized MLE and feature selection (Chamroukhi & Huynh, 2018, Chamroukhi & Huynh, 2019, Huynh & Chamroukhi, 2019), and to an l_1-oracle inequality focusing on the Lasso estimation properties rather than the model selection procedure (Nguyen et al., 2020c), see also in Section 4.2 for more details.

4.3.3 Proof of the oracle inequality

Sketch of the proof To work with conditional density estimation in the PSGaBloME regression models, in Section 4.3.3.1, we need to reuse again the general theorem for model selection, Theorem 3.3.4. It is worth mentioning that because of working on random subcollection, we have to use a model selection theorem for MLE among a random subcollection (*cf.,* Devijver, 2015b, Theorem 5.1 or Devijver & Gallopin, 2018, Theorem 7.3). This is the extension of Cohen & Le Pennec (2011, Theorem 2), which dealt with conditional density estimation but not with random subcollection, and of Massart (2007, Theorem 7.11), working only for density estimation. Then, we explain how we use Theorem 3.3.4 to get the oracle inequality, Theorem 4.3.2, in Section 4.3.3.1. To this end, our model collection has to satisfy some regularity assumptions, which are proved in Section 4.3.4. The main difficulties in proving our oracle inequality lies in bounding the bracketing entropy of the weights and means restricted on relevant variables as well as rank sparse models, and in particular with block-diagonal covariance matrices for PSGaBloME model. To overcome the former issue, we adapt the strategies from Montuelle et al. (2014), Devijver (2017a). For the second one, we make use of the recent novel result on block-diagonal covariance matrices in Devijver & Gallopin (2018) for Gaussian mixture models from Genovese & Wasserman (2000), Maugis & Michel (2011b).

In the next section, we show how Theorem 3.3.4 can be utilized to prove Theorem 4.3.2. In particular, the penalty can be chosen roughly proportional to the intrinsic dimension of the model, and thus of the order of the variance.

4.3.3.1 Proof of Theorem 4.3.2

It should be stressed that all we need is to verify that Assumption 3.2.1 (K), Assumption 3.2.2 (Sep) and Assumption 3.2.3 (H) hold for every $\mathbf{m} \in \mathcal{M}$. According to the result from Devijver (2015b, Section 5.3), Assumption 3.2.2 (Sep) holds when we consider Gaussian densities and the assumption defined by (3.3.7) is true if we assume further that the true conditional density s_0 is bounded and compactly supported. Furthermore, since we restricted to finite collection of models, it is true that there exists a family $(\xi_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ and $\Xi > 0$ such that Assumption 3.2.1 (K) is satisfied. Therefore, the only remaining step of the proof for Assumption 3.2.3 (H) is presented in Section 4.3.3.1. All technical results are deferred to Section 4.3.4.

Note that the definition of model complexity in Assumption 3.2.3 (H) is related to a classical entropy dimension of a compact set w.r.t. a Hellinger type divergence $\mathrm{d}^{\otimes \mathrm{n}}$, thanks to the following Proposition 4.3.4, which is established in (Cohen & Le Pennec, 2011, Proposition 2).

Proposition 4.3.4 (Proposition 2 from Cohen & Le Pennec (2011))**.** *If we have*

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ for any } \delta \in (0, \sqrt{2}], \text{ then the function}$$

$$\phi_{\mathbf{m}}(\delta) = \delta\sqrt{\dim(S_{\mathbf{m}})}\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min(\delta, 1)}\right)}\right)$$

satisfies Assumption 3.2.3 (H). Furthermore, the unique solution $\delta_{\mathbf{m}}$ of $\frac{1}{\delta}\phi_{\mathbf{m}}(\delta) = \sqrt{n}\delta$ satisfies

$$n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 \dim(S_{\mathbf{m}})}\right)_+\right).$$

Then, we calim that Proposition 4.3.4 implies Assumption 3.2.3 (H) because of the fact that

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \dim(S_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \tag{4.3.9}$$

where $C_{\mathbf{m}}$ is a constant depending on the model. Next, recall that the definition from (4.3.8) is defined

as follows:

$$S_{\mathbf{m}} = \left\{ (\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}}(\mathbf{y}|\mathbf{x}) =: s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \in \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \right\},$$

$$\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} = \left((\omega_{k\boldsymbol{\alpha}})_{k\in[K],\boldsymbol{\alpha}\in\mathcal{A}}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]} \right)_{k\in[K]}, (\boldsymbol{\Sigma}_k(\mathbf{B}_k))_{k\in[K]} \right)$$

$$\in \left(\mathbb{R}^L\right)^{K-1} \times \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})},$$

$$\boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} = \left\{ \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]}^{|\mathbf{J}|} \right)_{k\in[K]} \in \left(\mathbb{R}^{q\times 1} \times \left(\mathbb{R}^{q\times p}\right)^D \right)^K : \right.$$

$$\left. \forall k \in [K], \forall d \in [D], \operatorname{rank}\left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right) = R_{kd} \right\}. \tag{4.3.10}$$

We also require some additional definitions of the following sets:

$$\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto g_{\mathbf{w},k}(\mathbf{x}) = \frac{\exp(w_k(\mathbf{x}))}{\sum_{l=1}^K \exp(w_l(\mathbf{x}))}, \mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_{K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}}} \right\}, \tag{4.3.11}$$

$$\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})} = \{0\} \otimes \mathbf{W}_{\mathbf{J}_{\boldsymbol{\omega}}}^{K-1}, \mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]} = \left\{ \boldsymbol{\alpha} = (\alpha_t)_{t\in[p]} \in \mathcal{A} : \boldsymbol{\alpha}_j > 0, j \in [\mathbf{J}_{\boldsymbol{\omega}}] \right\}, \tag{4.3.12}$$

$$\mathbf{W}_{\mathbf{J}_{\boldsymbol{\omega}}} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto w(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \max_{\boldsymbol{\alpha}\in\mathcal{A}} |\boldsymbol{\omega}_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}} \right\},$$

$$\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} = \left\{ \mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto (\phi_q(\mathbf{x}; \boldsymbol{\nu}_k(\mathbf{y}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)))_{k\in[K]} : \boldsymbol{\nu} \in \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}, \boldsymbol{\Sigma}(\mathbf{B}) \in \mathbf{V}_K(\mathbf{B}) \right\}.$$

Moreover, given any $\mathbf{g}^+, \mathbf{g}^- \in \mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}$ and $\phi^+, \phi^- \in \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}$, let us define

$$d^2_{\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}}(\mathbf{g}^+, \mathbf{g}^-) = \mathbb{E}_{\mathbf{X}_{[n]}} \left[\frac{1}{n} \sum_{i=1}^n d_k^2 \left(\mathbf{g}^+(\mathbf{X}_i), \mathbf{g}^-(\mathbf{X}_i) \right) \right],$$

$$d^2_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}(\phi^+, \phi^-) = \mathbb{E}_{\mathbf{X}_{[n]}} \left[\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K d_{\mathbf{y}}^2 \left(\phi_k^+(\cdot\mathbf{X}_i), \phi_k^-(\cdot\mathbf{X}_i) \right) \right].$$

Then (4.3.9) can be established by first decomposing the entropy term between the softmax gating functions and the Gaussian experts via Lemma 4.3.5, which is immediately obtained from Montuelle et al. (2014, Lemma 6), an extension of the results in Genovese & Wasserman (2000, Theorem 2), Ghosal & van der Vaart (2001), Cohen & Le Pennec (2011, Lemma 7) and Cohen & Le Pennec (2013).

Lemma 4.3.5. *For all $\delta \in (0, \sqrt{2}]$, it holds that*

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq \mathcal{H}_{[\cdot],d_{\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})} \right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} \right).$$

We next define the metric entropy of the set $\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})})$, which measures the logarithm of the minimal number of balls of radius at most δ, according to a distance $d_{\|\sup\|_\infty}$, needed to cover $\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}$ where

$$d_{\|\sup\|_\infty}\left((\mathbf{s}_k)_{k\in[K]}, (\mathbf{t}_k)_{k\in[K]} \right) = \max_{k\in[K]} \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{s}_k(\mathbf{x}) - \mathbf{t}_k(\mathbf{x})\|_2, \tag{4.3.13}$$

for any K-tuples of functions $(\mathbf{s}_k)_{k\in[K]}$ and $(\mathbf{t}_k)_{k\in[K]}$. Here, $\mathbf{s}_k, \mathbf{t}_k : \mathcal{X} \ni \mathbf{x} \mapsto \mathbf{s}_k(\mathbf{x}), \mathbf{t}_k(\mathbf{x}) \in \mathbb{R}^p, \forall k \in [K]$, and given $\mathbf{x} \in \mathcal{X}, k \in [K]$, $\|\mathbf{s}_k(\mathbf{x}) - \mathbf{t}_k(\mathbf{x})\|_2$ is the Euclidean distance in $\mathbb{R}^p$.

Based on this metric, one can first relate the bracketing entropy of $\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}$ to $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})})$, and then obtain the upper bound for its entropy via Lemma 4.3.6, which is proved in Section 4.3.4.1. Note that Lemma 4.3.6 is a variation around the Lemma 4 from Montuelle et al. (2014) and is adapted for random subcollection on some relevant variables.

Lemma 4.3.6. *For all* $\delta \in (0, \sqrt{2}]$,

$$H_{[\cdot],d_{\mathcal{P}_{(K,L,\mathbf{J}_\omega)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,L,\mathbf{J}_\omega)}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}}, \mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right)$$
$$\leq \dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right)\left(C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right), \tag{4.3.14}$$

where $\dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right) = (K-1)L$, *and* $C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}} = \ln\left(\sqrt{2} + \frac{T_{\mathbf{W}}L}{3\sqrt{3}}\right)$.

Lemma 4.3.7 allows us to construct the Gaussian brackets to handle with the entropy metric for Gaussian experts, which is established in Section 4.3.4.2.

Lemma 4.3.7. *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{4.3.15}$$

Finally, (4.3.9) is proved via Lemmas 4.3.5 to 4.3.7. Indeed, with the fact that $\dim(S_{\mathbf{m}}) = \dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right) + \dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)$, it follows that

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, S_{\mathbf{m}}) \leq H_{[\cdot],d_{\mathcal{P}_{(K,L,\mathbf{J}_\omega)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,L,\mathbf{J}_\omega)}\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)$$
$$\leq \dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right)\left(C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right) + \dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right)$$
$$=: \dim(S_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ where}$$
$$C_{\mathbf{m}} = \frac{\dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}\right)}{\dim(S_{\mathbf{m}})}\left(C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}}\right)\right) + \frac{\dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}{\dim\left(S_{\mathbf{m}}\right)}$$
$$\leq C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}} + \ln\left(\frac{8\sqrt{K_{\max}-1}}{3\sqrt{3}}\right) + C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} := \mathfrak{C}.$$

It is interesting that the constant $\mathfrak{C}$ does not depend on the dimension of the model $\mathbf{m}$ thanks to the hypothesis that $C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)}}$ is common for every model $\mathbf{m}$ in the collection. Therefore, Proposition 4.3.4 implies that, given $C = 2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2$, the model complexity $\mathcal{D}_{\mathbf{m}}$ satisfies

$$\mathcal{D}_{\mathbf{m}} \equiv n\delta_{\mathbf{m}}^2 \leq \dim(S_{\mathbf{m}})\left(2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 \dim\left(S_{\mathbf{m}}\right)}\right)_+\right) \leq \dim(S_{\mathbf{m}})\left(C + \ln n\right).$$

To this end, Theorem 3.3.4 implies that to a collection of BLoME models $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ with the penalty functions satisfies $\mathrm{pen}(\mathbf{m}) \geq \kappa\left[\dim(S_{\mathbf{m}})\left(C + \ln n\right) + (1 \vee \tau)\xi_{\mathbf{m}}\right]$ with $\kappa > \kappa_0$ the oracle inequality of Theorem 4.3.2 holds.

4.3.4 Appendix: Lemma proofs

4.3.4.1 Proof of Lemma 4.3.6

We follow a technical proof of (Montuelle et al., 2014, Appendix B.2.2) with an adaption to random subcollection of relevant variables. Recall that we defined the following sets:

$$\mathcal{P}_{(K,L,\mathbf{J}_\omega)} = \left\{\mathcal{X} \ni \mathbf{x} \mapsto g_{\mathbf{w},k}(\mathbf{x}) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^K \exp\left(w_l(\mathbf{x})\right)}, \mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_{K,d_{\mathbf{W}},\mathbf{J}_\omega}\right\},$$
$$\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_\omega)} = \{0\} \otimes \mathbf{W}_{\mathbf{J}_\omega}^{K-1}, \mathcal{A}^{[\mathbf{J}_\omega]} = \left\{\boldsymbol{\alpha} = (\alpha_t)_{t\in[p]} \in \mathcal{A} : \boldsymbol{\alpha}_j > 0, j \in [\mathbf{J}_\omega]\right\},$$
$$\mathbf{W}_{\mathbf{J}_\omega} = \left\{\mathcal{X} \ni \mathbf{x} \mapsto w(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \boldsymbol{\omega}_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} \in \mathcal{A}^{[\mathbf{J}_\omega]}, \max_{\boldsymbol{\alpha}\in\mathcal{A}}|\boldsymbol{\omega}_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}}\right\}.$$

Following the same argument from the proof of (Montuelle et al., 2014, Lemma 4), it holds that

$$H_{[\cdot],d_{\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}}}\left(\frac{\delta}{2},\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}\right)\leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}},\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}\right).$$

Next, we need to find an upper bound of $H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}},\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}\right)$. Note that for all $\mathbf{w},\mathbf{v}\in\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}$, we obtain the following important inequality

$$\begin{aligned}d_{\|\sup\|_\infty}(\mathbf{w},\mathbf{v})&=\max_{k\in[K-1]}\sup_{\mathbf{x}\in\mathcal{X}}\left|\sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}}\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{w}}\mathbf{x}^{\boldsymbol{\alpha}}-\sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}}\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{v}}\mathbf{x}^{\boldsymbol{\alpha}}\right|\\&\leq\max_{k\in[K-1]}\sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}}\left|\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{w}}-\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{v}}\right|\sup_{\mathbf{x}\in\mathcal{X}}\mathbf{x}^{\boldsymbol{\alpha}}\leq\operatorname{card}\left(\mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}\right)\max_{k\in[K-1],\boldsymbol{\alpha}\in\mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}}\left|\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{w}}-\boldsymbol{\omega}_{k,\boldsymbol{\alpha}}^{\mathbf{v}}\right|.\end{aligned}$$

Therefore, given the fact that $\operatorname{card}\left(\mathcal{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}\right)=\binom{d_{\mathbf{W}}+\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})}{\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})}=L$, for all $\delta\in(0,\sqrt{2}]$, it holds that

$$\begin{aligned}H_{[\cdot],d_{\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}}}\left(\frac{\delta}{2},\mathcal{P}_{(K,L,\mathbf{J}_{\boldsymbol{\omega}})}\right)&\leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}},\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}\right)\\&\leq H_{\|\cdot\|_\infty}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}L},\left\{\boldsymbol{\omega}\in\mathbb{R}^{(K-1)L}:\|\boldsymbol{\omega}\|_\infty\leq T_{\mathbf{W}}\right\}\right)\\&\leq(K-1)L\ln\left(1+\frac{8\sqrt{K-1}T_{\mathbf{W}}L}{3\sqrt{3}\delta}\right)\\&=(K-1)L\left[\ln\left(\sqrt{2}+\frac{T_{\mathbf{W}}L}{3\sqrt{3}}\right)+\ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right]\\&=\dim\left(\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}\right)\left(C_{\mathbf{W}_{(K,d_{\mathbf{W}},\mathbf{J}_{\boldsymbol{\omega}})}}+\ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right).\end{aligned}$$

4.3.4.2 Proof of Lemma 4.3.7

It is worth mentioning that without restriction for relevant variables, rank sparse models on the means and structures on covariance matrices of Gaussian experts from the collection $\mathcal{M}$, the upper bound of bracketing entropy of Gaussian experts from Lemma 4.3.7 is implied immediately from the Proposition 2 of Montuelle et al. (2014) and arguments in Montuelle et al. (2014, Appendix B.2.3). However, in order to overcome the much more challenging problems with random subcollection based on relevant variables, rank sparse models on the means and block-diagonal covariance matrices, we have to reply on a much more constructive bracketing entropy in the spirits of works developed in Maugis & Michel (2011b), Montuelle et al. (2014), Devijver (2015b, 2017a), Devijver & Gallopin (2018).

Given any $k\in[K]$, we first define the following set and its corresponding distance:

$$\begin{aligned}\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}&=\left\{\mathcal{X}\times\mathcal{Y}\ni(\mathbf{x},\mathbf{y})\mapsto\phi_q\left(\mathbf{y};\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}),\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right):\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\in\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)},\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\in\mathbf{V}_k(\mathbf{B}_k)\right\},\\d^2_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\left(\phi_k^+,\phi_k^-\right)&=\mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n\mathrm{d}^2\left(\phi_k^+\left(\mathbf{X}_i,\cdot\right),\phi_k^-\left(\mathbf{X}_i,\cdot\right)\right)\right].\end{aligned}\tag{4.3.16}$$

Then, it follows that $\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}=\prod_{k=1}^K\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$, where $\prod$ stands for the cartesian product, and Lemma 4.3.8, established in Section 4.3.4.2.

Lemma 4.3.8. *Given $\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}=\prod_{k=1}^K\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$, it holds that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)\leq\sum_{k=1}^K\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2\sqrt{K}},\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right).$$

Next, we claim that Lemma 4.3.7 is implied immediately via Lemma 4.3.8 and the following important Lemma 4.3.9, which is proved in Section 4.3.4.2.

Lemma 4.3.9. *For all $\delta \in (0, \sqrt{2}]$ and $k \in [K]$, there exists a constant $C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$ such that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2}, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) \leq \dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)\left(C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{4.3.17}$$

To this end, by combining the previous two Lemmas 4.3.8 and 4.3.9, we have

$$\begin{aligned}\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) &\leq \sum_{k=1}^{K} \dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)\left(C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}} + \ln\left(\sqrt{K}\right) + \ln\left(\frac{1}{\delta}\right)\right) \\ &= \dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

Here, $\dim\left(\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) = \sum_{k=1}^{K} \dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)$, $\dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) = \dim\left(\mathbf{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right) + D_{\mathbf{B}_k}$, $C_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}} = \sum_{k=1}^{K} C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}} + \ln\left(\sqrt{K}\right)$, $D_{\mathbf{B}_k} = \dim\left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right) = \sum_{g=1}^{G_k} \frac{\operatorname{card}\left(d_k^{[g]}\right)\left(\operatorname{card}\left(d_k^{[g]}\right)+1\right)}{2}$.

Proof of Lemma 4.3.8

It is sufficient to verify that

$$\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}\right) \leq \prod_{k=1}^{K} \mathcal{N}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2\sqrt{K}}, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right).$$

By (3.2.23), for each $k \in [K]$, let $\left\{\left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1 \leq l \leq \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}$ be a minimal covering of δ_k-bracket for $d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$ of $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ with cardinality $\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\delta_k, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) =: \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$. By definition, we have

$$\forall l \in \left[\mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\right], d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\left(\phi_k^{l,-}, \phi_k^{l,+}\right) \leq \delta_k.$$

This leads to the set $\left\{\prod_{k=1}^{K}\left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1 \leq l \leq \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}$ is a covering of $\delta/2$-bracket for $d_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}$ of $\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}$ with cardinality $\prod_{k=1}^{K} \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$. Indeed, let any $\phi = (\phi_k)_{k\in[K]} \in \mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}$. Consequently, for each $k \in [K], \phi_k \in \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$, and there exists $l(k) \in \left[\mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\right]$, such that

$$\phi_k^{l(k),-} \leq \phi_k \leq \phi_k^{l(k),+}, d^2_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\left(\phi_k^{l(k),+}, \phi_k^{l(k),-}\right) \leq \left(\delta_k\right)^2.$$

Then, it follows that $\phi \in [\phi^-, \phi^+] \in \left\{\prod_{k=1}^{K}\left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1 \leq l \leq \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}$, with $\phi^- = \left(\phi_k^{l(k),-}\right)_{k\in[K]}, \phi^+ = \left(\phi_k^{l(k),+}\right)_{k\in[K]}$, which leads to $\left\{\prod_{k=1}^{K}\left[\phi_k^{l,-}, \phi_k^{l,+}\right]\right\}_{1 \leq l \leq \mathcal{N}_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}$ is a bracket covering of $\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}$.

Now, we want to verify that the size of this bracket is $\delta/2$ via choosing $\delta_k = \frac{\delta}{2\sqrt{K}}, \forall k \in [K]$. It holds that

$$\begin{aligned}d^2_{\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}}\left(\phi^-, \phi^+\right) &= \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathrm{d}^2\left(\phi_k^{l(k),-}\left(\mathbf{X}_i, \cdot\right), \phi_k^{l(k),+}\left(\mathbf{X}_i, \cdot\right)\right)\right] \\ &= \sum_{k=1}^{K} \mathbb{E}_{\mathbf{X}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{d}^2\left(\phi_k^{l(k),-}\left(\mathbf{X}_i, \cdot\right), \phi_k^{l(k),+}\left(\mathbf{X}_i, \cdot\right)\right)\right] \\ &= \sum_{k=1}^{K} d^2_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}\left(\phi_k^{l(k),-}, \phi_k^{l(k),+}\right) \leq K\left(\frac{\delta}{2\sqrt{K}}\right)^2 = \left(\frac{\delta}{2}\right)^2.\end{aligned}$$

Finally, Lemma 4.3.8 is followed by the definition of a minimal $\delta/2$-bracket covering number for $\mathcal{G}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})}$.

Proof of Lemma 4.3.9

We need to bound the bracketing entropy in (4.3.17). In particular, we make use of the ideas from Maugis & Michel (2011b), which is the extension to multidimensional Gaussian mixture of Genovese & Wasserman (2000), to define a net over the parameter space of Gaussian experts. Next, we aim to construct a bracket covering of $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ according to the tensorized Hellinger distance, $d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$ based on Gaussian dilatations. To this end, our technique is adapted from the results of Devijver (2015b, 2017a) to handle the means of Gaussian experts restricted on relevant variables and rank sparse models and of Devijver & Gallopin (2018) to deal with block-diagonal covariance matrices $\mathbf{V}_k\left(\mathbf{B}_k\right), k \in [K]$.

Step 1: Construction of a net for the block-diagonal covariance matrices. Firstly, for a given matrix $\mathbf{\Sigma}_k(\mathbf{B}_k) \in \mathbf{V}_k\left(\mathbf{B}_k\right), k \in [K]$, we denote by $\text{Adj}\left(\mathbf{\Sigma}_k(\mathbf{B}_k)\right)$ the adjacency matrix associated to the covariance matrix $\mathbf{\Sigma}_k(\mathbf{B}_k)$. Note that this matrix of size D^2 can be defined by a vector of concatenated upper triangular vectors. We are going to make use of the result from Devijver & Gallopin (2018) to handle the block-diagonal covariance matrices $\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)$, via its corresponding adjacency matrix. To do this, we need to construct a discrete space for $\{0,1\}^{q(q-1)/2}$, which is a one-to-one correspondence (bijection) with

$$\mathcal{A}_{\mathbf{B}_k} = \left\{\mathbf{A}_{\mathbf{B}_k} \in \mathcal{S}_q\left(\{0,1\}\right) : \exists \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k\left(\mathbf{B}_k\right) \text{ s.t } \text{Adj}\left(\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right) = \mathbf{A}_{\mathbf{B}_k}\right\},$$

where $\mathcal{S}_q\left(\{0,1\}\right)$ is the set of symmetric matrices of size D taking values on $\{0,1\}$.

Then, we want to deduce a discretization of the set of covariance matrices. Let h denotes Hamming distance on $\{0,1\}^{q(q-1)/2}$ defined by

$$d(z,z') = \sum_{i=1}^{n} \mathbb{I}\left\{z \neq z'\right\}, \text{ for all } z, z' \in \{0,1\}^{q(q-1)/2}.$$

Let $\{0,1\}^{q(q-1)/2}_{\mathbf{B}_k}$ be the subset of $\{0,1\}^{q(q-1)/2}$ of vectors for which the corresponding graph has structure $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$. Then, given any $\epsilon > 0$, Corollary 1 and Proposition 2 from Supplementary Material A of Devijver & Gallopin (2018) lead to that there exists some subset $\mathcal{R}$ of $\{0,1\}^{q(q-1)/2}$, as well as its equivalent $\mathcal{A}^{\text{disc}}_{\mathbf{B}_k}$ for adjacency matrices satisfy

$$\begin{aligned}
\left\|\mathbf{\Sigma}_k\left(\mathbf{B}_k\right) - \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2 \leq \frac{D_{\mathbf{B}_k}}{2} \wedge \epsilon^2, \forall \left(\mathbf{\Sigma}_k\left(\mathbf{B}_k\right), \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right) \in \left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right)^2 s.t.\ \mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \neq \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right), \\
\text{card}\left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right) \leq \left(\left\lfloor \frac{2\lambda_M}{\epsilon} \right\rfloor \frac{D\left(D-1\right)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}}, \qquad (4.3.18) \\
D_{\mathbf{B}_k} = \dim\left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right) = \sum_{g=1}^{G_k} \frac{\text{card}\left(d_k^{[g]}\right)\left(\text{card}\left(d_k^{[g]}\right) - 1\right)}{2}, \text{ where} \qquad (4.3.19)
\end{aligned}$$

$$\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon) = \left\{\mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathcal{S}_q^{++}(\mathbb{R}) : \text{Adj}\left(\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right) \in \mathcal{A}^{\text{disc}}_{\mathbf{B}_k}, \left[\mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right]_{i,j} = \sigma_{i,j}\epsilon, \sigma_{i,j} \in \left[\frac{-\lambda_M}{\epsilon}, \frac{\lambda_M}{\epsilon}\right] \bigcap \mathbb{Z}\right\}.$$

Therefore, by choosing $\epsilon^2 \leq \frac{D_{\mathbf{B}_k}}{2}$, given $\mathbf{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k\left(\mathbf{B}_k\right)$, there exists $\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right) \in \tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)$, such that

$$\left\|\mathbf{\Sigma}_k\left(\mathbf{B}_k\right) - \widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2 \leq \epsilon^2. \qquad (4.3.20)$$

Based on $\widetilde{\mathbf{\Sigma}}_k\left(\mathbf{B}_k\right)$, we can construct the following bracket covering of $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ via defining suitable nets for the means of Gaussian experts. More precisely, given any $\delta_{\mathbf{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} > 0$, we claim

that the set

$$\left\{ [l,u] \left| \begin{array}{l} l(\mathbf{x},\mathbf{y}) = (1+2\alpha)^{-D}\,\phi\left(\mathbf{y}; \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}), (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right), \\ u(\mathbf{x},\mathbf{y}) = (1+2\alpha)^{D}\,\phi\left(\mathbf{y}; \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}), (1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right), \\ \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)} \in G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right), \widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right) \in \tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon) \end{array} \right\},$$

is an $\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}$-brackets set over $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ where the constant $\alpha > 0$ and function $\mathcal{X} \ni \mathbf{x} \mapsto \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})$ and its corresponding space $G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$ will be specified later. Indeed, we consider any function $\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto f(\mathbf{x},\mathbf{y}) = \phi\left(\mathbf{y}; \boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}), \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right)$ that belongs to $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$, where $\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)} \in \boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}$ and $\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathbf{V}_k\left(\mathbf{B}_k\right)$. According to (4.3.20), there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right) \in \tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)$ such that

$$\left\|\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) - \widetilde{\boldsymbol{\Sigma}}_k\left(\mathbf{B}_k\right)\right\|_2^2 \leq \epsilon^2.$$

Step 2: Construction of a net for the mean functions. We claim that given any $\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} > 0$, any $\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)} \in \boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}$, there exist a minimal covering of δ_k-bracket $G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$ and a function $\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)} \in G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \left\|\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) - \boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})\right\|_2^2 \leq \delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}, \tag{4.3.21}$$

$$\text{card}\left(G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)\right) \leq \left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}. \tag{4.3.22}$$

To accomplish this, we use the singular value decomposition of $\boldsymbol{\beta}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} [\sigma_{kd}]_r \left[\mathbf{u}_{kd}\right]_{\bullet,r} \left[\mathbf{v}_{kd}^{\top}\right]_{r,\bullet}$, $k \in [K], d \in [D]$, with $[\sigma_{kd}]_r, r \in [R_{kd}]$, denote the singular values of $\boldsymbol{\beta}_{kd}^{R_{kd}}$, with corresponding orthogonal unit vectors $\left(\left[\mathbf{u}_{kd}\right]_{\bullet,r}\right)_{r\in[R_{kd}]}$ and $\left(\left[\mathbf{v}_{kd}^{\top}\right]_{r,\bullet}\right)_{r\in[R_{kd}]}$. Then, we construct $\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) = \widetilde{\boldsymbol{\beta}}_{k0} + \sum_{d=1}^{D} \widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} \mathbf{x}^d$, where $\widetilde{\boldsymbol{\beta}}_{k0}$ and $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} [\widetilde{\sigma}_{kd}]_r \left[\widetilde{\mathbf{u}}_{kd}\right]_{\bullet,r} \left[\widetilde{\mathbf{v}}_{kd}^{\top}\right]_{r,\bullet}$, $k \in [K], d \in [D]$, are determined so that (4.3.21) and (4.3.22) are satisfied. Note that for each $k \in [K], d \in [D]$, it holds that

$$\begin{aligned} \left\|\widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) - \boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})\right\|_2 &= \left\|\widetilde{\boldsymbol{\beta}}_{k0} - \boldsymbol{\beta}_{k0} + \sum_{d=1}^{D}\left(\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} - \boldsymbol{\beta}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2 \\ &\leq \left\|\widetilde{\boldsymbol{\beta}}_{k0} - \boldsymbol{\beta}_{k0}\right\|_2 + \sum_{d=1}^{D}\left\|\left(\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} - \boldsymbol{\beta}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2 \\ &\leq \sqrt{q}\left\|\widetilde{\boldsymbol{\beta}}_{k0} - \boldsymbol{\beta}_{k0}\right\|_\infty + p\sqrt{q}\sum_{d=1}^{D}\left\|\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} - \boldsymbol{\beta}_{kd}^{R_{kd}}\right\|_\infty \left\|\mathbf{x}^d\right\|_\infty \\ &\leq \sqrt{q}\left\|\widetilde{\boldsymbol{\beta}}_{k0} - \boldsymbol{\beta}_{k0}\right\|_\infty + p\sqrt{q}\sum_{d=1}^{D}\left\|\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} - \boldsymbol{\beta}_{kd}^{R_{kd}}\right\|_\infty, \end{aligned}$$

where we used the fact that for all $d \in [D]$, $\mathbf{x} \in \mathcal{X}$, $\left\|\mathbf{x}^d\right\|_\infty \leq 1$ as $\mathcal{X} = [0,1]^p$. Thus, (4.3.21) is immediately followed if we now choose $\widetilde{\boldsymbol{\beta}}_{k0}$ and $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}$ such that

$$\sqrt{q}\left\|\boldsymbol{\beta}_{k0} - \widetilde{\boldsymbol{\beta}}_{k0}\right\|_\infty \leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2}, \tag{4.3.23}$$

$$\left\|\boldsymbol{\beta}_{kd}^{R_{kd}} - \widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}\right\|_\infty \leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2Dp\sqrt{q}}. \tag{4.3.24}$$

Let us now see how to construct $\widetilde{\boldsymbol{\beta}}_{k0}$ to get (4.3.23). This task can be accomplished if for all $k \in [K]$, $z \in [q]$, we set

$$B = \mathbb{Z} \cap \left[\left\lfloor -A_{\mathbf{u},\mathbf{v}} \frac{2\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor, \left\lfloor A_{\mathbf{u},\mathbf{v}} \frac{2\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor \right],$$
$$\left[\widetilde{\boldsymbol{\beta}}_{k0}\right]_z = \arg\min_{b \in B} \left| [\boldsymbol{\beta}_{k0}]_z - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2\sqrt{q}} b \right|.$$

Next, let us now see how to construct $\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}$ to get (4.3.24). The boundedness assumption in (4.3.6) implies that

$$\begin{aligned}
\left|\left|\left| \boldsymbol{\beta}_{kd}^{R_{kd}} - \widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}} \right|\right|\right|_\infty &= \max_{z\in[q],j\in[p]} \left| \sum_{r=1}^{R_{kd}} \left([\sigma_{kd}]_r [\mathbf{u}_{kd}]_{z,r} \left[\mathbf{v}_{kd}^\top\right]_{r,j} - [\widetilde{\sigma}_{kd}]_r [\widetilde{\mathbf{u}}_{kd}]_{z,r} \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right) \right| \\
&= \max_{z\in[q],j\in[p]} \left| \sum_{r=1}^{R_{kd}} \Bigg(\left([\sigma_{kd}]_r - [\widetilde{\sigma}_{kd}]_r\right) [\mathbf{u}_{kd}]_{z,r} \left[\mathbf{v}_{kd}^\top\right]_{r,j} \right. \\
&\qquad - [\widetilde{\sigma}_{kd}]_r \left([\widetilde{\mathbf{u}}_{kd}]_{z,r} - [\mathbf{u}_{kd}]_{z,r} \right) \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \\
&\qquad \left. - [\widetilde{\sigma}_{kd}]_r [\mathbf{u}_{kd}]_{z,r} \left(\left[\mathbf{v}_{kd}^\top\right]_{r,j} - \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right) \Bigg) \right| \\
&\leq \max_{r\in[R_{kd}]} |[\sigma_{kd}]_r - [\widetilde{\sigma}_{kd}]_r| \max_{z\in[q],j\in[p]} \sum_{r=1}^{R_{kd}} \left| [\mathbf{u}_{kd}]_{z,r} \left[\mathbf{v}_{kd}^\top\right]_{r,j} \right| \\
&\quad + \max_{z\in[q],r\in[R_{kd}]} \left| [\widetilde{\mathbf{u}}_{kd}]_{z,r} - [\mathbf{u}_{kd}]_{z,r} \right| \max_{j\in[p]} \sum_{r=1}^{R_{kd}} \left| [\widetilde{\sigma}_{kd}]_r \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right| \\
&\quad + \max_{r\in[R_{kd}],j\in[p]} \left| \left[\mathbf{v}_{kd}^\top\right]_{r,j} - \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right| \max_{z\in[q]} \sum_{r=1}^{R_{kd}} \left| [\widetilde{\sigma}_{kd}]_r [\mathbf{u}_{kd}]_{z,r} \right| \\
&\leq R_{kd} A_{\mathbf{u},\mathbf{v}}^2 \max_{r\in[R_{kd}]} |[\sigma_{kd}]_r - [\widetilde{\sigma}_{kd}]_r| \\
&\quad + R_{kd} A_{\mathbf{u},\mathbf{v}} A_\sigma \left(\max_{z\in[q],r\in[R_{kd}]} \left| [\widetilde{\mathbf{u}}_{kd}]_{z,r} - [\mathbf{u}_{kd}]_{z,r} \right| + \max_{r\in[R_{kd}],j\in[p]} \left| \left[\mathbf{v}_{kd}^\top\right]_{r,j} - \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right| \right).
\end{aligned}$$

Therefore, (4.3.24) is immediately implied if we now choose $[\widetilde{\sigma}_{kd}]_r$, $[\widetilde{\mathbf{u}}_{kd}]_{z,r}$ and $\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j}$ such that

$$\max_{r\in[R_{kd}]} |[\sigma_{kd}]_r - [\widetilde{\sigma}_{kd}]_r| \leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}},$$
$$\max_{z\in[q],r\in[R_{kd}]} \left| [\widetilde{\mathbf{u}}_{kd}]_{z,r} - [\mathbf{u}_{kd}]_{z,r} \right| \leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}},$$
$$\max_{r\in[R_{kd}],j\in[p]} \left| \left[\mathbf{v}_{kd}^\top\right]_{r,j} - \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} \right| \leq \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}.$$

This task can be accomplished as follows: for all $r \in [R_{kd}]$, $j \in [p]$, $z \in [q]$, set

$$
\begin{aligned}
S &= \mathbb{Z} \cap \left[0, \left\lfloor A_\sigma \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor\right], \\
[\widetilde{\sigma}_{kd}]_r &= \arg\min_{\zeta \in S} \left| [\sigma_{kd}]_r - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}\zeta \right|, \\
U &= \mathbb{Z} \cap \left[\left\lfloor -A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor, \left\lfloor A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}} \right\rfloor\right], \\
[\widetilde{\mathbf{u}}_{kd}]_{z,r} &= \arg\min_{\mu \in U} \left| [\mathbf{u}_{kd}]_{z,r} - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}\mu \right|, \\
\left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} &= \arg\min_{\upsilon \in U} \left| \left[\mathbf{v}_{kd}^\top\right]_{r,j} - \frac{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}A_\sigma Dp\sqrt{q}}\upsilon \right|.
\end{aligned}
$$

Note that, according to Strang (2019, I.8), we only need to determine the vectors $\left(\left([\widetilde{\mathbf{u}}_{kd}]_{z,r}\right)_{z\in[q-r]}\right)_{r\in[R_{kd}]}$ and $\left(\left([\widetilde{\mathbf{v}}_{kd}]_{r,j}\right)_{j\in[\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})-r]}\right)_{r\in[R_{kd}]}$ since the remaining elements of such vectors belong to the the nullspace of $\boldsymbol{\beta}_{kd}^{R_{kd}}$ and $\boldsymbol{\beta}_{kd}^{R_{kd}\top}$. The number of total free parameters in the previous two vectors are

$$
\begin{aligned}
\sum_{r=1}^{R_{kd}} (q - r) &= R_{kd}\left(\frac{2q - R_{kd} - 1}{2}\right), \\
\sum_{r=1}^{R_{kd}} (\mathrm{card}\,(\mathbf{J}_{\boldsymbol{\omega}}) - r) &= R_{kd}\left(\frac{2\,\mathrm{card}\,(\mathbf{J}_{\boldsymbol{\omega}}) - R_{kd} - 1}{2}\right).
\end{aligned}
$$

To this end, for all $k \in [K]$, $d \in [D]$, and $z \in [q]$, we let

$$
\left[\widetilde{\boldsymbol{\beta}}_{kd}^{R_{kd}}\right]_{z,j} = \begin{cases} \sum_{r=1}^{R_{kd}} [\widetilde{\sigma}_{kd}]_r [\widetilde{\mathbf{u}}_{kd}]_{z,r} \left[\widetilde{\mathbf{v}}_{kd}^\top\right]_{r,j} & \text{if } j \in \mathbf{J}_{\boldsymbol{\omega}}, \\ 0 & \text{if } j \in \mathbf{J}_{\boldsymbol{\omega}}^C. \end{cases}
$$

In particular, (4.3.22) is proved by the following entropy controlling

$$
\begin{aligned}
&\mathrm{card}\left(G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)\right) \\
&\leq \left[\frac{4A_{\mathbf{u},\mathbf{v}}\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^q \prod_{d=1}^{D}\left[\frac{6R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{R_{kd}}\left[\frac{12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{R_{kd}(q+\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})-R_{kd}-1)} \\
&= \left[\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right]^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}, \text{ where} \\
&\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right) = q + \sum_{d=1}^{D} R_{kd}\left(q + \mathrm{card}\,(\mathbf{J}_{\boldsymbol{\omega}}) - R_{kd}\right),\ C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} = \frac{\ln\left(C_{(D,\mathbf{J},\mathbf{R}_k)}\right)}{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}, \\
&\text{and } C_{(D,\mathbf{J},\mathbf{R}_k)} = [4A_{\mathbf{u},\mathbf{v}}\sqrt{q}]^q \left[12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 Dp\sqrt{q}\right]^{\sum_{d=1}^{D} R_{kd}(q+\mathrm{card}(\mathbf{J}_{\boldsymbol{\omega}})-R_{kd})} 2^{-\sum_{d=1}^{D} R_{kd}}.
\end{aligned}
$$

Step 3: Upper bound of the number of the bracketing entropy for $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$. Next, in order to evaluate the ratio of two Gaussian densities, we make use of Lemma 4.3.10.

Lemma 4.3.10 (Proposition C.1 from Maugis & Michel (2011b)). *Let $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two Gaussian densities. If $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ is a positive definite matrix then for all $\mathbf{y} \in \mathbb{R}^q$,*

$$
\frac{\phi(\mathbf{y};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}{\phi(\mathbf{y};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)} \leq \sqrt{\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}} \exp\left[\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right].
$$

Then, Lemma 4.3.11 allows us to fulfill the assumptions of Lemma 4.3.10.

Lemma 4.3.11 (Similar to Lemma B.8 from Maugis & Michel (2011b)). *Assume that* $0 < \epsilon < \lambda_m^2/9$, *and set* $\alpha = 3\sqrt{\epsilon}/\lambda_m$. *Then, for every* $k \in [K]$, $(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)$ *and* $\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)$ *are both positive definite matrices. Moreover, for all* $\mathbf{y} \in \mathbb{R}^q$,

$$\mathbf{y}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{y} \geq \epsilon\,\|\mathbf{y}\|_2^2, \quad \mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{y} \geq \epsilon\,\|\mathbf{y}\|_2^2.$$

Proof of Lemma 4.3.11. For all $\mathbf{y} \neq \mathbf{0}$, since $\sup_{\lambda\in \mathrm{vp}\left(\boldsymbol{\Sigma}_k(\mathbf{B}_k) - \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right)} |\lambda| = \left\|\boldsymbol{\Sigma}_k(\mathbf{B}_k) - \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right\|_2 \leq \epsilon$, $-\epsilon \geq -\lambda_m/3$, and $\alpha = 3\epsilon/\lambda_m$, it follow that

$$\begin{aligned}
\mathbf{y}^\top\left[(1+\alpha)\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{y} &= (1+\alpha)\,\mathbf{y}^\top\left[\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right]\mathbf{y} + \alpha\mathbf{y}^\top\boldsymbol{\Sigma}_k(\mathbf{B}_k)\,\mathbf{y} \\
&\geq -(1+\alpha)\left\|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k) - \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right\|_2\|\mathbf{y}\|_2^2 + \alpha\lambda_m\|\mathbf{y}\|_2^2 \\
&\geq (\alpha\lambda_m - (1+\alpha)\,\epsilon)\,\|\mathbf{y}\|_2^2 = (\alpha\lambda_m - \alpha\epsilon - \epsilon)\,\|\mathbf{y}\|_2^2 \\
&\geq \left(\frac{2}{3}\alpha\lambda_m - \epsilon\right)\|\mathbf{y}\|_2^2 = \epsilon\,\|\mathbf{y}\|_2^2 > 0, \text{ and} \\
\mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - (1+\alpha)^{-1}\,\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{y} &= (1+\alpha)^{-1}\,\mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(\mathbf{B}_k) - \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right]\mathbf{y} + \left(1 - (1+\alpha)^{-1}\right)\mathbf{y}^\top\boldsymbol{\Sigma}_k(\mathbf{B}_k)\,\mathbf{y} \\
&\geq \left(\frac{\alpha\lambda_m - \epsilon}{1+\alpha}\right)\|\mathbf{y}\|_2^2 = \frac{2\epsilon}{1+\alpha}\|\mathbf{y}\|_2^2 \geq \epsilon\,\|\mathbf{y}\|_2^2 > 0.
\end{aligned}$$

□

By using Lemma 4.3.10 and the same argument as in the proof of Lemma B.9 from Maugis & Michel (2011b), given $0 < \epsilon < \lambda_m/3$, where ϵ is chosen later, and $\alpha = 3\epsilon/\lambda_m$, we obtain

$$\max\left\{\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}, \frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right\} \leq (1+2\alpha)^{-\frac{q}{2}}\exp\left(\frac{\left\|\boldsymbol{\upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}) - \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x})\right\|_2^2}{2\epsilon}\right). \tag{4.3.25}$$

Because $\ln(\cdot)$ is a non-decreasing function, $\ln(1+2\alpha) \geq \alpha, \forall\alpha \in [0,1]$. Combined with (4.3.21) where $\delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} = q\alpha\epsilon$, we conclude that

$$\max\left\{\ln\left(\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right), \ln\left(\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right)\right\} \leq -\frac{q}{2}\ln(1+2\alpha) + \frac{\delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2\epsilon} \leq -\frac{q}{2}\alpha + \frac{\delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}{2\epsilon} = 0.$$

This means that $l(\mathbf{x},\mathbf{y}) \leq f(\mathbf{x},\mathbf{y}) \leq u(\mathbf{x},\mathbf{y}), \forall(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$. Hence, it remains to bound the size of bracket $[l,u]$ w.r.t. $d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}$.

To this end, we aim to verify that $d^2_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}(l,u) \leq \frac{\delta}{2}$. To accomplish this, we make use of Lemma 4.3.12.

Lemma 4.3.12 (Proposition C.3 from Maugis & Michel (2011b)). *Let* $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ *and* $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ *be two Gaussian densities with full rank covariance. It holds that*

$$\begin{aligned}
&d^2\left(\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1), \phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right) \\
&= 2\left\{1 - 2^{q/2}\left|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right|^{-1/4}\left|\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right|^{-1/2}\exp\left[-\frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]\right\}.
\end{aligned}$$

Therefore, using the fact that $\cosh(t) = \frac{e^{-t}+e^{t}}{2}$, Lemma 4.3.12 leads to, for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\mathrm{d}^2(l(\mathbf{x},\cdot), u(\mathbf{x},\cdot)) &= \int_{\mathcal{Y}} \left[l(\mathbf{x},\mathbf{y}) + u(\mathbf{x},\mathbf{y}) - 2\sqrt{l(\mathbf{x},\mathbf{y})u(\mathbf{x},\mathbf{y})} \right] d\mathbf{y} \\
&= (1+2\alpha)^{-q} + (1+2\alpha)^{q} - 2 \\
&+ \mathrm{d}^2 \left(\phi\left(\cdot; \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}), (1+\alpha)^{-1} \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right), \phi\left(\cdot; \widetilde{\boldsymbol{\upsilon}}_{(D,\mathbf{J},\mathbf{R}_k)}(\mathbf{x}), (1+\alpha) \widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right)\right) \\
&= 2\cosh\left[q\ln(1+2\alpha)\right] - 2 \\
&\quad + 2\left[1 - 2^{q/2}\left[(1+\alpha)^{-1} + (1+\alpha)\right]^{-q/2} \left|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right|^{-1/2} \left|\widetilde{\boldsymbol{\Sigma}}_k(\mathbf{B}_k)\right|^{1/2}\right] \\
&= 2\cosh\left[q\ln(1+2\alpha)\right] - 2 + 2 - 2\left[\cosh\left(\ln(1+\alpha)\right)\right]^{-q/2} \\
&= 2g\left(q\ln(1+2\alpha)\right) + 2h\left(\ln(1+\alpha)\right),
\end{aligned}$$

where $g(t) = \cosh(t) - 1 = \frac{e^{-t}+e^{t}}{2} - 1$, and $h(t) = 1 - \cosh(t)^{-q/2}$. The upper bounds of terms g and h separately imply that, for all $\mathbf{y} \in \mathcal{Y}$,

$$\mathrm{d}^2(l(\mathbf{x},\cdot), u(\mathbf{x},\cdot)) \leq 2\left(2\cosh\left(\frac{1}{\sqrt{6}}\right)\alpha^2q^2 + \frac{1}{4}\alpha^2q^2\right) \leq 6\alpha^2q^2 = \frac{\delta^2}{4},$$

where we choose $\alpha = \frac{3\epsilon}{\lambda_m}, \epsilon = \frac{\delta\lambda_m}{6\sqrt{6q}}$, $\forall \delta \in (0,1], q \in \mathbb{N}^\star, \lambda_m > 0$, which appears in (4.3.25) and satisfies $\alpha = \frac{\delta}{2\sqrt{6q}}$ and $0 < \epsilon < \frac{\lambda_m}{3}$. Indeed, studying functions g and h yields

$$\begin{aligned}
\mathbf{g}'(t) &= \sinh(t), \mathbf{g}''(t) = \cosh(t) \leq \cosh(c), \forall t \in [0,c], c \in \mathbb{R}_+, \\
h'(t) &= \frac{q}{2}\cosh(t)^{-q/2-1}\sinh(t), \\
h''(t) &= \frac{q}{2}\left(-\frac{q}{2} - 1\right)\cosh(t)^{-q/2-2}\sinh^2(t) + \frac{q}{2}\cosh(t)^{-q/2} \\
&= \frac{q}{2}\left(1 - \left(\frac{q}{2}+1\right)\left(\frac{\sinh(t)}{\cosh(t)}\right)^2\right)\cosh(t)^{-q/2} \leq \frac{q}{2},
\end{aligned}$$

where we used the fact that $\cosh(t) \geq 1$. Then, since $g(0) = 0, \mathbf{g}'(0) = 0, h(0) = 0, h'(0) = 0$, by applying Taylor's Theorem, it is true that

$$\begin{aligned}
g(t) &= g(t) - g(0) - \mathbf{g}'(0)t = R_{0,1}(t) \leq \cosh(c)\frac{t^2}{2}, \forall t \in [0,c], \\
h(t) &= h(t) - h(0) - h'(0)t = R_{0,1}(t) \leq \frac{q}{2}\frac{t^2}{2} \leq \frac{q^2}{2}\frac{t^2}{2}, \forall t \geq 0.
\end{aligned}$$

We wish to find an upper bound for $t = q\ln(1+2\alpha)$, $q \in \mathbb{N}^\star$, $\alpha = \frac{\delta}{2\sqrt{6q}}$, $\delta \in (0,1]$. Since $\ln(\cdot)$ is an increasing function, then we have

$$t = q\ln\left(1 + \frac{\delta}{\sqrt{6}q}\right) \leq q\ln\left(1 + \frac{1}{\sqrt{6}q}\right) \leq q\frac{1}{\sqrt{6}q} = \frac{1}{\sqrt{6}}, \forall \delta \in (0,1],$$

since $\ln\left(1 + \frac{1}{\sqrt{6}q}\right) \leq \frac{1}{\sqrt{6}q}$, $\forall q \in \mathbb{N}^\star$. Then, since $\ln(1+2\alpha) \leq 2\alpha, \forall \alpha \geq 0$,

$$\begin{aligned}
g\left(q\ln(1+2\alpha)\right) &\leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{\left(q\ln(1+2\alpha)\right)^2}{2} \leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{q^2}{2}4\alpha^2, \\
h\left(\ln(1+\alpha)\right) &\leq \frac{q^2}{2}\frac{\left(\ln(1+\alpha)\right)^2}{2} \leq \frac{q^2\alpha^2}{4}.
\end{aligned}$$

Next, note that the set of $\delta/2$-brackets $[l,u]$ over $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ is totally defined by the parameter spaces $\tilde{S}^{\mathrm{disc}}_{\mathbf{B}_k}(\epsilon)$ and $G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)$. This leads to an upper bound of the $\delta/2$-bracketing

entropy of $\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}$ is evaluated from an upper bound of the two set cardinalities. Hence, given any $\delta > 0$, by choosing $\epsilon = \frac{\delta\lambda_m}{6\sqrt{6q}}$, $\alpha = \frac{3\epsilon}{\lambda_m} = \frac{\delta}{2\sqrt{6q}}$, and $\delta^2_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}} = q\alpha\epsilon = q\frac{\delta}{2\sqrt{6q}}\frac{\delta\lambda_m}{6\sqrt{6q}} = \frac{\delta^2\lambda_m}{72q}$, it holds that

$$
\begin{aligned}
&\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2}, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) \\
&\leq \text{card}\left(\tilde{S}^{\text{disc}}_{\mathbf{B}_k}(\epsilon)\right) \times \text{card}\left(G_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\left(\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)\right) \\
&\leq \left(\left\lfloor \frac{2\lambda_M}{\epsilon} \right\rfloor \frac{q\,(q-1)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)} \qquad \text{(using (4.3.19) and (4.3.22))} \\
&\leq \left(\frac{2\lambda_M 6\sqrt{6}q}{\delta\lambda_m}\frac{q\,(q-1)}{2D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2q}\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\delta\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)} \\
&= \left(\frac{6\sqrt{6}\lambda_M q^2\,(q-1)}{\lambda_m D_{\mathbf{B}_k}}\right)^{D_{\mathbf{B}_k}} \left(\frac{6\sqrt{2q}\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\sqrt{\lambda_m}}\right)^{\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)} \left(\frac{1}{\delta}\right)^{D_{\mathbf{B}_k}+\dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)}.
\end{aligned}
$$

To this end, note that $\dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) = D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)$, we obtain

$$
\begin{aligned}
&\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2}, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right) \\
&= \ln\left(\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}}}\left(\frac{\delta}{2}, \mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)\right) \\
&\leq D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M q^2\,(q-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)\ln\left(\frac{6\sqrt{2q}\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\sqrt{\lambda_m}}\right) \\
&\quad + \left(D_{\mathbf{B}_k} + \dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)\right)\ln\left(\frac{1}{\delta}\right) \\
&= \dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)\left(C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}} + \ln\left(\frac{1}{\delta}\right)\right),
\end{aligned}
$$

where $C_{\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}} = \dfrac{D_{\mathbf{B}_k}\ln\left(\frac{6\sqrt{6}\lambda_M q^2(q-1)}{\lambda_m D_{\mathbf{B}_k}}\right) + \dim\left(\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}\right)\ln\left(\frac{6\sqrt{2q}\exp\left(C_{\boldsymbol{\Upsilon}_{(D,\mathbf{J},\mathbf{R}_k)}}\right)}{\sqrt{\lambda_m}}\right)}{\dim\left(\mathcal{G}_{(D,\mathbf{B}_k,\mathbf{J},\mathbf{R}_k)}\right)}$.

4.3.5 The Lasso+l_2-MLE and Lasso+l_2-Rank procedures

Inspired by the ideas from Khalili (2010), Stadler et al. (2010), Devijver (2015b, 2017a,b), we propose Lasso+l_2-MLE and Lasso+l_2-Rank procedures for PSGaBloME regression models to deal with high-dimensional heterogeneous data. Note that the Lasso+l_2-MLE procedure takes advantage of the MLE, whereas the Lasso+l_2-Rank procedure takes advantage of the low-rank structures of regression coefficients $\mathbf{Q}_k\boldsymbol{\beta}_{kd}, k \in [K], d \in [D]$, where $\mathbf{Q}_k$ is defined by the Cholesky decomposition of the positive-definite matrix $\boldsymbol{\Sigma}_k^{-1}$, namely $\boldsymbol{\Sigma}_k^{-1} = \mathbf{Q}_k^\top\mathbf{Q}_k$.

These procedures are decomposed into three main steps. First, we construct a model collection, with models more or less sparse, with more or less mixture components and with more or less terms in polynomial of weights and means. Second, we refit estimations with the MLE or estimate the parameters by MLE under rank constraint on the restricted set of relevant columns. To this end, a model is selected thanks to the slope heuristic, which is a data-driven criterion based on non-asymptotic theory. In particular, this leads to a classification or clustering according to the MAP principle on the selected model.

It is important to emphasize that since we have to deal with multivariate responses in PSGaBloME regression models, we propose new penalty functions in (4.3.29) and the corresponding generalized EM algorithm in Section 4.3.6.

Note that in order to apply the finite-sample oracle inequality for relevant columns selected by the Group-Lasso estimator, we have to require further some specific conditions, *e.g.,* good discussions in Devijver (2017a, Section 4.2), Bunea et al. (2012, Section 2), and Lounici et al. (2011, results on the Group-Lasso estimators).

For the sake of simplicity, both the weights of softmax gating networks and the means of Gaussian experts are defined as the following simple polynomial functions:

$$\begin{aligned}
&\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}_{K-1},\\
&\mathbf{W}_{K-1} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto w_k(\mathbf{x}) = \omega_{k0} + \sum_{l=1}^{L} \boldsymbol{\omega}_{kl}^{\top}\mathbf{x}^l, \forall k \in [K-1] : \max_{l\in[L]} |\boldsymbol{\omega}_{kl}| \leq T_{\mathbf{W}} \right\},\\
&\boldsymbol{\Upsilon}_{K,D} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(\boldsymbol{\beta}_{k0} + \sum_{d=1}^{D} \boldsymbol{\beta}_{kd}\mathbf{x}^d \right)_{k\in[K]} : \max\left\{ \left\| \left| \boldsymbol{\beta}_{kd} \right| \right\|_{\infty} : k \in [K], d \in (\{0\} \cup [D]) \right\} \leq T_{\boldsymbol{\Upsilon}} \right\}.
\end{aligned}$$

However, our finite-sample oracle inequality still holds for a more general case when we utilize general polynomials, defined in (4.3.4), for weights of the gating networks. Furthermore, we consider 1-block-diagonal covariance matrices. This means that we do not need to select the potential hidden graph-structured interactions between variables $\mathbf{B}$.

Note that identifiability of a model is crucial for any valid statistical inference. Thus, motivated by the first results in the identifiability of MoE models from Jiang & Tanner (1999c), Hennig (2000) as well as recent works from Khalili (2010), Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019) for penalized cases, instead of considering an unrestricted parameterization of the MoE network from (4.3.2), we consider the gating parameters as follows: for all $k \in [K-1], \boldsymbol{\omega}_k^{\top} = \left(\omega_{k0}, \left(\boldsymbol{\omega}_{kl}^{\top}\right)_{l\in[L]}\right)$, $\boldsymbol{\omega}_K^{\top} = \left(\omega_{K0}, \left(\boldsymbol{\omega}_{Kl}^{\top}\right)_{l\in[L]}\right) = \left(0, (\mathbf{0}_l)_{l\in[L]}\right)$, $\boldsymbol{\omega} = \left(\boldsymbol{\omega}_k^{\top}\right)_{k\in[K-1]}$, $g_K(\mathbf{x};\boldsymbol{\omega}) = 1 - \sum_{j=1}^{K-1} g_j(\mathbf{x};\boldsymbol{\omega})$ with

$$g_k(\mathbf{x};\boldsymbol{\omega}) = \frac{\exp(w_k(\mathbf{x}))}{1 + \sum_{j=1}^{K-1}\exp(w_j(\mathbf{x}))}, w_k(\mathbf{x}) = \omega_{k0} + \sum_{l=1}^{L} \boldsymbol{\omega}_{kl}^{\top}\mathbf{x}^l. \tag{4.3.26}$$

4.3.5.1 Model collection construction

We firstly fix $K \in \mathcal{K}$, $L \in \mathcal{L}$ and $D \in \mathcal{D}$. To detect the relevant indices and construct the set $\mathbf{J} \in \mathcal{J}$, by generalizing the idea from Khalili (2010), Stadler et al. (2010), Devijver (2015b, 2017a,b), we utilize an l_2-penalized log-likelihood functions instead of the log-likelihood and combine with two l_1-penalties on the terms of polynomials from weights and the means. It is worth mentioning that in order to deal with PSGaBloME model, we must extend the results from Khalili (2010), Stadler et al. (2010) to multivariate response $\mathbf{Y} \in \mathbb{R}^q$ and the results from Devijver (2015b, 2017a,b) to mixture of polynomial experts with any arbitrarily degree of weights and mean functions. More precisely, we consider

$$\widehat{\boldsymbol{\psi}}^{\text{Lasso}+l_2}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}_{(K,L,D,\mathbf{J},\mathbf{R})}}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\psi}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) \right\}, \tag{4.3.27}$$

$$s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\omega})\, \phi_q\left(\mathbf{y}; \boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}), \boldsymbol{\Sigma}_k\right), \boldsymbol{\psi} = \left(\omega_{k0}, (\boldsymbol{\omega}_{kl})_{l\in[L]}, \boldsymbol{\beta}_{k0}, (\boldsymbol{\beta}_{kd})_{d\in[D]}, \boldsymbol{\Sigma}_k\right)_{k\in[K]}, \tag{4.3.28}$$

$$\text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \lambda_{kl}^{[1]} \|\boldsymbol{\omega}_{kl}\|_1 + \sum_{k=1}^{K}\sum_{d=1}^{D} \lambda_{kd}^{[2]} \|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\|_1 + \frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L} \|\boldsymbol{\omega}_{kl}\|_2^2, \quad \mathbf{Q}_k^{\top}\mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1}. \tag{4.3.29}$$

Here, $\boldsymbol{\lambda} = \left(\left(\lambda_{kl}^{[1]}\right)_{k\in[K],l\in[L]}, \left(\lambda_{kd}^{[2]}\right)_{k\in[K],d\in[D]}, \frac{\lambda^{[3]}}{2} \right)$ is a vector of non-negative regularization parameters, for any $k \in [K]$, $l \in [L]$, $d \in [D]$, $\|\boldsymbol{\omega}_{kl}\|_1 = \sum_{j=1}^{p} \left|[\boldsymbol{\omega}_{kl}]_j\right|$, $\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\|_1 = \sum_{j=1}^{p}\sum_{z=1}^{q} \left|[\mathbf{Q}_k\boldsymbol{\beta}_{kd}]_{z,j}\right|$, $\|\boldsymbol{\omega}_{kl}\|_2^2 = \sum_{j=1}^{p} [\boldsymbol{\omega}_{kl}]_j^2$ is the Euclidean norm in $\mathbb{R}^p$, and the Cholesky decomposition $\boldsymbol{\Sigma}_k^{-1} = \mathbf{Q}_k^\top \mathbf{Q}_k$ defines $\mathbf{Q}_k$ for all $k \in [K]$. Remark that the first two terms from (4.3.29) are the usual l_1-estimator, called the Lasso estimator, while the l_2 penalty function for the gating network is added to avoid wildly large positive and negative estimates of the regression coefficients corresponding to the mixing proportions. This behavior can be observed in logistic/multinomial regression when the number of potential features is large and highly correlated (*e.g.,* Park & Hastie (2008), Bunea et al. (2008)). However, this also affects the sparsity of the regularization model, which is confirmed from numerical experiments from Chamroukhi & Huynh (2018), Chamroukhi & Huynh (2019).

Computing those estimators leads to construct the relevant variables set. For a fixed number of mixture components $K \in \mathcal{K}$, fixed degrees $L \in \mathcal{L}$ and $D \in \mathcal{D}$ of polynomials from mean and weight functions, denote by $\mathbf{G}_{K,L,D}$ a candidate of grid of regularization parameters. Fixing a regularization parameter $\boldsymbol{\lambda} \in \mathbf{G}_{K,L,D}$, we could then use a generalized EM algorithm which is originally introduced by Dempster et al. (1977) and is extended for PSGaBloME models with univariate response, *e.g.,* Jordan & Jacobs (1994), Khalili (2010), Chamroukhi & Huynh (2018), Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019), to compute the Lasso $+l_2$ estimator, and construct the set of relevant variables $\mathbf{J}_{(K,L,D,\boldsymbol{\lambda})}$, saying the non-zero coefficients. We denote by $\mathcal{J}$ the random collection of all these sets,

$$\mathcal{J} = \bigcup_{K\in\mathcal{K}} \bigcup_{L\in\mathcal{L}} \bigcup_{D\in\mathcal{D}} \bigcup_{\boldsymbol{\lambda}\in\mathbf{G}_{K,L,D}} \mathbf{J}_{(K,L,D,\boldsymbol{\lambda})}. \tag{4.3.30}$$

4.3.5.2 Refitting

The Lasso$+l_2$-MLE procedure

The second step consists of approximating the MLE

$$\widehat{s}^{(K,L,D,\mathbf{J})} = \arg \min_{t\in\mathcal{S}_{(K,L,D,\mathbf{J})}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(t\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) \right\}, \tag{4.3.31}$$

which can be accomplished by using an EM algorithm for each model $(K, L, D, \mathbf{J}) \in \mathcal{K} \times \mathcal{L} \times \mathcal{D} \times \mathcal{J}$. Remark that we estimate all parameters, to reduce bias induced by the Lasso $+l_2$ estimator. The reason why we need to refit the Lasso $+l_2$ estimator can be referred to Devijver (2015b, Section 2.3).

The Lasso$+l_2$-Rank procedure

We use the generalized EM algorithm to estimate the parameters by MLE under rank constraint on the restricted set of relevant columns.

4.3.5.3 Model selection

The third step is devoted to model selection. We follow the framework from Devijver (2017b, Section 3) to select the refitted model rather than selecting the regularization parameter. Instead of using an asymptotic criterion, such as BIC or AIC, we use the slope heuristic, originally introduced by Birgé & Massart (2007) and recently reviewed by Baudry et al. (2012) and Arlot (2019), which is a data-driven non-asymptotic criterion for selecting a model among a collection of models. For an oracle inequality to only justify the penalty shape when using slope heuristic used here, see Section 4.3.2 for more details.

4.3.6 Generalized EM algorithm for the Lasso $+l_2$ estimator

The EM algorithm (Dempster et al., 1977, McLachlan & Krishnan, 1997) is most commonly known as a technique to produce MLEs in settings where the data under study is incomplete or when optimization

of the likelihood would be simplified if an additional set of variables were known. The iterative EM algorithm consists of an expectation (E) step followed by a maximization (M) step. Generally, during the E step the conditional expectation of the complete (i.e. observed and unobserved) data log-likelihood is computed, given the data and current parameter values. In the M step the expected log-likelihood is maximized with respect to the model parameters. The imputation of latent variables often makes maximization of the expected log-likelihood more feasible.

The log-likelihood function of the PSGaBloME model is

$$\mathrm{L}\left(\boldsymbol{\psi}\right) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} g_k\left(\mathbf{x}_i; \boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i; \boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i), \boldsymbol{\Sigma}_k\right) \right]. \tag{4.3.32}$$

It is difficult to directly obtain MLEs from this likelihood. In the EM framework, to alleviate this, the data are augmented by imputing for each incomplete observed-data vector $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$, the K-dimensional binary random variable $\mathbf{z}_i = (z_{ik})_{k\in[K]}$ (which is also called the latent (unobserved) random variable or the allocation variable in the mixture model context). This latent variable has a 1-of-K representation in which a particular element z_{ik} is equal to 1 and all other elements are equal to 0. More precisely, for any $i \in [n]$, $k \in [K]$, z_{ik} is an indicator binary-valued variable such that $z_{ik} = 1$ if the ith pair $(\mathbf{x}_i, \mathbf{y}_i)$ is generated from the kth expert component and $z_{ik} = 0$ otherwise. Here, for any $i \in [n]$, given the predictor $\mathbf{x}_i$, $\mathbf{z}_i$ are unobserved i.i.d. random variables following a multinomial distribution:

$$\mathbf{z}_i|\mathbf{x}_i \sim \mathrm{Mult}\left(1, \left(g_k\left(\mathbf{x}_i; \boldsymbol{\omega}\right)\right)_{k\in[K]}\right). \tag{4.3.33}$$

The EM algorithm for solving (4.3.31) firstly requires the construction of the penalized complete-data log-likelihood

$$\mathrm{PL_c}\left(\boldsymbol{\psi}, \mathbf{z}\right) = \mathrm{L_c}\left(\boldsymbol{\psi}, \mathbf{z}\right) - \mathrm{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}), \tag{4.3.34}$$

$$\mathrm{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \lambda_{kl}^{[1]} \left\|\boldsymbol{\omega}_{kl}\right\|_1 + \sum_{k=1}^{K}\sum_{d=1}^{D} \lambda_{kd}^{[2]} \left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1 + \frac{\lambda^{[3]}}{2} \sum_{k=1}^{K}\sum_{l=1}^{L} \left\|\boldsymbol{\omega}_{kl}\right\|_2^2, \quad \mathbf{Q}_k^\top\mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1},$$

via the standard complete-data log-likelihood

$$\mathrm{L_c}\left(\boldsymbol{\psi}, \mathbf{z}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln\left[g_k\left(\mathbf{x}_i; \boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i; \boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i), \boldsymbol{\Sigma}_k\right)\right]. \tag{4.3.35}$$

The generalized EM, or GEM, algorithm addresses the problem of an intractable M-step. Instead of aiming to maximize the conditional expectation of $\mathrm{PL_c}\left(\boldsymbol{\psi}\right)$ with respect to $\boldsymbol{\psi}$, it seeks instead to change the parameters in such a way as to increase its value. Then, the GEM algorithm for the PSGaBloME model in its general form runs as follows. After starting with an initial solution $\boldsymbol{\psi}^{(0)}$, it alternates between the following steps until convergence (*e.g.,* when there is no longer significant change in the relative variation of the regularized log-likelihood).

4.3.6.1 E-step

The E-step computes the conditional expectation of the penalized complete-data log-likelihood (4.3.34), given the observed data $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$ under the current parameter vector $\boldsymbol{\psi}^{(t)}$, t being the current iteration number of the EM algorithm:

$$\begin{aligned}
Q_{\mathrm{pen}}\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}\right) &= Q\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}\right) - \mathrm{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}), \text{ where} \\
Q\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}\right) &= \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y},\boldsymbol{\psi}^{(t)}}\left[\mathrm{L_c}\left(\boldsymbol{\psi}, \mathbf{z}\right) | \left(\mathbf{x}_i, \mathbf{y}_i\right)_{i\in[n]}, \boldsymbol{\psi}^{(t)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y},\boldsymbol{\psi}^{(t)}}\left[z_{ik}|\mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\psi}^{(t)}\right] \ln\left[g_k\left(\mathbf{x}_i; \boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i; \boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i), \boldsymbol{\Sigma}_k\right)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \ln\left[g_k\left(\mathbf{x}_i; \boldsymbol{\omega}\right) \phi_q\left(\mathbf{y}_i; \boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i), \boldsymbol{\Sigma}_k\right)\right].
\end{aligned} \tag{4.3.36}$$

Here,

$$\begin{aligned}\tau_{ik}^{(t)} &= \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y},\boldsymbol{\psi}^{(t)}}\left[z_{ik}|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\psi}^{(t)}\right] = \mathbb{P}\left(z_{ik}=1|\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\psi}^{(t)}\right) = \frac{\mathbb{P}\left(z_{ik}=1,\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\psi}^{(t)}\right)}{\mathbb{P}\left(\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\psi}^{(t)}\right)} \\ &= \frac{\mathbb{P}\left(z_{ik}=1|\mathbf{x}_i,\boldsymbol{\psi}^{(t)}\right)\mathbb{P}\left(\mathbf{y}_i|z_{ik}=1,\mathbf{x}_i,\boldsymbol{\psi}^{(t)}\right)}{\sum_{l=1}^{K}\mathbb{P}\left(z_{il}=1,\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\psi}^{(t)}\right)}\underbrace{\mathbb{P}\left(\mathbf{x}_i\right)}_{=1}\text{(since } \mathbf{x}_i \text{ are deterministic predictors)} \\ &= \frac{g_k\left(\mathbf{x}_i;\boldsymbol{\omega}^{(t)}\right)\phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}^{(t)}}(\mathbf{x}_i),\boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{l=1}^{K}g_l\left(\mathbf{x}_i;\boldsymbol{\omega}^{(t)}\right)\phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{l,\boldsymbol{\beta}^{(t)}}(\mathbf{x}_i),\boldsymbol{\Sigma}_l^{(t)}\right)}\end{aligned} \tag{4.3.37}$$

is the posterior probability that the data pair $(\mathbf{x}_i,\mathbf{y}_i)$ belongs to the kth expert. This step therefore only requires the computation of the conditional component probabilities $\tau_{ik}^{(t)}$ $(i\in[n])$ for each of the K experts.

4.3.6.2 Generalized M-step

The generalized M-step aims to update the parameters via improving the value of $Q_{\text{pen}}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right)$ w.r.t. $\boldsymbol{\psi}$, which can be written as

$$\begin{aligned}Q_{\text{pen}}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right) &= Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)+Q_{\text{pen}}\left(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{\psi}^{(t)}\right), \text{ where} \\ Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\ln\left[g_k\left(\mathbf{x}_i;\boldsymbol{\omega}\right)\right]-\sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{kl}^{[1]}\left\|\boldsymbol{\omega}_{kl}\right\|_1-\frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\boldsymbol{\omega}_{kl}\right\|_2^2, \\ Q_{\text{pen}}\left(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{\psi}^{(t)}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\ln\left[\phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i),\boldsymbol{\Sigma}_k\right)\right]-\sum_{k=1}^{K}\sum_{d=1}^{D}\lambda_{kd}^{[2]}\left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1, \mathbf{Q}_k^{\top}\mathbf{Q}_k=\boldsymbol{\Sigma}_k^{-1}.\end{aligned} \tag{4.3.38}$$

Note that in order to maximize $Q_{\text{pen}}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right)$ with respect to model parameters $\boldsymbol{\psi}$, in (4.3.38), we utilize the standard fact that $Q_{\text{pen}}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right)$ can be decomposed in terms of independent expressions for gate and expert models. In this way, the M-step can be performed independently for gate and expert parameters (Moerland, 1997, Peralta & Soto, 2014). In our problem, each of these optimizations has an additional term given by the respective regularization term which is similar to a regularized logistic regression in Lee et al. (2006).

The parameter $\boldsymbol{\omega}$ are therefore separated updated by maximizing the function

$$\begin{aligned}Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}w_k\left(\mathbf{x}_i\right)-\sum_{i=1}^{n}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k\left(\mathbf{x}_i\right)\right)\right]-\sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{kl}^{[1]}\left\|\boldsymbol{\omega}_{kl}\right\|_1 \\ &\quad-\frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\boldsymbol{\omega}_{kl}\right\|_2^2, \quad w_k(\mathbf{x}_i)=\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l, \forall k\in[K-1].\end{aligned} \tag{4.3.39}$$

Motivated by the recent novel works from Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019) for SGaME model with linear mean Gaussian experts and scalar responses, we propose and compare three approaches for maximizing (4.3.39) based on a majorization–minimization (MM) algorithm, a coordinate ascent algorithm and proximal Newton-type method. These approaches have some advantages since they do not use any approximate for the penalty function, and have a separate structure which avoid matrix inversion. Note that we extend the work from Chamroukhi & Huynh (2019), Huynh & Chamroukhi (2019) to devise a novel MM algorithm for the PSGaBloME model with polynomial mean of Gaussian functions and multivariate responses.

The task of determining the maximizers of (4.3.39) may be complicated by various factors that fall outside the scope of the traditional optimization. Such factors include the lack of differentiability of the objective functions, *e.g.*, $Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$, or difficulty in obtaining closed-form solutions to the first-order condition (FOC) equation $\nabla_{\boldsymbol{\omega}}Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)=\mathbf{0}$, where $\nabla_{\boldsymbol{\omega}}$ is the gradient operator

with respect to $\boldsymbol{\omega}$. To overcome such difficulties, De Leeuw (1977) presented an MM algorithm for multidimensional scaling contemporary with the classic Dempster et al. (1977) paper on EM algorithms, then Hunter & Lange (2000) proposed the MM algorithm framework to solve the quantile regression via iterative minimization of surrogate functions. MM algorithms are particularly attractive due to the monotonicity and thus stability of their objective sequences as well as global convergence of their limits, in general settings. A comprehensive treatment of the theory and implementation of MM algorithms for various problems can be found Hunter & Lange (2004), Lange (2016), Nguyen (2017).

MM algorithm for updating the gating network

Definition 4.3.13 (Philosophy of the MM Algorithm, *e.g.,* Hunter & Lange (2004), Nguyen (2017))**.** Let θ^s denote a fixed value of the parameter θ, and let $G\left(\theta;\theta^{(r)}\right)$ represent a real-value function of θ whose form depends on $\theta^{(r)}$. The function $G\left(\theta;\theta^{(r)}\right)$ is said to minorize $F(\theta)$ at the point $\theta^{(r)}$ if and only if for all θ, it holds that

$$F\left(\theta\right) \geq G\left(\theta;\theta^{(r)}\right), \quad F\left(\theta^{(r)}\right) \geq G\left(\theta;\theta^{(r)}\right). \tag{4.3.40}$$

In other words, the surface $\theta \mapsto G\left(\theta;\theta^{(r)}\right)$ lies below the surface $F(\theta)$ and is tangent to it at the point $\theta = \theta^{(r)}$. Suppose we wish to obtain

$$\hat{\theta} = \arg\max_{\theta\in\Theta} F(\theta), \tag{4.3.41}$$

for some difficulty to manipulate objective function F, where Θ is a subset of some Euclidean space. In the maximization step of the MM algorithm, we maximize the surrogate function $G\left(\theta;\theta^{(r)}\right)$, rather than the function $F(\theta)$ itself. Let $\theta^{(0)}$ be some initial value and $\theta^{(r)}$ be the rth iterate. We say that $\theta^{(r+1)}$ is the $(r+1)$th iterate of an MM algorithm if it satisfies

$$\theta^{(r+1)} = \arg\max_{\theta\in\Theta} G\left(\theta;\theta^{(r)}\right). \tag{4.3.42}$$

By Definition 4.3.13, we can deduce the monotonicity property of all MM algorithms. Indeed, we can show that the MM algorithm forces $F(\theta)$ uphill, because (4.3.42) and (4.3.40) imply that

$$F(\theta^{(r)}) = G\left(\theta^{(r)};\theta^{(r)}\right) \leq G\left(\theta^{(r+1)};\theta^{(r)}\right) \leq F(\theta^{(r)}). \tag{4.3.43}$$

If $G\left(\theta;\theta^{(r)}\right)$ is well constructed, then we can avoid matrix inversion when maximizing it.

Next, we devise the surrogate function for $Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ via Lemma 4.3.14.

Lemma 4.3.14. *The objective function $Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ is minorized at $\boldsymbol{\omega}^{(r)}$ by*

$$\begin{aligned} G\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right) = &\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)} w_k\left(\mathbf{x}_i\right) + H\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)}\right) - \sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{kl}^{[1]}\left\|\boldsymbol{\omega}_{kl}\right\|_1 \\ &- \frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\boldsymbol{\omega}_{kl}\right\|_2^2, \quad w_k(\mathbf{x}_i) = \omega_{k0} + \sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l, \forall k \in [K-1], \end{aligned} \tag{4.3.44}$$

where $H\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)}\right)$ minorizes $-\sum_{i=1}^{n}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k\left(\mathbf{x}_i\right)\right)\right]$ and is defined as follows:

$$\sum_{i=1}^{n}\left[-\sum_{k=1}^{K-1}\frac{g_k\left(\mathbf{x}_i;\boldsymbol{\omega}^{(r)}\right)\sum_{l=0}^{L}\sum_{j=1}^{p}\exp\left[(Lp+1)\left(\omega_{klj}-\omega_{klj}^{(r)}\right)x_{ij}^l\right]}{Lp+1} - \ln C_i^{(r)} + \frac{C_i^{(r)}-1}{C_i^{(r)}}\right].$$

Proof of Lemma 4.3.14. Firstly, we claim that if $\omega > 0$, then the function $-\ln(1+\omega)$ can be minorized by

$$-\ln\left(1+\omega^{(r)}\right) - \frac{\omega-\omega^{(r)}}{1+\omega^{(r)}}, \text{ at } \omega^{(r)} > 0. \tag{4.3.45}$$

Note that (4.3.45) is proved if, given any $\omega^{(r)} > 0$, we have

$$-\ln(1+\omega) \geq -\ln\left(1+\omega^{(r)}\right) - \frac{\omega-\omega^{(r)}}{1+\omega^{(r)}}, \forall \omega > 0, \text{ or equivalently}$$
$$\ln\left(\frac{1+\omega}{1+\omega^{(r)}}\right) \leq \frac{\omega-\omega^{(r)}}{1+\omega^{(r)}}, \forall \omega > 0. \tag{4.3.46}$$

Let $x = \frac{1+\omega}{1+\omega^{(r)}}, \omega > 0$. Then, (4.3.46) is obtained from the following standard logarithm inequality

$$\ln(x) \leq x - 1, \forall x > 0.$$

Indeed, if $x \geq 1$, let $f(x) = \ln(x) - x + 1$. Thus, $f'(x) = \frac{1-x}{x}$. So f is monotonically decreasing and $f(x) \leq f(1) = 0, \forall x \geq 1$. This means that $\ln(x) \leq x - 1, \forall x \geq 1$. Now, if $0 < x < 1$, we have $-\ln(x) = \int_x^1 \frac{dt}{t}$. Since $1 \leq \frac{1}{t} \leq \frac{1}{x}, \forall x \leq t \leq 1$, we have, $\int_x^1 dt \leq \int_x^1 \frac{dt}{t} \leq \int_x^1 \frac{dt}{x}$ or equivalently, $1 - x \leq -\ln(x) \leq \frac{1-x}{x}$. So,

$$\frac{x-1}{x} \leq \ln(x) \leq x - 1, \forall 0 < x < 1.$$

One of the virtues of applying inequality (4.3.45) in defining a surrogate function is that it eliminates the log terms w.r.t. model parameters. Then, (4.3.45) implies that $-\ln\left[1 + \sum_{k=1}^{K-1} \exp\left(w_k\left(\mathbf{x}_i\right)\right)\right]$ is minorized by

$$-\ln\left[1 + \sum_{k=1}^{K-1} \exp\left(w_k^{(r)}(\mathbf{x}_i)\right)\right] - \frac{\sum_{k=1}^{K-1}\left[\exp\left(w_k\left(\mathbf{x}_i\right)\right) - \exp\left(w_k^{(r)}(\mathbf{x}_i)\right)\right]}{1 + \sum_{k=1}^{K-1} \exp\left(w_k^{(r)}(\mathbf{x}_i)\right)}$$
$$= -\ln C_i^{(r)} - \sum_{k=1}^{K-1} \frac{\exp\left(w_k^{(r)}(\mathbf{x}_i)\right)\exp\left(w_k(\mathbf{x}_i) - w_k^{(r)}(\mathbf{x}_i)\right)}{C_i^{(r)}} + \frac{C_i^{(r)} - 1}{C_i^{(r)}}.$$

Here, $C_i^{(r)} = 1 + \sum_{k=1}^{K-1} \exp\left(w_k^{(r)}(\mathbf{x}_i)\right)$. Now we wish to apply the weighted arithmetic-geometric mean inequality to the exponential functions $\exp\left(w_k(\mathbf{x}_i) - w_k^{(r)}(\mathbf{x}_i)\right)$ to separate parameters. This feature is critically important in high-dimensional problems because it reduces optimization over $\mathbf{x}_i$ in potential large p-dimension to a sequence of one-dimensional optimizations over each component $x_{ij}, i \in [n], j \in [p]$.

Recall the weighted arithmetic-geometric mean inequality. Consider nonnegative numbers $x_1, \ldots, x_p$ and positive weights $\alpha_1, \ldots, \alpha_p$ with $\sum_{j=1}^p \alpha_j = 1$. Then, the weighted arithmetic-geometric mean inequality reads

$$\prod_{j=1}^p x_j^{\alpha_j} \leq \sum_{j=1}^p \alpha_j x_j, \text{ or equivalently with } y_j = \ln x_j, \exp\left(\sum_{j=1}^p \alpha_j y_j\right) \leq \sum_{j=1}^p \alpha_j \exp\left(y_j\right) \tag{4.3.47}$$

with equality if and only if all x_j are equal.

The weighted arithmetic-geometric mean inequality implies that

$$\exp\left(w_k(\mathbf{x}_i) - w_k^{(r)}(\mathbf{x}_i)\right) = \exp\left(\omega_{k0} - \omega_{k0}^{(r)} + \sum_{l=1}^L \left(\boldsymbol{\omega}_{kl}^\top - \boldsymbol{\omega}_{kl}^{(r)\top}\right)\mathbf{x}_i^l\right)$$
$$= \exp\left(\omega_{k0} - \omega_{k0}^{(r)} + \sum_{l=1}^L \sum_{j=1}^p \left(\omega_{klj} - \omega_{klj}^{(r)}\right) x_{ij}^l\right)$$
$$\leq \frac{\exp\left(Lp+1\right)}{Lp+1} \sum_{l=0}^L \sum_{j=1}^p \exp\left[\left(\omega_{klj} - \omega_{klj}^{(r)}\right) x_{ij}^l\right], \tag{4.3.48}$$

where $\omega_{k0j}^{(r)} = \omega_{k0}^{(r)}, \omega_{k0j} = \omega_{k0}, x_{ij}^0 = 1, \forall j \in [p]$ and the equality holds when $\left(\omega_{k0}, (\boldsymbol{\omega}_{kl})_{l\in[L]}\right) = \left(\omega_{k0}^{(r)}, \left(\boldsymbol{\omega}_{kl}^{(r)}\right)_{l\in[L]}\right)$.

Therefore, $-\sum_{i=1}^n \ln\left[1 + \sum_{k=1}^{K-1} \exp\left(w_k\left(\mathbf{x}_i\right)\right)\right]$ is minorized by $H\left(\boldsymbol{\omega}; \boldsymbol{\omega}^{(r)}\right)$, defined as follows:

$$\begin{aligned}
&\sum_{i=1}^n \left[-\sum_{k=1}^{K-1} \frac{\exp\left(w_k^{(r)}(\mathbf{x}_i)\right) \sum_{l=0}^L \sum_{j=1}^p \exp\left[(Lp+1)\left(\omega_{klj} - \omega_{klj}^{(r)}\right) x_{ij}^l\right]}{C_i^{(r)}\,(Lp+1)} - \ln C_i^{(r)} + \frac{C_i^{(r)} - 1}{C_i^{(r)}} \right] \\
&= \sum_{i=1}^n \left[-\sum_{k=1}^{K-1} \frac{g_k\left(\mathbf{x}_i; \boldsymbol{\omega}^{(r)}\right) \sum_{l=0}^L \sum_{j=1}^p \exp\left[(Lp+1)\left(\omega_{klj} - \omega_{klj}^{(r)}\right) x_{ij}^l\right]}{Lp+1} - \ln C_i^{(r)} + \frac{C_i^{(r)} - 1}{C_i^{(r)}} \right].
\end{aligned}$$

□

Lemma 4.3.14 allows us to maximize $Q_{\text{pen}}\left(\boldsymbol{\omega}; \boldsymbol{\psi}^{(t)}\right)$ via its surrogate function $G\left(\boldsymbol{\omega}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right)$, which benefits the elimination the log terms w.r.t. model parameters and avoiding matrix inversion in high-dimensional problems via separating of parameters. Next, we aim to decompose $G\left(\boldsymbol{\omega}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right)$ according to parameters as follows:

$$G\left(\boldsymbol{\omega}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = G\left(\omega_{k0}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) + \sum_{k=1}^K \sum_{l=1}^L \sum_{j=1}^p G\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) + I(\boldsymbol{\omega}^{(r)}), \tag{4.3.49}$$

where $I(\boldsymbol{\omega}^{(r)})$ is only function of $\boldsymbol{\omega}^{(r)}$. Here, for all $k \in [K], j \in [p], l \in \{0\} \cup [L]$, we have

$$G\left(\omega_{k0}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = \sum_{i=1}^n \tau_{ik}^{(t)} w_{k0} - \sum_{i=1}^n \frac{g_k\left(\mathbf{x}_i; \boldsymbol{\omega}^{(r)}\right) \exp\left[(Lp+1)\left(\omega_{k0} - \omega_{k0}^{(r)}\right)\right]}{Lp+1}, \tag{4.3.50}$$

and

$$\begin{aligned}
G\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = &\sum_{i=1}^n \tau_{ik}^{(t)} x_{ij}^l w_{klj} - \sum_{i=1}^n \frac{g_k\left(\mathbf{x}_i; \boldsymbol{\omega}^{(r)}\right) \exp\left[(Lp+1)\, x_{ij}^l \left(\omega_{klj} - \omega_{klj}^{(r)}\right)\right]}{Lp+1} \\
&- \lambda_{kl}^{[1]} |\omega_{klj}| - \frac{\lambda^{[3]}}{2} \omega_{klj}^2.
\end{aligned} \tag{4.3.51}$$

Then, by maximizing (4.3.50), we can update the ω_{k0} via solving the first-order condition $\nabla_{\omega_{k0}} G\left(\omega_{k0}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = 0$, where

$$\nabla_{\omega_{k0}} G\left(\omega_{k0}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = \sum_{i=1}^n \tau_{ik}^{(t)} - (Lp+1) \exp\left[(Lp+1)\left(\omega_{k0} - \omega_{k0}^{(r)}\right)\right] \frac{\sum_{i=1}^n g_k\left(\mathbf{x}_i; \boldsymbol{\omega}^{(r)}\right)}{Lp+1}.$$

Then, we obtain

$$\omega_{k0}^{(r+1)} = \omega_{k0}^{(r)} + \frac{1}{Lp+1} \ln\left[\frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{\sum_{i=1}^n g_k\left(\mathbf{x}_i; \boldsymbol{\omega}^{(r)}\right)}\right]. \tag{4.3.52}$$

Remark that $G\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right)$ is a concave and univariate function w.r.t. w_{klj}. Therefore, we can maximize it globally w.r.t. each coefficient w_{klj} separately and then avoid matrix inversion. Indeed, note that

$$G\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) = U\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) - \lambda_{kl}^{[1]} |\omega_{klj}| = \begin{cases} U\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) - \lambda_{kl}^{[1]} \omega_{klj}, & \text{if } \omega_{klj} > 0, \\ U\left(0; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right), & \text{if } \omega_{klj} = 0, \\ U\left(\omega_{klj}; \boldsymbol{\omega}^{(r)}, \boldsymbol{\psi}^{(t)}\right) + \lambda_{kl}^{[1]} \omega_{klj}, & \text{if } \omega_{klj} < 0, \end{cases}$$

where

$$U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)=\sum_{i=1}^{n}\tau_{ik}^{(t)}x_{ij}^{l}w_{klj}-\sum_{i=1}^{n}\frac{g_k\left(\mathbf{x}_i;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(Lp+1\right)x_{ij}^{l}\left(\omega_{klj}-\omega_{klj}^{(r)}\right)\right]}{Lp+1}-\frac{\lambda^{[3]}}{2}\omega_{klj}^{2}.$$

Remark that $G\left(\cdot;\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)$ is a smooth concave function on both $\mathbb{R}^{+}$ and $\mathbb{R}^{-}$. We therefore can use one-dimensional generalized Newton-Raphson (GNR) algorithm to find the global maximizers of these functions and compare with $G\left(0;\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)$ so that we have

$$\omega_{klj}^{(r+1)}=\arg\max_{\omega_{klj}}G\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right). \tag{4.3.53}$$

After starting from an initial value $s=0$, $\omega_{klj}^{(0)}=\omega_{klj}^{(r)}$, at each iteration s of the GNR, according to the following updating rule:

$$\omega_{klj}^{(s+1)}=\omega_{klj}^{(s)}-\left(\frac{\partial^2 G\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}}\right)^{-1}\Bigg|_{\omega_{klj}^{(s)}}\frac{\partial G\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}\Bigg|_{\omega_{klj}^{(s)}}. \tag{4.3.54}$$

Here, the scalar gradient and Hessian are respectively given by:

$$\begin{aligned}\frac{\partial G\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}&=\begin{cases}\frac{\partial U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}-\lambda_{kl}^{[1]}, & \text{if } \omega_{klj}>0,\\ \frac{\partial U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}+\lambda_{kl}^{[1]}, & \text{if } \omega_{klj}<0,\end{cases}\\ \frac{\partial^2 G\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}}&=\frac{\partial^2 U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}},\ \text{if } \omega_{klj}\neq 0.\end{aligned} \tag{4.3.55}$$

Note that we have

$$\begin{aligned}\frac{\partial U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}&=\sum_{i=1}^{n}\tau_{ik}^{(t)}x_{ij}^{l}-\sum_{i=1}^{n}x_{ij}^{l}g_k\left(\mathbf{x}_i;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(Lp+1\right)x_{ij}^{l}\left(\omega_{klj}-\omega_{klj}^{(r)}\right)\right]-\lambda^{[3]}\omega_{klj},\\ \frac{\partial^2 U\left(\omega_{klj};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}}&=-\left(Lp+1\right)\sum_{i=1}^{n}x_{ij}^{l2}g_k\left(\mathbf{x}_i;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(Lp+1\right)x_{ij}^{l}\left(\omega_{klj}-\omega_{klj}^{(r)}\right)\right]-\lambda^{[3]}.\end{aligned}$$

Remark 4.3.15. Although one of the virtues of the MM algorithm in high-dimensional problems is that it allows us to update the parameters separately and then avoids matrix inversion, we have to deal with some drawbacks.

Coordinate ascent algorithm for updating the gating network

Motivated by Tseng (1988, 2001), we aim to use the coordinate ascent algorithm to update the parameters $\boldsymbol{\omega}=\left(\omega_{k0},\left(\boldsymbol{\omega}_{kl}\right)_{l\in[L]}\right)_{k\in[K]}$ of the gating networks.

We first use a univariate Newton-Raphson algorithm to update ω_{k0}. By starting with initial value $r=0,\omega_{k0}^{(0)}=\omega_{k0}^{(t)}$, we can use one-dimensional generalized Newton-Raphson (GNR) algorithm to approximate the global maximizers of a univariate concave function $Q_{\text{pen}}\left(\boldsymbol{\omega}_{k0};\boldsymbol{\psi}^{(t)}\right)$ in order to update $\omega_{k0}^{(r)}$ by

$$\begin{aligned}\omega_{k0}^{(r+1)}&=\arg\max_{\omega_{k0}}Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right),\ \text{where}\\ Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right)&=\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^{l}\right)-\sum_{i=1}^{n}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^{l}\right)\right].\end{aligned} \tag{4.3.56}$$

Here, r denotes the r loop of the coordinate ascent algorithm. After starting from an initial value $s=0$, $\omega_{k0}^{(0)}=\omega_{k0}^{(r)}$, at each iteration s of the GNR, according to the following updating rule:

$$\omega_{k0}^{(s+1)}=\omega_{k0}^{(s)}-\left(\frac{\partial^2 Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{k0}}\right)^{-1}\left|_{\omega_{k0}^{(s)}}\right.\frac{\partial Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{k0}}\left|_{\omega_{k0}^{(s)}}\right.. \tag{4.3.57}$$

Here s denotes the inner GNR iteration number, the scalar gradient and Hessian are respectively given by

$$\begin{aligned}
\frac{\partial Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{k0}}&=\sum_{i=1}^{n}\tau_{ik}^{(t)}-\sum_{i=1}^{n}\frac{\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l\right)}{C_i\left(\omega_{k0}\right)},\\
\frac{\partial^2 Q_{\text{pen}}\left(\omega_{k0};\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{k0}}&=-\sum_{i=1}^{n}\frac{\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l\right)\left[C_i\left(\omega_{k0}\right)-\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l\right)\right]}{C_i\left(\omega_{k0}\right)^2},\\
C_i\left(\omega_{k0}\right)&=1+\sum_{u=1}^{K-1}\exp\left(\omega_{u0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{ul}^{\top}\mathbf{x}_i^l\right).
\end{aligned}$$

With the same idea, each coefficient $\boldsymbol{\omega}_{kl}, l\neq 0$, is updated at each time in a cyclic way, while fixing the other parameters to their previous values. With this setting, we have

$$Q_{\text{pen}}\left(\boldsymbol{\omega}_{kl};\boldsymbol{\psi}^{(t)}\right)=U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)-\lambda_{kl}^{[1]}\left|\boldsymbol{\omega}_{klj}\right|=\begin{cases}U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)-\lambda_{kl}^{[1]}\omega_{klj}, & \text{if } \omega_{klj}>0,\\ U\left(0;\boldsymbol{\psi}^{(t)}\right), & \text{if } \omega_{klj}=0,\\ U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)+\lambda_{kl}^{[1]}\omega_{klj}, & \text{if } \omega_{klj}<0,\end{cases}$$

where

$$U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)=\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l\right)-\sum_{i=1}^{n}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l\right)\right]-\frac{\lambda^{[3]}}{2}\boldsymbol{\omega}_{klj}^2.$$

Note that $U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)-\lambda_{kl}^{[1]}\omega_{klj}$ and $U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)+\lambda_{kl}^{[1]}\omega_{klj}$ are both smooth concave functions on both $\mathbb{R}^+$ and $\mathbb{R}^-$. We therefore can use one-dimensional generalized Newton-Raphson (GNR) algorithm with initial value $\omega_{klj}^{(0)}=\omega_{klj}^{(t)}$ to find the global maximizers of these functions and compare with $G\left(0;\boldsymbol{\psi}^{(t)}\right)$ in order to update $\omega_{klj}^{(r)}$ by

$$\omega_{klj}^{(r+1)}=\arg\max_{\omega_{klj}} G\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right). \tag{4.3.58}$$

Here, r denotes the r loop of the coordinate ascent algorithm. After starting from an initial value $s=0$, $\omega_{klj}^{(0)}=\omega_{klj}^{(r)}$, at each iteration s of the GNR, according to the following updating rule:

$$\omega_{klj}^{(s+1)}=\omega_{klj}^{(s)}-\left(\frac{\partial^2 G\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}}\right)^{-1}\left|_{\omega_{klj}^{(s)}}\right.\frac{\partial G\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}\left|_{\omega_{klj}^{(s)}}\right., \tag{4.3.59}$$

Here s denotes the inner GNR iteration number, the scalar gradient and Hessian are respectively given by:

$$\begin{aligned}
\frac{\partial G\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}&=\begin{cases}\frac{\partial U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}-\lambda_{kl}^{[1]}, & \text{if } \omega_{klj}>0,\\ \frac{\partial U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}+\lambda_{kl}^{[1]}, & \text{if } \omega_{klj}<0,\end{cases}\\
\frac{\partial^2 G\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}}&=\frac{\partial^2 U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial^2\omega_{klj}},\ \text{if } \omega_{klj}\neq 0.
\end{aligned} \tag{4.3.60}$$

Note that we have

$$\frac{\partial U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial\omega_{klj}}=\sum_{i=1}^{n}\tau_{ik}^{(t)}x_{ij}^{l}-\sum_{i=1}^{n}\frac{x_{ij}^{l}\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_{i}^{l}\right)}{C_{i}\left(\omega_{klj}\right)}-\lambda^{[3]}\omega_{klj},$$

$$\frac{\partial^{2}U\left(\omega_{klj};\boldsymbol{\psi}^{(t)}\right)}{\partial^{2}\omega_{klj}}=-\sum_{i=1}^{n}\frac{x_{ij}^{l2}\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_{i}^{l}\right)\left[C_{i}\left(\omega_{klj}\right)-\exp\left(\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_{i}^{l}\right)\right]}{C_{i}\left(\omega_{klj}\right)^{2}}-\lambda^{[3]},$$

$$C_{i}\left(\omega_{klj}\right)=1+\sum_{u=1}^{K-1}\exp\left(\omega_{u0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{ul}^{\top}\mathbf{x}_{i}^{l}\right).$$

For other parameters, we fix their previous values $\omega_{abc}^{(r+1)}=\omega_{abc}^{(r)}, a\in[K]\backslash\{k\}, b\in[L]\backslash\{l\}, c\in[p]\backslash\{j\}$.

Remark 4.3.16. The main virtue of the coordinate ascent algorithm in high-dimensional problems is that it allows us to update the parameters separately and then avoids matrix inversion. Furthermore, it holds that the parameter $\boldsymbol{\omega}_{klj}$ may change during the algorithm even after they shrink to zero at an earlier stage of the algorithm, which overcome this weakness of the MM algorithm.

Proximal Newton-type procedure for updating the gating network

Definition 4.3.17 (Proximal Newton-type methods, *e.g.,* Lee et al. (2014))**.** Assume that we want to solve an optimization problem given by

$$\arg\max_{x\in\mathbb{R}^{n}}f(x)=g(x)+h(x),\tag{4.3.61}$$

with a composite function f where g is a concave function, continuously differentiable loss function, and h is a concave but not necessarily differentiable penalty function or regularizer (*e.g.,* Lasso, elastic net, ...). Proximal Newton-type methods approximate only the smooth part g with a local quadratic function of the form:

$$\hat{f}_{s}(x)=g(x_{s})+\nabla g\left(x_{s}\right)^{\top}\left(x-x_{s}\right)+\frac{1}{2}\left(x-x_{s}\right)^{\top}H_{s}\left(x-x_{s}\right)+h(x),\tag{4.3.62}$$

where $\nabla g\left(x_{s}\right)$ is the gradient vector of g at x_s and H_s is an approximation to the Hessian matrix $\nabla^2 g(x_s)$. If we choose $H_s=\nabla_g^2\left(x_s\right)$, we obtain the *proximal Newton method*. In this method, one uses an iterative algorithm with initial value x_0 and in which at step s minimizes the proximal function $\hat{f}_s(x)$ instead of f and searches for the next value x_{s+1} based on the solution of (4.3.62) that will improve the value of f, *i.e.,* $f\left(x_{s+1}\right)<f\left(x_s\right)$ by using a back tracking line research until the algorithm convergences. A generic proximal Newton-type method can be found in Algorithm 2.

Algorithm 2 A generic proximal Newton-type procedure

1: Starting point $x_0\in\operatorname{dom}f$.
2: **while** Stopping condition is satisfied **do**
3: Choose H_s, a positive definite approximation to the Hessian.
4: Solve the subproblem for a search direction:

$$\nabla x_{s}\leftarrow\arg\min_{d}\nabla g\left(x_{s}\right)^{\top}d+\frac{1}{2}dH_{s}d+h\left(x_{s}+d\right).$$

5: Select t_s with a backtracking line search.
6: Update: $x_{s+1}\leftarrow x_s+t_s\nabla x_s$.
7: **end while**

In this part, we propose two approaches for updating the gating network parameters $\boldsymbol{\omega}$ via maximizing $Q_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ based on the proximal Newton and the proximal Newton-type methods defined in Definition 4.3.17.

Firstly, the proximal Newton method is used to approximate only the smooth part of (4.3.39) given by

$$I\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}w_k(\mathbf{x}_i)-\sum_{i=1}^{n}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k(\mathbf{x}_i)\right)\right],$$
$$w_k(\mathbf{x}_i)=\omega_{k0}+\sum_{l=1}^{L}\boldsymbol{\omega}_{kl}^{\top}\mathbf{x}_i^l,\forall k\in[K-1]. \tag{4.3.63}$$

Then, its Taylor expansion at current s iteration is provided by

$$\widetilde{I}\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(s)},\boldsymbol{\psi}^{(t)}\right)=I\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)+\nabla_{\boldsymbol{\omega}}I\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)^{\top}\left(\boldsymbol{\omega}-\boldsymbol{\omega}^{(s)}\right)$$
$$+\frac{1}{2}\left(\boldsymbol{\omega}-\boldsymbol{\omega}^{(s)}\right)^{\top}\nabla_{\boldsymbol{\omega}}^{2}I\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)^{\top}\left(\boldsymbol{\omega}-\boldsymbol{\omega}^{(s)}\right), \tag{4.3.64}$$

where $\nabla_{\boldsymbol{\omega}}I\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)$ and $\nabla_{\boldsymbol{\omega}}^{2}I\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)$ are the gradient vector and the Hessian matrix of $I\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ at $\boldsymbol{\omega}^{(s)}$, respectively. Then, let us define the proximal function as follows:

$$\widetilde{Q}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)=\widetilde{I}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)-\sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{kl}^{[1]}\left\|\boldsymbol{\omega}_{kl}\right\|_1-\frac{\lambda^{[3]}}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\boldsymbol{\omega}_{kl}\right\|_2^2. \tag{4.3.65}$$

Next, one uses an iterative algorithm with initial value $\boldsymbol{\omega}^{(0)}$ and in which at step s minimizes the proximal function $\widetilde{Q}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ instead of $Q\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(t)}\right)$ and searches for the next value $\boldsymbol{\omega}^{(s+1)}$ based on the solution of (4.3.65) that will improve the value of Q, *i.e.*, $Q\left(\boldsymbol{\omega}^{(s)};\boldsymbol{\psi}^{(t)}\right)<Q\left(\boldsymbol{\omega}^{(s+1)};\boldsymbol{\psi}^{(t)}\right)$ by using a back tracking line research until the algorithm convergences. Note that we can effectively solve the local quadratic form from (4.3.65) via several good algorithms such as coordinate ascent.

Updating the Gaussian expert networks

For the penalized MoE models with univariate response variables, performing the update for the Gaussian experts network parameters corresponds to solving K separated weighted Lasso problems (see Chamroukhi & Huynh (2019, Section 3.3.3) for more details). However, for multivariate case, we propose a new method to deal with the following complex objective function

$$\begin{aligned}
&Q_{\text{pen}}\left(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{\psi}^{(t)}\right)\\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\ln\left[\phi_q\left(\mathbf{y}_i;\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i),\boldsymbol{\Sigma}_k\right)\right]-\sum_{k=1}^{K}\sum_{d=1}^{D}\lambda_{kd}^{[2]}\left\|\mathbf{Q}_k\boldsymbol{\beta}_{kd}\right\|_1,\mathbf{Q}_k^{\top}\mathbf{Q}_k=\boldsymbol{\Sigma}_k^{-1},\boldsymbol{\Gamma}_{kd}=\mathbf{Q}_k\boldsymbol{\beta}_{kd},\\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\ln\left[\frac{1}{(2\pi)^{q/2}\det\left(\boldsymbol{\Sigma}_k\right)^{1/2}}\exp\left(-\frac{\left(\mathbf{y}_i-\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i)\right)^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_i-\boldsymbol{\upsilon}_{k,\boldsymbol{\beta}}(\mathbf{x}_i)\right)}{2}\right)\right]\\
&\quad-\sum_{k=1}^{K}\sum_{d=1}^{D}\lambda_{kd}^{[2]}\left\|\boldsymbol{\Gamma}_{kd}\right\|_1\\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\ln\left[\frac{\det\left(\mathbf{Q}_k\right)}{(2\pi)^{q/2}}\exp\left(-\frac{\left(\mathbf{Q}_k\mathbf{y}_i-\sum_{d=0}^{D}\boldsymbol{\Gamma}_{kd}\mathbf{x}^d\right)^{\top}\left(\mathbf{Q}_k\mathbf{y}_i-\sum_{d=0}^{D}\boldsymbol{\Gamma}_{kd}\mathbf{x}^d\right)}{2}\right)\right]\\
&\quad-\sum_{k=1}^{K}\sum_{d=1}^{D}\lambda_{kd}^{[2]}\left\|\boldsymbol{\Gamma}_{kd}\right\|_1=:Q_{\text{pen}}\left(\boldsymbol{\Gamma},\mathbf{Q};\boldsymbol{\psi}^{(t)}\right).
\end{aligned} \tag{4.3.66}$$

Since (4.3.66), optimizing $Q_{\text{pen}}\left(\boldsymbol{\beta},\boldsymbol{\Sigma};\boldsymbol{\psi}^{(t)}\right)$ w.r.t. $(\boldsymbol{\beta},\boldsymbol{\Sigma})$ is equivalent to maximize $Q_{\text{pen}}\left(\boldsymbol{\Gamma},\mathbf{Q};\boldsymbol{\psi}^{(t)}\right)$ w.r.t. $(\boldsymbol{\Gamma},\mathbf{Q})=(\boldsymbol{\Gamma}_{kd},\mathbf{Q}_k)_{k\in[K],d\in[D]\cup\{0\}}$.

Motivated by Tseng (1988, 2001), we aim to use the block coordinate ascent algorithm to update the parameters $(\boldsymbol{\Gamma}, \mathbf{Q}) = (\boldsymbol{\Gamma}_k, \mathbf{Q}_k)_{k\in[K]} = (\boldsymbol{\Gamma}_{kd}, \mathbf{Q}_k)_{k\in[K],d\in[D]\cup\{0\}}$ of the expert networks. A simple calculation shows that $Q_{\text{pen}}\left(\boldsymbol{\Gamma}, \mathbf{Q}; \boldsymbol{\psi}^{(t)}\right)$ can be decoupled for each components into k distinct optimization problem of the form

$$Q_{\text{pen}}\left(\boldsymbol{\Gamma}_k, \mathbf{Q}_k; \boldsymbol{\psi}^{(t)}\right) = \sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{j=1}^{q} \ln\left([\mathbf{Q}_k]_{j,j}\right) - \frac{1}{2}\sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{j=1}^{q} \left([\mathbf{Q}_k]_{j,j}\, \mathbf{y}_{ij} - \sum_{d=0}^{D} [\boldsymbol{\Gamma}_{kd}]_{j,\cdot}\, \mathbf{x}_i^d\right)^2 - \sum_{d=1}^{D} \lambda_{kd}^{[2]} \left\|\boldsymbol{\Gamma}_{kd}\right\|_1 . \tag{4.3.67}$$

Note that $\mathbf{Q}_k$ is a diagonal matrix of size $q \times q$, defined by the Cholesky decomposition of a diagonal matrix $\boldsymbol{\Sigma}_k^{-1}$.

Chapter 5

Conclusion and Perspectives

Contents

5.1 Approximation capabilities of the mixtures of experts models

Using recent results mixture model approximation results Nguyen et al. (2020b) and Nguyen et al. (2020d), and the indicator approximation theorem of Jiang & Tanner (1999b) (cf. Section 2.3.2), in Section 2.3, we have proved two approximation theorems (Theorems 2.3.1 and 2.3.2) regarding the class of softmax gated MoE models with experts arising from arbitrary location-scale families of conditional density functions. Via an equivalence result (Lemma 2.3.3), the results of Theorems 2.3.1 and 2.3.2 also extend to the setting of Gaussian gated MoE models (Corollary 2.3.4), which can be seen as a generalization of the softmax gated MoE models.

Although we explicitly make the assumption that $\mathbb{X} = [0,1]^d$, for the sake of mathematical argument (so that we can make direct use of Lemma 2.3.6), a simple shift-and-scale argument can be used to generalize our result to cases where $\mathbb{X}$ is any generic compact domain. The compactness assumption regarding the input domain is common in the MoE and mixture of regression models literature, as per the works of Jiang & Tanner (1999b), Norets et al. (2010), Montuelle et al. (2014), Pelenis (2014), Devijver (2015a), Devijver (2015b), and Nguyen et al. (2020c).

Open Problem 5.1.1. *The assumption permits the application of the result to the settings where the inputs $\boldsymbol{X}$ is assumed to be non-random design vectors that take value on some compact set $\mathbb{X}$. This is often the case when there is only a finite number of possible design vector elements for which $\boldsymbol{X}$ can take. Otherwise, the assumption also permits the scenario where $\boldsymbol{X}$ is some random element with compactly supported distribution, such as uniformly distributed, or beta distributed inputs. Unfortunately, the case of random $\boldsymbol{X}$ over an unbounded domain (e.g., if $\boldsymbol{X}$ has multivariate Gaussian distribution) is not covered under our framework. An extension to such cases would require a more general version of Lemma 2.3.6, which we believe is a nontrivial direction for future work.*

Like the input, we also assume that the output domain is restricted to a compact set $\mathbb{Y}$. However, the output domain of the approximating class of MoE models is unrestricted to $\mathbb{Y}$ and thus the functions (i.e., we allow ψ to be a PDF over $\mathbb{R}^q$). The restrictions placed on $\mathbb{Y}$ is also common in the mixture approximation literature, as per the works of Zeevi & Meir (1997), Li & Barron (1999), and Rakhlin et al. (2005), and is also often made in the context of nonparametric regression (see, e.g.,

Györfi et al., 2002 and Cucker & Zhou, 2007). Here, our use of the compactness of $\mathbb{Y}$ is to bound the integral of v_n, in (2.3.10).

Open Problem 5.1.2. *A more nuanced approach, such as via the use of a generalized Lebesgue spaces (see e.g., Castillo & Rafeiro, 2016 and Cruz-Uribe & Fiorenza, 2013), may lead to result for unbounded $\mathbb{Y}$. This is another exciting future direction of our research program.*

A trivial modification to the proof of Lemma 2.3.8 allows us to replace the assumption that f is a PDF with a sub-PDF assumption (i.e., $\int_{\mathbb{Y}} f \mathrm{d}\lambda \leq 1$), instead. This in turn permits us to replace the assumption that $f(\cdot|\boldsymbol{x})$ is a conditional PDF in Theorems 2.3.1 and 2.3.2 with sub-PDF assumptions as well (i.e., for each $\boldsymbol{x} \in \mathbb{X}$, $\int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\lambda(\boldsymbol{y}) \leq 1$). Thus, in this modified form, we have a useful interpretation for situations when the input $\boldsymbol{Y}$ is unbounded. That is, when $\boldsymbol{Y}$ is unbounded, we can say that the conditional PDF f can be arbitrarily well approximated in $\mathcal{L}_p$ norm by a sequence $\left\{m_K^{\psi}\right\}_{K \in \mathbb{N}^\star}$ of either softmax or Gaussian gated MoEs over any compact subdomain $\mathbb{Y}$ of the unbounded domain of $\boldsymbol{Y}$. Thus, although we cannot provide guarantees of the entire domain of $\boldsymbol{Y}$, we are able to guarantee arbitrary approximate fidelity over any arbitrarily large compact subdomain. This is a useful result in practice since one is often not interested in the entire domain of $\boldsymbol{Y}$, but only on some subdomain where the probability of $\boldsymbol{Y}$ is concentrated. This version of the result resembles traditional denseness results in approximation theory, such as those of Cheney & Light (2000, Ch. 20).

Finally, our results can be directly applied to provide approximation guarantees for a large number of currently used models in applied statistics and machine learning research. Particularly, our approximation guarantees are applicable to the recent MoE models of Ingrassia et al. (2012), Chamroukhi et al. (2013b), Ingrassia et al. (2014), Deleforge et al. (2015c,b), Chamroukhi (2017, 2016a), Kalliovirta et al. (2016), and Perthame et al. (2018), among many others. Here, we may guarantee that the underlying data generating processes, if satisfying our assumptions, can be adequately well approximated by sufficiently complex forms of the models considered in each of the aforementioned work. Furthermore, establishing the convergence rates of the MLE for full Gaussian MoE, including SGaME, GLoME, and BLoMPE models, is an interesting and important question.

Open Problem 5.1.3. *Exploiting the connection between the algebraic independence and a certain class of partial differential equations (Nguyen, 2013, Ho & Nguyen, 2016, 2019) maybe allows us to extend the convergence rates and minimax lower bounds for parameter estimation in the work of Ho et al. (2019). Note that such extension from an over-specified Gaussian mixtures of experts with covariate-free gating networks (or often called mixture of Gaussian regression models) to full Gaussian MoE is not trivial. This future work requires an appropriate generalization of the transportation distance to capture the variation of parameters from the gating networks.*

5.2 Universal approximation for mixture of experts models in approximate Bayesian computation

In Section 2.4, the issue of choosing summary statistics was revisited. We built on the seminal work of Fearnhead & Prangle (2012) and their semi-automatic ABC by replacing the approximate posterior expectations with functional statistics; namely approximations of the posterior distributions. These surrogate posterior distributions were obtained in a preliminary learning step, based on an inverse regression principle. This is original with respect to most standard regression procedures, which usually provide only point-wise predictions, *i.e.* first order moments. So doing, we not only could compute approximate posterior moments of higher orders as summary statistics but, more generally, approximate full posterior distributions. More specifically, this learning step was based on the so-called GLLiM model, which provides surrogate posteriors in the parametric family of Gaussian mixtures. Preliminary experiments showed that although the posterior moments provided by GLLiM were not always leading to better results than that provided by semi-automatic ABC, the use of the full surrogate posteriors was always an improvement.

To handle distributions as summary statistics, our procedure required appropriate distances. We investigated an L_2 and a Wassertein-based distance (MW_2), which are both tractable for mixtures of Gaussians. No significant differences between the two distances have been observed in our experiments but the MW_2 distance appeared to be more robust in the sense of being less sensitive to small variations in the compared distributions.

Open Problem 5.2.1. *Among aspects that have not been thoroughly investigated in this work, we could refine the way to choose this tolerance level ϵ or combine GLLiM with more sophisticated ABC schemes than the simple rejection scheme. In particular, using other ways to choose ϵ is needed to investigate more in future research,* e.g., *cross-validation, hold-out.*

In this current work, our proposal applies to the ABC settings, where, for a given parameter value, only one observation (that is possibly multi-dimensional) is available at a time. Such settings are of practical importance as they are typical of inverse problems, where many observations are measured but for different parameter values, due to experimental limitations or costs. In addition, even when more than one observation is available, it is common to use summary statistics. For instance, in their g-and-k distribution experiment, Fearnhead & Prangle (2012) consider, for a true given parameter, a sample of 10^4 observations, but reduce it to 100 features to apply the regression step of their semi-automatic procedure. Similarly, Drovandi & Pettitt (2011) reduce their sample of 10^4 observations to a vector of 7 octiles. So doing their analyses imply the one observation scenario, that we consider.

In contrast, methods using discrepancies (Bernton et al., 2019, Jiang et al., 2018) can handle samples directly and bypass the need for summary statistics. However, they require a relatively large number of generally *i.i.d.* observations for both the true and simulated parameters.

Open Problem 5.2.2. *The current implementation of GLLiM is not adapted to the multiple observation case but a straightforward modification of the underlying EM algorithm would allow to extend this work to the case of i.i.d. samples. For computational reasons, as for the semi-automatic procedure, the preliminary regression step in standard GLLiM is not adapted to the multiple observation case. Therefore, an important future direction is to extend this work to the case of i.i.d. samples.*

This requires the modification of the standard GLLiM procedure to maintain its approximation quality and computational efficiency. With this in mind, an important feature of GLLiM, not illustrated in this paper, is to allow the application of ABC procedures in high dimensional settings and to address the curse of dimensionality that is usually encountered in standard summary statistics based ABC. The rest of our proposal would then be easily adapted.

Another interesting perspective would be to investigate the use of GLLiM in the context of synthetic likelihood (SL) approaches. When used in a Bayesian framework, SL techniques can be viewed as alternatives to ABC in which the intractable likelihood is replaced by an estimator of the likelihood (Price et al., 2018). Since the seminal work of Wood (2010), several estimators have been proposed (e.g. Ong et al., 2018, An et al., 2019, 2020, Frazier & Drovandi, 2021), often derived from auxiliary models (Drovandi et al., 2015).

Open Problem 5.2.3. *In the ABC framework of Section 2.4, GLLiM was used to provide approximate posteriors but these posteriors are themselves coming from approximate likelihoods that could lead to new SL procedures. Investigating more in this potential direction will be an important question for future research.*

At last, in principle, any other method that is able to provide approximate surrogate posteriors could be used in place of GLLiM to produce the functional summaries.

Open Problem 5.2.4. *Besides the family of mixture of experts models which are similar to GLLiM, mixture density networks (Bishop, 1994) or normalizing flows (Dinh et al., 2015, Kruse et al., 2021) are potential candidates. To our knowledge, other common neural networks, like most regression techniques, would not be appropriate as they only focus on point-wise predictions.*

Open Problem 5.2.5. *It would be interesting to refine our consistency results by looking at the rate of convergence; either towards the posterior distribution with the same spirit as in Frazier et al. (2018), or directly by studying the statistical properties of the GLLiM-ABC algorithm. We now add this as our priority open question problem.*

5.3 Model selection in the Gaussian-gated localized mixture of polynomial experts regression model

In Chapter 3, we have studied the PMLEs for GLoME and BLoMPE regression models. Our main contributions are non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds on the penalties hold true. Furthermore, aside from important theoretical issues regarding the tightness of the bounds of the PMLE, we hope that our contribution helps to popularize GLoME models, as well as GLLiM models and slope heuristics, by giving some theoretical foundations for model selection technique in this area and demonstrating some interesting numerical schemes and experiments.

Open Problem 5.3.1. *Recall that the main methods to account for the model selection procedures are cross-validation and hold-out (see,* e.g., *Arlot & Celisse, 2010, Maillard, 2020 for the complete bibliography), or penalized criteria. In this thesis, we focused on the PMLEs for high-dimensional GLoME and BLoMPE regression models. However, one may ask whether we can establish such similar weak oracle inequalities for hold-out and cross-validation procedure in the context of MoE regression models. We wish to resolve this interesting question in future research.*

Note that some recent attempts have been made to develop the estimation of non-standard MoEs, the theory regarding their approximation capacity, as well as their applications in functional data analysis and signal processing. For a recent account of the theory, we refer the reader to the works of estimation methodology for non-standard MoE models with Laplace, Student-t, and skew-t experts, see *e.g.,* Nguyen & McLachlan (2016), Perthame et al. (2018), Chamroukhi (2016a,b, 2017), respectively.

Open Problem 5.3.2. *In particular, we aim to provide an extension of the finite-sample oracle inequality, Theorems 3.2.3 and 3.3.2, to a more general framework where Gaussian experts are replaced by another distributions,* e.g., *Student t-distributions, elliptical distributions, in the future work.*

Open Problem 5.3.3. *In Theorems 3.2.3 and 3.3.2, we can only obtain weak oracle inequalities,* i.e., $C_1 > 1$. *We aim to provide "exact" oracle inequalities with $C_1 = 1$ in future work. To our knowledge, this issue has not been solved in PMLEs with Kullback-Leibler loss but only with L_2 norm or aggregation of a finite number of densities as in Rigollet (2012), Dalalyan & Sebbar (2018).*

Open Problem 5.3.4. *Note that in Chapter 3, we only aim to construct an upper bound and do not focus on the important question of the existence of a corresponding lower bound. To the best of our knowledge, providing a minimax analysis of our proposed estimator is still an open question. Furthermore, it is not trivial to extend the lower bound result regarding Gaussian mixtures, as presented in Maugis-Rabusseau & Michel (2013, Theorem 2.8), to the context of GLoME models. However, we wish to provide such minimax analysis in future research.*

Note that few theoretical confidence intervals have been studied for predicting a new response from a predictor in the era of high-dimensional data, in particular for MoE regression models. For Lasso based estimators, the standard works on deriving confidence regions for slope coefficient and statistical testing of sparsity for linear model are Javanmard & Montanari (2014, approximate inverse of the Gram matrix), van de Geer et al. (2014, desparsifying Lasso), Zhang & Zhang (2014, relaxed projection), Janková & van de Geer (2015, extensions for generalised linear model with subdifferential loss), Meinshausen (2015, for groups of variables), Stucky & van de Geer (2018, linear regression models with structured sparsity), Lee et al. (2016, exact post-selection inference) among others. those results rely on strong assumptions on the design and those results rely on strong assumptions on the design and remain difficult to be implemented.

Open Problem 5.3.5. *To overcome this problem, Devijver & Perthame (2020) proposed an explicit asymptotic prediction regions for the response to address the linear regression problem for elliptical distributions under an inverse regression approach rather than sparse regression. Therefore, we aim to extend this model to generalized linear model by considering other distributions of the noise of the inverse model or by our GLoME and BLoME models for future work.*

5.4 Joint rank and variable selection in the softmax-gated block-diagonal mixture of experts regression model

In Section 4.2, we have studied an l_1-regularization estimator for finite mixtures of Gaussian experts regression models with softmax gating functions. Our main contribution is the proof of the l_1-oracle inequality that provides the lower bound on the regularization of the Lasso that ensures non-asymptotic theoretical control on the Kullback-Leibler loss of the estimator. Other than some remaining questions regarding the tightness of the bounds and the form of penalization functions, we believe that our contribution helps to further popularize mixtures of Gaussian experts regression models by providing a theoretical foundation for their application in high-dimensional problems.

Open Problem 5.4.1. *First of all, the Theorem 4.3.2 of Section 4.3 was announced without supporting of numerical experiments. Therefore, in future work, we aim to test our algorithms in Section 4.3.5 and apply and evaluate our methods both on simulated and real datasets to understand how they work in practice.*

Open Problem 5.4.2. *Next, in Section 4.2, we focused on a simplified but standard setting in which the means of the experts are linear functions, with respect to explanatory variables. Although simplified, this model captures the core of the MoE regression problem, which is the interactions among the different mixture components. We believe that the general techniques that we develop here can be extended to more general experts, such as Gaussian experts with polynomial means* (e.g., *Mendes & Jiang, 2012) or even with hierarchical MoE for exponential family regression models in Jiang & Tanner (1999a). But we leave such nontrivial developments for future work.*

Open Problem 5.4.3. *In Lemma 2.3.3, we established the connection between the softmax-gated, $\mathcal{G}_S^K$, and Gaussian gating network, $\mathcal{G}_G^K$, namely $\mathcal{G}_S^K \subset \mathcal{G}_G^K$. Therefore, one may conjecture that the l_1-oracle inequality as in Theorem 4.2.2 for GLoME model instead of SGaME model still hold or not. However, we leave this interesting problem for future research.*

As pointed out by Arlot (2019), model-selection performance can be improved empirically by overpenalizing a bit the penalty function, see *e.g.,* Arlot & Baudry (2002), Arlot (2007, Chapter 11), Arlot (2009, Section 6.3.2, in the regression setting), Arlot & Lerasle (2016, Figure 3, in least-squares density estimation)

Open Problem 5.4.4. *In particular, for histogram selection in density estimation, Saumard & Navarro (2021) (see also Saumard, 2019 and Saumard, 2010 for a deeper comparison between their overpenalization strategy and slope heuristics as well as AIC for small sample sizes) proposed a new corrected version of the AIC criterion based on a natural way to overpenalize automatically. Therefore, we aim to investigate the choosing from data an appropriate overpenalization factor in the framework of MoE regression models in future research.*

Open Problem 5.4.5. *Furthermore, we would like to mathematically and fully justify the slope heuristic in MoE regression models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively Birgé & Massart (2007), Arlot & Massart (2009), Arlot & Bach (2009), Arlot (2019).*

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. (Cited on pages 12, 16, 54, 114, 144, 257, and 261.)

Aliprantis, C. D. & Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media. (Cited on pages 110 and 111.)

An, Z., Nott, D. J., & Drovandi, C. (2020). Robust Bayesian Synthetic Likelihood via a Semi-Parametric Approach. *Statistics and Computing*, 30, 543–557. (Cited on page 225.)

An, Z., South, L., Nott, D. J., & Drovandi, C. (2019). Accelerating Bayesian Synthetic Likelihood With the Graphical Lasso. *Journal of Computational and Graphical Statistics*, 28(2), 471–475. (Cited on page 225.)

Anderson, C. W., Stolz, E. A., & Shamsunder, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3), 277–286. (Cited on pages 24, 25, 196, and 284.)

Anderson, D. & Burnham, K. (2002). *Model Selection and Multi-model Inference*. A Pratical Information-Theoretic Approach. Springer-Verlag, New York, 2 edition. (Cited on pages 12, 16, 144, 257, and 261.)

Arlot, S. (2007). *Rééchantillonnage et Sélection de modèles. Resampling and Model selection*. Theses, Université Paris Sud - Paris XI. (Cited on page 227.)

Arlot, S. (2009). Model selection by resampling penalization. *Electronic Journal of Statistics*, 3(none), 557–624. (Cited on page 227.)

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3), 1–106. (Cited on pages 12, 17, 18, 51, 114, 121, 144, 212, 227, 257, and 263.)

Arlot, S. & Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, volume 22: Curran Associates, Inc. (Cited on pages 121 and 227.)

Arlot, S. & Baudry, J.-P. (2002). *Sélection de modèles*. Master 1 report, ENS Paris. (Cited on page 227.)

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), 40–79. (Cited on pages 15, 226, and 260.)

Arlot, S. & Lerasle, M. (2016). Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation. *Journal of Machine Learning Research*, 17(208), 1–50. (Cited on page 227.)

Arlot, S. & Massart, P. (2009). Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research*, 10(10), 245–279. (Cited on pages 121 and 227.)

Arlot, S., Vincent, B., Baudry, J.-P., Maugis, C., & Michel, B. (2016). capushe: CAlibrating Penalities Using Slope HEuristics. *R package version 1.1.1*, 1(1). (Cited on pages 18, 123, and 263.)

Arridge, S., Maass, P., Öktem, O., & Schönlieb, C.-B. (2019). Solving Inverse Problems Using Data-Driven Models. *Acta Numerica*, 28, 1–174. (Cited on page 93.)

Bach, F. R. (2008). Bolasso: Model Consistent Lasso Estimation through the Bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (pp. 33–40). New York, NY, USA: Association for Computing Machinery. (Cited on pages 24, 196, and 283.)

Bacharoglou, A. (2010). Approximation of probability distributions by convex mixtures of Gaussian measures. *Proceedings of the American Mathematical Society*, 138(7), 2619–2628. (Cited on pages 59, 60, 77, 268, and 269.)

Bahadur, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya: The Indian Journal of Statistics*, (pp. 207–210). (Cited on pages 14 and 258.)

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3), 930–945. (Cited on page 77.)

Barron, A. R., Huang, C., Li, J., & Luo, X. (2008). The MDL principle, penalized likelihoods, and statistical risk. *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, (pp. 33–63). (Cited on pages 30 and 130.)

Bartle, R. G. (1995). *The Elements Of Integration And Lebesgue Measure*. Wiley. (Cited on pages 63, 64, 78, 85, and 272.)

Baudry, J.-P. (2009). *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Université Paris-Sud XI. (Cited on page 122.)

Baudry, J.-P., Maugis, C., & Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2), 455–470. (Cited on pages 12, 18, 51, 55, 114, 122, 123, 144, 212, 257, and 263.)

Beal, M. J., Jojic, N., & Attias, H. (2003). A Graphical Model for Audiovisual Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7), 828–836. (Cited on page 102.)

Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., & Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research: Planets*, 114(E6). (Cited on pages 64, 272, and 273.)

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 235–269. (Cited on pages 65, 66, 68, 96, 97, 99, 105, 225, 273, 274, 275, and 276.)

Birgé, L. & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1), 113–150. (Cited on pages 14 and 258.)

Birgé, L. & Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3), 203–268. (Cited on pages 16, 17, 261, and 262.)

Birgé, L. & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1), 33–73. (Cited on pages 12, 17, 51, 54, 55, 114, 121, 122, 144, 212, 227, 257, and 262.)

Birgé, L., Massart, P., et al. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3), 329–375. (Cited on pages 19, 119, and 278.)

Bishop, C. M. (1994). *Mixture density networks*. Technical report, Aston University. (Cited on page 225.)

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. (Cited on page 139.)

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2), 189–208. (Cited on pages 65 and 273.)

Bock, A. S. & Fine, I. (2014). Anatomical and functional plasticity in early blind individuals and the mixture of experts architecture. *Frontiers in human neuroscience*, 8, 971. (Cited on pages 5 and 251.)

Bouchard, G. (2003). Localised Mixtures of Experts for Mixture of Regressions. In M. Schader, W. Gaul, & M. Vichi (Eds.), *Between Data Science and Applied Data Analysis* (pp. 155–164). Berlin, Heidelberg: Springer Berlin Heidelberg. (Cited on pages 5, 114, and 251.)

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press. (Cited on page 179.)

Boucheron, S. & Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3), 405–433. (Cited on pages 20 and 279.)

Brezis, H. (2010). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer. (Cited on pages 79 and 84.)

Brinkman, N. D. (1981). Ethanol Fuel—Single—Cylinder Engine Study of Efficiency and Exhaust Emissions. *SAE transactions*, (pp. 1410–1424). (Cited on page 129.)

Buchholz, A. & Chopin, N. (2019). Improving Approximate Bayesian Computation via Quasi-Monte Carlo. *Journal of Computational and Graphical Statistics*, 28(1), 205–219. (Cited on pages 64 and 273.)

Bunea, F. et al. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics*, 2, 1153–1194. (Cited on pages 51, 161, and 212.)

Bunea, F., She, Y., & Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5), 2359–2388. (Cited on pages 195 and 211.)

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3), 368–385. (Cited on pages 135, 136, and 138.)

Castillo, R. E. & Rafeiro, H. (2016). *An introductory course in Lebesgue spaces*. Springer. (Cited on page 224.)

Cavanaugh, J. E. & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3), e1460. (Cited on pages 16 and 261.)

Celeux, G. & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, 28(5), 781–793. (Cited on pages 21, 116, and 280.)

Chamroukhi, F. (2016a). Robust mixture of experts modeling using the t distribution. *Neural Networks*, 79, 20–36. (Cited on pages 224 and 226.)

Chamroukhi, F. (2016b). Skew-normal Mixture of Experts. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 3000–3007). (Cited on page 226.)

Chamroukhi, F. (2017). Skew t mixture of experts. *Neurocomputing*, 266, 390–408. (Cited on pages 224 and 226.)

Chamroukhi, F., Glotin, H., & Samé, A. (2013a). Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112, 153–163. (Cited on pages 5 and 251.)

Chamroukhi, F. & Huynh, B. T. (2018). Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). (Cited on pages 28, 51, 55, 161, 165, 198, and 212.)

Chamroukhi, F. & Huynh, B.-T. (2019). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *Journal de la Société Française de Statistique*, 160(1), 57–85. (Cited on pages 28, 51, 52, 55, 161, 165, 198, 211, 212, 214, and 221.)

Chamroukhi, F., Mohammed, S., Trabelsi, D., Oukhellou, L., & Amirat, Y. (2013b). Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120, 633–644. (Cited on page 224.)

Chamroukhi, F., Samé, A., Govaert, G., & Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6), 593–602. (Cited on pages 5 and 251.)

Chamroukhi, F., Samé, A., Govaert, G., & Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9), 1210–1221. (Cited on page 122.)

Chen, Y., Georgiou, T. T., & Tannenbaum, A. (2019). Optimal Transport for Gaussian Mixture Models. *IEEE Access*, 7, 6269–6278. (Cited on pages 68, 95, 103, and 276.)

Cheney, W. & Light, W. (2000). *A Course in Approximation Theory*. Pacific Grove: Brooks/Cole. (Cited on pages 58, 60, 61, 74, 75, 78, 80, 224, 267, 269, and 270.)

Cohen, S. & Le Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *Technical report, INRIA*. (Cited on pages 12, 18, 19, 20, 30, 31, 32, 33, 37, 38, 54, 114, 119, 120, 121, 128, 130, 133, 134, 144, 148, 149, 199, 200, 257, 263, 278, and 279.)

Cohen, S. X. & Le Pennec, E. (2013). Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, 17, 672–697. (Cited on pages 20, 30, 33, 200, and 279.)

Cohen, S. X. & Le Pennec, E. (2014). Unsupervised segmentation of spectral images with a spatialized gaussian mixture model and model selection. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, 69(2), 245–259. (Cited on pages 5 and 251.)

Conway, J. B. (2012). *A Course in Abstract Analysis*, volume 141. American Mathematical Society. (Cited on page 78.)

Cook, R. D. & Forzani, L. (2019). Partial Least Squares Prediction in High-Dimensional Regression. *The Annals of Statistics*, 47(2), 884–908. (Cited on page 93.)

Craven, P. & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4), 377–403. (Cited on pages 16 and 261.)

Cruz-Uribe, D. V. & Fiorenza, A. (2013). *Variable Lebesgue spaces: Foundations and harmonic analysis*. Springer Science & Business Media. (Cited on page 224.)

Csillery, K., Francois, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*. (Cited on page 99.)

Cucker, F. & Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press. (Cited on page 224.)

Dalalyan, A. S. & Sebbar, M. (2018). Optimal Kullback–Leibler aggregation in mixture density estimation by maximum likelihood. *Mathematical Statistics and Learning*, 1(1), 1–35. (Cited on pages 120 and 226.)

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer New York. (Cited on pages 58, 61, 266, 267, and 270.)

Dayton, C. M. & Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, 83(401), 173–178. (Cited on pages 5 and 251.)

De Branges, L. (1959). The Stone-Weierstrass theorem. *Proceedings of the American Mathematical Society*, 10(5), 822–824. (Cited on pages 58 and 266.)

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. *Recent developments in statistics*. (Cited on page 215.)

Del Moral, P., Doucet, A., & Jasra, A. (2012). An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5), 1009–1020. (Cited on pages 64 and 273.)

Deleforge, A., Forbes, F., Ba, S., & Horaud, R. (2015a). Hyper-Spectral Image Analysis With Partially Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing*, 9(6), 1037–1048. (Cited on pages 9, 64, 144, 256, and 273.)

Deleforge, A., Forbes, F., & Horaud, R. (2015b). Acoustic space learning for sound-source separation and localization on binaural manifolds. *International journal of neural systems*, 25(01), 1440003. (Cited on page 224.)

Deleforge, A., Forbes, F., & Horaud, R. (2015c). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5), 893–911. (Cited on pages 4, 5, 6, 7, 8, 9, 33, 65, 93, 95, 99, 114, 117, 122, 137, 144, 145, 146, 224, 250, 251, 252, 253, 254, 256, 273, and 275.)

Delon, J. & Desolneux, A. (2020). A Wasserstein-Type Distance in the Space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2), 936–970. (Cited on pages 68, 95, 103, 104, and 276.)

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38. (Cited on pages 51, 212, and 215.)

DeSarbo, W. S. & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2), 249–282. (Cited on pages 5 and 251.)

Devijver, E. (2015a). An l_1-oracle inequality for the Lasso in multivariate finite mixture of multivariate Gaussian regression models. *ESAIM: PS*, 19, 649–670. (Cited on pages 13, 26, 28, 52, 53, 54, 55, 56, 114, 122, 144, 160, 161, 163, 165, 167, 223, and 257.)

Devijver, E. (2015b). Finite mixture regression: a sparse variable selection by model selection for clustering. *Electronic journal of statistics*, 9(2), 2642–2674. (Cited on pages 13, 18, 24, 37, 38, 41, 46, 50, 51, 114, 144, 148, 149, 152, 162, 195, 196, 199, 202, 204, 210, 211, 212, 223, 257, 258, 263, and 283.)

Devijver, E. (2017a). Joint rank and variable selection for parsimonious estimation in a high-dimensional finite mixture regression model. *Journal of Multivariate Analysis*, 157, 1–13. (Cited on pages 13, 24, 37, 38, 46, 50, 114, 144, 162, 195, 196, 199, 202, 204, 210, 211, 257, and 283.)

Devijver, E. (2017b). Model-based regression clustering for high-dimensional data: application to functional data. *Advances in Data Analysis and Classification*, 11(2), 243–279. (Cited on pages 13, 24, 46, 50, 51, 144, 195, 196, 210, 211, 212, 257, and 283.)

Devijver, E. & Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional gaussian graphical models. *Journal of the American Statistical Association*, 113(521), 306–314. (Cited on pages 9, 13, 18, 22, 25, 37, 38, 41, 42, 43, 144, 145, 146, 148, 152, 154, 162, 195, 197, 199, 202, 204, 254, 257, 258, 263, 281, and 285.)

Devijver, E., Gallopin, M., & Perthame, E. (2017). Nonlinear network-based quantitative trait prediction from transcriptomic data. *arXiv preprint arXiv:1701.07899*. (Cited on pages 4, 5, 6, 8, 9, 22, 38, 114, 117, 143, 144, 145, 146, 147, 250, 251, 253, 254, 256, and 281.)

Devijver, E. & Perthame, E. (2020). Prediction regions through Inverse Regression. *Journal of Machine Learning Research*, 21(113), 1–24. (Cited on pages 135 and 226.)

Devroye, L. & Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer Science & Business Media. (Cited on page 58.)

Ding, P. (2016). On the Conditional Distribution of the Multivariate t Distribution. *The American Statistician*, 70(3), 293–295. (Cited on page 139.)

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. (Cited on page 225.)

Donahue, M. J., Darken, C., Gurvits, L., & Sontag, E. (1997). Rates of convex approximation in non-hilbert spaces. *Constructive Approximation*, 13(2), 187–220. (Cited on pages 60, 77, 78, and 269.)

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455. (Cited on pages 14 and 259.)

Drovandi, C., Pettitt, T., & Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, 30(1), 72–95. (Cited on page 225.)

Drovandi, C. C. & Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55, 2541–2556. (Cited on page 225.)

Duistermaat, J. J. & Kolk, J. A. (2004). *Multidimensional real analysis I: differentiation*, volume 86. Cambridge University Press. (Cited on pages 36, 141, 142, and 192.)

Eavani, H., Hsieh, M. K., An, Y., Erus, G., Beason-Held, L., Resnick, S., & Davatzikos, C. (2016). Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging. *Neuroimage*, 125, 498–514. (Cited on pages 5 and 251.)

Everitt, B. S. & Hand, D. J. (1981). Finite Mixture Distributions. *Monographs on Applied Probability and Statistics*. (Cited on pages 56 and 265.)

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348–1360. (Cited on pages 160 and 161.)

Fan, J., Zhang, C., & Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, 29(1), 153–193. (Cited on pages 20 and 279.)

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric Multivariate And Related Distributions*. Chapman and Hall. (Cited on page 135.)

Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3), 419–474. (Cited on pages 65, 93, 94, 95, 99, 224, 225, and 273.)

Feichtinger, H. G. (1977). A characterization of Wiener's algebra on locally compact groups. *Archiv der Mathematik*, 29(1), 136–140. (Cited on pages 59 and 267.)

Feller, W. (1957). *An introduction to probability theory and its applications, Vol. 1*. John Wiley. (Cited on pages 23, 195, and 283.)

Folland, G. B. (1999). *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons. (Cited on pages 78, 84, and 92.)

Forbes, F., Nguyen, H. D., Nguyen, T. T., & Arbel, J. (2021). Approximate Bayesian computation with surrogate posteriors. *Preprint hal-03139256*. (Cited on pages 5, 65, 69, 71, 101, 103, 251, 265, and 273.)

Forbes, F. & Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6), 971–984. (Cited on pages 60 and 269.)

Frahm, G. (2004). *Generalized Elliptical Distributions : Theory and Applications*. Universität zu Köln. (Cited on pages 135 and 136.)

Frazier, D. T. & Drovandi, C. (2021). Robust Approximate Bayesian Inference With Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, (pp. 1–19). (Cited on page 225.)

Frazier, D. T., Martin, G. M., Robert, C. P., & Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3), 593–607. (Cited on page 225.)

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302. (Cited on pages 25, 196, and 284.)

Friston, K. J., Preller, K. H., Mathys, C., Cagnan, H., Heinzle, J., Razi, A., & Zeidman, P. (2019). Dynamic causal modelling revisited. *NeuroImage*, 199, 730–744. (Cited on pages 25, 196, and 284.)

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Science & Business Media. (Cited on pages 57 and 265.)

Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of Mixture Analysis*. CRC Press. (Cited on pages 3, 4, 57, 249, 250, and 265.)

Genovese, C. R. & Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4), 1105–1127. (Cited on pages 3, 34, 38, 42, 140, 154, 199, 200, 204, and 249.)

Geyer, C. J. & Jonhson, L. T. (2020). *mcmc: Markov chain Monte Carlo*. R package version 0.9-7. (Cited on page 103.)

Ghosal, S. & van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5), 1233–1263. (Cited on page 200.)

Golub, G. H. & Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press. (Cited on page 191.)

Golub, G. H. & Van Loan, C. F. (2013). *Matrix computations*, volume 3. JHU press. (Cited on pages 24, 195, and 283.)

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439), 531–537. (Cited on pages 9, 144, and 256.)

Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, 28(2), 411–425. (Cited on pages 65, 68, 99, 273, and 276.)

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., & Others (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer. (Cited on page 224.)

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press. (Cited on pages 24, 196, and 284.)

He, H., Boyd-Graber, J., Kwok, K., & Daumé III, H. (2016). Opponent Modeling in Deep Reinforcement Learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research* (pp. 1804–1813). New York, New York, USA: PMLR. (Cited on pages 5 and 251.)

Hennig, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. *Journal of Classification*, 17(2), 273–296. (Cited on pages 20, 211, and 278.)

Ho, N. & Nguyen, X. (2016). Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions. *arXiv preprint arXiv:1609.02655*. (Cited on page 224.)

Ho, N. & Nguyen, X. (2019). Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions. *SIAM Journal on Mathematics of Data Science*, 1(4), 730–758. (Cited on page 224.)

Ho, N., Nguyen, X., et al. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6), 2726–2755. (Cited on pages 3 and 249.)

Ho, N., Nguyen, X., et al. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1), 271–307. (Cited on pages 3 and 249.)

Ho, N., Yang, C.-Y., & Jordan, M. I. (2019). Convergence rates for gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*. (Cited on pages 3, 13, 147, 198, 224, 249, and 258.)

Horn, R. A. & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press. (Cited on page 193.)

Hospedales, T. M. & Vijayakumar, S. (2008). Structure inference for Bayesian multisensory scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12), 2140–2157. (Cited on page 102.)

Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T., & Wilinska, M. E. (2004). Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4), 905–920. (Cited on pages 64 and 272.)

Hult, H. & Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3), 587–608. (Cited on pages 135 and 136.)

Hunter, D. R. & Lange, K. (2000). Quantile Regression via an MM Algorithm. *Journal of Computational and Graphical Statistics*, 9(1), 60–77. (Cited on page 215.)

Hunter, D. R. & Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician*, 58(1), 30–37. (Cited on page 215.)

Huynh, B. T. & Chamroukhi, F. (2019). Estimation and feature selection in mixtures of generalized linear experts models. *arXiv preprint arXiv:1907.06994*. (Cited on pages 51, 52, 162, 198, 211, 212, and 214.)

Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, 71, 159–182. (Cited on page 224.)

Ingrassia, S., Minotti, S. C., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*, 29(3), 363–401. (Cited on pages 5, 93, 114, 137, 224, and 251.)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1), 79–87. (Cited on pages 1, 3, 5, 11, 62, 115, 145, 163, 194, 248, 250, 251, 256, and 271.)

Janková, J. & van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1), 1205–1229. (Cited on page 226.)

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1), 2869–2909. (Cited on page 226.)

Jiang, B., Wu, T.-Y., C., Z., & Wong, W. H. (2017). Learning summary statistics for Approximate Bayesian Computation via Deep Neural Network. *Statistica Sinica*, (pp. 1595–1618). (Cited on pages 65, 94, 99, and 273.)

Jiang, B., Wu, T.-Y., & Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. (Cited on pages 68, 100, 225, and 276.)

Jiang, W. & Tanner, M. A. (1999a). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, (pp. 987–1011). (Cited on pages 3, 13, 63, 85, 147, 198, 227, 249, 258, and 271.)

Jiang, W. & Tanner, M. A. (1999b). On the Approximation Rate of Hierarchical Mixtures-of-Experts for Generalized Linear Models. *Neural Computation*, 11(5), 1183–1198. (Cited on pages 63, 85, 86, 92, 223, and 271.)

Jiang, W. & Tanner, M. A. (1999c). On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9), 1253–1258. (Cited on pages 20, 211, and 278.)

Jones, L. K. (1992). A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *The Annals of Statistics*, 20(1), 608–613. (Cited on page 77.)

Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181–214. (Cited on pages 1, 51, 163, 212, and 248.)

Kalliovirta, L., Meitz, M., & Saikkonen, P. (2016). Gaussian mixture vector autoregression. *Journal of Econometrics*, 192(2), 485–498. (Cited on page 224.)

Kelker, D. (1970). Distribution Theory of Spherical Distributions and a Location-Scale Parameter Generalization. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 32(4), 419–430. (Cited on page 136.)

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4), 519–539. (Cited on pages 28, 46, 50, 51, 54, 55, 147, 161, 165, 198, 210, 211, and 212.)

Khalili, A. & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479), 1025–1038. (Cited on pages 55, 160, and 161.)

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114. Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. (Cited on pages 13, 31, and 258.)

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media. (Cited on pages 30 and 130.)

Kotz, S. & Nadarajah, S. (2004). *Multivariate T-Distributions and Their Applications*. Cambridge: Cambridge University Press. (Cited on page 139.)

Kristan, M., Leonardis, A., & Skočaj, D. (2011). Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recognition*, 44(10-11), 2630–2642. (Cited on pages 68 and 276.)

Kruse, J., Ardizzone, L., Rother, C., & Kothe, U. (2021). Benchmarking Invertible Architectures on Inverse Problems. *arXiv preprint arXiv:2101.10763*. (Cited on page 225.)

Krzyzak, A. & Schafer, D. (2005). Nonparametric regression estimation by normalized radial basis function networks. *IEEE Transactions on Information Theory*, 51(3), 1003–1010. (Cited on pages 63, 85, and 271.)

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. (Cited on pages 57 and 266.)

Lange, K. (2016). *MM Optimization Algorithms*. Philadelphia, PA: Society for Industrial and Applied Mathematics. (Cited on page 215.)

Lathuilière, S., Juge, R., Mesejo, P., Muñoz-Salinas, R., & Horaud, R. (2017). Deep mixture of linear inverse regressions applied to head-pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4817–4825). (Cited on pages 5, 114, and 251.)

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), Article 35. (Cited on pages 9, 144, and 256.)

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), 907–927. (Cited on page 226.)

Lee, J. D., Sun, Y., & Saunders, M. A. (2014). Proximal Newton-Type Methods for Minimizing Composite Functions. *SIAM Journal on Optimization*, 24(3), 1420–1443. (Cited on page 220.)

Lee, S.-I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient L_1 regularized logistic regression. In *AAAI*, volume 6 (pp. 401–408). (Cited on page 214.)

Lee, S. X. & McLachlan, G. J. (2016). Finite mixtures of canonical fundamental skew

t

-distributions. *Statistics and Computing*, 26(3), 573–589. (Cited on pages 60 and 269.)

Lee, Y.-S. & Cho, S.-B. (2014). Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data. *Neurocomputing*, 126, 106–115. (Cited on pages 5 and 251.)

Lemasson, B., Pannetier, N., Coquery, N., Boisserand, L. S. B., Collomb, N., Schuff, N., Moseley, M., Zaharchuk, G., Barbier, E. L., & Christen, T. (2016). MR Vascular Fingerprinting in Stroke and Brain Tumors Models. *Scientific Reports*, 6, 37071. (Cited on pages 64 and 273.)

Li, F., Villani, M., & Kohn, R. (2011). Modelling Conditional Densities Using Finite Smooth Mixtures. In *Mixtures: Estimation and Applications* chapter 6, (pp. 123–144). Wiley Online Library. (Cited on pages 5 and 251.)

Li, J. & Barron, A. (1999). Mixture Density Estimation. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in Neural Information Processing Systems*, volume 12: MIT Press. (Cited on pages 57 and 223.)

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414), 316–327. (Cited on pages 5, 93, 143, and 251.)

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. (Cited on pages 56 and 265.)

Lloyd-Jones, L. R., Nguyen, H. D., & McLachlan, G. J. (2018). A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics & Data Analysis*, 119, 19–38. (Cited on page 160.)

Lounici, K., Pontil, M., van de Geer, S., & Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4), 2164–2204. (Cited on page 211.)

Ma, D., Gulani, V., Seiberlich, N., Liu, K., Sunshine, J. L., Duerk, J. L., & Griswold, M. A. (2013). Magnetic Resonance Fingerprinting. *Nature*, 495(7440), 187–192. (Cited on pages 64 and 272.)

Magnus, J. R. & Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons. (Cited on page 190.)

Maillard, G. (2020). *Hold-out and Aggregated hold-out*. PhD thesis, Université Paris-Saclay. (Cited on pages 15, 226, and 260.)

Makarov, B. & Podkorytov, A. (2013). *Real Analysis: Measures, Integrals and Applications*. Springer Nature. (Cited on pages 75, 78, 82, 107, and 108.)

Mallows, C. L. (1973). Some Comments on CP. *Technometrics*, 15(4), 661–675. (Cited on pages 12, 16, 144, 257, and 261.)

Mansuripur, M. (1987). *Introduction to information theory*. Prentice-Hall, Inc. (Cited on pages 28 and 165.)

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computation methods. *Statistics and Computing*, 22, 1167–1180. (Cited on pages 100 and 101.)

Masoudnia, S. & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2), 275–293. (Cited on pages 2 and 248.)

Massart, P. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer. (Cited on pages 12, 14, 15, 18, 19, 30, 33, 37, 38, 53, 54, 114, 119, 130, 144, 148, 162, 167, 179, 199, 257, 259, 263, and 278.)

Massart, P. & Meynet, C. (2011). The Lasso as an l_1-ball model selection procedure. *Electronic Journal of Statistics*, 5, 669–687. (Cited on pages 12, 114, 144, 160, and 257.)

Maugis, C. & Michel, B. (2011a). Data-driven penalty calibration: A case study for Gaussian mixture model selection. *ESAIM: PS*, 15, 320–339. (Cited on pages 13, 144, and 257.)

Maugis, C. & Michel, B. (2011b). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15, 41–68. (Cited on pages 13, 33, 38, 41, 42, 44, 45, 122, 144, 152, 154, 155, 156, 199, 202, 204, 207, 208, and 257.)

Maugis-Rabusseau, C. & Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM: Probability and Statistics*, 17, 698–724. (Cited on pages 120, 167, and 226.)

Mazumder, R. & Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, 13(1), 781–794. (Cited on pages 26, 197, and 285.)

McLachlan, G. J. & Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, volume 38. Marcel Dekker. (Cited on pages 56 and 265.)

McLachlan, G. J. & Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley. (Cited on pages 51 and 212.)

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons. (Cited on pages 2, 57, 122, 248, and 265.)

Meinshausen, N. (2015). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 77(5), 923–945. (Cited on page 226.)

Mendes, E. F. & Jiang, W. (2012). On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11), 3025–3051. (Cited on pages 13, 63, 85, 147, 195, 198, 227, 258, and 271.)

Mengersen, K. L., Robert, C., & Titterington, M. (2011). *Mixtures: Estimation and Applications*, volume 896. John Wiley & Sons. (Cited on pages 57 and 265.)

Mesejo, P., Saillet, S., David, O., Bénar, C., Warnking, J. M., & Forbes, F. (2016). A Differential Evolution-Based Approach for Fitting a Nonlinear Biophysical Model to fMRI BOLD Data. *IEEE Journal of Selected Topics in Signal Processing*, 10(2), 416–427. (Cited on pages 64 and 272.)

Meynet, C. (2013). An l_1-oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM: Probability and Statistics*, 17, 650–671. (Cited on pages 13, 26, 28, 52, 53, 54, 55, 56, 114, 122, 144, 160, 161, 163, 165, 167, 180, and 257.)

Moerland, P. (1997). *Some methods for training mixtures of experts*. Technical report, IDIAP Research Institute. (Cited on page 214.)

Moerland, P. (1999). Classification using localized mixture of experts. In *Ninth International Conference on Artificial Neural Networks*, volume 2 (pp. 838–843). (Cited on pages 5, 114, and 251.)

Montuelle, L., Le Pennec, E., et al. (2014). Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1), 1661–1695. (Cited on pages 11, 13, 19, 21, 23, 26, 30, 31, 32, 33, 38, 39, 40, 42, 43, 55, 56, 114, 115, 116, 120, 122, 123, 129, 130, 133, 134, 135, 139, 142, 143, 144, 147, 149, 150, 152, 154, 155, 161, 162, 163, 194, 195, 198, 199, 200, 201, 202, 223, 256, 257, 258, 278, 280, and 282.)

Muandet, K., Fukumizu, K., Dinuzzo, F., & Scholkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems* (pp. 10–18). (Cited on pages 68 and 276.)

Nataraj, G., Nielsen, J.-F., Scott, C., & Fessler, J. A. (2018). Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels. *IEEE Trans. Med. Imaging*, 37(9), 2103–2114. (Cited on page 93.)

Neath, A. A. & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, 4(2), 199–203. (Cited on pages 16 and 261.)

Nestoridis, V. & Papadimitropoulos, C. (2005). Abstract theory of universal series and an application to Dirichlet series. *Comptes Rendus Mathematique*, 341(9), 539–543. (Cited on pages 58 and 267.)

Nestoridis, V., Schmutzhard, S., & Stefanopoulos, V. (2011). Universal series induced by approximate identities and some relevant applications. *Journal of Approximation Theory*, 163(12), 1783–1797. (Cited on pages 58, 59, 60, 74, 77, 267, 268, and 269.)

Nestoridis, V. & Stefanopoulos, V. (2007). *Universal series and approximate identities*. Technical report, Technical Report TR-28-2007, Department of Mathematics and Statistics. (Cited on pages 58, 59, 60, 74, 267, 268, and 269.)

Nguyen, D. V. & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1), 39–50. (Cited on pages 9, 144, and 256.)

Nguyen, H. D. (2017). An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery*, 7(2), e1198. (Cited on page 215.)

Nguyen, H. D., Arbel, J., Lü, H., & Forbes, F. (2020a). Approximate Bayesian Computation Via the Energy Statistic. *IEEE Access*, 8, 131683–131698. (Cited on pages 68, 94, 100, and 276.)

Nguyen, H. D. & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1246. (Cited on pages 2, 161, 163, and 248.)

Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366, 208–214. (Cited on pages 3, 5, 56, 63, 85, 93, 114, 120, 249, 251, and 271.)

Nguyen, H. D., Lloyd-Jones, L. R., & McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12), 2585–2593. (Cited on pages 3, 13, 63, 85, 147, 198, 249, 258, and 271.)

Nguyen, H. D. & McLachlan, G. (2019). On approximations via convolution-defined mixture models. *Communications in Statistics-Theory and Methods*, 48(16), 3945–3955. (Cited on pages 57, 265, and 266.)

Nguyen, H. D. & McLachlan, G. J. (2014). Asymptotic inference for hidden process regression models. In *2014 IEEE Workshop on Statistical Signal Processing (SSP)* (pp. 256–259).: IEEE. (Cited on pages 5 and 251.)

Nguyen, H. D. & McLachlan, G. J. (2016). Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93, 177–191. (Cited on page 226.)

Nguyen, H. D., Nguyen, T., Chamroukhi, F., & McLachlan, G. J. (2021a). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1), 13. (Cited on pages 3, 13, 56, 69, 71, 93, 120, 147, 198, 249, 258, and 264.)

Nguyen, T., Chamroukhi, F., Nguyen, H. D., & McLachlan, G. J. (2020b). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *arXiv preprint arXiv:2008.09787. To appear. Communications in Statistics - Theory and Methods.* (Cited on pages 3, 56, 61, 63, 69, 71, 79, 86, 93, 223, 249, 264, 270, and 271.)

Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2020c). An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. *arXiv preprint arXiv:2009.10622.* (Cited on pages 13, 28, 29, 70, 114, 121, 122, 144, 147, 159, 166, 198, 223, 257, and 282.)

Nguyen, T. T., Chamroukhi, F., Nguyen, H. D., & Forbes, F. (2021b). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *arXiv preprint arXiv:2104.08959.* (Cited on pages 3, 9, 13, 22, 26, 69, 95, 113, 114, 198, 250, 254, 257, 277, and 281.)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2021c). A non-asymptotic penalization criterion for model selection in mixture of experts models. *arXiv preprint arXiv:2104.02640.* (Cited on pages 3, 5, 13, 21, 26, 38, 69, 95, 113, 114, 115, 144, 145, 147, 148, 149, 150, 151, 198, 250, 251, 257, 258, 277, 280, and 282.)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2020d). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1), 1750861. (Cited on pages 3, 56, 60, 63, 68, 71, 72, 223, 249, 264, 269, and 271.)

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1), 370–400. (Cited on pages 3, 224, and 249.)

Norets, A. et al. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of statistics*, 38(3), 1733–1766. (Cited on pages 3, 63, 85, 223, 249, and 271.)

Norets, A. & Pati, D. (2017). Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4), 980–1012. (Cited on pages 5, 114, and 251.)

Norets, A. & Pelenis, J. (2014). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, 30(3), 606–646. (Cited on pages 5, 63, 85, 114, 251, and 271.)

Nunes, M. A. & Prangle, D. (2015). *abctools: An R package for tuning Approximate Bayesian Computation analyses.* R package version 1.1.3. (Cited on page 99.)

Oden, J. T. & Demkowicz, L. (2010). *Applied Functional Analysis.* CRC Press. (Cited on page 88.)

Ong, V., Nott, D., Tran, M.-N., Sisson, S., & Drovandi, C. (2018). Likelihood-free inference in high dimensions with synthetic likelihood. *Computational Statistics and Data Analysis*, 128. (Cited on page 225.)

Park, M., Jitkrittum, W., & Sejdinovic, D. (2016). K2-ABC: approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. (Cited on pages 68 and 276.)

Park, M. Y. & Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1), 30–50. (Cited on pages 51, 161, and 212.)

Peel, D. & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4), 339–348. (Cited on pages 60 and 269.)

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, 178, 624–638. (Cited on page 223.)

Peralta, B. & Soto, A. (2014). Embedded local feature selection within mixture of experts. *Information Sciences*, 269, 176–187. (Cited on page 214.)

Perthame, E., Forbes, F., & Deleforge, A. (2018). Inverse regression approach to robust nonlinear high-to-low dimensional mapping. *Journal of Multivariate Analysis*, 163, 1–14. (Cited on pages 6, 8, 117, 139, 144, 224, 226, 251, and 253.)

Perthame, E., Forbes, F., Deleforge, A., Devijver, E., & Gallopin, M. (2017). *xLLiM: An R package for High Dimensional Locally-Linear Mapping*. R package version 2.2. (Cited on page 99.)

Prangle, D., Everitt, R. G., & Kypraios, T. (2018). A rare event approach to high-dimensional approximate Bayesian computation. *Statistics and Computing*, 28, 819–834. (Cited on pages 66, 96, and 274.)

Price, L. F., Drovandi, C. C., Lee, A., & Nott, D. J. (2018). Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, 27(1), 1–11. (Cited on page 225.)

Quandt, R. E. (1972). A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, 67(338), 306–310. (Cited on pages 5 and 251.)

R Core Team (2020). R: A language and environment for statistical computing. *Vienna, Austria*. (Cited on page 122.)

Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: PS*, 9, 220–229. (Cited on pages 3, 57, 99, 108, 109, 223, and 249.)

Ramamurti, V. & Ghosh, J. (1996). Structural adaptation in mixture of experts. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 4 (pp. 704–708 vol.4). (Cited on pages 5, 114, and 251.)

Ramamurti, V. & Ghosh, J. (1998). Use of localized gating in mixture of experts networks. In *Proc.SPIE*, volume 3390. (Cited on pages 5, 114, and 251.)

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2), 195–239. (Cited on page 122.)

Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2), 639–665. (Cited on pages 120 and 226.)

Rubio, F. J. & Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7, 1632–1654. (Cited on pages 66, 96, and 274.)

Rudin, W. (1976). *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York. (Cited on pages 75 and 80.)

Rudin, W. (1987). *Real and complex analysis*. New York: McGraw-Hill. (Cited on page 84.)

Samé, A., Chamroukhi, F., Govaert, G., & Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5(4), 301–321. (Cited on pages 5 and 251.)

Sandberg, I. W. (2001). Gaussian radial basis functions and inner product spaces. *Circuits, Systems and Signal Processing*, 20(6), 635–642. (Cited on pages 58 and 266.)

Sato, M. & Ishii, S. (2000). On-line EM algorithm for the normalized gaussian network. *Neural computation*, 12(2), 407–432. (Cited on pages 5, 114, and 251.)

Saumard, A. (2010). Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. *hal-00512310*. (Cited on page 227.)

Saumard, A. (2019). Discussion on "Minimal penalties and the slope heuristic: a survey" by Sylvain Arlot. *Journal de la société française de statistique*, 160(3), 154–157. (Cited on page 227.)

Saumard, A. & Navarro, F. (2021). Finite Sample Improvement of Akaike's Information Criterion. *IEEE Transactions on Information Theory*, (pp.1̃). (Cited on page 227.)

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*. Springer. (Cited on pages 57 and 265.)

Schmidt, F. & Fernando, J. (2015). Realistic Uncertainties on Hapke Model Parameters from Photometric Measurements. *Icarus*, 260, 73–93 (IF 2,84). (Cited on pages 64 and 272.)

Schuhmacher, D., Bahre, B., Gottschlich, C., Hartmann, V., Heinemann, F., & Schmitzer, B. (2020). *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. R package version 0.12-2. (Cited on page 104.)

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464. (Cited on pages 12, 16, 54, 114, 144, 257, and 261.)

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538 ICLR 2017*. (Cited on pages 5 and 251.)

Shi, Y., Siddharth, N., Paige, B., & Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems* (pp. 15718–15729). (Cited on pages 5 and 251.)

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231–245. (Cited on pages 24, 196, and 284.)

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press. (Cited on pages 64 and 272.)

Soubeyrand, S., Carpentier, F., Guiton, F., & Klein, E. K. (2013). Approximate Bayesian computation with functional statistics. *Statistical Applications in Genetics and Molecular Biology*, 12(1), 17–37. (Cited on pages 68 and 277.)

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Scholkopf, B., & Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11, 1517–1561. (Cited on pages 68 and 276.)

Stadler, N., Buhlmann, P., & van de Geer, S. (2010). l_1-penalization for mixture regression models. *TEST*, 19, 209–256. (Cited on pages 46, 50, 55, 122, 160, 210, and 211.)

Stone, M. H. (1948). The Generalized Weierstrass Approximation Theorem. *Mathematics Magazine*, 21(4), 167–184. (Cited on pages 58 and 266.)

Strang, G. (2019). *Linear algebra and learning from data.* Wellesley-Cambridge Press Cambridge. (Cited on pages 25, 49, 197, 207, and 284.)

Stucky, B. & van de Geer, S. (2018). Asymptotic Confidence Regions for High-Dimensional Structured Sparsity. *IEEE Transactions on Signal Processing*, 66(8), 2178–2190. (Cited on page 226.)

Tao, T. (2011). *An Introduction to Measure Theory*, volume 126. American Mathematical Society Providence, RI. (Cited on page 108.)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. (Cited on pages 13, 24, 114, 160, 196, 257, and 283.)

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley. (Cited on pages 2, 56, 57, 248, and 265.)

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. *LIDS-P-1840. Technical Report.* (Cited on pages 218 and 222.)

Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 109(3), 475–494. (Cited on pages 218 and 222.)

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Science & Business Media, 1st edition. (Cited on page 108.)

Tu, C.-C., Forbes, F., Lemasson, B., & Wang, N. (2019). Prediction with high dimensional regression via hierarchically structured Gaussian mixtures and latent variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5), 1485–1507. (Cited on pages 5 and 251.)

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press. (Cited on pages 30 and 130.)

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202. (Cited on page 226.)

Van Der Vaart, A. & Wellner, J. (1996). Weak convergence and empirical processes: With applications to statistics springer series in statistics. *Springer*, 58, 59. (Cited on pages 30, 130, and 179.)

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data (Springer Series in Statistics).* Springer-Verlag. (Cited on pages 53, 54, and 167.)

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press. (Cited on pages 15, 162, 191, and 259.)

Walker, J. L. & Ben-Akiva, M. (2011). Advances in Discrete Choice: Mixture Models. In *A Handbook of Transport Economics* chapter 8. Edward Elgar Publishing. (Cited on pages 57 and 265.)

Wang, D. & Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley-IEEE Press. (Cited on page 102.)

Wang, F., Syeda-Mahmood, T., Vemuri, B. C., Beymer, D., & Rangarajan, A. (2009). Closed-form Jensen-Renyi divergence for mixture of Gaussians and applications to group-wise shape registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 648–655).: Springer. (Cited on pages 68 and 276.)

Wang, L. X. & Mendel, J. M. (1992). Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE Transactions on Neural Networks*, 3(5), 807–814. (Cited on pages 63, 85, and 271.)

Wang, P., Puterman, M. L., Cockburn, I., & Le, N. (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*, 52(2), 381–400. (Cited on pages 5 and 251.)

Wang, X., Yu, F., Dunlap, L., Ma, Y.-A., Wang, R., Mirhoseini, A., Darrell, T., & Gonzalez, J. E. (2020). Deep mixture of experts via shallow embedding. In *Uncertainty in Artificial Intelligence* (pp. 552–562).: PMLR. (Cited on pages 5 and 251.)

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1), 1–25. (Cited on pages 20, 128, and 278.)

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62. (Cited on pages 20 and 279.)

Williams, C. K. & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA. (Cited on page 193.)

Wiqvist, S., Mattei, P.-A., Picchini, U., & Frellsen, J. (2019). Partially Exchangeable Networks and Architectures for Learning Summary Statistics in Approximate Bayesian Computation. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 (pp. 6798–6807). Long Beach, California, USA. (Cited on pages 65, 94, and 273.)

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310), 1102–1104. (Cited on page 225.)

Xu, L., Jordan, M., & Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in Neural Information Processing Systems*, volume 7: MIT Press. (Cited on pages 3, 62, 114, 115, 145, 250, and 271.)

Xu, Y., Light, W. A., & Cheney, E. W. (1993). Constructive methods of approximation by ridge functions and radial functions. *Numerical Algorithms*, 4(2), 205–223. (Cited on pages 58, 60, 267, and 269.)

Yona, G. (2010). *Introduction to Computational Proteomics*. CRC Press. (Cited on pages 57 and 265.)

Young, D. S. (2014). Mixtures of regressions with changepoints. *Statistics and Computing*, 24(2), 265–281. (Cited on page 129.)

Yu, Q., Kavitha, M. S., & Kurita, T. (2021). Mixture of experts with convolutional and variational autoencoders for anomaly detection. *Applied Intelligence*, 51(6), 3241–3254. (Cited on pages 5 and 251.)

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67. (Cited on pages 24, 196, and 283.)

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8), 1177–1193. (Cited on pages 2 and 248.)

Zeevi, A. J. & Meir, R. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks*, 10(1), 99–109. (Cited on pages 57, 63, 85, 106, 223, and 271.)

Zeevi, A. J., Meir, R., & Maiorov, V. (1998). Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory*, 44(3), 1010–1025. (Cited on pages 63, 85, and 271.)

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 217–242. (Cited on page 226.)

Résumé long en français

Résumé

Les modèles de mélanges d'experts (MoE) sont omniprésents dans l'analyse de données hétérogènes dans de nombreux domaines, notamment en statistique, bioinformatique, reconnaissance des formes, économie et médecine, entre autres. Ils fournissent des constructions conditionnelles pour la régression dans lesquelles les poids du mélange, ainsi que les densités de ses composants, sont expliqués par les prédicteurs. Ceci qui permet une meilleur flexibilité dans la modélisation de données provenant de processus générateurs complexes. Dans cette thèse, nous étudions les capacités d'approximation et les propriétés d'estimation et de sélection de modèle d'un large éventail de distributions mélange, avec un accent particulier sur une riche famille de modèles MoE dans un cadre de grande dimension; Cela inclut les modèles MoE avec des experts gaussiens et des poids de mélanges (appelés *gating network*) modélisés par des fonctions softmax ou gaussiennes normalisées, qui sont les choix les plus populaires et sont des outils puissants pour modéliser des relations non linéaires complexes entre les réponses et les prédicteurs qui proviennent de différentes sous-populations. Nous considérons à la fois les aspects théoriques, statistiques et méthodologiques, et les outils numériques, liés à la conception de ces modèles, ainsi qu'à leur estimation à partir de données et à la sélection du meilleur modèle.

Plus précisément, dans cette thèse, nous passons d'abord en revue les propriétés d'approximation universelles des mélanges de densités classiques afin de préparer le cadre théorique et de clarifier certaines affirmations vagues et peu claires dans la littérature, avant de les considérer dans le contexte des modèles MoE. En particulier, nous prouvons que, à un degré de précision arbitraire, les mélanges de translatées-dilatées d'une fonction de densité de probabilité (FDP) continue peuvent approximer toute FDP continue, uniformément, sur un ensemble compact; et les mélanges de translatées dilatées d'une FDP essentiellement bornée peuvent approximer toute FDP dans les espaces de Lebesgue. Ensuite, après avoir apporté des améliorations aux résultats d'approximation dans le contexte des mélanges inconditionnels, nous étudions les capacités d'approximation universelles des modèles MoE dans une variété de contextes, y compris en approximation de densité conditionnelle et en calcul bayésien approximatif (ABC). Étant donné des variables d'entrée et de sortie toutes deux à support compact, nous prouvons que les MoE pour les FDP conditionnelles sont denses dans les espaces de Lebesgue. Ensuite, nous prouvons que la distribution quasi-postérieure résultant de l'ABC avec des postérieurs de substitution construits à partir de mélanges gaussiens finis en utilisant une approche de régression inverse, converge vers la vraie distribution, dans des conditions standard. Enfin, nous nous concentrons sur les prédicteurs et les réponses de grande dimension. Par la suite, nous établissons des résultats non asymptotiques de sélection de modèle dans des scénarios de régression à grande dimension, pour une variété de modèles de régression MoE, y compris GLoME et SGaME, en s'appuyant sur une stratégie de régression inverse ou une pénalisation Lasso, respectivement. Ceux-ci incluent des résultats pour la sélection du nombre de composantes du mélange d'experts, ainsi que pour la sélection jointe de variable et des rangs des matrices de covariances. En particulier, ces résultats fournissent garanties théoriques fortes: une inégalité d'oracle en échantillon fini satisfaite par l'estimateur de maximum de vraisemblance pénalisé avec une perte de type Jensen-Kullback-Leibler et une justification théorique de la forme de la pénalité pour utiliser l'heuristique de pente, par rapport aux critères asymptotiques classiques. Cela permet de calibrer les fonctions de pénalité, connues seulement à une constante multiplicative près, étant donnés complexité de la (sous-)collection aléatoire considérée de modèles MoE, y compris le nombre de composantes du mélange, le degré de sparsité (les coefficients et les niveaux de sparsité des rangs des matrices de covariances), le degré des fonctions moyennes polynomiales, et

les structures potentielles de diagonales par bloc cachées des matrices de covariance du prédicteur ou de la réponse multivariée. Enfin, pour étayer nos résultats théoriques et l'étude statistique de la sélection non asymptotique de modèles dans une variété de modèles MoE, nous réalisons des études numériques en considérant des données simulées et réelles, qui mettent en évidence la performance de nos résultats, y compris celles d'inégalités d'oracle à échantillon fini.

Mots-clés

Mélange d'experts; modèles de mélange; approximation universelle; maximum de vraisemblance pénalisé; sélection de variables, sélection de modèle non asymptotique; statistiques à grande dimension; Lasso; régularisation l_1; distance de Wasserstein; algorithme EM; algorithme MM; proximal-Newton; clustering; classification; régression inverse; calcul bayésien approximatif.

Contexte scientifique

Les modèles de mélange d'experts (MoE), initialement introduits dans Jacobs et al. (1991) et Jordan & Jacobs (1994), sont des modèles flexibles qui généralisent les modèles classiques de mélange fini ainsi que les modèles de mélange fini pour la régression (McLachlan & Peel, 2000, Section 5.13), et sont largement utilisés en statistique et en apprentissage automatique, grâce à leur flexibilité et à l'abondance d'outils d'estimation statistique et de sélection de modèles applicables. Leur flexibilité provient du fait que les poids des mélanges (appelés gating network) peuvent dépendre des variables explicatives, ainsi que des densités des composants (ou des experts). Cela permet de modéliser des données issues de processus de génération de données plus complexes que les mélanges finis classiques et les modèles de mélange fini pour la régression, dont les paramètres de mélange sont indépendants des covariables. En raison de leur flexibilité, les MoE peuvent être utilisés dans de nombreux problèmes statistiques, notamment pour regrouper ou classer des données, pour estimer des densités conditionnelles, pour effectuer des analyses de régression et pour analyser les résultats des régressions. Des examens détaillés des aspects pratiques et théoriques des modèles MoE sont disponibles dans Yuksel et al. (2012), Masoudnia & Ebrahimpour (2014) et Nguyen & Chamroukhi (2018).

Sur le plan statistique, nous considérons un cadre de régression et visons à capturer la relation non linéaire potentielle entre une réponse multivariée $\mathbf{Y}$ et un vecteur de covariables $\mathbf{X}$. Ici et par la suite, $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := \left(\mathbf{X}_i, \mathbf{Y}_i\right)_{i\in[n]}$, $[n] = \{1, \ldots, n\}, n \in \mathbb{N}^\star$, désigne un échantillon aléatoire, et $\mathbf{x}$ et $\mathbf{y}$ représentent les valeurs observées des variables aléatoires $\mathbf{X}$ et $\mathbf{Y}$, respectivement. Nous supposons que la variable de réponse $\mathbf{Y}$ dépend de la variable explicative $\mathbf{X}$ à travers un modèle de type régression. La variable explicative porte différents noms, tels que covariable, prédicteur, variable indépendante, caractéristique, ou parfois simplement variable. La variable de réponse est souvent appelée variable de sortie ou variable dépendante. Tout au long de cette thèse, nous utiliserons tous ces termes de manière interchangeable.

Tout d'abord, nous nous concentrons sur un point de départ naturel pour mettre en place un modèle de MoE plus probabiliste, souvent appelé "application directe" de la modélisation des mélanges; voir Titterington et al., 1985 pour plus de détails. Ensuite, nous préciserons une autre perspective sur la modélisation des mélanges dans (5.4.4) et (5.4.15) qui fournissent une vue "analytique", complémentaire à cette vue "synthétique". En ce qui concerne la première perspective, supposons que la population à partir de laquelle nous échantillonnons est hétérogène, c'est-à-dire qu'étant donné une entrée $\mathbf{X} = \mathbf{x}$, il existe de multiples groupes (qui peuvent être interprétés comme des clusters), indexés par $k \in [K]$, présents dans la population dans des proportions $g_k\left(\mathbf{x}; \boldsymbol{\omega}\right), k \in [K]$, où $\boldsymbol{\omega}$ est un vecteur de paramètres définissant la fonction de proportion dépendant de l'entrée $g_k\left(\mathbf{x}; \cdot\right)$, avec $g_k\left(\mathbf{x}; \boldsymbol{\omega}\right) > 0$, et $\sum_{k=1}^{K} g_k\left(\mathbf{x}; \boldsymbol{\omega}\right) = 1$. Ainsi, il existe une représentation de variable aléatoire latente non observée du modèle de mélange (généralement appelée *latent* ou *variables d'allocation*), impliquant l'appartenance latente de chaque observation à un groupe, désignée par $Z \in [K], K \in \mathbb{N}^\star$, où $Z = k$ si l'observation $\mathbf{y}$ étant donné l'entrée $\mathbf{x}$ appartient au groupe k pour $k \in [K]$. Ensuite, la relation conditionnelle entre Z et l'entrée $\mathbf{X}$ peut être caractérisée par

$$p\left(Z = k | \mathbf{X} = \mathbf{x}\right) = g_k\left(\mathbf{x}; \boldsymbol{\omega}\right). \tag{5.4.1}$$

Alors, nous pouvons caractériser la relation entre la sortie $\mathbf{y}$ et l'entrée $\mathbf{X}$ par

$$p\left(\mathbf{y}|\mathbf{X}=\mathbf{x}, Z=k\right)=\phi_k\left(\mathbf{y};\boldsymbol{\theta}_k\left(\mathbf{x}\right)\right), \tag{5.4.2}$$

où $\boldsymbol{\theta}_k\left(\mathbf{x}\right)$ est un vecteur de paramètres et, étant donné une entrée $\mathbf{x}$, $\phi_k\left(\cdot;\boldsymbol{\theta}_k\left(\mathbf{x}\right)\right)$ est une fonction de densité de probabilité (FDP). Nous pouvons imaginer qu'étant donné une entrée $\mathbf{X}=\mathbf{x}$, la sortie observée $\mathbf{y}$, tirée de la population, est générée en deux étapes: premièrement, le groupe Z est tiré d'une distribution multinomiale avec un seul essai et des probabilités égales à $g_k\left(\mathbf{x};\boldsymbol{\omega}\right)$; et deuxièmement, étant donné $\mathbf{X}=\mathbf{x}, Z=k$, la sortie $\mathbf{y}$ est tirée de $\phi_k\left(\mathbf{y};\boldsymbol{\theta}_k\left(\mathbf{x}\right)\right)$.

Notez que cet échantillonnage en deux étapes donne exactement les mêmes modèles en vue analytique, voir (5.4.4) et (5.4.15) pour plus de détails, pour la distribution conditionnelle de $\mathbf{y}|\mathbf{x}$. En effet, via les caractérisations (5.4.1) et (5.4.2), et en utilisant la loi de probabilité totale, nous pouvons caractériser la relation marginale entre la réponse et l'entrée, inconditionnelle sur Z, via l'expression

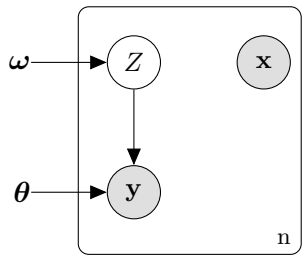
$$\begin{aligned} p\left(\mathbf{y}|\mathbf{X}=\mathbf{x}\right) &= \sum_{k=1}^{K} p\left(\mathbf{y}|\mathbf{X}=\mathbf{x}, Z=k\right) p\left(Z=k|\mathbf{X}=\mathbf{x}\right) \\ &= \sum_{k=1}^{K} g_k\left(\mathbf{x};\boldsymbol{\omega}\right)\phi_k\left(\mathbf{y};\boldsymbol{\theta}_k\left(\mathbf{x}\right)\right) =: s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}), \end{aligned} \tag{5.4.3}$$

où $\boldsymbol{\psi}=(\boldsymbol{\omega},\boldsymbol{\theta})$ est le vecteur de tous les éléments de paramètre qui sont nécessaires pour caractériser (5.4.3).

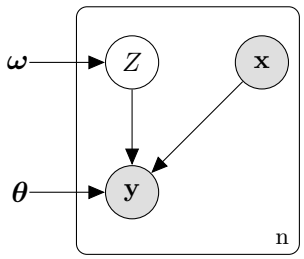
Pour une meilleure comparaison entre un modèle de régression standard MoE (dans lequel tous les paramètres du modèle sont des fonctions des covariables) et les cas particuliers, où certains des paramètres du modèle ne dépendent pas des covariables, les quatre modèles du cadre MoE sont présentés dans Figure 5.1; voir également Fruhwirth-Schnatter et al. (2019, Chapitre 12) pour plus de détails.

L'idée principale de la MoE est un principe de division et de conquête qui propose de diviser un problème complexe en un ensemble de sous-problèmes plus simples, puis d'affecter un ou plusieurs outils de résolution de problèmes spécialisés, ou experts, à chacun de ces sous-problèmes. Dans le contexte de la régression, les modèles MoE avec des experts gaussiens et des poids de mélanges (appelés gating network) modélisés par des fonctions softmax ou gaussiennes normalisées, qui sont les choix les plus populaires et sont des outils puissants pour modéliser des relations non linéaires complexes entre les réponses et les prédicteurs qui proviennent de différentes sous-populations. Ceci est largement étudié en raison de leurs propriétés d'approximation universelles, voir Chapter 2 pour plus de détails, qui ont été largement étudiées non seulement pour les modèles de mélange fini (Genovese & Wasserman, 2000, Rakhlin et al., 2005, Nguyen, 2013, Ho et al., 2016a,b, Nguyen et al., 2020d,b) mais aussi les densités conditionnelles des modèles MoE (Jiang & Tanner, 1999a, Norets et al., 2010, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2019, 2021a).

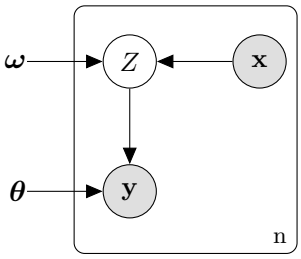
Plus précisément, dans Chapter 2, nous passons d'abord en revue les propriétés d'approximation universelles des mélanges de densités classiques afin de préparer le cadre théorique et de clarifier certaines affirmations vagues et peu claires dans la littérature, avant de les considérer dans le contexte des modèles MoE. En particulier, nous prouvons que, à un degré de précision arbitraire, les mélanges de translatées-dilatées d'une fonction de densité de probabilité (FDP) continue peuvent approximer toute FDP continue, uniformément, sur un ensemble compact; et les mélanges de translatées dilatées d'une FDP essentiellement bornée peuvent approximer toute FDP dans les espaces de Lebesgue. Ensuite, après avoir apporté des améliorations aux résultats d'approximation dans le contexte des mélanges inconditionnels, nous étudions les capacités d'approximation universelles des modèles MoE dans une variété de contextes, y compris en approximation de densité conditionnelle et en calcul bayésien approximatif (ABC). Étant donné des variables d'entrée et de sortie toutes deux à support compact, nous prouvons que les MoE pour les FDP conditionnelles sont denses dans les espaces de Lebesgue. Ensuite, nous prouvons que la distribution quasi-postérieure résultant de l'ABC avec des postérieurs de substitution construits à partir de mélanges gaussiens finis en utilisant une approche de régression inverse, converge vers la vraie distribution, dans des conditions standard.



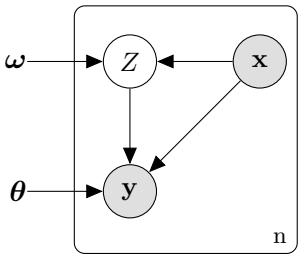
(a) Modèle de mélange



(b) Modèle de régression du MoE



(c) Modèle de régression simple du MoE



(d) Modèle de régression standard du MoE

Figure 5.1: Selon la représentation graphique du modèle, à savoir la présence ou l'absence de bords entre les covariables $\mathbf{x}$ et la variable latente Z et la variable réponse $\mathbf{y}$, il existe quatre cas particuliers de modèles de régression à mélange d'experts. Plus précisément, dans Figure 5.1a, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|Z)p(Z)$; dans Figure 5.1b, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, Z)p(Z)$; dans la Figure 5.1c, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|Z)p(Z|\mathbf{x})$; et dans la Figure 5.1d, $p(\mathbf{y}, Z|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, Z)p(Z|\mathbf{x})$. Cette figure est inspirée du Fruhwirth-Schnatter et al. (2019, Chapitre 12, Fig. 12.2).

Dans cette thèse, nous souhaitons tout d'abord étudier les modèles MoE avec des experts gaussiens et des poids de mélanges modélisés par gaussiennes normalisées pour le clustering et la régression, introduits pour la première fois par Xu et al. (1995), qui ont étendu les modèles MoE originaux de Jacobs et al. (1991). En nous basant sur les travaux de Nguyen et al. (2021c,b), nous désignons ces modèles sous le nom de modèles de MoE localisés gaussiens (≪Gaussian-gated localized MoE≫, GLoME) et de modèles de mélange localisé bloc-diagonal d'experts (≪block-diagonal covariance for localized mixture of experts≫, BLoME), qui seront développés dans Chapter 3. Il est intéressant de souligner que les modèles BLoME généralisent les modèles GLoME en utilisant une structure de covariance parcimonieuse, via des structures bloc-diagonales pour les matrices de covariance dans les experts gaussiens. Il est également intéressant de souligner que les modèles supervisés de cartographie gaussienne localement linéaire (≪Gaussian locally-linear mapping≫, GLLiM) et de covariance bloc-diagonale pour la cartographie gaussienne localement linéaire (≪block-diagonal covariance for Gaussian locally-linear mapping≫, BLLiM) dans Deleforge et al. (2015c) et Devijver et al. (2017) sont des instances affines des modèles GLoME et BLoME, respectivement, où la combinaison linéaire de fonctions bornées est considérées au lieu d'affines pour les fonctions moyennes des experts gaussiens.

Ensuite, le Chapter 4 est consacré à l'étude des modèles de MoE avec des fonctions de gating softmax. Dans Chapter 4, nous obtenons ce que nous appellerons combinaison linéaire de fonctions bornées modèles de régression à mélange d'experts bloc-diagonal à fonctions de gating softmax (≪linear-combination-of-bounded-functions softmax-gated block-diagonal mixture of experts≫, LinBoSGaBloME).En particulier, nous nous référons simplement aux instances affines des modèles LinBoSGaBloME en tant que des modèles de régression MoE à fonctions de gating softmax (≪softmax-gated mixture of experts≫, SGaME). L'un des principaux inconvénients des modèles SGaME est

la difficulté d'appliquer un algorithme EM, qui nécessite une procédure d'optimisation numérique itérative interne (*e.g.,* algorithme MM, moindres carrés pondérés par itération, procédure proximale de type Newton, algorithme de Newton-Raphson) pour mettre à jour les paramètres de la softmax. Les modèles GLoME et BLoME surmontent ce problème en utilisant le réseau de gating gaussien qui nous permet de lier GLoME à des mélanges finis de modèles gaussiens. Étant donné son fondement de modèle de mélange, la maximisation par rapport aux paramètres du réseau de déclenchement peut être résolue analytiquement dans le cadre de l'algorithme EM, ce qui réduit la complexité de calcul de la routine d'estimation. En outre, nous pouvons également utiliser des résultats théoriques bien établis pour les modèles de mélange finis.

Malgré le fait que la nomenclature des MoE trouve son origine dans la littérature sur l'apprentissage automatique (Jacobs et al., 1991), les modèles SGaME ont été largement appliqués à de nombreux domaines scientifiques, technologiques et commerciaux, pour les tâches de classification, de clustering et de régression; modèles de régression par commutation (Quandt, 1972), modèles de classe latente à variables concomitantes (Dayton & Macready, 1988), modèles de régression à classes latentes (DeSarbo & Cron, 1988), modèles mixtes (Wang et al., 1996), analyse des données fonctionnelles et traitement du signal (Chamroukhi et al., 2009, Samé et al., 2011, Chamroukhi et al., 2013a), mélanges lisses finis (Li et al., 2011), classification d'images et segmentation sémantique tâches (Wang et al., 2020), modélisation de la connectivité neuronale (Bock & Fine, 2014), segmentation d'images spectrales (Cohen & Le Pennec, 2014), modélisation des changements climatiques (Nguyen & McLachlan, 2014), reconnaissance de l'activité téléphonique (Lee & Cho, 2014), modélisation de l'hétérogénéité dans les données de connectivité neuronale (Eavani et al., 2016), l'apprentissage par renforcement (He et al., 2016), les tâches de modélisation du langage et de traduction automatique (Shazeer et al., 2017), les modèles génératifs profonds multimodaux sur différents ensembles de modalités, y compris un ensemble de données image-langage difficile (Shi et al., 2019), la détection d'anomalies (Yu et al., 2021), pour n'en citer que quelques-uns.

Il est important de noter ici que les modèles GLoME et BLoME ont également fait l'objet d'études approfondies dans la littérature sur les statistiques et l'apprentissage automatique et que leurs formes apparaissent sous de nombreuses formes différentes, notamment les modèles MoE localisés (Ramamurti & Ghosh, 1996, 1998, Moerland, 1999, Bouchard, 2003), les réseaux gaussiens normalisés (Sato & Ishii, 2000), modélisation MoE des prieurs dans la régression non paramétrique bayésienne (Norets & Pelenis, 2014, Norets & Pati, 2017), modélisation cluster-weighted (Ingrassia et al., 2012), GLLiM dans la régression inverse (Deleforge et al., 2015c), modèle BLLiM (Devijver et al., 2017), mélange profond de régressions inverses linéaires (Lathuilière et al., 2017), modèle de mélange structuré de cartographie gaussienne localement linéaire hiérarchique (HGLLiM) (Tu et al., 2019), mélange gaussien à sorties multiples mélange d'experts linéaires (Nguyen et al., 2019), et calcul bayésien approximatif avec des postérieurs de substitution à l'aide de GLLiM (Forbes et al., 2021).

À partir de maintenant, nous nous intéressons à l'estimation de la loi de la variable aléatoire $\mathbf{Y}$ conditionnellement à $\mathbf{X}$. Les hypothèses suivantes seront nécessaires tout au long de la thèse. Nous supposons que les covariables $\mathbf{X}$ sont indépendantes mais pas nécessairement distribuées de manière identique. Les hypothèses sur les réponses $\mathbf{Y}$ sont plus fortes: conditionnellement à $\mathbf{X}_{[n]}$, les $\mathbf{Y}_i, i \in [n]$, sont indépendants, et chaque $\mathbf{Y}$ suit une loi avec une FDP vraie (mais inconnue) $s_0(\cdot|\mathbf{X} = \mathbf{x})$, qui est approximée via un modèle GLoME, BLoME ou SGaME.

Modèles GLoME et BLoME

Motivé par une stratégie de régression inverse où les rôles des variables prédicteurs et des variables réponses doivent être échangés de telle sorte que $\mathbf{Y} = (\mathbf{Y}_j)_{j\in[L]}, [L] = \{1, \ldots, L\}$, devient l'entrée et $\mathbf{X} = (\mathbf{X}_j)_{j\in[D]}$ joue le rôle d'une sortie multivariée, nous considérons le modèle GLoME suivant, défini par (5.4.4) (voir aussi dans Nguyen et al., 2021c). Cette construction remonte aux travaux de Li (1991), Deleforge et al. (2015c), et Perthame et al. (2018). De cette manière, nous définissons la

FDP conditionnelle correspondante comme suit:

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y};\boldsymbol{\omega}\right)\phi_D\left(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}),\boldsymbol{\Sigma}_k\right), \tag{5.4.4}$$

$$g_k\left(\mathbf{y};\boldsymbol{\omega}\right) = \frac{\pi_k\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K}\pi_j\phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}. \tag{5.4.5}$$

Ici, $g_k(\cdot;\boldsymbol{\omega})$ et $\phi_D\left(\cdot;\boldsymbol{\upsilon}_{k,d}(\cdot),\boldsymbol{\Sigma}_k\right)$, $k\in[K]$, $K\in\mathbb{N}^\star$, $d\in\mathbb{N}^\star$, sont appelés respectivement gaussiennes normalisées et experts gaussiens. En outre, nous décomposons les paramètres du modèle comme suit: $\boldsymbol{\Psi}_{K,d} = (\boldsymbol{\omega},\boldsymbol{\upsilon}_d,\boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K\times\boldsymbol{\Upsilon}_{K,d}\times\mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi},\boldsymbol{c},\boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1}\times\mathbf{C}_K\times\boldsymbol{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$, $\boldsymbol{c} = (\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d = (\boldsymbol{\upsilon}_{k,d})_{k\in[K]} \in \boldsymbol{\Upsilon}_{K,d}$, et $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in[K]} \in \mathbf{V}_K$. Remarquons que $\boldsymbol{\Pi}_{K-1} = \left\{(\pi_k)_{k\in[K]}\in\left(\mathbb{R}^+\right)^K, \sum_{k=1}^K\pi_k = 1\right\}$ est un simplex de probabilité de dimension $K-1$, $\mathbf{C}_K$ est un ensemble de K-tuples de vecteurs moyens de taille $L\times 1$, $\boldsymbol{V}'_K$ est un ensemble de K-tuples d'éléments dans $\mathcal{S}_L^{++}$, où $\mathcal{S}_L^{++}$ désigne la collection de matrices symétriques définies positives sur $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ est un ensemble de K-tuples de fonctions moyennes de $\mathbb{R}^L$ à $\mathbb{R}^D$ dépendant d'un degré d (*e.g.,* un degré de polynômes), et $\mathbf{V}_K$ est un ensemble contenant K-tuples de $\mathcal{S}_D^{++}$.

Ensuite, nous décrivons une caractérisation des modèles GLLiM, une instance affine des modèles GLoME, qui est particulièrement utile pour les données de régression à grande dimension.

Un modèle GLLiM, tel que présenté à l'origine dans Deleforge et al. (2015c), est utilisé pour capturer la relation non linéaire entre la réponse et l'ensemble des covariables dans des données de régression de grande dimension, typiquement dans le cas où $D \gg L$, par des K mappings localement affines:

$$\mathbf{Y} = \sum_{k=1}^{K}\mathbb{I}\left(Z=k\right)\left(\mathbf{A}_k^*\mathbf{X}+\mathbf{b}_k^*+\mathbf{E}_k^*\right). \tag{5.4.6}$$

Ici, $\mathbb{I}$ est une fonction indicatrice et Z est une variable latente capturant une relation de grappe, telle que $Z=k$ si $\mathbf{Y}$ provient de la grappe $k\in[K]$. Les transformations affines spécifiques aux clusters sont définies par les matrices $\mathbf{A}_k^*\in\mathbb{R}^{L\times D}$ et les vecteurs $\mathbf{b}_k^*\in\mathbb{R}^L$. De plus, $\mathbf{E}_k^*$ sont des termes d'erreur capturant à la fois l'erreur de reconstruction due aux approximations affines locales et le bruit d'observation dans $\mathbb{R}^L$.

Suivant l'hypothèse commune que $\mathbf{E}_k^*$ est un vecteur gaussien de moyenne nulle avec une matrice de covariance $\boldsymbol{\Sigma}_k^*\in\mathbb{R}^{L\times L}$, il s'avère que

$$p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x},Z=k;\boldsymbol{\psi}_K^*\right) = \phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right), \tag{5.4.7}$$

où nous désignons par $\boldsymbol{\psi}_K^*$ le vecteur des paramètres du modèle et ϕ_L est la FDP d'une distribution gaussienne de dimension L. Afin d'imposer que les transformations affines soient locales, $\mathbf{X}$ est défini comme un mélange de K composantes gaussiennes comme suit:

$$p\left(\mathbf{X}=\mathbf{x}|Z=k;\boldsymbol{\psi}_K^*\right) = \phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right), p\left(Z=k;\boldsymbol{\psi}_k^*\right) = \boldsymbol{\pi}_k^*, \tag{5.4.8}$$

où $\mathbf{c}_k^*\in\mathbb{R}^D, \boldsymbol{\Gamma}_k^*\in\mathbb{R}^{D\times D}$, $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_k^*)_{k\in[K]}\in\boldsymbol{\Pi}_{K-1}^*$, et $\boldsymbol{\Pi}_{K-1}^*$ est le simplex de probabilité de dimension $K-1$. Ensuite, selon les formules pour les variables gaussiennes multivariées conditionnelles et la décomposition hiérarchique suivante

$$p\left(\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x};\boldsymbol{\psi}_K^*\right) = \sum_{k=1}^{K} p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x},Z=k;\boldsymbol{\psi}_K^*\right)p\left(\mathbf{X}=x|Z=k;\boldsymbol{\psi}_K^*\right)p\left(Z=k;\boldsymbol{\psi}_K^*\right),$$

$$= \sum_{k=1}^{K}\boldsymbol{\pi}_k^*\phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right)\phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right),$$

nous obtenons la *densité conditionnelle directe* suivante (Deleforge et al., 2015c):

$$p\left(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x};\boldsymbol{\psi}_K^*\right)=\sum_{k=1}^{K}\frac{\boldsymbol{\pi}_k^*\phi_D\left(\mathbf{x};\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*\right)}{\sum_{j=1}^{K}\boldsymbol{\pi}_j^*\phi_D\left(\mathbf{x};\mathbf{c}_j^*,\boldsymbol{\Gamma}_j^*\right)}\phi_L\left(\mathbf{y};\mathbf{A}_k^*\mathbf{x}+\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*\right), \tag{5.4.9}$$

où $\boldsymbol{\psi}_K^*=(\boldsymbol{\pi}^*,\boldsymbol{\theta}_K^*)\in\Pi_{K-1}\times\boldsymbol{\Theta}_K^*=:\boldsymbol{\Psi}_K^*$. Ici, $\boldsymbol{\theta}_K^*=(\mathbf{c}_k^*,\boldsymbol{\Gamma}_k^*,\mathbf{A}_k^*,\mathbf{b}_k^*,\boldsymbol{\Sigma}_k^*)_{k\in[K]}$ et

$$\boldsymbol{\Theta}_K^*=\left(\mathbb{R}^D\times\mathcal{S}_D^{++}(\mathbb{R})\times\mathbb{R}^{L\times D}\times\mathbb{R}^L\times\mathcal{S}_L^{++}(\mathbb{R})\right)^K.$$

Sans rien supposer de plus sur la structure des paramètres, la dimension du modèle (désignée par $\dim(\cdot)$), est définie comme le nombre total de paramètres qui doivent être estimés, comme suit:

$$\dim\left(\boldsymbol{\Psi}_K^*\right)=K\left(1+D(L+1)+\frac{D(D+1)}{2}+\frac{L(L+1)}{2}+L\right)-1.$$

Il convient de mentionner que $\dim\left(\boldsymbol{\Psi}_K\right)$ peut être très grand par rapport à la taille de l'échantillon (voir, par exemple, Deleforge et al., 2015c, Devijver et al., 2017, Perthame et al., 2018 pour plus de détails dans leurs ensembles de données réelles) lorsque D est grand et $D\gg L$. En outre, il est plus réaliste de faire des hypothèses sur les matrices de covariance résiduelle $\boldsymbol{\Sigma}_k^*$ des vecteurs d'erreur $\mathbf{E}_k^*$ plutôt que sur $\boldsymbol{\Gamma}_k^*$ (cf. Deleforge et al., 2015c, Section 3). Cela justifie l'utilisation de l'astuce de régression inverse de Deleforge et al. (2015c), qui conduit à une réduction drastique du nombre de paramètres à estimer.

Plus précisément, dans (5.4.9), les rôles des variables d'entrée et de réponse doivent être échangés de sorte que $\mathbf{Y}$ devienne les covariables et que $\mathbf{X}$ joue le rôle de la réponse multivariée. Par conséquent, sa *densité conditionnelle inverse* correspondante est définie comme suit un modèle de cartographie gaussienne localement linéaire (GLLiM), basé sur le modèle de mélange gaussien hiérarchique précédent, comme suit:

$$p\left(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y},Z=k;\boldsymbol{\psi}_K\right)=\phi_D\left(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k\right), \tag{5.4.10}$$

$$p\left(\mathbf{Y}=\mathbf{y}|Z=k;\boldsymbol{\psi}_K\right)=\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right),p\left(Z=k;\boldsymbol{\psi}_k\right)=\pi_k, \tag{5.4.11}$$

$$p\left(\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}_K\right)=\sum_{k=1}^{K}\frac{\pi_k\phi_L\left(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K}\pi_j\phi_L\left(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j\right)}\phi_D\left(\mathbf{x};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Sigma}_k\right), \tag{5.4.12}$$

où $\boldsymbol{\Sigma}_k$ est une structure de covariance $D\times D$ (généralement diagonale, choisie pour réduire le nombre de paramètres) apprise automatiquement à partir des données, et $\boldsymbol{\psi}_K$ est l'ensemble des paramètres, noté $\boldsymbol{\psi}_K$. de paramètres, désigné par $\boldsymbol{\psi}_K=(\boldsymbol{\pi},\boldsymbol{\theta}_K)\in\boldsymbol{\Pi}_{K-1}\times\boldsymbol{\Theta}_K=:\boldsymbol{\Psi}_K$. Une caractéristique intrigante du modèle GLLiM est décrite dans Lemma 5.4.1, qui est prouvée dans Section 3.2.5.1.

Lemma 5.4.1. *Le paramètre $\boldsymbol{\psi}_K^*$ dans la FDP conditionnelle avant, défini dans (5.4.9), peut alors être déduit de $\boldsymbol{\psi}_K$ dans (5.4.12) via la correspondance biunivoque suivante:*

$$\boldsymbol{\theta}_K=\begin{pmatrix}\mathbf{c}_k\\ \boldsymbol{\Gamma}_k\\ \mathbf{A}_k\\ \mathbf{b}_k\\ \boldsymbol{\Sigma}_k\end{pmatrix}_{k\in[K]}\mapsto\begin{pmatrix}\mathbf{c}_k^*\\ \boldsymbol{\Gamma}_k^*\\ \mathbf{A}_k^*\\ \mathbf{b}_k^*\\ \boldsymbol{\Sigma}_k^*\end{pmatrix}_{k\in[K]}=\begin{pmatrix}\mathbf{A}_k\mathbf{c}_k+\mathbf{b}_k\\ \boldsymbol{\Sigma}_k+\mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top\\ \boldsymbol{\Sigma}_k^*\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\\ \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1}\mathbf{c}_k-\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k)\\ \left(\boldsymbol{\Gamma}_k^{-1}+\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1}\end{pmatrix}_{k\in[K]}\in\boldsymbol{\Theta}_K^*, \tag{5.4.13}$$

avec la remarque que $\boldsymbol{\pi}^*\equiv\boldsymbol{\pi}$.

Nous souhaitons fournir quelques exemples simulés de modèles de régression GLoME sur des ensembles de données à 1 de dimension, c'est-à-dire avec $L=D=1$. Nous construisons des ensembles de données simulées suivant deux scénarios: un cas *bien spécifié* (WS) cas dans lequel la véritable densité conditionnelle avant s_0^*, qui peut être estimée via une instance de moyenne gaussienne affine

de GLoME, à savoir le modèle GLLiM, basé sur la FDP conditionnelle inverse $s_{\boldsymbol{\psi}_{K,d}}$ en utilisant une stratégie de régression inverse, appartient à la classe des modèles proposés:

$$s_0^*(y|x) = \frac{\phi(x;0.2,0.1)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi(y;-5x+2,0.09) + \frac{\phi(x;0.8,0.15)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi(y;0.1x,0.09),$$

et un cas *mal spécifié* (MS), dans lequel une telle hypothèse n'est pas vraie:

$$s_0^*(y|x) = \frac{\phi(x;0.2,0.1)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi\left(y;x^2-6x+1,0.09\right) + \frac{\phi(x;0.8,0.15)}{\phi(x;0.2,0.1)+\phi(x;0.8,0.15)}\phi\left(y;-0.4x^2,0.09\right).$$

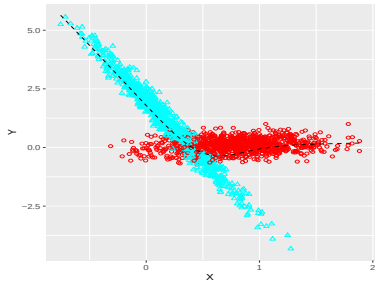
Nous supposons ici que le vrai nombre de composantes du mélange $K_0 = 2$. Les figures 5.2a et 5.2e montrent quelques réalisations typiques de 2000 points de données provenant des scénarios WS et MS. Notez qu'en utilisant GLLiM, notre estimateur l'estimateur de maximum de vraisemblance pénalisé de GLLiM, tel qu'introduit dans Chapter 3, est performant dans le cadre du WS (Figures 5.2b à 5.2d). Dans le cas de l'EM, nous attendons notre procédure, à savoir l'algorithme GLLiM-EM introduit dans Section 3.2.3, utilisant l'heuristique de pente, voir Section 5.4 pour plus de détails, pour équilibrer automatiquement le biais du modèle et sa variance (Figures 5.2f à 5.2h), ce qui conduit au choix d'un modèle complexe, avec 4 de composantes de mélange.

Dans le modèle BLoME, nous souhaitons utiliser les structures bloc-diagonales en remplaçant $\boldsymbol{\Sigma}_k$ et $\mathbf{V}_K$ par $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$ et $\mathbf{V}_K(\mathbf{B})$, définies respectivement dans (5.4.14), (voir, *e.g.,* Devijver et al., 2017, Devijver & Gallopin, 2018, Nguyen et al., 2021b). Ces structures bloc-diagonales pour les matrices de covariance ne sont pas seulement utilisées pour un compromis entre la complexité et la sparsité, mais sont également motivées par certaines applications réelles, où nous voulons effectuer une prédiction sur des ensembles de données avec des observations hétérogènes et des interactions cachées structurées en graphe entre les covariables; par exemple, pour les ensembles de données d'expression génique dans lesquels, conditionnellement à la réponse phénotypique, les gènes interagissent uniquement avec quelques autres gènes, c'est-à-dire qu'il existe de petits modules de gènes corrélés (voir Devijver et al., 2017, Devijver & Gallopin, 2018 pour plus de détails). Pour être plus précis, pour $k \in [K]$, on décompose $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$ en G_k blocs, $G_k \in \mathbb{N}^\star$, et nous désignons par $d_k^{[g]}$ l'ensemble des variables dans le $g$ième groupe, pour $g \in [G_k]$, et par $\text{card}\left(d_k^{[g]}\right)$ le nombre de variables dans l'ensemble correspondant. Ensuite, nous définissons $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$ comme une structure de blocs pour le cluster k, et $\mathbf{B} = (\mathbf{B}_k)_{k\in[K]}$ comme les indices de covariables dans chaque groupe pour chaque cluster. De cette façon, pour construire les matrices de covariance diagonales par blocs, jusqu'à une permutation, nous faisons la définition suivante: $\mathbf{V}_K(\mathbf{B}) = (\mathbf{V}_k(\mathbf{B}_k))_{k\in[K]}$, pour chaque $k \in [K]$,

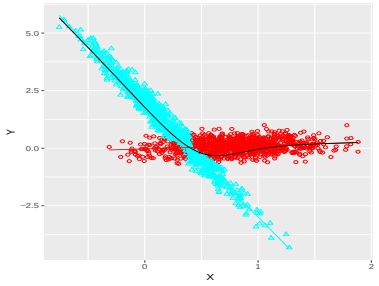
$$\mathbf{V}_k(\mathbf{B}_k) = \left\{ \boldsymbol{\Sigma}_k(\mathbf{B}_k) \in \mathcal{S}_D^{++} \;\middle|\; \begin{array}{l} \boldsymbol{\Sigma}_k(\mathbf{B}_k) = \mathbf{P}_k \begin{pmatrix} \boldsymbol{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \boldsymbol{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\text{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}, \tag{5.4.14}$$

où $\mathbf{P}_k$ correspond à la permutation conduisant à une matrice bloc-diagonale dans le cluster k. Il est utile de préciser qu'en dehors des blocs, tous les coefficients de la matrice sont des zéros et nous autorisons également le réordonnancement des blocs: *e.g.,* $\{(1,3);(2,4)\}$ est identique à $\{(2,4);(1,3)\}$, et la permutation à l'intérieur des blocs: *e.g.,* la partition de 4 variables en blocs $\{(1,3);(2,4)\}$ est la même que la partition $\{(3,1);(4,2)\}$.

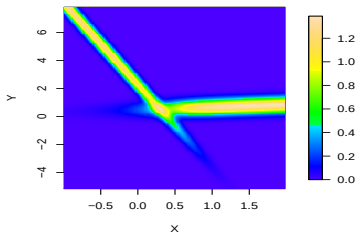
Il est intéressant de souligner que les modèles GLLiM et BLLiM dans Deleforge et al. (2015c), Devijver et al. (2017) sont des instances affines des modèles GLoME et BLoME, respectivement, où



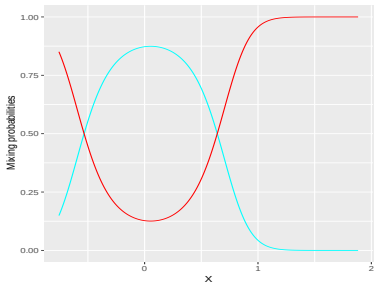
(a) Réalisation typique d'un exemple de WS



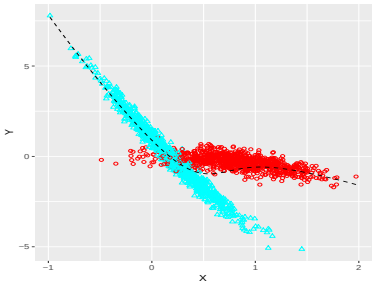
(b) Clustering par GLoME



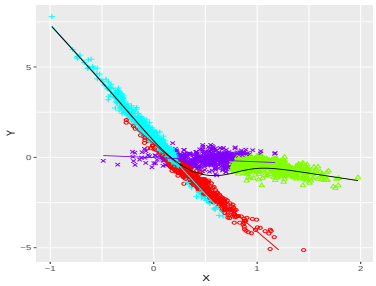
(c) Vue en 2D de la densité conditionnelle résultante avec les 2 composantes de régression



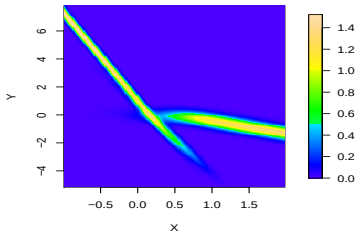
(d) Probabilités du réseau de portes



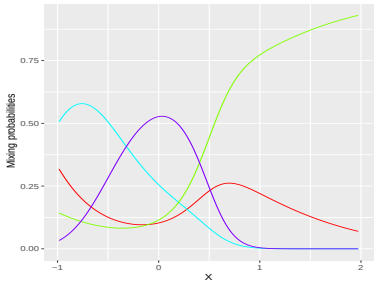
(e) Réalisation typique d'un exemple de MS



(f) Clustering par GLoME



(g) Vue en 2D de la densité conditionnelle résultante avec les 4 composantes de régression



(h) Probabilités du réseau de portes

Figure 5.2: Clustering déduit de la densité conditionnelle estimée de GLoME par un principe MAP avec 2000 points de données des exemples des scénarios WS et MS. Les courbes noires en tirets et pleines présentent les fonctions moyennes réelles et estimées.

la combinaison linéaire de fonctions bornées (*e.g.,* polynômes) est considérée au lieu de fonctions moyennes affines pour les experts gaussiens. Le cadre BLLiM vise à modéliser un échantillon de données de régression de grande dimension provenant d'une population hétérogène avec une interaction cachée structurée en graphe entre les covariables. En particulier, le modèle BLLiM est considéré comme un bon candidat pour effectuer un clustering basé sur le modèle et pour prédire la réponse dans des situations affectées par le phénomène de "malédiction de la dimensionnalité", où le nombre de

paramètres pourrait être plus grand que la taille de l'échantillon. En effet, pour traiter les problèmes de régression à grande dimension, le modèle BLLiM est basé sur une stratégie de régression inverse, qui inverse le rôle du prédicteur à grande dimension et de la réponse multivariée. Par conséquent, le nombre de paramètres à estimer est considérablement réduit. Plus précisément, BLLiM utilise GLLiM, décrit dans Deleforge et al. (2015a,c), en conjonction avec une hypothèse de structure bloc-diagonale sur les matrices de covariance résiduelles pour faire un compromis entre la complexité et la sparsité.

Ce modèle de prédiction est entièrement paramétrique et hautement interprétable. Par exemple, il pourrait être utile pour l'analyse des données transcriptomiques en biologie moléculaire pour classer les observations ou prédire les états phénotypiques, comme par exemple la maladie par rapport à la non-maladie ou la tumeur par rapport à la normale (Golub et al., 1999, Nguyen & Rocke, 2002, Lê Cao et al., 2008). En effet, si les variables prédictives sont des données d'expression génique mesurées par des microarrays ou par les technologies RNA-seq et que la réponse est une variable phénotypique, les situations affectées par le BLLiM ne fournissent pas seulement des clusters d'individus basés sur la relation entre les données d'expression génique et le phénotype, mais implique également un réseau de régulation génique spécifique à chaque groupe d'individus (voir Devijver et al., 2017 pour plus de détails).

Modèles SGaME et LinBoSGaBloME

Nous allons considérer les cadres statistiques dans lesquels nous modélisons un échantillon de données de régression de grande dimension issu d'une population hétérogène via le modèle SGaBloME. Nous soulignons que la dimension de la variable d'entrée $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ et/ou de la variable de sortie $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^q$ est typiquement beaucoup plus élevée que la taille de l'échantillon n. Dans cette thèse, en se basant sur les modèles MoE originaux de Jacobs et al. (1991), nous cherchons à établir un modèle MoE avec des fonctions softmax aussi générique que possible afin qu'il puisse être utilisé pour traiter des ensembles de données de régression à haute dimension et pour étudier les inégalités d'oracle. Pour ce faire, nous définissons d'abord $s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x})$ comme une FDP conditionnelle du modèle MoE comme suit :

$$s_{\boldsymbol{\psi}_K}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} g_{\mathbf{w},k}(\mathbf{x})\, \phi_q\left(\mathbf{y}; \boldsymbol{\upsilon}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), \text{ où,} \tag{5.4.15}$$

$$g_{\mathbf{w},k}(\mathbf{x}) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\mathbf{x})\right)}, \mathbf{w}(\mathbf{x}) = \left(w_k(\mathbf{x})\right)_{k\in[K]}. \tag{5.4.16}$$

Ici, $g_{\mathbf{w},k}(\cdot)$ et $\phi_q\left(\cdot; \boldsymbol{\upsilon}_k(\cdot), \boldsymbol{\Sigma}_k(\mathbf{B}_k)\right), k \in [K]$, sont appelés respectivement fonctions softmax et experts gaussiens. Notez que pour chaque $\mathbf{x} \in \mathcal{X}$, $\left(g_{\mathbf{w},k}(\mathbf{x})\right)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$. En outre, nous décomposons l'ensemble des paramètres du modèle comme suit: $\boldsymbol{\psi}_K = (\mathbf{w}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \mathbf{W}_K \times \boldsymbol{\Upsilon}_K \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_K$, $\mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_K$, $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}_k)_{k\in[K]} \in \boldsymbol{\Upsilon}_K$, et $\boldsymbol{\Sigma}(\mathbf{B}) = \left(\boldsymbol{\Sigma}_k(\mathbf{B}_k)\right)_{k\in[K]} \in \boldsymbol{\Sigma}_k(\mathbf{B}_k)$. Il convient de noter que $\mathbf{W}_K$ et $\boldsymbol{\Upsilon}_K$ sont des ensembles de K-tuples de poids et de fonctions moyennes de $\mathbb{R}^p$ à $\mathbb{R}^+$ et de $\mathbb{R}^p$ à $\mathbb{R}^q$, respectivement; et $\boldsymbol{\Sigma}_k(\mathbf{B}_k)$ est un ensemble contenant K-tuples de $\mathcal{S}_q^{++}$ avec les structures bloc-diagonales définies dans (5.4.14), où $\mathcal{S}_q^{++}$ désigne la collection de matrices définies positives symétriques sur $\mathbb{R}^q$. Puisque nous devons limiter la complexité du modèle en utilisant la dimension du modèle, nous devons restreindre notre attention aux modèles LinBoSGaBloME, où $\mathcal{W}_K$ et $\boldsymbol{\Upsilon}_K$ sont définis comme la combinaison linéaire d'un ensemble fini de fonctions bornées dont les coefficients appartiennent à un ensemble compact. Lorsque la dimension des entrées et des sorties n'est pas trop grande, nous n'avons pas besoin de sélectionner des variables pertinentes. Nous pouvons alors travailler sur les modèles LinBoSGaBloME précédents avec des structures générales pour les moyennes, les poids et les matrices de covariance multi-blocs-diagonales. Dans certaines situations, nous n'avons pas besoin de prendre en compte le compromis entre complexité et sparsité pour les matrices de covariance, dans les modèles LinBoSGaBloME, nous pouvons considérer des matrices de covariance 1-block-diagonales, ce qui est bien étudié dans Montuelle et al. (2014) et sera appelé ≪linear-combination-of-bounded-functions softmax-gated mixture of experts≫. (LinBoSGaME). Cependant, pour traiter des données à grande dimension et pour simplifier l'interprétation de la sparsité, dans le modèle LinBoSGaBloME, nous proposons d'utiliser des polynômes pour les poids des fonctions de

softmax et des moyennes d'experts gaussiennes, qui seront appelés ≪polynomial softmax-gated block-diagonal mixture of experts≫. (PSGaBloME). En particulier, nous appelons simplement les instances affines des modèles LinBoSGaBloME des modèles de régression ≪mélange d'experts softmax-gated≫ (SGaME).

Sélection du modèle dans les modèles de régression de mélanges d'experts

Il convient de souligner que plusieurs hyperparamètres doivent être estimés pour construire des modèles de régression BLoME et SGaBloME, notamment le nombre de composantes du mélange, le degré de sparsité (les coefficients et les niveaux de sparsité des rangs des matrices de covariances), le degré des fonctions moyennes polynomiales, et les structures potentielles de diagonales par bloc cachées des matrices de covariance du prédicteur ou de la réponse multivariée. Les choix d'hyperparamètres d'algorithmes d'apprentissage basés sur les données appartiennent à la classe de problèmes de sélection de modèles, qui a attiré beaucoup d'attention en statistique et en apprentissage automatique au cours des 50 dernières années: (Akaike, 1974, Mallows, 1973, Anderson & Burnham, 2002, Massart, 2007). Il s'agit d'une instance particulière du problème de sélection d'un estimateur (ou d'un modèle): étant donné une famille d'estimateurs, comment choisir, à l'aide des données, l'un d'entre eux dont le risque est le plus faible possible? Notez que la pénalisation est l'une des principales stratégies proposées pour la sélection de modèles. Elle suggère de choisir l'estimateur qui minimise la somme de son risque empirique et de certains termes de pénalité correspondant à la façon dont le modèle s'ajuste aux données, tout en évitant le surajustement.

Dans cette thèse, nous nous intéressons au contrôle et à la prise en compte de la complexité du modèle lors de la sélection du meilleur nombre de composantes de mélange d'un modèle. En général, la sélection de modèles est souvent effectuée à l'aide du critère d'information d'Akaike (AIC; Akaike, 1974) ou du critère d'information bayésien (BIC; Schwarz et al., 1978). Une limitation importante de ces critères, cependant, est qu'ils ne sont valables qu'asymptotiquement. Cela implique qu'il n'y a pas de garantie d'échantillon fini lorsqu'on utilise l'AIC ou le BIC, pour choisir entre différents niveaux de complexité. Leur utilisation dans des contextes de petits échantillons est donc ad hoc. Pour surmonter ces difficultés, Birgé & Massart (2007) a proposé une nouvelle approche, appelée heuristique de pente, soutenue par une inégalité oracle non-asymptotique. Cette méthode conduit à un choix optimal, basé sur les données des constantes multiplicatives pour les pénalités. L'heuristique de pente de Birgé & Massart (2007), soutenue par une inégalité oracle non asymptotique, est une méthode qui permet l'inférence par échantillons finis au lieu de l'AIC et du BIC. Des revues récentes et des questions pratiques concernant l'heuristique de pente peuvent être trouvées dans Baudry et al. (2012), Arlot (2019), et les références qui y sont données.

Il convient de souligner qu'un résultat général de sélection de modèles, établi à l'origine par Massart (2007, Theorem 7.11), garantit qu'un critère pénalisé conduit à une bonne sélection de modèles et que la pénalité n'est connue que jusqu'à des constantes multiplicatives et proportionnelles aux dimensions des modèles. En particulier, de telles constantes multiplicatives peuvent être calibrées par l'approche heuristique de la pente dans un cadre d'échantillon fini. Ensuite, dans l'esprit des méthodes basées sur l'inégalité de concentration développées dans Massart (2007), Massart & Meynet (2011), et Cohen & Le Pennec (2011), un certain nombre de résultats d'oracle à échantillon fini ont été établis pour l'opérateur de sélection et de rétrécissement le plus faible possible. (LASSO), (Tibshirani, 1996) et les estimateurs généraux de maximum de vraisemblance pénalisés (PMLE). Ces résultats incluent les travaux pour les modèles graphiques gaussiens de grande dimension (Devijver & Gallopin, 2018), la sélection de modèles à mélange gaussien (Maugis & Michel, 2011b,a), les modèles de régression à mélange fini (Meynet, 2013, Devijver, 2015a,b, 2017b,a), modèles SGaME sans tenir compte de la grande dimension (Montuelle et al., 2014).

Aucune tentative n'a été faite dans la littérature pour développer une inégalité d'oracle à échantillon fini pour le cadre des modèles de régression MoE pour les données de grande dimension. Dans cette thèse, à notre connaissance, nous sommes les premiers à fournir des inégalités oracle à échantillon fini pour plusieurs modèles de régression MoE de grande dimension, y compris le modèle GLoME (Nguyen et al., 2021c, Section 3.2), modèle BLoME (Nguyen et al., 2021b, Section 3.3), modèle SGaME utilisant LASSO (Nguyen et al., 2020c, Section 4.2), et modèle SGaBloME (Section 4.3). En particulier, notre

stratégie de preuve utilise de nouvelles approches récentes comprenant un théorème de sélection de modèle pour l'estimateur du maximum de vraisemblance (MLE) parmi une sous-collection aléatoire (Devijver, 2015b), un résultat de sélection de modèle non asymptotique pour la détection d'une bonne structure bloc-diagonale dans les grands modèles graphiques (Devijver & Gallopin, 2018) et une astuce de reparamétrisation pour limiter l'entropie métrique de l'espace des paramètres de déclenchement gaussien dans les modèles GLoME (Nguyen et al., 2021c), voir également Section 3.2 pour plus de détails. Notez que pour les paramètres gaussiens normalisés, la technique de traitement des poids logistiques dans les modèles SGaME de Montuelle et al. (2014) n'est pas directement applicable au cadre GLoME ou BLoME, en raison de la forme quadratique du lien canonique. Par conséquent, nous proposons un *reparameterization trick*[1] pour limiter l'entropie métrique de l'espace des paramètres gaussiens normalisés; voir Equation (3.2.25) et Section 3.2.5.2 pour plus de détails. En outre, dans Nguyen et al. (2021c, Theorem 3.2.3), voir également Section 3.2, nous étendons les résultats de Montuelle et al. (2014, Theorem 1) lorsqu'on utilise des fonctions linéaires à la forme quadratique du lien canonique des gating networks.

Parmi les principales contributions de cette thèse figurent d'importants résultats théoriques: des inégalités d'oracle à échantillon fini qui fournissent des limites non-asymptotiques sur les risques, et des limites inférieures sur les fonctions de pénalité qui assurent des contrôles théoriques non-asymptotiques sur les estimateurs sous la perte de Jensen–Kullback–Leibler. Ces inégalités d'oracle fournissent également des justifications théoriques solides pour les formes de pénalité lors de l'utilisation de l'heuristique de pente pour les modèles GLLiM, GLoME, BLLiM, BLoME, SGaME et SGaBloME. Nous soulignons que, bien que les inégalités d'oracle à échantillon fini comparent les performances de nos estimateurs avec le meilleur modèle de la collection, elles nous permettent également de bien approximer une classe riche de densités conditionnelles si nous prenons suffisamment de degrés de polynômes de moyennes d'experts gaussiens (appartient à $\mathcal{D}_{\boldsymbol{\Upsilon}}$) et/ou suffisamment de clusters (parmi l'ensemble $\mathcal{K}$) dans le contexte du mélange d'experts gaussiens (Jiang & Tanner, 1999a, Mendes & Jiang, 2012, Nguyen et al., 2016, Ho et al., 2019, Nguyen et al., 2021a). Cela conduit à ce que les bornes supérieures des risques soient petites, pour $\mathcal{D}_{\boldsymbol{\Upsilon}}$ et $\mathcal{K}$ bien choisis.

En particulier, en dehors des questions théoriques importantes concernant la rigueur des bornes, la manière d'intégrer l'information a priori et l'analyse minimax de notre PMLE proposé, nous espérons que nos inégalités d'oracle à échantillon fini et les expériences numériques intéressantes correspondantes aideront à répondre partiellement aux deux questions importantes suivantes soulevées dans le domaine des modèles de régression MoE : (1) Quel nombre de composantes de mélange K devrait être choisi, étant donné la taille de l'échantillon n, et (2) S'il est préférable d'utiliser quelques experts complexes ou de combiner plusieurs experts simples, étant donné le nombre total de paramètres. Notez que, de tels problèmes sont considérés dans le travail de Mendes & Jiang (2012, Proposition 1), où les auteurs ont fourni quelques aperçus qualitatifs et ont seulement suggéré une méthode pratique pour choisir K et d impliquant une pénalité de complexité ou une validation croisée. En outre, leur modèle ne concerne qu'une estimation de maximum de vraisemblance non régularisée et ne convient donc pas au cadre à haute dimension.

Dans cette thèse, nous considérerons le problème de sélection des paramètres comme un problème de sélection de modèles, en construisant une collection de modèles, avec plus ou moins de clusters, des experts complexes ou simples contrôlant via les ordres de polynômes des poids et des experts gaussiens, des modèles sparse de rang élevé ou faible, et des coefficients plus ou moins actifs. Il restera ensuite à choisir un modèle parmi cette collection. De manière générale, désignons par $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ la collection de modèles que nous considérons, indexée par $\mathcal{M}$. Il est utile de souligner que, contrairement à ce que l'on pourrait penser, avoir une collection de modèles trop importante peut être préjudiciable, par exemple en sélectionnant des estimateurs incohérents (Bahadur, 1958) ou sous-optimaux (Birgé & Massart, 1993). C'est ce qu'on appelle le paradigme de la sélection de modèles.

Avant de discuter des inégalités d'oracle à échantillon fini pour la sélection de modèle par pénalisation dans les modèles de régression MoE, nous passons en revue quelques faits standard concernant

[1]Notez que nous utilisons cette nomenclature uniquement pour effectuer un changement de variables de l'espace des paramètres gaussiens normalisés des modèles GLoME via les poids logistiques des modèles SGaME. Cette astuce de reparamétrisation ne correspond pas à celle, bien connue, des auto-encodeurs variationnels (VAE) dans la littérature sur l'apprentissage profond (voir Kingma & Welling, 2013, pour plus de détails).

l'estimation par minimisation de contraste.

Estimation par minimisation du contraste

La méthode d'estimation par minimisation du contraste repose sur l'existence d'une fonction de contraste, notée γ, remplissant la propriété fondamentale que le FDP conditionnel inconnu satisfait

$$s_0 = \arg\min_{t \in \mathcal{S}} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right].$$

De cette manière, nous obtenons ce que nous appellerons la fonction de perte associée, notée l, qui est définie par

$$l\left(s_0, t\right) = \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right], \quad \forall t \in \mathcal{S}.$$

Définissons un certain contraste empirique γ_n (basé sur l'observation $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := \left(\mathbf{X}_i, \mathbf{Y}_i\right)_{i \in [n]}$) tel que

$$\forall t \in \mathcal{S}, \quad \gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma\left(t, \mathbf{X}_i, \mathbf{Y}_i\right).$$

Pour le modèle $\mathbf{m}$, un *estimateur de contraste minimal* $\widehat{s}_{\mathbf{m}}$ de s_0 est un minimiseur du contraste empirique γ_n sur $S_{\mathbf{m}}$, *i.e.,* $\widehat{s}_{\mathbf{m}} = \arg\min_{t \in S_{\mathbf{m}}} \gamma_n(t)$. L'idée est que, dans des conditions raisonnables, $\gamma_n(t)$ converge vers $\mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\gamma(t, \mathbf{X}, \mathbf{Y})\right]$, et qu'il y a un certain espoir d'obtenir un estimateur sensible de s_0, du moins si s_0 appartient (ou est assez proche) au modèle $S_{\mathbf{m}}$. Pour mesurer la qualité d'un tel estimateur, nous faisons usage de la *risque* suivante $\mathcal{R}\left(\widehat{s}_{\mathbf{m}}\right) = \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$.

Par exemple, dans le cadre de l'estimation de la densité, l'estimateur populaire du maximum de vraisemblance est un estimateur du contraste minimum. En effet, on suppose que l'échantillon $\left(\mathbf{X}_i, \mathbf{Y}_i\right)_{i \in [n]}$ a la densité s_0 w.r.t. une mesure μ et on considère une autre densité t w.r.t. la même mesure. Alors, le log-vraisemblance négatif $-\ln\left[t\left(\mathbf{y}|\mathbf{x}\right)\right]$ est le contraste de vraisemblance maximum, et la fonction de perte correspondante est la divergence de Kullback–Leibler définie par $\mathrm{KL}\left(s_0, t\right) = \int s_0 \ln\left(\frac{s_0}{t}\right) d\mu$. Pour un traitement plus complet concernant d'autres exemples de contraste pour la régression, la classification et le bruit blanc gaussien, nous renvoyons le lecteur à Massart (2007).

Le paradigme du choix du modèle

Le but est de sélectionner le "meilleur" estimateur parmi la collection $\left(\widehat{s}_{\mathbf{m}}\right)_{\mathbf{m} \in \mathcal{M}}$. Soit $S_{\widehat{\mathbf{m}}}$ le modèle sélectionné par une procédure de sélection de modèle donnée. Nous désignerons par $\widehat{s}_{\widehat{\mathbf{m}}}$ l'estimateur sélectionné et soulignerons que tant $\widehat{s}_{\mathbf{m}}$ (pour tout $\mathbf{m}$) que $\widehat{\mathbf{m}}$ sont construits à partir du même échantillon $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right)$. Cette procédure a été bien étudiée tant d'un point de vue asymptotique que non asymptotique.

Idéalement, pour un n donné et un ensemble de données donné, on aimerait considérer $\mathbf{m}^*$ minimisant le risque $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$, par rapport à $\mathbf{m} \in \mathcal{M}$. En d'autres termes,

$$\mathbf{m}^* \in \arg\min_{\mathbf{m} \in \mathcal{M}} l\left(s_0, \widehat{s}_{\mathbf{m}}\right). \tag{5.4.17}$$

L'estimateur de contraste minimal $\widehat{s}_{\mathbf{m}^*}$ sur le modèle correspondant $S_{\mathbf{m}^*}$ est appelé un *oracle*. Cette terminologie a été introduite précédemment par Donoho & Johnstone (1994). Malheureusement, puisque la perte $l\left(s_0, \widehat{s}_{\mathbf{m}}\right)$ dépend de la distribution d'échantillon inconnue s_0, il en va de même pour $\mathbf{m}^*$ et l'oracle $\widehat{s}_{\mathbf{m}^*}$ ne devrait pas être un estimateur de s_0. Cependant, cet oracle peut servir de référence pour construire toute procédure de sélection pilotée par les données parmi la collection d'estimateurs $\left(\widehat{s}_{\mathbf{m}}\right)_{m \in \mathcal{M}}$. Il est maintenant naturel de considérer des critères pilotés par les données pour sélectionner un estimateur qui tend à imiter un oracle. En d'autres termes, nous voudrions que le risque de l'estimateur sélectionné $\widehat{s}_{\widehat{\mathbf{m}}}$, *i.e.,* $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right]$, pour être aussi proche que possible du risque d'un oracle, *i.e.,* $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[l(s_0, \widehat{s}_{\mathbf{m}^*})\right]$.

Il convient de souligner que l'approche non-asymptotique (voir, *e.g.,* Massart, 2007, Wainwright, 2019 pour la bibliographie complète) diffère du point de vue asymptotique habituel dans le sens où

le nombre ainsi que les dimensions des modèles dans $\mathcal{M}$ peuvent dépendre de n. Nous souhaitons construire une procédure de sélection de modèles telle que le modèle sélectionné $S_{\widehat{\mathbf{m}}}$ soit *optimal*. Par exemple, il remplit l'inégalité d'oracle suivante

$$l\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq C_1 l\left(s_0, \widehat{s}_{\mathbf{m}^*}\right) + \frac{C_2}{n} \tag{5.4.18}$$

avec C_1 aussi proche de 1 que possible et C_2/n un terme résiduel. L'inégalité de l'oracle est dite exacte si $C_1 = 1$. Nous nous attendons à ce que cette inégalité tienne en valeur attendue ou avec une forte probabilité. En particulier, lorsque de tels résultats sont trop difficiles à obtenir, il suffit d'obtenir une forme plus faible:

$$\mathbb{E}_{\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}}\left[l\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m} \in \mathcal{M}} \mathbb{E}_{\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}}\left[l\left(s_0, \widehat{s}_{\mathbf{m}^*}\right)\right] + \frac{C_2}{n}. \tag{5.4.19}$$

Motivation pourquoi l'asymptotique échoue.

Sélection de modèle via la pénalisation

Décrivons maintenant comment sélectionner un modèle via la minimisation d'un critère pénalisé, pour atteindre un compromis biais/variance. En effet, nous pouvons décomposer la perte en une approximation et une estimation une partie biais et une partie variance comme suit:

$$\begin{aligned} l\left(s_0, \widehat{s}_{\mathbf{m}}\right) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] + \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] \\ &= \underbrace{l\left(s_0, s_{\mathbf{m}}\right)}_{\text{biais}} + \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y}) - \gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right]}_{\text{variance}}, \end{aligned}$$

où $s_{\mathbf{m}} = \arg\min_{t \in S_{\mathbf{m}}} \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(t, \mathbf{X}, \mathbf{Y})\right]$ est l'une des meilleures approximations de s_0 dans $S_{\mathbf{m}}$. Il convient de souligner que pour minimiser le biais, nous avons besoin d'un modèle complexe, qui s'ajuste très étroitement aux données; et pour minimiser la variance, nous ne devons pas considérer des modèles trop complexes, afin d'éviter un surajustement des données.

Les principales méthodes pour tenir compte de ces procédures de sélection de modèles sont la validation croisée et le hold-out (voir, *e.g.,* Arlot & Celisse, 2010, Maillard, 2020 pour la bibliographie complète), ou les critères pénalisés. Il est souligné que la principale difficulté dans l'exécution de la validation croisée et du hold-out est la complexité du temps, en particulier dans un cadre à grande dimension. Par conséquent, le choix de critères de pénalisation semble le mieux adapté à nos modèles de régression MoE à grande dimension.

Décrivons la méthode plus en détail. La procédure *sélection de modèle par pénalisation* consiste à considérer une *fonction de pénalité* pen: $\mathcal{M} \to \mathbb{R}_+$ et à prendre $\widehat{\mathbf{m}}$ qui minimise le *critère pénalisé*, défini comme $\gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m})$ sur $\mathcal{M}$. Cela signifie que nous choisissons

$$\widehat{\mathbf{m}} = \arg\min_{\mathbf{m} \in \mathcal{M}} \left\{\gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m})\right\}. \tag{5.4.20}$$

En d'autres termes, dans le contexte de l'estimateur du maximum de vraisemblance pour le cas de la régression, pour un choix donné de pen($\mathbf{m}$), le *meilleur modèle* $S_{\widehat{\mathbf{m}}}$. est choisi comme celui dont l'indice est un η'-almost minimiseur de la somme de la log-vraisemblance négative (NLL) et de cette pénalité:

$$\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\widehat{\mathbf{m}}}\right)\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m} \in \mathcal{M}} \left(\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \text{pen}(\mathbf{m})\right) + \eta'. \tag{5.4.21}$$

Ici, $\widehat{s}_{\mathbf{m}}$ est défini comme le η-minimiseur de la NLL:

$$\sum_{i=1}^{n} -\ln\left(s_{\widehat{s}_{\mathbf{m}}}\right)\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) \leq \inf_{s_{\mathbf{m}} \in S_{\mathbf{m}}} \sum_{i=1}^{n} -\ln\left(s_{\mathbf{m}}\left(\mathbf{y}_i|\mathbf{x}_i\right)\right) + \eta, \tag{5.4.22}$$

où le terme d'erreur η est nécessaire lorsque l'infimum peut ne pas être unique ou même ne pas être atteint. Notez que $\widehat{s}_{\widehat{\mathbf{m}}}$ est alors appelé l'estimateur de vraisemblance pénalisé par η' et dépend à la fois des termes d'erreur η et η'. À partir de maintenant, le terme *le meilleur modèle ou estimation basé sur les données* sont tous deux utilisés pour indiquer qu'il satisfait (5.4.21).

Nous soulignons que le choix de la pénalité est délicat mais évidemment nécessaire. La construction de telles fonctions dans le contexte de l'estimateur du maximum de vraisemblance remonte aux travaux d'Akaike et de Schwarz, voir respectivement Akaike (1974) et Schwarz et al. (1978). Ils ont proposé les désormais classiques critères AIC et BIC, les deux outils les plus connus et les plus outils les plus utilisés dans la sélection de modèles statistiques, où la pénalité est respectivement établie comme suit:

$$\begin{aligned}\text{pen}_{\text{AIC}}(\mathbf{m}) &= D_{\mathbf{m}},\\ \text{pen}_{\text{BIC}}(\mathbf{m}) &= \frac{\ln(n) D_{\mathbf{m}}}{2},\end{aligned}$$

où $D_{\mathbf{m}}$ est la dimension du modèle $\mathbf{m}$, et n est la taille de l'échantillon considéré. Ces critères pénalisés bien connus ont été largement étudiés (voir, *e.g.,* Anderson & Burnham, 2002) et sont basés sur des approximations asymptotiques. Par conséquent, ces critères peuvent être erronés dans un contexte non asymptotique. Plus précisément, l'AIC et le BIC sont basés sur le théorème de Wilks et une approche bayésienne, voir, respectivement, *e.g.,* Cavanaugh & Neath (2019) et Neath & Cavanaugh (2012), pour des revues récentes sur les fondements conceptuels et théoriques. Parallèlement, Mallows (1973), et plus tard Craven & Wahba (1978) ont proposé d'autres critères pénalisés célèbres: le C_p de Mallows et la validation croisée généralisée (GCV), respectivement, dans le contexte de la régression linéaire. Mathématiquement, Mallows a obtenu

$$\text{pen}_{\text{Mallows}}(\mathbf{m}) = \frac{2D_{\mathbf{m}}\sigma^2}{n},$$

où σ^2 est le niveau de bruit du vrai modèle de régression qui est inconnu (s'il existe) et σ^2 est donc difficile à estimer. De même, la solution proposée par la méthode GCV est basée sur la validation croisée pour choisir le paramètre de réglage inconnu (dont la meilleure valeur est en fait σ^2). Ainsi, une fois encore, nous devons estimer un paramètre inconnu.

Heureistique de pente

Motivé par certains travaux récents sur les inégalités de concentration, Birgé & Massart (2001) a introduit l'heuristique de pente, qui est une méthodologie non-asymptotique permettant de sélectionner un modèle parmi une collection de modèles. Cette heuristique de pente nous permet de choisir une pénalité optimale à partir de données qui sont connues jusqu'à une constante multiplicative κ. Décrivons les idées de cette heuristique. Dans ce cadre, la forme de la pénalité est alors adaptée comme $\text{pen}_{\text{shape}}(\cdot)$ et il existe une constante inconnue κ_{opt} telle que

$$\text{pen}_{\text{opt}} : \mathbf{m} \in \mathcal{M} \mapsto \kappa_{\text{opt}} \text{pen}_{\text{shape}}(\mathbf{m})$$

est une pénalité optimale. Afin de sélectionner le modèle d'oracle en utilisant (5.4.17) et (5.4.20), nous recherchons une pénalité proche de la fonction de pénalité suivante:

$$\mathcal{M} \ni \mathbf{m} \mapsto \text{pen}(\mathbf{m}) = l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}).$$

Cependant, comme s_0 est inconnu en pratique, nous allons essayer d'approcher cette quantité en la décomposant en:

$$\begin{aligned} l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right] - \gamma_n(\widehat{s}_{\mathbf{m}}) \\ &= \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(\widehat{s}_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right]}_{\nu_{\mathbf{m}}} + \underbrace{\left[\gamma_n(s_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}})\right]}_{\widehat{\nu}_{\mathbf{m}}} \\ &+ \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})\right] - \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\gamma(s_0, \mathbf{X}, \mathbf{Y})\right]}_{(1)} - \underbrace{\left[\gamma_n(s_{\mathbf{m}}) - \gamma_n(s_0)\right]}_{(2)} - \gamma_n(s_0). \end{aligned} \tag{5.4.23}$$

Ici, $\nu_{\mathbf{m}}$ est un terme "erreur d'estimation", $\widehat{\nu}_{\mathbf{m}}$ est un terme "erreur d'estimation" empirique. Nous désignerons par $\Delta_n(s_{\mathbf{m}}) = (1) - (2)$, ce qui correspond à la différence entre le terme "biais" et sa version empirique. Notez que $\gamma_n(s_0)$ ne dépend pas de $\mathbf{m}$, l'idée principale est d'estimer la *pénalité idéale* suivante, définie comme $\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}})$, à partir des données afin de construire une fonction de pénalité optimale. Ensuite, (5.4.23) implique que

$$l(s_0, \widehat{s}_{\mathbf{m}}) - \gamma_n(\widehat{s}_{\mathbf{m}}) = \text{pen}^*(\mathbf{m}) - \gamma_n(s_0), \quad \forall \mathbf{m} \in \mathcal{M}. \tag{5.4.24}$$

Ensuite, nous souhaitons prouver l'inégalité d'oracle suivante

$$l(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) + [\text{pen}(\widehat{\mathbf{m}}) - \text{pen}^*(\widehat{\mathbf{m}})] \leq \inf_{\mathbf{m} \in \mathcal{M}} \{l(s_0, \widehat{s}_{\mathbf{m}}) + [\text{pen}(\mathbf{m}) - \text{pen}^*(\mathbf{m})]\}. \tag{5.4.25}$$

En effet, par définition de la pénalité idéale par (5.4.24), et le fait que $\widehat{m} \in \mathcal{M}$, il s'avère que pour tous les $\mathbf{m} \in \mathcal{M}$,

$$\begin{aligned} l(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) + [\text{pen}(\widehat{\mathbf{m}}) - \text{pen}^*(\widehat{\mathbf{m}})] &= \gamma_n(\widehat{s}_{\widehat{\mathbf{m}}}) + \text{pen}(\widehat{\mathbf{m}}) - \gamma_n(s_0) \\ &\leq \gamma_n(\widehat{s}_{\mathbf{m}}) + \text{pen}(\mathbf{m}) - \gamma_n(s_0) \text{ (using (5.4.20))} \\ &= l(s_0, \widehat{s}_{\mathbf{m}}) + [\text{pen}(\mathbf{m}) - \text{pen}^*(\mathbf{m})] \text{ (using (5.4.24))}. \end{aligned}$$

Le point important à noter ici est que la forme de (5.4.25) nous incite à rechercher une pénalité proche de la pénalité idéale pour obtenir une inégalité oracle. Selon l'expression de la pénalité idéale $\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}})$, tant $\nu_{\mathbf{m}}$ que $\Delta_n(s_{\mathbf{m}})$ dépendent du FDP conditionnel inconnu s_0. Par conséquent, il est naturel d'essayer de relier la fonction de pénalité au terme d'erreur d'estimation empirique $\widehat{\nu}_{\mathbf{m}}$.

Pour accomplir cette tâche, d'un point de vue théorique, Birgé & Massart (2001, 2007) ont proposé et prouvé pour la première fois la méthode heuristique de pente dans le contexte de la régression par moindres carrés homoscédastiques gaussiens avec plan fixe. Ils prouvent qu'il existe une *peine minimale*, $\text{pen}_{\min}(\mathbf{m}) = \widehat{\nu}_{\mathbf{m}}$, à savoir telle que la dimension et le risque des modèles sélectionnés avec des pénalités plus légères deviennent très grands, alors que des pénalités plus élevées devraient sélectionner des modèles d'une complexité "raisonnable". En outre, ils montrent que si l'on considère une pénalité égale à *deux fois cette pénalité minimale* permet de sélectionner un modèle proche du modèle de l'oracle en termes de risque.

Plus précisément, étant donné la pénalité choisie comme $\text{pen}(\mathbf{m}) = \kappa\widehat{\nu}_{\mathbf{m}}$, le critère pénalisé peut être écrit comme suit

$$\text{crit}(\mathbf{m}) = (1-\kappa)\gamma_n(\widehat{s}_{\mathbf{m}}) + \kappa\gamma_n(s_{\mathbf{m}}).$$

Par conséquent, trois cas se présentent:

- si $\kappa = 1$ alors $\text{crit}(\mathbf{m}) = \gamma_n(s_{\mathbf{m}})$, qui se concentre autour de son espérance $\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\gamma(s_{\mathbf{m}}, \mathbf{X}, \mathbf{Y})] = \underbrace{l(s_0, s_{\mathbf{m}})}_{\text{biais}} + \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\gamma(s_0, \mathbf{X}, \mathbf{Y})]$ pour de grands n:cette procédure sélectionne un modèle minimisant le biais et ne prend pas en compte la variance, ce qui conduit à un tel critère a une forte probabilité de sélectionner un modèle trop complexe;
- si $\kappa < 1$ ensuite, lorsque la complexité augmente, le critère diminue toujours car les deux termes de $\text{crit}(\mathbf{m})$ sont en chute libre: les modèles sélectionnés sont toujours parmi les plus complexes;
- si $\kappa > 1$ puis le critère augmente avec la complexité des modèles les plus complexes du fait de l'élimination des termes de biais correspondants (ces modèles ont presque le même biais): la dimension des modèles sélectionnés sera plus raisonnable.

Le premier point de l'heuristique de pente est $\widehat{\nu}_{\mathbf{m}} \approx \nu_{\mathbf{m}}$ puisque $\widehat{\nu}_{\mathbf{m}}$ est la contrepartie empirique de $\nu_{\mathbf{m}}$. En particulier, on s'attend à pouvoir contrôler la fluctuation de $\Delta_n(s_{\mathbf{m}})$ autour de son espérance zéro grâce aux résultats de la concentration. Par conséquent, nous pouvons approximer la pénalité idéale comme étant le double de la pénalité minimale en raison du fait que

$$\text{pen}^*(\mathbf{m}) = \nu_{\mathbf{m}} + \widehat{\nu}_{\mathbf{m}} + \Delta_n(s_{\mathbf{m}}) \approx 2\widehat{\nu}_{\mathbf{m}}.$$

Ainsi, en pratique, le principal problème restant est de déterminer la pénalité minimale $\widehat{\nu}_{\mathbf{m}}$. À cette fin, sur l'ensemble de données $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]}$; soit on recherche le plus grand saut de complexité du modèle sélectionné en fonction de la constante multiplicative κ de la pénalité, soit on regarde la pente asymptotique d'une régression linéaire entre la forme de la pénalité $\text{pen}_{\text{shape}}(\cdot)$ et la valeur de contraste $\gamma_n(s_{\mathbf{m}})$ pour les modèles les plus complexes, que l'on appellera *saut de dimension* ou *estimation de la pente dirigée par les donnees*, respectivement. De cette façon, nous obtenons la pénalité minimale, et nous la multiplions par deux pour obtenir la pénalité optimale. Pour une discussion plus approfondie du principe des heuristiques de pente, nous renvoyons le lecteur à Baudry et al. (2012), Arlot (2019) et aux références qui y sont données. Figures 5.3a and 5.3b illustrent ces idées.
Nous soulignons que, d'un point de vue pratique, nous utilisons la méthode dite CAPUSHE (CAlibrating Penalty Using Slope HEuristics) package in R (Arlot et al., 2016, Baudry et al., 2012) pour mettre en œuvre les approches de saut de dimension et d'estimation de la pente basée sur les données. En pratique, l'utilisation de l'heuristique de pente est efficace lorsqu'une pénalité optimale $\text{pen}_{\text{opt}}(\cdot) = \kappa_{\text{opt}}\text{pen}_{\text{shape}}(\cdot)$ est connue jusqu'à un facteur multiplicatif. Il convient de souligner que la clé de voûte de l'heuristique de pente est que $\frac{\kappa_{\text{opt}}}{2}\text{pen}_{\text{shape}}(\mathbf{m})$ est une bonne estimation de $\widehat{s}_{\mathbf{m}}$ et fournit une pénalité minimale. De manière générale, le $\text{pen}_{\text{shape}}(\cdot)$ peut être choisi comme mesure de complexité, lorsque sa définition n'est pas évidente a priori. Cette mesure de complexité est généralement la dimension du modèle $D_{\mathbf{m}}$ ou le nombre de paramètres libres nécessaires à l'estimation.

D'un point de vue théorique, dans cette thèse, nous apportons plusieurs inégalités d'oracle non-asymptotiques, Theorems 1.2.2 and 1.2.3, qui fournissent des limites non asymptotiques sur les risques, et des limites inférieures sur les fonctions de pénalité qui assurent des contrôles théoriques non asymptotiques sur les estimateurs sous la perte de Jensen–Kullback–Leibler. Ces inégalités d'oracle fournissent également certaines justifications théoriques des formes de pénalité lors de l'utilisation de l'heuristique de pente pour les modèles de régression MoE correspondants. Plus précisément, des critères de vraisemblance pénalisés sont proposés dans Chapters 3 and 4 pour sélectionner les meilleurs modèles de régression MoE basés sur des données parmi une collection spécifique de modèles. Ces critères dépendent de constantes inconnues qui peuvent être calibrées dans des situations pratiques par une heuristique de pente. En particulier, afin de travailler avec la FDP conditionnelle dans plusieurs modèles de régression MoE, nous souhaitons faire usage d'un théorème de sélection de modèle pour MLE parmi une sous-collection aléatoire (cf. Devijver, 2015b, Théorème 5.1 et Devijver & Gallopin, 2018, Théorème 7. 3), qui est une extension de toute une collection de densités conditionnelles de Cohen & Le Pennec (2011, Théorème 2), et de Massart (2007, Théorème 7.11), fonctionnant uniquement pour l'estimation de densité.

Dans la section suivante, nous résumons les contributions de la thèse.

Contributions de la thèse

Le reste du manuscrit est organisé comme suit.

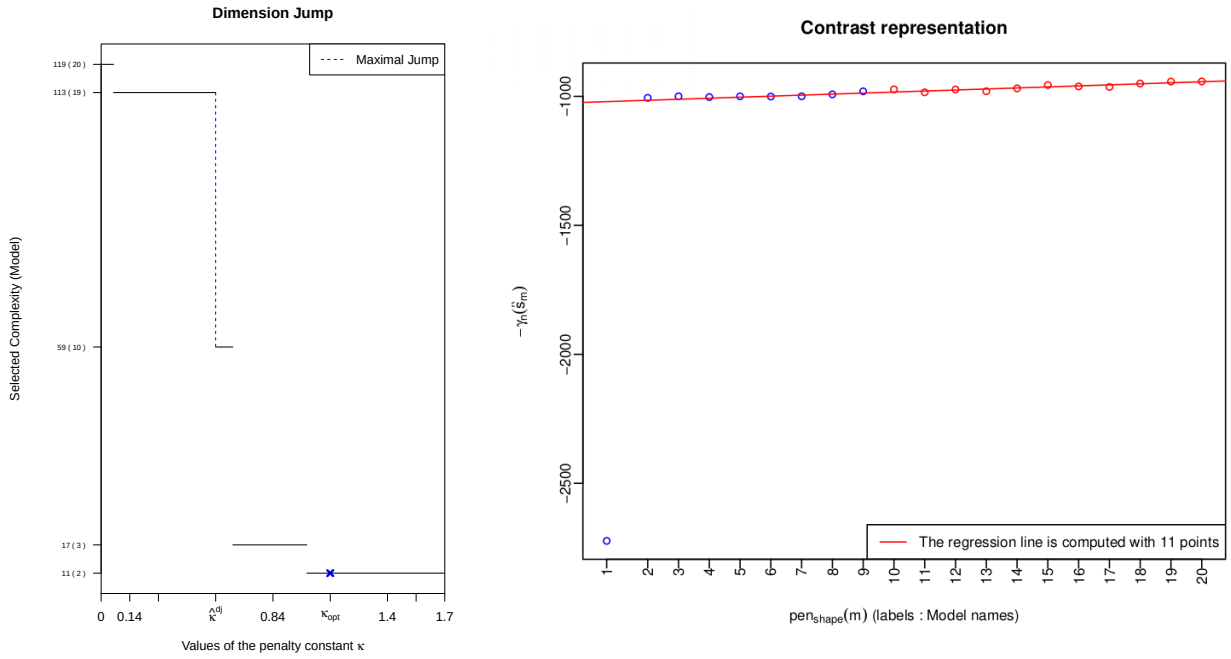
Contribution du Chapitre 1

Chapter 1 est consacré à l'état de l'art. En outre, nous soulignons également les principales contributions dans les autres chapitres de notre thèse.

Contribution du Chapitre 2

Dans le Chapter 2, nous présentons nos premières contributions principales en établissant des résultats d'approximation théorique des modèles de mélanges d'experts sur la plus large classe de FDP et de FDP conditionnels, sous le plus faible ensemble d'hypothèses, à partir des travaux:

(C1) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *Approximation by finite mixtures of continuous density functions that vanish at infinity.* Cogent Mathematics & Statistics, volume 7, page 1750861. Cogent OA, 2020.



(a) Saut de dimension

(b) Estimation de la pente dirigée par les donnees

Figure 5.3: Illustration de l'heuristique de pente avec 2000 points de données des exemples du scénarior WS. Dans Figure 5.3a nous estimons κ en utilisant $\widehat{\kappa}^{\mathrm{dj}}$ le plus grand saut de complexité. Nous choisissons ensuite un modèle qui minimise la log-vraisemblance pénalisée par $\kappa_{\mathrm{opt}} = 2\widehat{\kappa}^{\mathrm{dj}}$. Dans Figure 5.3b, nous estimons κ en recherchant la pente asymptotique d'une régression linéaire entre la forme de la pénalité $\mathrm{pen}_{\mathrm{shape}}(\cdot)$ et la valeur de contraste $\gamma_n(s_{\mathbf{m}})$ pour les modèles les plus complexes.

Link: https://www.tandfonline.com/doi/full/10.1080/25742558.2020.1750861
(Nguyen et al., 2020d).

(C2) Hien Duy Nguyen, **TrungTin Nguyen**, Faicel Chamroukhi, and Geoffrey McLachlan. *Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models.* Journal of Statistical Distributions and Applications, 8(1), 13, 2021.
Link: https://doi.org/10.1186/s40488-021-00125-0
(Nguyen et al., 2021a).

(C3) **TrungTin Nguyen**, Faicel Chamroukhi, Hien D Nguyen, and Geoffrey J McLachlan. *Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces.* arXiv preprint arXiv:2008.09787. To appear, Communications in Statistics - Theory and Methods, 2021.
Link: https://arxiv.org/pdf/2008.09787.pdf
(Nguyen et al., 2020b).

Plus précisément, dans Chapter 2, nous passons d'abord en revue les propriétés d'approximation universelles des mélanges de densités classiques afin de préparer le cadre théorique et de clarifier certaines affirmations vagues et peu claires dans la littérature, avant de les considérer dans le contexte des modèles MoE. En particulier, nous prouvons que, à un degré de précision arbitraire, les mélanges de translatées-dilatées d'une fonction de densité de probabilité (FDP) continue peuvent approximer toute FDP continue, uniformément, sur un ensemble compact; et les mélanges de translatées dilatées d'une FDP essentiellement bornée peuvent approximer toute FDP dans les espaces de Lebesgue. Ensuite, après avoir apporté des améliorations aux résultats d'approximation dans le contexte des mélanges inconditionnels, nous étudions les capacités d'approximation universelles des modèles MoE dans une variété de contextes, y compris en approximation de densité conditionnelle et en calcul bayésien approximatif (ABC). Étant donné des variables d'entrée et de sortie toutes deux à support

compact, nous prouvons que les MoE pour les FDP conditionnelles sont denses dans les espaces de Lebesgue Dans une autre contribution du sujet de cette thèse au sens large, nous avons considéré les modèles MoE dans le cadre bayésien. Ensuite, nous prouvons que la distribution quasi-postérieure résultant de l'ABC avec des postérieurs de substitution construits à partir de mélanges gaussiens finis en utilisant une approche de régression inverse, converge vers la vraie distribution, dans des conditions standard via les travaux suivants:

(C4) Florence Forbes, Hien Duy Nguyen, **TrungTin Nguyen**, and Julyan Arbel. *Approximate Bayesian computation with surrogate posteriors.* hal-03139256. Under Review, Statistics and Computing, February 2021.
Link: https://hal.archives-ouvertes.fr/hal-03139256v2/document
(Forbes et al., 2021).

Plus précisément, nous cherchons à résumer les principaux théorèmes de ce chapitre.

Capacités d'approximation des modèles de mélanges finis

On définit $(\mathbb{E}, \|\cdot\|_{\mathbb{E}})$ comme un espace vectoriel normé (NVS), et soit $x \in \left(\mathbb{R}^d, \|\cdot\|_2\right)$, pour un certain nombre de $d \in \mathbb{N}^\star$, où $\|\cdot\|_2$ est la norme euclidienne. Soient $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction satisfaisant $f \geq 0$ et $\int f \mathrm{d}\lambda = 1$, où λ est la mesure de Lebesgue. Nous disons que f est une fonction de FDP sur le domaine $\mathbb{R}^d$ (que nous que nous omettrons pour des raisons de brièveté, à partir de maintenant). Soit $g : \mathbb{R}^d \to \mathbb{R}$ est un autre FDP et définissons la classe fonctionnelle $\mathcal{M}^g = \bigcup_{m\in\mathbb{N}^\star} \mathcal{M}_m^g$, où

$$\mathcal{M}_m^g = \left\{ h_m^g : h_m^g(\cdot) = \sum_{i=1}^m \frac{c_i}{\sigma_i^d} g\left(\frac{\cdot - \mu_i}{\sigma_i}\right), \mu_i \in \mathbb{R}^d, \sigma_i \in \mathbb{R}_+, c \in \Pi_{m-1}, i \in [m] \right\},$$

$c^\top = (c_1, \ldots, c_m)$, $\mathbb{R}_+ = (0, \infty)$, un simplex de probabilité défini par

$$\Pi_{m-1} = \left\{ (\pi_i)_{i\in[m]} \in \mathbb{R}^m | \forall i \in [m], \pi_i > 0, \sum_{i=1}^m \pi_i = 1 \right\}, \tag{5.4.26}$$

$[m] = \{1, \ldots, m\}, m \in \mathbb{N}^\star$, et $(\cdot)^\top$ est l'opérateur de transposition de matrice.

Nous disons que tout $h_m^g \in \mathcal{M}_m^g$ est un mélange fini à m-composant de translatées dilatées de la FDP g.

L'étude des FDP dans la classe $\mathcal{M}_m^g$ est un domaine de recherche appliquée et technique toujours d'actualité, en particulier dans le domaine de la recherche appliquée et technique, en statistique. Nous renvoyons le lecteur intéressé aux nombreux ouvrages complets sur le sujet, tels que Everitt & Hand (1981), Titterington et al. (1985), McLachlan & Basford (1988), Lindsay (1995), McLachlan & Peel (2000), Frühwirth-Schnatter (2006), Schlattmann (2009), Mengersen et al. (2011), and Fruhwirth-Schnatter et al. (2019).

Une grande partie de la popularité des modèles de mélange fini provient du théorème populaire, qui stipule que pour toute densité f, il existe un $h \in \mathcal{M}_m^g$, pour un nombre suffisamment grand de composantes $m \in \mathbb{N}^\star$, de sorte que h se rapproche de f de manière arbitraire, dans un certain sens. Des exemples de ce théorème populaire apparaissent dans des déclarations telles que: "à condition que le nombre de densités composantes n'est pas borné ci-dessus, certaines formes de de mélange peuvent être utilisées pour fournir une approximation arbitrairement proche d'une distribution de probabilité donnée" (Titterington et al., 1985, p. 50), "les formes de modèles [le mélange] peuvent s'ajuster à n'importe quelle distribution et augmenter significativement l'ajustement du modèle" (Walker & Ben-Akiva, 2011, p. 173), et "un modèle de modèle de mélange peut approcher presque n'importe quelle distribution" (Yona, 2010, p. 500). D'autres déclarations exprimant le même sentiment sont rapportées dans Nguyen & McLachlan (2019). Il y a un sentiment de flou dans les déclarations rapportées, et la nature technique du théorème populaire n'est jamais clairement établie.

Afin de poursuivre, nous avons besoin des définitions suivantes. Nous disons que f est supporté de manière compacte sur le sous-ensemble $\mathbb{K} \subset \mathbb{R}^d$, si $\mathbb{K}$ est compact et si $\mathbf{1}_{\mathbb{K}^\complement} f = 0$, où $\mathbf{1}_{\mathbb{X}}$ est la fonction indicatrice qui prend valeur 1 lorsque $x \in \mathbb{X}$ et 0, ailleurs, et $(\cdot)^\complement$ est l'opérateur de complémentation

d'ensemble (c'est-à-dire, $\mathbb{X}^{\complement} = \mathbb{R}^d \backslash \mathbb{X}$). Ici, $\mathbb{X}$ est un sous-ensemble générique de $\mathbb{R}^d$. De plus, nous disons que $f \in \mathcal{L}_p(\mathbb{X})$ pour tout $1 \leq p < \infty$, si

$$\|f\|_{\mathcal{L}_p(\mathbb{X})} = \left(\int |\mathbf{1}_{\mathbb{X}} f|^p \, \mathrm{d}\lambda \right)^{1/p} < \infty,$$

et pour $p = \infty$, si

$$\|f\|_{\mathcal{L}_\infty(\mathbb{X})} = \inf \{a \geq 0 : \lambda(\{x \in \mathbb{X} : |f(x)| > a\}) = 0\} < \infty,$$

où l'on appelle $\|\cdot\|_{\mathcal{L}_p(\mathbb{X})}$ la $\mathcal{L}_p$-norm sur $\mathbb{X}$. Lorsque $\mathbb{X} = \mathbb{R}^d$, nous écrirons $\|\cdot\|_{\mathcal{L}_p(\mathbb{R}^d)} = \|\cdot\|_{\mathcal{L}_p}$. Dénotons la classe de toutes les fonctions bornées sur $\mathbb{X}$ par

$$\mathcal{B}(\mathbb{X}) = \{f \in \mathcal{L}_\infty(\mathbb{X}) : \exists a \in [0, \infty), \text{ such that } |f(x)| \leq a, \forall x \in \mathbb{X}\}$$

et on écrit

$$\|f\|_{\mathcal{B}(\mathbb{X})} = \sup_{x \in \mathbb{X}} |f(x)|.$$

Par souci de concision, nous écrirons $\mathcal{B}(\mathbb{R}^d) = \mathcal{B}$, et $\|f\|_{\mathcal{B}(\mathbb{R}^d)} = \|f\|_{\mathcal{B}}$.

En outre, nous définissons la divergence dite de Kullback–Leibler, voir Kullback & Leibler (1951), entre deux FDP quelconques f et g sur $\mathbb{X}$, comme suit

$$\mathrm{KL}_{\mathbb{X}}(f, g) = \int \mathbf{1}_{\mathbb{X}} f \log\left(\frac{f}{g}\right) \mathrm{d}\lambda.$$

Soit $g : \mathbb{R}^d \to \mathbb{R}$ à nouveau un FDP. Alors, pour chaque $m \in \mathbb{N}^\star$, nous définissons

$$\mathcal{N}_m^g = \left\{ h : h(x) = \sum_{i=1}^m c_i \frac{1}{\sigma_i^d} g\left(\frac{x - \mu_i}{\sigma_i}\right), \, \mu_i \in \mathbb{R}^d, \, \sigma_i \in \mathbb{R}_+, \, c_i \in \mathbb{R}, \, i \in [m] \right\},$$

que nous appelons l'ensemble des m-composant combinaisons linéaires de translatées dilatées de la FDP g. Dans le passé, les résultats concernant les approximations des FDP f par l'intermédiaire de fonctions $\eta \in \mathcal{N}_m^g$ ont été plus nombreux. Par exemple, dans le cas de $g = \phi$, où

$$\phi(x) = (2\pi)^{-n/2} \exp\left(-\|x\|_2^2 / 2\right), \tag{5.4.27}$$

est la FDP normale standard. Nous désignons la classe des fonctions continues et uniformément continues par $\mathcal{C}$ et $\mathcal{C}^u$, respectivement. Les classes de fonctions continues bornées sont désignées par $\mathcal{C}_b = \mathcal{C} \cap \mathcal{B}$.

Nous avons le résultat que pour tout FDP f, ensemble compact $\mathbb{K} \subset \mathbb{R}^d$, et $\epsilon > 0$, il existe un $m \in \mathbb{N}^\star$ et $h \in \mathcal{N}_m^\phi$, tels que $\|f - h\|_{\mathcal{L}_\infty(\mathbb{K})} < \epsilon$. (Sandberg, 2001, Lem. 1). De plus, en définissant l'ensemble des fonctions continues qui disparaissent à l'infini par

$$\mathcal{C}_0 = \left\{ f \in \mathcal{C} : \forall \epsilon > 0, \exists \text{ une compacte } \mathbb{K} \subset \mathbb{R}^d, \text{ tel que } \|f\|_{\mathcal{L}_\infty(\mathbb{K}^{\complement})} < \epsilon \right\},$$

nous avons également le résultat suivant: pour chaque FDP $f \in \mathcal{C}_0$ et $\epsilon > 0$, il existe un $m \in \mathbb{N}^\star$ et $h \in \mathcal{N}_m^\phi$, de telle sorte que $\|f - h\|_{\mathcal{L}_\infty} < \epsilon$ (Sandberg, 2001, Thm. 2). Les deux résultats de Sandberg (2001) sont des implications simples du célèbre théorème de Stone–Weierstrass (cf. Stone (1948) et De Branges (1959)). Dans Nguyen & McLachlan (2019), l'approximation de FDPs f par la classe $\mathcal{M}_m^g$ a été explorée dans un cadre restrictif. Soit $\{h_m^g\}$ une séquence de fonctions qui tirent des éléments de la séquence imbriquée d'ensembles $\{\mathcal{M}_m^g\}$ (c'est-à-dire, $h_1^g \in \mathcal{M}_1^g, h_2^g \in \mathcal{M}_2^g, \ldots$).

A notre connaissance, la revendication la plus forte qui est disponible concernant le théorème populaire, dans un contexte probabiliste ou statistique, est celle de (DasGupta, 2008, Thm. 33.2). Soit $\{\eta_m^g\}$ est une séquence de fonctions qui tirent des éléments de la séquence imbriquée d'ensembles $\{\mathcal{N}_m^g\}$, de la même façon que $\{h_m^g\}$. Nous paraphrasons l'affirmation sans perte de fidélité, comme suit.

Claim 5.4.2. Si $f, g \in \mathcal{C}$ sont des FDP et que $\mathbb{K} \subset \mathbb{R}^d$ est compact, alors il existe une séquence $\{\eta_m^g\}$, telle que

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

Malheureusement, la preuve de Claim 5.4.2 n'est pas fournie dans DasGupta (2008). La seule référence du résultat est à un endroit non divulgué dans le document Cheney & Light (2000), qui, après enquête enquête, on peut en déduire qu'il s'agit du théorème 5 du (Cheney & Light, 2000, Ch. 20). Il est également à noter qu'aucune preuve n'est fournie pour ce théorème. Au lieu de cela, il est indiqué que la preuve est similaire à celle du Théorème 1 dans (Cheney & Light, 2000, Ch. 24), qui est une reproduction de la preuve de (Xu et al., 1993, Lem. 3.1).

Il y a un problème majeur à appliquer la technique de preuve de (Xu et al., 1993, Lem. 3.1) afin de prouver Claim 5.4.2. La preuve de (Xu et al., 1993, Lem. 3.1) dépend de façon critique de l'affirmation selon laquelle "il n'y a pas de perte de généralité en supposant que $f(x) = 0$ pour $x \in \mathbb{R}^d \backslash 2\mathbb{K}$". Ici, pour $a \in \mathbb{R}_+$, $a\mathbb{K} = \left\{x \in \mathbb{R}^d : x = ay,\, y \in \mathbb{K}\right\}$. Cette hypothèse est nécessaire afin d'écrire toute convolution avec f et une fonction continue arbitraire comme une intégrale sur un domaine compact, et ensuite d'utiliser une somme de Riemann pour approximer une telle intégrale. Par la suite, une telle technique de preuve ne fonctionne pas en dehors de la classe des fonctions continues qui sont supportées de manière compacte sur $a\mathbb{K}$. Ainsi, on ne peut pas vérifier Claim 5.4.2 à partir des matériaux de Xu et al. (1993), Cheney & Light (2000), et DasGupta (2008), seuls.

Certains résultats récents dans l'esprit de Claim 5.4.2 ont été été obtenus par Nestoridis & Stefanopoulos (2007) et Nestoridis et al. (2011), en utilisant des méthodes issues de l'étude des séries universelles (voir par exemple dans Nestoridis & Papadimitropoulos (2005)).

Soit

$$\mathcal{W} = \left\{ f \in \mathcal{C}_0 : \sum_{y\in\mathbb{W}^d} \sup_{x\in[0,1]^d} |f(x+y)| < \infty \right\}$$

désignent l'algèbre dite de Wiener (voir, par exemple, Feichtinger (1977)) et laissons

$$\mathcal{V} = \left\{ f \in \mathcal{C}_0 : \forall x \in \mathbb{R}^d,\ |f(x)| \leq \beta \left(1 + \|x\|_2\right)^{-d-\theta},\ \beta, \theta \in \mathbb{R}_+ \right\}$$

soit une classe de fonctions dont la queue se désintègre à un rythme plus rapide que $o\left(\|x\|_2^d\right)$. Dans Nestoridis et al. (2011), il est noté que $\mathcal{V} \subset \mathcal{W}$. En outre, soit

$$\mathcal{C}_c = \left\{ f \in \mathcal{C} : \exists \text{ a compact set } \mathbb{K}, \text{ such that } \mathbf{1}_{\mathbb{K}^\complement} f = 0 \right\},$$

dénote l'ensemble des fonctions continues à support compact. La Theorem 5.4.3 suivante a été prouvée dans Nestoridis & Stefanopoulos (2007).

Theorem 5.4.3 (Nestoridis & Stefanopoulos, 2007, Thm. 3.2)**.** *Si $g \in \mathcal{V}$, alors les affirmations suivantes sont vraies.*

(a) Pour tout $f \in \mathcal{C}_c$, il existe une séquence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), telle que

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_1} + \|f - \eta_m^g\|_{\mathcal{L}_\infty} = 0.$$

(b) Pour tout $f \in \mathcal{C}_0$, il existe une séquence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), telle que

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty} = 0.$$

(c) Pour tout $1 \leq p < \infty$ et $f \in \mathcal{L}_p$, il y a existe une séquence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), telle que

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_p} = 0.$$

(d) Pour tout mesurable f, il existe une séquence $\{\eta_m^g\}$ ($\eta_m^g \in \mathcal{N}_m^g$), telle que

$$\lim_{m\to\infty} \eta_m^g = f, \text{ presque partout.}$$

(e) *Si ν est une σ-finite mesure de Borel sur $\mathbb{R}^d$, alors pour tout ν-measurable f, il existe une séquence $\{\eta_m^g\}$ $(\eta_m^g \in \mathcal{N}_m^g)$, telle que*

$$\lim_{m\to\infty} \eta_m^g = f,$$

presque partout, par rapport à ν.

Ce résultat a ensuite été amélioré, dans Nestoridis et al. (2011), où l'espace plus général $\mathcal{W}$ a été pris comme un remplacement pour $\mathcal{V}$, dans Theorem 5.4.3. On désigne la classe des fonctions continues bornées par $\mathcal{C}_b = \mathcal{C} \cap \mathcal{L}_\infty$. Le théorème suivant a été prouvé dans Nestoridis et al. (2011).

Theorem 5.4.4 (Nestoridis et al., 2011, Thm. 3.2)**.** *Si $g \in \mathcal{W}$, alors les affirmations suivantes sont vraies.*

(a) *La conclusion de Theorem 5.4.3 (a) est vraie, avec $\mathcal{C}_c$ remplacé par $\mathcal{C}_0 \cap \mathcal{L}_1$.*

(b) *Les conclusions de Theorem 5.4.3 (b)–(e) sont valables.*

(c) *Pour tout $f \in \mathcal{C}_b$ et tout $\mathbb{K} \subset \mathbb{R}^d$ compact, il existe une séquence $\{\eta_m^g\}$, telle que*

$$\lim_{m\to\infty} \|f - \eta_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

En utilisant les techniques de Nestoridis & Stefanopoulos (2007), Bacharoglou (2010) a prouvé un ensemble de résultats similaires à ceux de Theorem 5.4.3, sous la restriction que f est une fonction non-négative avec support $\mathbb{R}$, en utilisant $g = \phi$ (c'est-à-dire que g a la forme (5.4.27), où $d = 1$) et en prenant $\left\{h_m^\phi\right\}$ comme la séquence d'approximation, au lieu de $\{\eta_m^g\}$. C'est-à-dire qu'on obtient le résultat suivant est obtenu.

Theorem 5.4.5 (Bacharoglou, 2010, Cor. 2.5)**.** *Si $f : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$, alors les affirmations suivantes sont vraies.*

(a) *Pour tout FDP $f \in \mathcal{C}_c$, il existe une séquence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, telle que*

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_1} + \left\|f - h_m^\phi\right\|_{\mathcal{L}_\infty} = 0.$$

(b) *Pour tout $f \in \mathcal{C}_0$, tel que $\|f\|_{\mathcal{L}_1} \leq 1$, il existe une séquence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, telle que*

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_\infty} = 0.$$

(c) *Pour tout $1 < p < \infty$ et $f \in \mathcal{C} \cap \mathcal{L}_p$, tel que $\|f\|_{\mathcal{L}_1} \leq 1$, il existe une séquence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, de telle sorte que*

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_p} = 0.$$

(d) *Pour tout mesurable f, il existe une séquence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, de telle sorte que*

$$\lim_{m\to\infty} h_m^\phi = f, \text{ presque partout.}$$

(e) *Pour tout FDP $f \in \mathcal{C}$, il existe une séquence $\left\{h_m^\phi\right\}$ $(h_m^\phi \in \mathcal{M}_m^\phi)$, de telle sorte que*

$$\lim_{m\to\infty} \left\|f - h_m^\phi\right\|_{\mathcal{L}_1} = 0.$$

Le Theorem 5.4.5 est restrictif de deux manières. Premièrement, il ne permet pas la caractérisation de l'approximation via la classe $\mathcal{M}_m^g$ pour tout g sauf la normale FDP ϕ. Bien que ϕ soit traditionnellement le choix le plus courant pour g dans la pratique, la littérature moderne sur les modèles de mélange a vu l'utilisation de nombreuses densités de probabilité de composantes plus exotiques, telles que la densité de probabilité de student-t et ses variantes obliques et modifiées (voir, par exemple, Peel & McLachlan, 2000, Forbes & Wraith, 2014, et Lee & McLachlan, 2016). Ainsi, son utilisation est quelque peu limitée dans le contexte moderne. En outre, les applications modernes ont tendance à exiger $d > 1$, ce qui limite encore plus l'impact du résultat en tant que rempart théorique pour la modélisation des mélanges finis dans la pratique. Une remarque dans Bacharoglou (2010) indique que le résultat peut être généralisé au cas où $g \in \mathcal{V}$ au lieu de $g = \phi$. Cependant, aucune suggestion n'a été proposée, concernant la généralisation de Theorem 5.4.5 au cas de $d > 1$.

Dans Section 2.1, nous prouvons un nouvel ensemble de résultats qui généralisent largement Theorem 5.4.5. En utilisant des techniques inspirées de Donahue et al. (1997) et Cheney & Light (2000), nous sommes en mesure d'obtenir un ensemble de résultats concernant la capacité d'approximation de la classe des modèles de mélanges m-composant $\mathcal{M}_m^g$, lorsque $g \in \mathcal{C}_0$, ou $g \in \mathcal{V}$, et pour tout $d \in \mathbb{N}^\star$. Par définition de $\mathcal{V}$, la majorité de nos résultats s'étendent au-delà des généralisations possibles proposées de Theorem 5.4.5.

Motivé par les preuves incomplètes de Xu et al. (1993, Lem 3.1) et du Théorème 5 de Cheney & Light (2000, Chapitre 20), ainsi que par les résultats restreints de Nestoridis & Stefanopoulos (2007), Bacharoglou (2010), et Nestoridis et al. (2011), dans Section 2.1, voir aussi dans Nguyen et al. (2020d), nous établissons et prouvons Theorem 5.4.6 concernant les suites de FDPs $\{h_m^g\}$ à partir de $\mathcal{M}^g$.

Theorem 5.4.6 (Nguyen et al., 2020d, Théorème 5)**.** *Si nous supposons que f et g sont des FDP et que $g \in \mathcal{C}_0$, alors les affirmations suivantes sont vraies.*

(a) Pour tout $f \in \mathcal{C}_0$, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), de telle sorte que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty} = 0.$$

(b) Pour tout $f \in \mathcal{C}_b$ et tout $\mathbb{K} \subset \mathbb{R}^d$ compact, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), telle que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_\infty(\mathbb{K})} = 0.$$

(c) Pour tout $1 < p < \infty$ et $f \in \mathcal{L}_p$, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), telle que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

(d) Pour tout mesurable f, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), de telle sorte que

$$\lim_{m\to\infty} h_m^g = f, \text{ presque partout.}$$

(e) Si ν est une σ-finite mesure de Borel sur $\mathbb{R}^d$, alors pour tout ν-measurable f, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), telle que

$$\lim_{m\to\infty} h_m^g = f,$$

presque partout, par rapport à ν.

Si nous supposons plutôt que $g \in \mathcal{V}$, alors l'affirmation suivante est également vraie.

(f) Pour tout $f \in \mathcal{C}$, il existe une séquence $\{h_m^g\}$ ($h_m^g \in \mathcal{M}_m^g$), de telle sorte que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_1} = 0.$$

De plus, dans Section 2.2, voir aussi dans Nguyen et al. (2020b), nous établissons Theorem 5.4.7 qui améliore Theorem 5.4.6 de plusieurs façons. Plus précisément, alors que les énoncés (a), (c), (d) et (e) sont toujours valables sous les mêmes hypothèses que dans Theorem 5.4.6; l'affirmation (b) de Theorem 5.4.6 est améliorée pour s'appliquer à une plus grande classe de fonction cible $f \in \mathcal{C}$, pour plus de détails voir l'affirmation (a) de Theorem 5.4.7; et l'énoncé (f) de Theorem 5.4.6 est drastiquement amélioré pour s'appliquer à tout $f \in \mathcal{L}_1$ et $g \in \mathcal{L}_\infty$, voir plus dans l'énoncé (b) de Theorem 5.4.7. Le but de Section 2.2 est de rechercher l'ensemble d'hypothèses le plus faible afin d'établir des résultats théoriques d'approximation sur la classe la plus large possible de problèmes de densité de probabilité. Nous notons en particulier que notre amélioration par rapport à l'affirmation (b) de Theorem 5.4.6 donne exactement le résultat du théorème 5 de Cheney & Light (2000, Chapitre 20), qui a été prouvé de manière incorrecte (voir aussi DasGupta, 2008, Théorème 33.2).

Theorem 5.4.7 (Nguyen et al., 2020b, Théorème 2)**.** *Détachons $h_m^g \in \mathcal{M}^g$ comme un m-composante d'un mélange fini FDP. Si nous supposons que f et g sont des FDP, alors les affirmations suivantes sont vraies.*

(a) Si $f, g \in \mathcal{C}$ et que $\mathbb{K} \subset \mathbb{R}^d$ est un ensemble compact, alors il existe une séquence $\{h_m^g\}_{m=1}^\infty \subset \mathcal{M}^g$, tel que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{B}(\mathbb{K})} = 0.$$

(b) Pour $p \in [1, \infty)$, si $f \in \mathcal{L}_p$ et $g \in \mathcal{L}_\infty$, alors il existe une séquence $\{h_m^g\}_{m=1}^\infty \subset \mathcal{M}^g$, tel que

$$\lim_{m\to\infty} \|f - h_m^g\|_{\mathcal{L}_p} = 0.$$

Modèle de mélange d'experts

Soit $\mathbb{W} = \mathbb{Y} \times \mathbb{X}$, où $\mathbb{X} \subseteq \mathbb{R}^d$ et $\mathbb{Y} \subseteq \mathbb{R}^q$, pour $d, q \in \mathbb{N}^\star$. Supposons que les variables aléatoires d'entrée et de sortie, $\boldsymbol{X} \in \mathbb{X}$ et $\boldsymbol{Y} \in \mathbb{Y}$, soient liées via la FDP conditionnelle $f(\boldsymbol{y}|\boldsymbol{x})$ dans la classe fonctionnelle:

$$\mathcal{F} = \left\{ f : \mathbb{W} \to [0, \infty) \,|\, \int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{y}) = 1, \forall \boldsymbol{x} \in \mathbb{X} \right\},$$

où λ désigne la mesure de Lebesgue. L'approche MoE cherche à d'approximer la FDP conditionnelle inconnue de la cible f par une fonction de la forme MoE:

$$m(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{Gate}_k(\boldsymbol{x}) \, \mathrm{Expert}_k(\boldsymbol{y}),$$

où $\mathbf{Gate} = (\mathrm{Gate}_k)_{k\in[K]} \in \mathcal{G}^K$ ($[K] = \{1, \ldots, K\}$), $\mathrm{Expert}_1, \ldots, \mathrm{Expert}_K \in \mathcal{E}$, et $K \in \mathbb{N}^\star$. Nous disons ici que m est un modèle K-composant de MoE avec des portes issues de la classe $\mathcal{G}^K$ et des experts issus de la classe $\mathcal{E}$, où $\mathcal{E}$ est une classe de FDPs avec support $\mathbb{Y}$.

Les choix les plus populaires pour $\mathcal{G}^K$ sont les classes paramétriques softmax et gaussienne:

$$\mathcal{G}_S^K = \left\{ \mathbf{Gate} = (\mathrm{Gate}_k(\cdot;\boldsymbol{\gamma}))_{k\in[K]} \,|\forall k \in [K], \mathrm{Gate}_k(\cdot;\boldsymbol{\gamma}) = \frac{\exp\left(a_k + \boldsymbol{b}_k^\top \cdot\right)}{\sum_{l=1}^K \exp\left(a_l + \boldsymbol{b}_l^\top \cdot\right)}, \boldsymbol{\gamma} \in \mathbb{G}_S^K \right\}$$

et

$$\mathcal{G}_G^K = \left\{ \mathbf{Gate} = (\mathrm{Gate}_k(\cdot;\boldsymbol{\gamma}))_{k\in[K]} \,|\forall k \in [K], \mathrm{Gate}_k(\cdot;\boldsymbol{\gamma}) = \frac{\pi_k \phi(\cdot;\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \phi(\cdot;\boldsymbol{\nu}_l, \boldsymbol{\Sigma}_l)}, \boldsymbol{\gamma} \in \mathbb{G}_G^K \right\},$$

respectivement, où

$$\mathbb{G}_S^K = \left\{ \boldsymbol{\gamma} = (a_1, \ldots, a_K, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K) \in \mathbb{R}^K \times \left(\mathbb{R}^d\right)^K \right\}$$

et

$$\mathbb{G}_G^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K) \in \Pi_{K-1} \times \left(\mathbb{R}^d\right)^K \times \mathbb{S}_d^K \right\}.$$

Ici,

$$\phi\left(\cdot;\boldsymbol{\nu},\boldsymbol{\Sigma}\right)=\left|2\pi\boldsymbol{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\cdot-\boldsymbol{\nu}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\cdot-\boldsymbol{\nu}\right)\right]$$

est la fonction de densité normale multivariée avec un vecteur moyen $\boldsymbol{\nu}$ et une matrice de covariance $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}^{\top}=(\pi_1,\ldots,\pi_K)$ est un vecteur de poids dans le simplex de probabilité de $K-1$ Π_{K-1}, défini dans (5.4.26), et $\mathbb{S}_d$ est la classe des matrices définies positives symétriques $d\times d$. Les classes de gating softmax et gaussien ont été introduites respectivement par Jacobs et al. (1991) et Xu et al. (1995). En général, on choisit des experts qui proviennent d'une certaine classe de translatées dilatées:

$$\mathcal{E}_{\psi}=\left\{g_{\psi}\left(\cdot;\boldsymbol{\mu},\sigma\right):\mathbb{Y}\rightarrow[0,\infty)\,|g_{\psi}\left(\cdot;\boldsymbol{\mu},\sigma\right)=\frac{1}{\sigma^{q}}\psi\left(\frac{\cdot-\boldsymbol{\mu}}{\sigma}\right),\boldsymbol{\mu}\in\mathbb{R}^{q},\sigma\in(0,\infty)\right\},$$

où ψ est une FDP, par rapport à $\mathbb{R}^q$ au sens où que $\psi:\mathbb{R}^q\rightarrow[0,\infty)$ et $\int_{\mathbb{R}^q}\psi\left(\boldsymbol{y}\right)\mathrm{d}\lambda\left(\boldsymbol{y}\right)=1$.

Nous dirons que $f\in\mathcal{L}_p\left(\mathbb{W}\right)$ pour tout $p\in[1,\infty)$ si

$$\left\|f\right\|_{p,\mathbb{W}}=\left(\int_{\mathbb{W}}\left|\mathbf{1}_{\mathbb{W}}f\right|^{p}\mathrm{d}\lambda\left(\boldsymbol{z}\right)\right)^{1/p}<\infty,$$

On appellera $\left\|\cdot\right\|_{p,\mathbb{W}}$ la la norme $\mathcal{L}_p$ sur $\mathbb{W}$, pour $p\in[0,\infty]$, et lorsque le contexte est évident, nous laisserons tomber la référence à $\mathbb{W}$.

Supposons que la FDP conditionnelle cible f soit dans la classe $\mathcal{F}_p=\mathcal{F}\cap\mathcal{L}_p$. Nous abordons le problème de l'approximation de f, par rapport à la classe norme $\mathcal{L}_p$, à l'aide de modèles MoE dans les classes softmax et gaussiennes,

$$\begin{aligned}\mathcal{M}_S^{\psi}=\Bigg\{m_K^{\psi}:\mathbb{W}\rightarrow[0,\infty)\,|m_K^{\psi}\left(\boldsymbol{y}|\boldsymbol{x}\right)=\sum_{k=1}^{K}\mathrm{Gate}_k\left(\boldsymbol{x}\right)g_{\psi}\left(\boldsymbol{y};\boldsymbol{\mu}_k,\sigma_k\right),\\ g_{\psi}\in\mathcal{E}_{\psi}\cap\mathcal{L}_{\infty},\mathbf{Gate}\in\mathcal{G}_S^K,\boldsymbol{\mu}_k\in\mathbb{Y},\sigma_k\in(0,\infty),k\in[K],K\in\mathbb{N}^{\star}\Bigg\},\end{aligned}\tag{5.4.28}$$

et

$$\begin{aligned}\mathcal{M}_G^{\psi}=\Bigg\{m_K^{\psi}:\mathbb{W}\rightarrow[0,\infty)\,|m_K^{\psi}\left(\boldsymbol{y}|\boldsymbol{x}\right)=\sum_{k=1}^{K}\mathrm{Gate}_k\left(\boldsymbol{x}\right)g_{\psi}\left(\boldsymbol{y};\boldsymbol{\mu}_k,\sigma_k\right),\\ g_{\psi}\in\mathcal{E}_{\psi}\cap\mathcal{L}_{\infty},\mathbf{Gate}\in\mathcal{G}_G^K,\boldsymbol{\mu}_k\in\mathbb{Y},\sigma_k\in(0,\infty),k\in[K],K\in\mathbb{N}^{\star}\Bigg\},\end{aligned}\tag{5.4.29}$$

en montrant que $\mathcal{M}_S^{\psi}$ et $\mathcal{M}_G^{\psi}$ sont denses dans la classe $\mathcal{F}_p$, lorsque $\mathbb{X}=[0,1]^d$ et $\mathbb{Y}$ est un sous-ensemble compact de $\mathbb{R}^q$. Nos résultats de densité sont rendus possibles par le résultat d'approximation de fonction indicatrice de Jiang & Tanner (1999b), et les théorèmes de densité de modèle de mélange fini de Nguyen et al. (2020b) et Nguyen et al. (2020d). Nos Theorems 5.4.8 and 5.4.9, Lemma 5.4.10, and Corollary 5.4.11 dans Section 2.3 contribuent à une continuité durable de l'intérêt porté aux capacités d'approximation des modèles MoE. En relation avec nos résultats, des contributions concernant les capacités d'approximation de la fonction d'espérance conditionnelle des classes $\mathcal{M}_S^{\psi}$ et $\mathcal{M}_G^{\psi}$, voir les définitions dans (5.4.28) et (5.4.29), respectivement, (Jiang & Tanner, 1999b, Krzyzak & Schafer, 2005, Mendes & Jiang, 2012, Nguyen et al., 2016, 2019, Wang & Mendel, 1992, Zeevi et al., 1998) et les capacités d'approximation des sous-classes de $\mathcal{M}_S^{\psi}$ et $\mathcal{M}_G^{\psi}$, par rapport à la divergence de Kullback–Leibler (Jiang & Tanner, 1999a, Norets et al., 2010, Norets & Pelenis, 2014). Nos résultats peuvent être considérés comme des compléments aux théorèmes d'approximation de Kullback–Leibler de Norets et al. (2010) et Norets & Pelenis (2014), par la relation entre la divergence de Kullback-Leibler et la norme $\mathcal{L}_2$ (Zeevi & Meir, 1997). C'est-à-dire que lorsque $f>1/\kappa$, pour tout $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{W}$ et une certaine constante $\kappa>0$, nous avons que la divergence de Kullback–Leibler conditionnelle intégrée considérée par Norets & Pelenis (2014):

$$\int_{\mathbb{X}}D\left(f\left(\cdot|\boldsymbol{x}\right)\|m_K^{\psi}\left(\cdot|\boldsymbol{x}\right)\right)\mathrm{d}\lambda\left(\boldsymbol{x}\right)=\int_{\mathbb{X}}\int_{\mathbb{Y}}f\left(\boldsymbol{y}|\boldsymbol{x}\right)\log\frac{f\left(\boldsymbol{y}|\boldsymbol{x}\right)}{m_K^{\psi}\left(\boldsymbol{y}|\boldsymbol{x}\right)}\mathrm{d}\lambda\left(\boldsymbol{y}\right)\mathrm{d}\lambda\left(\boldsymbol{x}\right)$$

satisfait à

$$\int_{\mathbb{X}} D\left(f\left(\cdot|\boldsymbol{x}\right)\|m_K^{\psi}\left(\cdot|\boldsymbol{x}\right)\right)\mathrm{d}\lambda\left(\boldsymbol{x}\right) \leq \kappa^2\left\|f-m_K^{\psi}\right\|_{2,\mathbb{W}}^2,$$

et donc une bonne approximation de la divergence de Kullback-Leibler intégrée est garantie si l'on peut trouver une bonne approximation de la norme $\mathcal{L}_2$, ce qui est garanti par nos principaux résultats.

Theorem 5.4.8. *Supposons que $\mathbb{X}=[0,1]^d$. pour $d\in\mathbb{N}^{\star}$. Pour tout $f\in\mathcal{F}\cap\mathcal{C}$, tout $p\in[1,\infty)$, et tout ensemble compact $\mathbb{Y}\subset\mathbb{R}^q$, $q\in\mathbb{N}^{\star}$, il existe une suite $\left\{m_K^{\psi}\right\}_{K\in\mathbb{N}^{\star}}\subset\mathcal{M}_S^{\psi}$, où $\psi\in\mathcal{C}\left(\mathbb{R}^q\right)$ est un PDF sur support $\mathbb{R}^q$, telle que $\lim_{K\to\infty}\left\|f-m_K^{\psi}\right\|_p=0$.*

Puisque la convergence dans les espaces de Lebesgue n'implique pas de modes de convergence ponctuels de convergence, le résultat suivant est également utile et intéressant dans certains scénarios restreints. Ici, nous notons que le mode de convergence est presque uniforme, ce qui implique une convergence presque partout et une convergence en mesure (cf. Bartle 1995, Lem 7.10 et Thm. 7.11). La convergence presque uniforme de $\left\{m_K^{\psi}\right\}_{K\in\mathbb{N}^{\star}}$ vers f dans le résultat suivant est à comprendre au sens de Bartle (1995, Def. 7.9). C'est-à-dire que pour chaque $\delta>0$, il existe un ensemble $\mathbb{E}_{\delta}\subset\mathbb{W}$ avec $\lambda\left(\mathbb{W}\right)<\delta$, tel que $\left\{m_K^{\psi}\right\}_{K\in\mathbb{N}^{\star}}$ converge vers f, uniformément sur $\mathbb{W}\backslash\mathbb{E}_{\delta}$.

Theorem 5.4.9. *Supposons que $\mathbb{X}=[0,1]$. Pour tout $f\in\mathcal{F}\cap\mathcal{C}$, et tout ensemble compact $\mathbb{Y}\subset\mathbb{R}^q$, $q\in\mathbb{N}^{\star}$, il existe une suite $\left\{m_K^{\psi}\right\}_{K\in\mathbb{N}^{\star}}\subset\mathcal{M}_S^{\psi}$, où $\psi\in\mathcal{C}\left(\mathbb{R}^q\right)$ est un PDF sur support $\mathbb{R}^q$, tel que $\lim_{K\to\infty}m_K^{\psi}=f$, presque uniformément.*

Le résultat suivant établit la connexion entre les classes de gating classes $\mathcal{G}_S^K$ et $\mathcal{G}_G^K$.

Lemma 5.4.10. *Pour chaque $K\in\mathbb{N}^{\star}$, $\mathcal{G}_S^K\subset\mathcal{G}_G^K$. De plus, si nous définissons la classe des vecteurs de gaussiennes normalisées avec des matrices de covariance égales:*

$$\mathcal{G}_E^K=\left\{\mathbf{Gate}=\left(\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right)\right)_{k\in[K]}|\forall k\in[K],\mathrm{Gate}_k\left(\cdot;\boldsymbol{\gamma}\right)=\frac{\pi_k\phi\left(\cdot;\boldsymbol{\nu}_k,\boldsymbol{\Sigma}\right)}{\sum_{l=1}^K\pi_l\phi\left(\cdot;\boldsymbol{\nu}_l,\boldsymbol{\Sigma}\right)},\boldsymbol{\gamma}\in\mathbb{G}_E^K\right\},$$

où

$$\mathbb{G}_E^K=\left\{\boldsymbol{\gamma}=\left(\boldsymbol{\pi},\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_K,\boldsymbol{\Sigma}\right)\in\Pi_{K-1}\times\left(\mathbb{R}^d\right)^K\times\mathbb{S}_d\right\},$$

then $\mathcal{G}_E^K\subset\mathcal{G}_S^K$.

Nous pouvons appliquer directement Lemma 5.4.10 pour établir le suivant corollaire à Theorem 5.4.8 et 5.4.9, concernant la capacité d'approximation de la classe $\mathcal{M}_G^{\psi}$.

Corollary 5.4.11. *Theorem 5.4.8 et 5.4.9 tiennent lorsque $\mathcal{M}_S^{\psi}$ est remplacé par $\mathcal{M}_G^{\psi}$ dans leurs énoncés.*

Une approximation universelle pour les modèles de mélange d'experts dans le calcul bayésien approximatif

Le calcul bayésien approximatif (ABC) (voir, *e.g.,* Sisson et al. 2018) apparaît comme un candidat naturel pour traiter les problèmes, où il y a un manque de disponibilité ou de tractabilité de la vraisemblance. De tels cas se produisent lorsque le modèle direct ou le processus de génération de données n'est pas disponible, ou de manière analytique, mais est disponible en tant que procédure de simulation; *e.g.,* lorsque le processus de génération de données est caractérisé comme une série d'équations différentielles ordinaires, comme dans Mesejo et al. (2016), Hovorka et al. (2004). En outre, les caractéristiques ou contraintes typiques qui peuvent se produire dans la pratique sont les suivantes: (1) les observations $\mathbf{y}$ sont de grande dimension, car elles représentent des signaux dans le temps ou des spectres, comme dans Schmidt & Fernando (2015), Bernard-Michel et al. (2009), Ma et al. (2013); et (2) le paramètre $\boldsymbol{\theta}$, à estimer, est lui-même multidimensionnel avec des dimensions

corrélées de sorte que la prédiction indépendante de ses composants est sous-optimale; *e.g.,* lorsqu'il existe des contraintes connues, comme lorsque les éléments du paramètre sont des concentrations ou des probabilités dont la somme est égale à un (Deleforge et al., 2015a, Lemasson et al., 2016, Bernard-Michel et al., 2009).

L'idée fondamentale de l'ABC est de générer des propositions de paramètres $\boldsymbol{\theta}$ dans un espace de paramètres Θ en utilisant une distribution a priori $\pi(\boldsymbol{\theta})$ et d'accepter une proposition si les données simulées $\mathbf{z}$ pour cette proposition sont similaires aux données observées $\mathbf{y}$, toutes deux dans un espace d'observation $\mathcal{Y}$. Cette similarité est généralement mesurée à l'aide d'une mesure de distance ou de discrimination D et un échantillon simulé $\mathbf{z}$ est retenu si $D(\mathbf{z}, \mathbf{y})$ est inférieur à un seuil donné ϵ. Sous cette forme simple, la procédure est généralement appelée rejet ABC. D'autres variantes sont possibles et souvent recommandées, par exemple l'utilisation de MCMC ou de procédures séquentielles. (*e.g.*, Del Moral et al., 2012, Buchholz & Chopin, 2019), mais nous nous concentrerons sur la version de rejet pour les besoins de Section 2.4.

Dans le cas d'un algorithme de rejet, les échantillons sélectionnés sont tirés de ce que l'on appelle le quasi-postérieur ABC, qui est une approximation du vrai postérieur $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. Dans des conditions similaires à celles de Bernton et al. (2019), concernant l'existence d'une FDP $f_{\boldsymbol{\theta}}(\mathbf{z})$ pour la vraisemblance, le quasi-postérieur ABC dépend de D et d'un seuil ϵ, et peut être écrit comme suit

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \, . \tag{5.4.30}$$

Plus précisément, la similarité entre $\mathbf{z}$ et $\mathbf{y}$ est généralement évaluée sur la base de deux éléments: le choix de statistiques sommaires $s(\cdot)$ pour rendre compte des données de manière plus robuste, et le choix d'une distance pour comparer les statistiques sommaires. Autrement dit, $D(\mathbf{y}, \mathbf{z})$ dans (1.4.1) devrait alors être remplacé par $D(s(\mathbf{y}), s(\mathbf{z}))$, après quoi nous surchargeons D pour qu'il désigne également la distance entre les statistiques sommaires $s(\cdot)$.

Cependant, il n'existe pas de règle générale pour construire de bonnes statistiques sommaires pour les modèles complexes et si une statistique sommaire ne capture pas les caractéristiques importantes des données, l'algorithme ABC est susceptible de produire des échantillons d'une (Blum et al., 2013, Fearnhead & Prangle, 2012, Gutmann et al., 2018) incorrecte. Les travaux des auteurs suivants ont permis de mieux comprendre la situation Fearnhead & Prangle (2012), qui a introduit le cadre *semi-automatique*. Cette espérance conditionnelle ne peut pas être calculée analytiquement mais peut être estimée par régression en utilisant un ensemble de données d'apprentissage avant la procédure ABC elle-même.

Dans Fearnhead & Prangle (2012), il est suggéré qu'un simple modèle de régression peut suffire pour approximer $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$, mais cela a depuis été contredit, par exemple par Jiang et al. (2017) et Wiqvist et al. (2019), qui montrent que la qualité de l'approximation peut avoir de l'importance en pratique. Toujours en se concentrant sur les moyennes postérieures comme statistiques sommaires, ils utilisent des réseaux neuronaux profonds qui capturent des relations non linéaires complexes et présentent de bien meilleurs résultats que les approches de régression standard. Toutefois, les réseaux neuronaux profonds restent des outils très coûteux en termes de calcul, tant en ce qui concerne la taille requise des données d'apprentissage que le nombre de paramètres et d'hyperparamètres à estimer et à régler.

Dans Section 2.4, voir aussi Forbes et al. (2021), notre première contribution est d'étudier une autre manière efficace de construire des statistiques sommaires, dans la même veine que l'ABC semi-automatique, mais basée sur les moments postérieurs, sans se limiter aux moyennes postérieures. Bien que cette extension naturelle ait déjà été proposée dans Jiang et al. (2017), elle nécessite la disponibilité d'un modèle de régression flexible et traçable, capable de capturer des relations non linéaires complexes et de fournir des moments postérieurs, de manière directe. Par conséquent, Jiang et al. (2017) n'a pas envisagé une mise en œuvre de la procédure. À cette fin, la méthode GLLiM (Deleforge et al., 2015c), que nous rappelons dans Section 2.4.1, apparaît comme un bon candidat, avec des propriétés qui s'équilibrent entre les réseaux neuronaux coûteux en calcul et les techniques de régression standard simples.

Contrairement à la plupart des méthodes de régression qui ne fournissent que des prédictions ponctuelles, GLLiM fournit, à faible coût, une estimation paramétrique des véritables distributions

postérieures complètes. En particulier, nous prouvons des théorèmes universels selon lesquels la distribution quasi-postérieure résultant de ABC avec des postérieurs de substitution construits à partir de GLLiM converge vers la vraie, dans des conditions standard, voir plus dans Section 2.4.3. En utilisant un ensemble de couples de paramètres et d'observations, GLLiM apprend une famille de mélanges gaussiens finis dont les paramètres dépendent analytiquement de l'observation à inverser. Pour toute donnée observée, la vraie postérieure peut être approximée comme un mélange gaussien, dont les moments sont facilement calculés en forme fermée et transformés en statistiques sommaires pour la sélection ultérieure d'échantillons ABC.

Plus précisément, nous fournissons deux types de résultats, ci-dessous. Dans le premier résultat (Theorem 5.4.12), le vrai postérieur est utilisé pour comparer les échantillons $\mathbf{y}$ et $\mathbf{z}$. Ce résultat vise à donner un aperçu de la formulation quasi-postérieure proposée et à illustrer ses avantages potentiels. Dans le deuxième résultat (Theorem 5.4.13), un postérieur de substitution est appris et utilisé pour comparer les échantillons. Les conditions sous lesquelles le quasi-postérieur ABC résultant converge vers le vrai postérieur sont spécifiées.

Convergence du quasi-postérieur ABC

Dans cette section, nous supposons un observateur donné fixe $\mathbf{y}$ et la dépendance à l'égard de $\mathbf{y}$ est omise de la notation, lorsqu'il n'y a pas de confusion.

Rappelons d'abord la forme standard du quasi-postérieur ABC, en omettant les statistiques sommaires dans la notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z}\ . \tag{5.4.31}$$

Si D est une distance et $D(\mathbf{y},\mathbf{z})$ est continu dans $\mathbf{z}$, on peut montrer que le postérieur ABC dans (5.4.31) a la propriété souhaitable de converger vers le vrai postérieur lorsque ϵ tend vers 0 (see Prangle et al., 2018).

La preuve repose sur le fait que lorsque ϵ tend vers 0, en raison de la propriété de la distance D, l'ensemble $\{\mathbf{z}\in\mathcal{Y} : D(\mathbf{y},\mathbf{z})\leq\epsilon\}$, définissant la fonction indicatrice dans (5.4.31), tend vers le singleton $\{\mathbf{y}\}$ de sorte que, par conséquent, $\mathbf{z}$ dans la vraisemblance peut être remplacé par le $\mathbf{y}$ observé, ce qui conduit alors à un quasi-postérieur ABC proportionnel à $\pi(\boldsymbol{\theta})f_{\boldsymbol{\theta}}(\mathbf{y})$ et donc au vrai postérieur comme souhaité (voir aussi Rubio & Johansen, 2013, Bernton et al., 2019). Il est intéressant de noter que cette preuve est basée sur le travail sur le terme sous l'intégrale seulement et utilise l'égalité, à la convergence, de $\mathbf{z}$ à $\mathbf{y}$, qui est en fait une hypothèse plus forte que nécessaire pour que le résultat tienne. Sinon, si nous réécrivons d'abord (5.4.31) en utilisant le théorème de Bayes, il s'ensuit que

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbf{1}_{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta})\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z}\ \propto \int_{\mathcal{Y}} \mathbf{1}_{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}\ . \tag{5.4.32}$$

C'est-à-dire, en tenant compte de la constante de normalisation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{D(\mathbf{y},\mathbf{z})\leq\epsilon\}}\ \pi(\mathbf{z})\ d\mathbf{z}}\ . \tag{5.4.33}$$

En utilisant cette formulation équivalente, nous pouvons alors remplacer $D(\mathbf{y},\mathbf{z})$ par $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, avec D désignant maintenant une distance sur les densités, et obtenir le même résultat de convergence lorsque ϵ tend vers 0. Plus précisément, nous pouvons montrer le résultat général suivant. Définissons notre quasi-postérieur ABC comme,

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z},$$

qui peut s'écrire comme suit

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}}\pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right)\pi\left(\mathbf{z}\right)d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}}\pi\left(\mathbf{z}\right)d\mathbf{z}}\ . \tag{5.4.34}$$

Le théorème suivant montre que $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ converge vers $\pi\left(\cdot \mid \mathbf{y}\right)$ en variation totale, pour un $\mathbf{y}$ fixe. La preuve est détaillée dans Section 2.4.6.1.

Theorem 5.4.12. *Pour chaque $\epsilon > 0$, soit $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})) \leq \epsilon\}$. Supposons ce qui suit:*

(A1) $\pi(\boldsymbol{\theta} \mid \cdot)$ est continue pour tout $\boldsymbol{\theta} \in \Theta$, et $\sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \mathbf{y}) < \infty$;

(A2) Il existe un $\gamma > 0$ de telle sorte que $\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty$;

(A3) $D(\cdot, \cdot) : \Pi \times \Pi \to \mathbb{R}_+$ est une métrique sur la classe fonctionnelle $\Pi = \{\pi(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}\}$;

(A4) $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$ est continue, par rapport à $\mathbf{z}$.

Sous (A1)–(A4), $q_\epsilon(\cdot \mid \mathbf{y})$ in (5.4.34) converge en variation totale vers $\pi(\cdot \mid \mathbf{y})$, pour fixe $\mathbf{y}$, comme $\epsilon \to 0$.

Il apparaît que l'important n'est pas de choisir des $\mathbf{z}$ proches (et à la limite égaux) aux $\mathbf{y}$ observés mais de choisir des $\mathbf{z}$ tels que la postérieure $\pi(\cdot \mid \mathbf{z})$ (le terme apparaissant dans l'intégrale dans (5.4.32)) est proche (et à la limite égal) à $\pi(\cdot \mid \mathbf{y})$. Et cette dernière propriété est moins exigeante que $\mathbf{z} = \mathbf{y}$. Potentiellement, il peut y avoir plusieurs $\mathbf{z}$ satisfaisant $\pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})$, mais cela n'est pas problématique lorsque l'on utilise (5.4.32), alors que cela est problématique lorsque l'on suit la preuve standard comme dans Bernton et al. (2019).

Convergence du quasi-postérieur ABC avec des postérieurs de substitution

Dans la plupart des contextes ABC, basés sur la divergence des données ou les statistiques sommaires, la considération et le résultat ci-dessus ne sont pas utiles car la vraie postérieure est inconnue par construction et ne peut pas être utilisée pour comparer des échantillons. Cependant, ce principe devient utile dans notre cadre, qui est basé sur des postérieurs de substitution. Bien que le résultat précédent puisse être considéré comme une sorte d'oracle, il est plus intéressant en pratique d'étudier si un résultat similaire est valable lors de l'utilisation de postérieurs de substitution dans la vraisemblance ABC. C'est l'objectif de Theorem 5.4.13 ci-dessous, que nous prouvons pour une classe restreinte de distribution cible et de postérieurs de substitution qui sont appris comme des mélanges. Une preuve détaillée est fournie dans Section 2.4.6.2.

Nous supposons maintenant que $\mathcal{X} = \Theta \times \mathcal{Y}$ est un ensemble compact et considérons la classe suivante $\mathcal{H}_\mathcal{X}$ de gaussiennes isotropes sur $\mathcal{X}$, $\mathcal{H}_\mathcal{X} = \{g_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \Psi\}$, avec contraintes sur les paramètres, Ψ étant un ensemble de paramètres bornés. De plus, les densités dans $\mathcal{H}_\mathcal{X}$ sont supposées satisfaire pour tout $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$, il existe des scalaires positifs arbitraires a, b et B tels que

$$\text{pour tout } \mathbf{x} \in \mathcal{X},\ a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \text{ et } \sup_{\mathbf{x} \in \mathcal{X}} |\log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x})| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1 .$$

Il est montré dans Deleforge et al. (2015c) qu'un modèle GLLiM est un mélange gaussien conjoint sur $\Theta \times \mathcal{Y}$ sous une paramétrisation spécifique. Nous désignons par p^K un mélange à K composantes de distributions de $\mathcal{H}_\mathcal{X}$ et défini pour tous les $\mathbf{y} \in \mathcal{Y}$, $p^{K,N}(\cdot \mid \mathbf{y})$ comme suit:

$$\forall \boldsymbol{\theta} \in \Theta, \quad p^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p^K\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right),$$

avec $\boldsymbol{\phi}^*_{K,N}$ l'estimateur du maximum de vraisemblance (MLE) pour l'ensemble de données $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$, généré à partir de la vraie distribution conjointe $\pi(\cdot, \cdot)$:

$$\boldsymbol{\phi}^*_{K,N} = \arg\max_{\boldsymbol{\phi} \in \Phi} \sum_{n=1}^{N} \log\left(p^K(\boldsymbol{\theta}_n, \mathbf{y}_n; \boldsymbol{\phi})\right).$$

De plus, pour chaque $\epsilon > 0$, soit $A^{K,N}_{\epsilon,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) \leq \epsilon\right\}$ et $q^{K,N}_\epsilon$ désignent le quasi-postérieur ABC défini avec $p^{K,N}$ par

$$q^{K,N}_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_\mathcal{Y} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z} . \tag{5.4.35}$$

Theorem 5.4.13. *Supposons ce qui suit: $\mathcal{X} = \Theta \times \mathcal{Y}$ est un ensemble compact et*

(B1) *Pour une densité jointe π, il existe G_π une mesure de probabilité sur Ψ telle que, avec $g_{\boldsymbol{\varphi}} \in \mathcal{H}_{\mathcal{X}}$, $\pi(\mathbf{x}) = \int_\Psi g_{\boldsymbol{\varphi}}(\mathbf{x})\, G_\pi(d\boldsymbol{\varphi})$;*

(B2) *La densité postérieure vraie $\pi(\cdot \mid \cdot)$ est continue à la fois par rapport à $\boldsymbol{\theta}$ et à $\mathbf{y}$;*

(B3) *$D(\cdot,\cdot) : \Pi \times \Pi \to \mathbb{R}_+ \cup \{0\}$ est une métrique sur une classe fonctionnelle Π, qui contient la classe $\left\{p^{K,N}(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}^\star, N \in \mathbb{N}^\star\right\}$. En particulier, $D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) = 0$, si et seulement si $p^{K,N}(\cdot \mid \mathbf{y}) = p^{K,N}(\cdot \mid \mathbf{z})$;*

(B4) *Pour chaque $\mathbf{y} \in \mathcal{Y}$, $\mathbf{z} \mapsto D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right)$ est une fonction continue sur $\mathcal{Y}$.*

Alors, sous (B1)–(B4), la distance de Hellinger $D_{\mathrm{H}}\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converge vers 0 dans une certaine mesure λ, par rapport à $\mathbf{y} \in \mathcal{Y}$ et en probabilité, par rapport à l'échantillon $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$. C'est-à-dire que pour tout $\alpha > 0, \beta > 0$, il s'avère que

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1. \tag{5.4.36}$$

Dans Section 2.4, notre deuxième contribution est de proposer de comparer directement les distributions postérieures de substitution complètes fournies par GLLiM, sans les réduire à leurs moments. Ce faisant, nous introduisons l'idée de statistiques sommaires fonctionnelles, qui nécessitent également une notion différente des mesures de distance ou de divergence habituelles. Les développements récents des distances optimales basées sur le transport et conçues pour les mélanges gaussiens (Delon & Desolneux, 2020, Chen et al., 2019) répondent parfaitement à ce besoin via la distance dite de *Mixture-Wasserstein* telle que mentionnée dans Delon & Desolneux (2020), et dénotée dans le texte comme MW_2. Il existe d'autres distances entre les mélanges qui sont traitables, et parmi elles la distance L_2 est également considérée dans ce travail.

Comme alternative à l'ABC semi-automatique, dans les travaux de Nguyen et al. (2020a), Jiang et al. (2018), Bernton et al. (2019), Park et al. (2016), Gutmann et al. (2018), les difficultés associées à la recherche de statistiques sommaires efficaces ont été contournées en adoptant, respectivement, la distance énergétique, un estimateur de divergence de Kullback–Leibler, la distance de Wasserstein, la divergence moyenne maximale (MMD) et la précision de la classification pour fournir une mesure de divergence des données. Ces approches comparent les données simulées et les données observées en les considérant comme des échantillons *i.i.d.* de distributions, respectivement liées au paramètre simulé et au paramètre vrai, à l'exception de Bernton et al. (2019) et Gutmann et al. (2018) qui ont proposé des solutions pour traiter également les séries temporelles. Nous soupçonnons que pour être efficaces, ces méthodes nécessitent que les données observées et simulées contiennent chacune un nombre modérément élevé d'échantillons. Typiquement, elles ne peuvent pas être appliquées si nous n'observons qu'un seul échantillon limité lié au paramètre à récupérer. C'est une différence majeure avec l'approche que nous proposons.

Nous proposons de ne pas comparer les échantillons des distributions, mais de comparer directement les distributions par leurs substituts en utilisant les distances entre les distributions. Il est toujours possible d'utiliser les écarts de données précédents en simulant les premiers échantillons des distributions à comparer, mais cela risque d'être sous-optimal sur le plan informatique. Nous pouvons à la place utiliser les mêmes divergences de Wasserstein, Kullback–Leibler, *etc.*, mais dans leurs versions *population* plutôt que dans leurs versions empiriques. À titre d'exemple une distance basée sur Wasserstein peut être calculée entre des mélanges de gaussiens, grâce aux travaux récents de Delon & Desolneux (2020) et Chen et al. (2019). Notez qu'il ne s'agit pas à proprement parler de la distance de Wasserstein, mais d'une distance basée sur Wasserstein. Des expressions sous forme fermée existent également pour la distance L_2, pour la MMD avec un noyau RBF gaussien ou un noyau polynomial (see Sriperumbudur et al., 2010, Muandet et al., 2012) et pour la divergence de Jensen–Rényi de degré deux (see Wang et al., 2009). Kristan et al. (2011) ont également proposé un algorithme basé sur ce que l'on appelle la transformée ascendante afin de calculer la distance de Hellinger entre deux mélanges gaussiens, bien que la complexité de cet algorithme ne soit pas claire.

Pour souligner la différence avec les résumés plus standard, nous nous référons à nos a posteriori de substitution comme à des statistiques sommaires fonctionnelles. Le terme a déjà été utilisé par

Soubeyrand et al. (2013) dans le contexte ABC dans leurs tentatives de caractériser les structures spatiales (*e.g.* processus ponctuels spatiaux) en utilisant des statistiques qui sont des fonctions (*e.g.* corrélogrammes ou variogrammes). Leur approche est différente dans l'esprit car elle n'aborde pas la question du choix des statistiques sommaires. Étant donné certaines statistiques fonctionnelles dont la définition et la nature peuvent changer pour chaque modèle considéré, leur objectif est d'optimiser les distances pour les comparer afin d'extraire la meilleure information sur les paramètres d'intérêt. Soubeyrand et al. (2013) proposent une distance L_2 pondérée pour comparer de telles statistiques. Dans notre proposition, les statistiques fonctionnelles sont des distributions de probabilité. Elles apparaissent comme un moyen de contourner le choix des statistiques sommaires, mais dans ce travail, nous utilisons des métriques existantes pour les comparer, sans optimisation.

Contribution du Chapitre 3

Dans le Chapter 3, nous fournissons une sélection de modèle non-asymptotique pour une variété de modèles de régression MoE, dans des scénarios de grande dimension, basée sur une stratégie de régression inverse. En particulier, ces résultats fournissent une garantie théorique solide: une inégalité d'oracle en échantillon fini satisfaite par l'estimateur du maximum de vraisemblance pénalisé avec une perte de type Jensen-Kullback-Leibler, pour soutenir le critère heuristique de pente dans un cadre d'échantillon fini, par rapport aux critères asymptotiques classiques. Cela permet de calibrer les fonctions de pénalité, connues jusqu'à une constante multiplicative, et à la complexité de la (sous-)collection aléatoire considérée de modèles MoE, y compris le nombre de composantes du mélange, le degré des fonctions moyennes polynomiales et les structures diagonales en bloc cachées potentielles des matrices de covariance de la variable prédicteur ou de la variable réponse multivariée.

En particulier, les travaux de Chapter 3 constituent nos deuxièmes contributions principales pour la sélection non-asymptotique de modèles dans un modèle de régression GLoME et un modèle de régression BLoME des travaux:

(C5) **TrungTin Nguyen**, Hien Duy Nguyen, Faicel Chamroukhi, and Florence Forbes. *A non-asymptotic penalization criterion for model selection in mixture of experts models.*
arXiv preprint arXiv:2104.02640. Under revision, Electronic Journal of Statistics, 2021.
Link: https://arxiv.org/pdf/2104.02640.pdf
(Nguyen et al., 2021c).

(C6) **TrungTin Nguyen**, Faicel Chamroukhi, Hien Duy Nguyen, and Florence Forbes. *Non-asymptotic model selection in block-diagonal mixture of polynomial experts models.*
arXiv preprint arXiv:2104.08959. Under revision, Journal of Multivariate Analysis, 2021.
Link: https://arxiv.org/pdf/2104.08959.pdf
(Nguyen et al., 2021b).

Ici et par la suite, afin d'établir nos inégalités d'oracle, nous devons supposer que l'espace d'entrée est un ensemble limité et expliciter certaines conditions classiques de limitation sur l'espace des paramètres.

Le but de la section suivante était d'expliquer plus en détail pourquoi nous devrions accorder plus d'attention à la borne supérieure non asymptotique en présentant les inconvénients de l'analyse asymptotique d'un modèle paramétrique.

Analyse asymptotique d'un modèle paramétrique

Nous prouvons qu'il est naturel d'essayer de relier la procédure de sélection de modèle non-asymptotique, en particulier l'heuristique de pente, aux inégalités d'oracle.

Nous devons maintenant spécifier nos critères de *goodness*. Dans l'approche du maximum de vraisemblance, la divergence de Kullback–Leibler est la fonction de perte la plus naturelle, qui est définie pour deux densités s et t par

$$\mathrm{KL}(s,t) = \begin{cases} \int_{\mathbb{R}^D} \ln\left(\frac{s(\mathbf{y})}{t(\mathbf{y})}\right) s(\mathbf{y}) d\mathbf{y} & \text{si } s d\mathbf{y} \text{ est absolument continue w.r.t. } t d\mathbf{y}, \\ +\infty & \text{autrement.} \end{cases}$$

Cependant, pour prendre en compte la structure des densités conditionnelles inverses inverses et des covariables aléatoires $\mathbf{Y}_{[n]}$, nous considérons la *divergence de Kullback-Leibler tensorisée* $\mathrm{KL}^{\otimes \mathrm{n}}$, définie comme:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right], \tag{5.4.37}$$

si *sdy* est absolument continu w.r.t. *tdy*, et $+\infty$ sinon. Notez que si les prédicteurs sont fixes, cette divergence est la divergence classique de type plan fixe dans laquelle il n'y a pas d'espérance. Nous appelons notre résultat une *inégalité d'oracle faible*, car son énoncé est basé sur une divergence plus petite, comparée à $\mathrm{KL}^{\otimes \mathrm{n}}$, à savoir la *divergence de Jensen-Kullback-Leibler tensorisée*:

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), (1-\rho)\, s\left(\cdot|\mathbf{Y}_i\right) + \rho t\left(\cdot|\mathbf{Y}_i\right)\right)\right], \tag{5.4.38}$$

avec $\rho \in (0,1)$. Nous notons que $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ a été utilisé pour la première fois dans Cohen & Le Pennec (2011). Cependant, une version de cette divergence apparaît explicitement avec $\rho = \frac{1}{2}$ dans Massart (2007), et on la trouve aussi implicitement dans Birgé et al. (1998). Cette perte est toujours bornée par $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ mais se comporte comme $\mathrm{KL}^{\otimes \mathrm{n}}$, lorsque t est proche de s. Les principaux outils de la preuve d'une telle inégalité d'oracle faible sont les inégalités de déviation pour les sommes de variables aléatoires et leurs suprêmes. Ces outils nécessitent une hypothèse de bornage sur les fonctions contrôlées qui n'est pas satisfaite par $-\ln\frac{s_{\mathbf{m}}}{s_0}$, et donc pas non plus satisfaite par $\mathrm{KL}^{\otimes \mathrm{n}}$. Par conséquent, nous considérons plutôt l'utilisation de $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$. En particulier, en général, il est vrai que $C_\rho\, \mathrm{d}^{2\otimes \mathrm{n}} \leq \mathrm{JKL}_{\rho}^{\otimes \mathrm{n}} \leq \mathrm{KL}^{\otimes \mathrm{n}}$, où $C_\rho = \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho}, 1\right)\left(\ln\left(1+\frac{\rho}{1-\rho}\right) - \rho\right)$ (voir Cohen & Le Pennec 2011, Prop. 1) et $\mathrm{d}^{2\otimes \mathrm{n}}$ est une extension tensorisée de la distance de Hellinger au carré $\mathrm{d}^{2\otimes \mathrm{n}}$, définie par

$$\mathrm{d}^{2\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}^2\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right].$$

De plus, si nous supposons que, pour tout $\mathbf{m} \in \mathcal{M}$ et tout $s_{\mathbf{m}} \in S_{\mathbf{m}}, s_0 d\lambda \ll s_{\mathbf{m}} d\lambda$, alors (voir Montuelle et al., 2014, Cohen & Le Pennec, 2011)

$$\frac{C_\rho}{2 + \ln\left\|s_0/s_{\mathbf{m}}\right\|_\infty}\,\mathrm{KL}^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}) \leq \mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s_0, s_{\mathbf{m}}). \tag{5.4.39}$$

Nous considérerons un modèle paramétrique de FDP conditionnelles inverses auquel la vraie FDP conditionnelle inverse s_0 n'appartient pas nécessairement comme suit:

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{\mathbf{m}}}(\mathbf{x}|\mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}} \subset \mathbb{R}^{\dim(S_{\mathbf{m}})}\right\}.$$

Cette construction de *modèle mal spécifié, i.e.,* $s_0 \notin S_{\mathbf{m}}$, remonte aux travaux de White (1982) pour l'estimation de densité. Pour un traitement d'un cas plus général pour les FDP conditionnels, nous renvoyons le lecteur à Cohen & Le Pennec (2011). Section 5.4 contient un bref résumé de ces résultats classiques sans preuves via Theorem 5.4.14.

Theorem 5.4.14 (White, 1982, Cohen & Le Pennec, 2011)**.** *Supposons que le modèle $S_{\mathbf{m}}$ soit identifiable (pour les modèles de régression MoE, voir,* e.g., *Jiang & Tanner, 1999c, Hennig, 2000) et qu'il existe les* $\dim(S_{\mathbf{m}}) \times \dim(S_{\mathbf{m}})$ *matrices* $\mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}})$ *et* $\mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}})$ *définies par:*

$$\left[\mathbf{A}(\boldsymbol{\psi}_{\mathbf{m}})\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{-\partial^2 \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},k}\partial\psi_{\mathbf{m},l}}(\mathbf{x}|\mathbf{Y}_i)\, s_0(\mathbf{x}|\mathbf{Y}_i)\, d\lambda\right],$$

$$\left[\mathbf{B}(\boldsymbol{\psi}_{\mathbf{m}})\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},k}}(\mathbf{x}|\mathbf{Y}_i)\frac{\partial \ln s_{\boldsymbol{\psi}_{\mathbf{m}}}}{\partial\psi_{\mathbf{m},l}}(\mathbf{x}|\mathbf{Y}_i)\, s_0(\mathbf{x}|\mathbf{Y}_i)\, d\lambda\right].$$

Nous définissons $\boldsymbol{\psi}_{\mathbf{m}}^{*}$ *comme étant les éléments de* $\arg\min_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}}\right)$. *Alors, sous certaines hypothèses de régularité forte sur* $\boldsymbol{\psi}_{\mathbf{m}} \mapsto s_{\boldsymbol{\psi}_{\mathbf{m}}}$, $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ *est asymptotiquement équivalent à*

$$\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}^{*}}\right) + \frac{1}{2n} tr\left(\mathbf{B}\left(\boldsymbol{\psi}_{\mathbf{m}}^{*}\right)\mathbf{A}\left(\boldsymbol{\psi}_{\mathbf{m}}^{*}\right)^{-1}\right).$$

En particulier, lorsque $s_0 \in S_{\mathbf{m}}$, *il s'avère que* $s_0 = s_{\boldsymbol{\psi}_{\mathbf{m}}^{*}}, \mathbf{A}\left(\boldsymbol{\psi}_{\mathbf{m}}^{*}\right) = \mathbf{B}\left(\boldsymbol{\psi}_{\mathbf{m}}^{*}\right)$. *Par conséquent, l'équivalent asymptotique précédent de* $\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ *devient l'équivalent paramétrique classique, c'est-à-dire,*

$$\underbrace{\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}^{*}}\right)}_{=0} + \frac{1}{2n}\dim\left(S_{\mathbf{m}}\right).$$

Theorem 5.4.14 dépend fortement de la normalité asymptotique de $\sqrt{n}\left(\widehat{\boldsymbol{\psi}}_{\mathbf{m}} - \boldsymbol{\psi}_{\mathbf{m}}^{*}\right)$. On peut se demander si cela est toujours vrai si cette normalité ne tient pas. Plusieurs travaux sont consacrés à l'étude de la normalité non-asymptotique: extension au cas non paramétrique ou modèle non identifiable, souvent appelé phénomène de Wilk (voir Wilks, 1938 pour plus de détails); généralisation du "Chi-Square goodness-of-fit test" correspondant (Fan et al., 2001); écart en échantillon fini de la quantité empirique correspondante dans un cadre de perte bornée (Boucheron & Massart, 2011). Motivés par les travaux de Cohen & Le Pennec (2011, 2013), avec aussi peu d'hypothèses que possible sur la collection de FDP conditionnelles $S_{\mathbf{m}}$, nous sommes initialement intéressés par la recherche d'une borne supérieure non asymptotique de type

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq \left(\inf_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}}\right) + \frac{1}{2n}\dim\left(S_{\mathbf{m}}\right)\right) + C_2\frac{1}{n}.$$

Cependant, en réalité, nous avons obtenu la borne supérieure plus faible suivante (inégalités d'oracle)

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq C_1\left(\inf_{\boldsymbol{\psi}_{\mathbf{m}}\in\boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\boldsymbol{\psi}_{\mathbf{m}}}\right) + \frac{\kappa}{n}\mathfrak{D}_{\mathbf{m}}\right) + C_2\frac{1}{n}.$$

En effet, par problème technique, nous devons remplacer la divergence de gauche $\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)$ par une divergence plus petite $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)$ et la constante $C_1 = 1+\epsilon, \epsilon > 0$, ne peut pas être égale à 1. De plus, κ est une constante qui dépend de ϵ et le terme de complexité du modèle $\mathfrak{D}_{\mathbf{m}}$ remplace le terme de dimension $\dim\left(S_{\mathbf{m}}\right)$. Cependant, ce résultat nous permet d'avoir la bonne saveur du compromis biais/variance et de récupérer les propriétés minimax habituelles des estimateurs spécifiques.

Ici et par la suite, afin d'établir nos inégalités d'oracle, nous devons supposer que l'espace d'entrée est un ensemble borné et rendre explicites certaines conditions de bornage classiques sur l'espace des paramètres.

Inégalité d'Oracle pour les modèles GLoME

Dans les modèles de régression GLoME, nous choisissons le degré des polynômes d et le nombre de composantes K parmi des ensembles finis $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ et $\mathcal{K} = [K_{\max}]$, respectivement, où $d_{\max} \in \mathbb{N}^{\star}$ et $K_{\max} \in \mathbb{N}^{\star}$ peuvent dépendre de la taille de l'échantillon n. Nous souhaitons estimer la densité conditionnelle inverse inconnue s_0 par des densités conditionnelles appartenant à la collection suivante de modèles inverses $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \{(K,d) : K \in \mathcal{K}, d \in\in \mathcal{D}_{\boldsymbol{\Upsilon}}\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\right\}. \tag{5.4.40}$$

Ici, $\widetilde{\boldsymbol{\Omega}}_K$ sont des vecteurs de paramètres de gaussiennes normalisées bornés, $\boldsymbol{\Upsilon}_{K,d}$ est défini comme une combinaison linéaire d'un ensemble fini de fonctions bornées dont les coefficients appartiennent à un ensemble compact, et $\mathbf{V}_K$ sont des matrices de covariance définies positives bornées, voir (5.4.41), (5.4.43) (ou plus généralement (5.4.42)), et (5.4.44), respectivement, pour plus de détails.

Plus précisément, supposons qu'il existe des constantes positives déterministes $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}}, A_{\boldsymbol{\Gamma}}$, $\widetilde{\boldsymbol{\Omega}}_K$ est défini par

$$\widetilde{\boldsymbol{\Omega}}_K = \{\boldsymbol{\omega} \in \boldsymbol{\Omega}_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}_k) \leq M(\boldsymbol{\Gamma}_k) \leq A_{\boldsymbol{\Gamma}}, a_{\boldsymbol{\pi}} \leq \pi_k\}, \tag{5.4.41}$$

où $m(\mathbf{A})$ et $M(\mathbf{A})$ représentent, respectivement, le module de la plus petite et de la plus grande valeur propre de toute matrice $\mathbf{A}$. Suivant la même structure pour les moyennes des experts gaussiens de Montuelle et al. (2014), l'ensemble $\boldsymbol{\Upsilon}_{K,d}$ sera choisi comme un produit tensoriel d'ensembles compacts de dimension modérée (*e.g.,* un ensemble de polynômes de degré inférieur à d, dont les coefficients sont plus petits en valeur absolue que $T_{\boldsymbol{\Upsilon}}$). Plus précisément, $\boldsymbol{\Upsilon}_{K,d} = \otimes_{k\in[K]} \boldsymbol{\Upsilon}_{k,d} =: \boldsymbol{\Upsilon}_{k,d}^K$, où $\boldsymbol{\Upsilon}_{k,d} = \boldsymbol{\Upsilon}_{b,d}$, $\forall k \in [K]$, et

$$\boldsymbol{\Upsilon}_{b,d} = \left\{ \mathbf{y} \mapsto \left(\sum_{i=1}^{d} \boldsymbol{\alpha}_i^{(j)} \varphi_{\boldsymbol{\Upsilon},i}(\mathbf{y}) \right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}} \right\}. \tag{5.4.42}$$

Ici, $d \in \mathbb{N}^\star, T_{\boldsymbol{\Upsilon}} \in \mathbb{R}^+$, et $(\varphi_{\boldsymbol{\Upsilon},i})_{i\in[d]}$ est une collection de fonctions bornées sur $\mathcal{Y}$. En particulier, nous nous concentrons sur le cas de $\mathcal{Y}$ borné et supposons que $\mathcal{Y} = [0,1]^L$, sans perte de généralité. Dans ce cas, les $\varphi_{\boldsymbol{\Upsilon},i}$ peuvent être choisis comme des monômes de degré maximal (non négatif) d: $\mathbf{y}^{\mathbf{r}} = \prod_{l=1}^{L} \mathbf{y}_l^{\mathbf{r}_l}$. Rappelons qu'un multi-index $\mathbf{r} = (\mathbf{r}_l)_{l\in[L]}, \mathbf{r}_l \in \mathbb{N}^\star \cup \{0\}, \forall l \in [L]$, est un L-tuple d'entiers non négatifs. Nous définissons $|\mathbf{r}| = \sum_{l=1}^{L} \mathbf{r}_l$ et le nombre $|\mathbf{r}|$ est appelé l'ordre ou le degré de $\mathbf{y}^{\mathbf{r}}$. Alors, $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$, où

$$\boldsymbol{\Upsilon}_{p,d} = \left\{ \mathbf{y} \mapsto \left(\sum_{|\mathbf{r}|=0}^{d} \boldsymbol{\alpha}_{\mathbf{r}}^{(j)} \mathbf{y}^{\mathbf{r}} \right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}} \right\}. \tag{5.4.43}$$

Notez que toute matrice de covariance $\boldsymbol{\Sigma}_k$ peut être décomposée sous la forme $B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top$, telle que $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$ est un scalaire positif correspondant au volume, $\mathbf{P}_k$ est la matrice des vecteurs propres de $\boldsymbol{\Sigma}_k$ et $\mathbf{A}_k$ la matrice diagonale des valeurs propres normalisées de $\boldsymbol{\Sigma}_k$; $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$, $\mathcal{A}(\lambda_-, \lambda_+)$ est un ensemble de matrices diagonales $\mathbf{A}_k$, telles que $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$; et $SO(D)$ est le groupe orthogonal spécial de dimension D. De cette façon, nous obtenons ce que l'on appelle les ensembles classiques de matrices de covariance décrits par Celeux & Govaert (1995) pour les modèles gaussiens de clustering parcimonieux, définis par

$$\mathbf{V}_K = \left\{ \left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top \right)_{k\in[K]} : \forall k \in [K], B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}(\lambda_-, \lambda_+) \right\}. \tag{5.4.44}$$

La Theorem 5.4.15 suivante fournit une borne inférieure sur la fonction de pénalité, $\text{pen}(\mathbf{m})$, qui garantit que le PMLE pour les modèles GLoME sélectionne un modèle qui est presque aussi performant que le "meilleur" modèle. Les preuves détaillées apparaîtront dans Section 3.2.4, voir aussi Nguyen et al. (2021c).

Theorem 5.4.15 (Inégalité Oracle pour les modèles GLoME)**.** *Supposons que nous observions* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, provenant d'une densité conditionnelle inconnue* s_0*. Étant donné une collection de modèles GLoME,* $\mathcal{S} = (S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$*, définie par* (5.4.40)*, il existe une constante* C *telle que pour toute* $\rho \in (0,1)$*, pour toute* $\mathbf{m} \in \mathcal{M}$*,* $z_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$ *et tout* $C_1 > 1$*, il existe une constante* κ_0 *dépendant uniquement de* ρ *et de* C_1*, telle que si pour tout indice* $\mathbf{m} \in \mathcal{M}$*,* $\text{pen}(\mathbf{m}) \geq \kappa \left[(C + \ln n) \dim(S_{\mathbf{m}}) + z_{\mathbf{m}} \right]$ *avec* $\kappa > \kappa_0$*, alors l'estimateur de vraisemblance pénalisé par* η' *est le suivant* $\widehat{s}_{\widehat{\mathbf{m}}}$*, défini dans* (5.4.21)*, satisfait à*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}} \left[\text{JKL}_\rho^{\otimes \text{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}}) \right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}} \left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \text{KL}^{\otimes \text{n}}(s_0, s_{\mathbf{m}}) + \frac{\text{pen}(\mathbf{m})}{n} \right) + \frac{\kappa_0 C_1 \Xi}{n} + \frac{\eta + \eta'}{n}. \tag{5.4.45}$$

Inégalité d'Oracle pour les modèles BLoME

Dans le cadre des modèles BLoME, nous choisissons le degré des polynômes d et le nombre de composantes K parmi des ensembles finis $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ et $\mathcal{K} = [K_{\max}]$, respectivement, où $d_{\max} \in \mathbb{N}^\star$ et $K_{\max} \in \mathbb{N}^\star$ peuvent dépendre de la taille de l'échantillon n. De plus, $\mathbf{B}$ est sélectionné parmi une liste de structures candidates $(\mathcal{B}_k)_{k\in[K]} \equiv (\mathcal{B})_{k\in[K]}$, où $\mathcal{B}$ désigne l'ensemble de toutes les partitions possibles des covariables indexées par $[D]$, pour chaque cluster d'individus. Nous souhaitons estimer la densité conditionnelle inconnue s_0 par des densités conditionnelles appartenant à la collection de modèles suivante: $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \left\{(K, d, \mathbf{B}) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}, \mathbf{B} \in (\mathcal{B})_{k\in[K]}\right\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}(\mathbf{B})) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K(\mathbf{B})\right\}, \tag{5.4.46}$$

où $\widetilde{\boldsymbol{\Omega}}_K$, $\boldsymbol{\Upsilon}_{K,d}$, et $\mathbf{V}_K(\mathbf{B})$ sont définis dans (5.4.41), (5.4.43) (ou plus généralement (5.4.42)), et (5.4.14), respectivement.

Pour les covariances bloc-diagonales des experts gaussiens, nous supposons qu'il existe certaines constantes positives λ_m et λ_M telles que, pour chaque $k \in [K]$,

$$0 < \lambda_m \leq m(\boldsymbol{\Sigma}_k(\mathbf{B}_k)) \leq M(\boldsymbol{\Sigma}_k(\mathbf{B}_k)) \leq \lambda_M. \tag{5.4.47}$$

Notez qu'il s'agit d'une hypothèse assez générale et qu'elle est également utilisée dans la sélection de covariance bloc-diagonale pour les modèles graphiques gaussiens de Devijver & Gallopin (2018).

En théorie, nous aimerions considérer toute la collection de modèles $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$. Cependant, la cardinalité de $\mathcal{B}$ est grande; sa taille est un nombre de Bell. Même pour un nombre modéré de variables D, il n'est pas possible d'explorer l'ensemble $\mathcal{B}$, de manière exhaustive. Nous limitons notre attention à une sous-collection aléatoire $\mathcal{B}^R$ de taille modérée. Par exemple, nous pouvons considérer la procédure BLLiM de Devijver et al. (2017, Section 2.2).

Notez que la collection construite de modèles avec des structures bloc-diagonales pour chaque groupe d'individus est conçue, par exemple, par la procédure BLLiM de Devijver et al. (2017), où chaque collection de partition est triée par niveau de sparsité. Néanmoins, notre inégalité d'oracle à échantillon fini, Theorem 5.4.16, tient toujours pour toute sous-collection aléatoire de $\mathcal{M}$, qui est construite par certains outils appropriés dans le cadre des modèles de régression BLoME. Pour les preuves, nous renvoyons le lecteur à Section 3.3.3, voir aussi Nguyen et al. (2021b).

Theorem 5.4.16 (Inégalité Oracle pour les modèles BLoME)**.** *Soit $(\mathbf{x}_{[n]}, \mathbf{y}_{[n]})$ les observations provenant d'une densité conditionnelle inconnue s_0. Pour chaque $\mathbf{m} = (K, d, \mathbf{B}) \in (\mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} \times \mathcal{B}) \equiv \mathcal{M}$, laissez $S_{\mathbf{m}}$ être défini par (5.4.46). Supposons qu'il existe $\tau > 0$ et $\epsilon_{KL} > 0$ tels que, pour tout $\mathbf{m} \in \mathcal{M}$, on peut trouver $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$, tel que*

$$\mathrm{KL}^{\otimes \mathrm{n}}(s_0, \bar{s}_{\mathbf{m}}) \leq \inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + \frac{\epsilon_{KL}}{n}, \text{ et } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0.$$

Ensuite, nous construisons une sous-collection aléatoire $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ en laissant $\widetilde{\mathcal{M}} \equiv (\mathcal{K} \times \mathcal{D}_{\boldsymbol{\Upsilon}} \times \mathcal{B}^R) \subset \mathcal{M}$ tel que $\mathcal{B}^R$ est une sous-collection aléatoire $\mathcal{B}$, de taille modérée. Considérons la collection $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ de η-log likelihood minimizers satisfaisant (5.4.22) pour tous les $\mathbf{m} \in \widetilde{\mathcal{M}}$. Alors, il existe une constante C telle que pour tout $\rho \in (0, 1)$, et tout $C_1 > 1$, il existe deux constantes κ_0 et C_2 dépendant uniquement de ρ et de C_1 telles que, pour tout indice, $\mathbf{m} \in \mathcal{M}$, $\xi_{\mathbf{m}} \in \mathbb{R}^+$, $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$ et

$$\mathrm{pen}(\mathbf{m}) \geq \kappa\left[(C + \ln n)\dim(S_{\mathbf{m}}) + (1 \vee \tau)\xi_{\mathbf{m}}\right],$$

avec $\kappa > \kappa_0$, l'estimateur de vraisemblance pénalisé $\widehat{s}_{\widehat{\mathbf{m}}}$, défini comme dans (5.4.21) sur le sous-ensemble $\widetilde{\mathcal{M}} \subset \mathcal{M}$, satisfait à

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_\rho^{\otimes \mathrm{n}}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(s_0, t) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta' + \eta}{n}.$$

Contribution du Chapitre 4

Dans le Chapter 4, nous établissons des résultats non asymptotiques de sélection de modèle dans des scénarios de régression à grande dimensio pour SGaME et SGaBloME, en s'appyuant sur une une pénalisation Lasso. Ceux-ci incluent des résultats pour la sélection du nombre de composantes du mélange d'experts, ainsi que pour la sélection jointe de variable et des rangs des matrices de covariances. En particulier, ces résultats fournissent une garantie théorique forte: une inégalité d'oracle en échantillon fini satisfaite par l'estimateur de maximum de vraisemblance pénalisé avec une perte de type Jensen-Kullback-Leibler, pour soutenir l'heuristique de pente dans un cadre d'échantillon fini, par rapport aux critères asymptotiques classiques. Cela permet de calibrer les fonctions de pénalité, connues seulement à une constante multiplicative près, étant donnés complexité de la (sous-)collection aléatoire considérée de modèles MoE, y compris le nombre de composantes du mélange, le degré de sparsité (les coefficients et les niveaux de sparsité des rangs des matrices de covariances), le degré des fonctions moyennes polynomiales, et les structures potentielles de diagonales par bloc cachées des matrices de covariance du prédicteur ou de la réponse multivariée.

Enfin, nos troisièmes contributions principales pour les résultats non asymptotiques pour la sélection jointe de variable et des rangs des matrices de covariances dans un modèle de régression SGaBloME sont fournies via Chapter 4 des travaux:

(C7) **TrungTin Nguyen**, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. *An l_1-oracle inequality for the lasso in mixture of experts regression models.* arXiv preprint arXiv:2009.10622. Under revision, ESAIM: Probability and Statistics, 2020. Link: https://arxiv.org/pdf/2009.10622.pdf (Nguyen et al., 2020c).

(C8) *Joint rank and variable selection by a non-asymptotic model selection in mixture of polynomial experts models.* Working paper.

Notez que ces modèles sont utiles pour les données hétérogènes de grande dimension, où le nombre de variables explicatives peut être beaucoup plus grand que la taille de l'échantillon et où il existe des interactions potentielles cachées de type graphe structuré entre les variables.

Inégalité d'Oracle pour les modèles SGaBloME

Combinaison linéaire de fonctions bornées pour les poids et les moyennes

Nous suivons l'idée de Montuelle et al. (2014) de restreindre notre attention sur un ensemble fini de fonctions bornées dont les coefficients appartiennent à un ensemble compact. Il convient de mentionner que ce cadre assez général inclut la base polynomiale lorsque les prédicteurs sont bornés, les dictionnaires d'ondelettes renormalisés appropriés ainsi que la base de Fourier sur un intervalle. Plus précisément, nous définissons d'abord les deux collections suivantes de fonctions bornées pour les poids et les moyennes: $\mathcal{X} \ni \mathbf{x} \mapsto \left(\theta_{\mathbf{W},d}\left(\mathbf{x}\right)\right)_{d\in[d_{\mathbf{W}}]} \in [-1,1]^{d_{\mathbf{W}}}$ et $\mathcal{X} \ni \mathbf{x} \mapsto \left(\theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right)\right)_{d\in[d_{\boldsymbol{\Upsilon}}]} \in [-1,1]^{d_{\boldsymbol{\Upsilon}}}$, où $d_{\mathbf{W}} \in \mathbb{N}^{\star}$ et $d_{\boldsymbol{\Upsilon}} \in \mathbb{N}^{\star}$ indiquent ses degrés, respectivement. Ensuite, en utilisant ces collections, nous sommes en mesure de définir les espaces bornés souhaités correspondants via des constructions tensorielles comme suit:

$$\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}^{K-1}, \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{d=1}^{d_{\mathbf{W}}} \omega_d \theta_{\mathbf{W},d}\left(\mathbf{x}\right) \in \mathbb{R} : \max_{d\in[d_{\mathbf{W}}]} |\omega_d| \leq T_{\mathbf{W}} \right\},$$

$$\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}} = \boldsymbol{\Upsilon}^{K}, \boldsymbol{\Upsilon} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(\sum_{d=1}^{d_{\boldsymbol{\Upsilon}}} \beta_d^{(z)} \theta_{\boldsymbol{\Upsilon},d}\left(\mathbf{x}\right) \right)_{z\in[q]} : \max_{d\in[d_{\boldsymbol{\Upsilon}}],z\in[q]} \left|\beta_d^{(z)}\right| \leq T_{\boldsymbol{\Upsilon}} \right\}.$$

Lorsque p et q ne sont pas trop grands, nous n'avons pas besoin de sélectionner les variables pertinentes et/ou d'utiliser des modèles à rangs épars. Nous pouvons alors travailler sur les structures précédentes pour les moyennes et les poids comme cela a été fait dans Montuelle et al. (2014), Nguyen et al. (2021c), voir aussi Theorems 5.4.15 and 5.4.16. Cependant, afin de traiter des données de grande

dimension et de simplifier l'interprétation de la sparsité, nous proposons d'utiliser des monômes pour les poids et des modèles de régression polynomiaux pour les fonctions softmax et les moyennes des experts gaussiens, *i.e.*,

$$\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}^{K-1}, \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \max_{\boldsymbol{\alpha} \in \mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}} \right\},$$

$$\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(\boldsymbol{\beta}_{k0} + \sum_{d=1}^{d_{\boldsymbol{\Upsilon}}} \boldsymbol{\beta}_{kd} \mathbf{x}^d \right)_{k \in [K]} : \max \left\{ |||\boldsymbol{\beta}_{kd}|||_\infty : k \in [K], d \in (\{0\} \cup [d_{\boldsymbol{\Upsilon}}]) \right\} \leq T_{\boldsymbol{\Upsilon}} \right\}.$$

Notons ici que le multi-indice $\boldsymbol{\alpha} = (\alpha_t)_{t \in [p]}, \alpha_t \in \mathbb{N}^\star \bigcup \{0\} =: \mathbb{N}, \forall t \in [p]$, est un p-tuple d'entiers non négatifs qui satisfait à $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^p x_j^{\alpha_j}$ et $|\boldsymbol{\alpha}| = \sum_{t=1}^p \alpha_t$. Alors, pour tout $l \in [d_{\mathbf{W}}]$, on définit $\mathcal{A} = \bigcup_{l=0}^{d_{\mathbf{W}}} \mathcal{A}_l$, $\mathcal{A}_l = \left\{ \boldsymbol{\alpha} = (\alpha_t)_{t \in [p]} \in \mathbb{N}^p, |\boldsymbol{\alpha}| = l \right\}$. Le nombre $\boldsymbol{\alpha}$ est appelé l'ordre ou le degré des monômes $\mathbf{x}^{\boldsymbol{\alpha}}$. En utilisant les méthodes bien connues des étoiles et des barres, *e.g.*, Feller (1957, Chapitre 2), la cardinalité de l'ensemble $\mathcal{A}$, noté par card $(\mathcal{A})$, est égale à $\binom{d_{\mathbf{W}}+p}{p}$. Notez que, pour tout $d \in [d_{\boldsymbol{\Upsilon}}]$, on définit $\mathbf{x}^d$ comme $\left(x_j^d\right)_{j \in [p]}$ pour les moyennes, qui sont souvent utilisées pour les modèles de régression polynomiale. De plus, étant donné toute matrice $\mathbf{A} \in \mathbb{R}^{q \times p}$, les notations suivantes sont utilisées pour les normes de matrice: la *norme de médicale* $|||\mathbf{A}|||_\infty = \max_{i \in [q], j \in [p]} \left|[\mathbf{A}]_{i,j}\right|$, la 2-*norm* $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \left|\mathbf{x}^\top \mathbf{A} \mathbf{x}\right| = \sup_{\lambda \in \text{vp}(\mathbf{A})} |\lambda|$, où vp$(\mathbf{A})$ désigne le spectre de $\mathbf{A}$, et la *norme de Frobenius* $\|\mathbf{A}\|_F^2 = \sum_{i=1}^q \sum_{j=1}^p \left|[\mathbf{A}]_{i,j}\right|^2$. Alors, il tient que $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$, $\|\mathbf{A}\|_2 \leq \sqrt{qp} |||\mathbf{A}|||_\infty$, et pour tout $\mathbf{x} \in \mathbb{R}^p$, $\|\mathbf{x}\|_2 \leq \sqrt{p} \|\mathbf{x}\|_\infty$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2$, *e.g.*, Golub & Van Loan (2013, Chapitre 2).

Sélection de variables via la sélection de variables pertinentes

L'estimateur Lasso, établi à l'origine par Tibshirani (1996), est un choix classique pour la sélection de variables et a été étendu pour traiter les modèles de régression multivariés multiples pour la sparsité des colonnes à l'aide de l'estimateur Group-Lasso (Yuan & Lin, 2006). Notez que la pénalité Group-Lasso peut être utilisée pour sélectionner un sous-ensemble de variables pour un choix de paramètre de régularisation paramètre de régularisation dans la procédure Lasso-Rank, comme fait, (Devijver, 2015b, 2017a,b) ou pour obtenir un classement des variables, comme fait, par exemple, dans Bach (2008).

Rappelons que, pour tout $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $\boldsymbol{\beta}_{kd}$ est la matrice du d-ième terme des coefficients de régression, $\boldsymbol{\Sigma}_k$ $(\mathbf{B}_k)$ est la matrice de covariance dans la composante du mélange k, et le g_k est la proportion du mélange k dont le terme de $\boldsymbol{\alpha}$-ième ordre de ses monômes est $\omega_{k\boldsymbol{\alpha}}$. De plus, étant donné un régresseur $\mathbf{x}$, pour tout $k \in [K]$, pour tout $d \in [d_{\boldsymbol{\Upsilon}}]$ et pour tout $z \in [q]$, $\left[\boldsymbol{\beta}_{kd} \mathbf{x}^d\right]_z = \sum_{j=1}^p \left[\boldsymbol{\beta}_{kd}\right]_{z,j} x_j^d$ est la z-ième composante des d-ième termes de moyennes pour les composantes du mélange k. En particulier, pour tous les $l \in [d_{\mathbf{W}}]$, $j \in [p]$, nous définissons $\boldsymbol{\omega}_k^{[j,l]} = \left\{ \omega_{k\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} = (\alpha_t)_{t \in [p]} \mathcal{A}_l, \boldsymbol{\alpha}_j > 0 \right\}$.

Nous devons traiter des données de grande dimension où nous estimons de nombreux coefficients tout en ayant un petit nombre de variables cibles. Par conséquent, nous devons nous concentrer sur la sélection des variables pertinentes via la notion d'indices non pertinents dans Definition 5.4.17.

Definition 5.4.17 (Variables pertinentes dans les modèles SGaBloME)**.** . Un couple $(\mathbf{Y}_z, \mathbf{X}_j)$ et ses indices correspondants $(z, j) \in [q] \times [p]$ sont dits *irrelevant* si, pour tout $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, $l \in [d_{\mathbf{W}}]$, $[\boldsymbol{\beta}_{kd}]_{z,j} = 0, \boldsymbol{\omega}_k^{[j,l]} = \mathbf{0}$. Cela signifie que la variable $\mathbf{X}_j$ n'explique pas la variable $\mathbf{Y}_z$ pour les modèles de régression. Un couple et ses indices correspondants sont pertinents s'ils ne sont pas non pertinents. On dit d'un modèle qu'il est clairsemé s'il y a peu de variables pertinentes. Nous désignons par $\mathbf{J}$ l'ensemble des indices (z, j) des couples pertinents $(\mathbf{Y}_z, \mathbf{X}_j)$. Ensuite, nous définissons l'ensemble des variables pertinentes (colonnes) comme $\mathbf{J}_{\boldsymbol{\omega}} = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$. Nous désignons par $\mathbf{A}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ et $\mathbf{b}^{[\mathbf{J}_{\boldsymbol{\omega}}]}$ la matrice et le vecteur avec des vecteurs $\mathbf{0}$ sur les colonnes indexées par l'ensemble $\mathbf{J}_{\boldsymbol{\omega}}^C$ et des valeurs 0 sur l'ensemble $\mathbf{J}_{\boldsymbol{\omega}}^C$, respectivement. Ici, $\mathbf{J}_{\boldsymbol{\omega}}^C$ est le complément de l'ensemble $\mathbf{J}_{\boldsymbol{\omega}}$.

Remarquez que $\mathbf{J} \subset \mathcal{P}([q] \times [p])$ and $\mathbf{J}_{\boldsymbol{\omega}} \subset \mathcal{P}([q])$, où $\mathcal{P}([q] \times [p])$ contient tous les sous-ensembles de $[q] \times [p]$.

Dans notre contexte, nous nous concentrons sur l'estimateur de Group-Lasso pour détecter les variables pertinentes, où les groupes correspondent aux colonnes. Par conséquent, si pour tous les $k \in [K]$, $d \in [d_{\boldsymbol{\Upsilon}}]$, une matrice $\boldsymbol{\beta}_{kd}$ possède $\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})$ colonnes pertinentes, il y a $q \operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})$ coefficients à estimer au lieu de qp par groupes et matrices de coefficients. Le nombre de paramètres à estimer est alors considérablement réduit lorsque $\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}}) \ll p$. De plus, une telle sparsité de colonnes peut améliorer l'interprétation puisque les réponses sont décrites par seulement quelques colonnes pertinentes. Pour construire la régularisation des coefficients des fonctions polynomiales, nous pouvons considérer l'estimateur clairsemé de Group-Lasso de Simon et al. (2013) et Hastie et al. (2015, Chapitre 4).

Modélisation à faible rang et éparse

Cette approche est basée sur les modèles à rangs épars, introduits par Anderson et al. (1998). Plus précisément, si les matrices de régression ont un rang faible ou du moins peuvent être bien approximées par des matrices de faible rang, alors les modèles de régression correspondants sont dits de rang clairsemé. Dans le modèle SGaBloME, pour chaque $k \in [K]$, $d \in [L]$, la matrice $\boldsymbol{\beta}_{kd}$ est entièrement déterminée par $R_{kd}(p + q - R_{kd})$ coefficients si elle a un rang R_{kd}. Cet avantage sera très utile car le total des paramètres à estimer peut être inférieur à la taille de l'échantillon nq. Il convient de noter que cette estimation de rang faible généralise l'analyse classique en composantes principales pour réduire la dimension des données multivariées et apparaît dans de nombreuses applications: *e.g.,* Friston et al. (2003, 2019, analyse des données d'image IRMf), Anderson et al. (1998, analyse du décodage des données EEG).

En combinant la sparsité de rang et de colonne précédente, on considère les matrices de coefficients de régression $\boldsymbol{\beta}_{kd}$ de rang R_{kd} et un vecteur de rangs $\mathbf{R} = (R_{kd})_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}]}$ appartient à $[\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{d_{\boldsymbol{\Upsilon}}K}$, où en général, $a \wedge b = \min(a, b)$ et $a \vee b = \max(a, b)$.

Nous décrivons plus en détail la collection de modèles SGaBloME avec des variables pertinentes et des modèles à rangs épars dans la suite.

Collection de modèles

Pour simplifier les notations, L et D représentent $\binom{d_{\mathbf{W}}+\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})}{\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}})}$ et $d_{\boldsymbol{\Upsilon}}$, qui sont liées aux dimensions de $\mathbf{W}_{K,d_{\mathbf{W}}}$ et $\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}}$, respectivement. En combinant toutes les structures précédentes définies dans étant donné $\mathbf{m} = (K, L, D, \mathbf{B}, \mathbf{J}, \mathbf{R}) \in \mathbb{N}^\star \times \mathbb{N}^\star\mathbb{N}^\star \times (\mathcal{B}_k)_{k\in[K]} \times \mathcal{P}([q] \times [p]) \times [\operatorname{card}(\mathbf{J}_{\boldsymbol{\omega}}) \wedge q]^{DK}$, quelques constantes positives réelles $A_{\mathbf{u},\mathbf{v}} > 0$, $A_\sigma > 0$, on obtient le modèle suivant:

$$
\begin{aligned}
S_{\mathbf{m}} &= \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}}(\mathbf{y}|\mathbf{x}) =: s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \in \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} \right\}, \\
\boldsymbol{\psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \left((\omega_{k\boldsymbol{\alpha}})_{k\in[K],\boldsymbol{\alpha}\in\mathcal{A}}^{[\mathbf{J}_{\boldsymbol{\omega}}]}, \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]} \right)_{k\in[K]}, (\boldsymbol{\Sigma}_k(\mathbf{B}_k))_{k\in[K]} \right) \\
&\in \left(\mathbb{R}^L\right)^{K-1} \times \boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} \times \mathbf{V}_K(\mathbf{B}) =: \boldsymbol{\Psi}_{(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R})}, \\
\boldsymbol{\Upsilon}_{(K,D,\mathbf{B},\mathbf{J},\mathbf{R})} &= \Bigg\{ \left(\boldsymbol{\beta}_{k0}, \left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right)_{d\in[D]}^{[\mathbf{J}_{\boldsymbol{\omega}}]} \right)_{k\in[K]} \in \left(\mathbb{R}^{q\times 1} \times \left(\mathbb{R}^{q\times p}\right)^D \right)^K : \forall k \in [K], \forall d \in [D], \\
&\qquad \boldsymbol{\beta}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} [\sigma_{kd}]_r [\mathbf{u}_{kd}]_{\bullet,r} \left[\mathbf{v}_{kd}^\top \right]_{r,\bullet}, \operatorname{rank}\left(\boldsymbol{\beta}_{kd}^{R_{kd}} \right) = R_{kd}, \forall r \in [R_{kd}], [\sigma_{kd}]_r < A_\sigma, \\
&\qquad \max_{k\in[K],d\in[d_{\boldsymbol{\Upsilon}}],r\in[R_{kd}]} \left\{ \|\boldsymbol{\beta}_{k0}\|_\infty, \left\| [\mathbf{u}_{kd}]_{\bullet,r} \right\|_\infty, \left\| \left[\mathbf{v}_{kd}^\top \right]_{r,\bullet} \right\|_\infty \right\} \leq A_{\mathbf{u},\mathbf{v}} \Bigg\}. \qquad (5.4.48)
\end{aligned}
$$

Dans ce qui précède, pour $k \in [K], d \in [D]$, $[\sigma_{kd}]_r, r \in [R_{kd}]$, désignent les valeurs singulières de $\boldsymbol{\beta}_{kd}^{R_{kd}}$, avec les vecteurs unitaires orthogonaux correspondants $\left([\mathbf{u}_{kd}]_{\bullet,r} \right)_{r\in[R_{kd}]}$ et $\left(\left[\mathbf{v}_{kd}^\top \right]_{r,\bullet} \right)_{r\in[R_{kd}]}$ (Strang,

2019, I. 8). La dimension de $S_{\mathbf{m}}$ est

$$\dim\left(S_{\mathbf{m}}\right)=\left(K-1\right)L+qK+\sum_{k=1}^{K}\sum_{d=1}^{D}R_{kd}\left(\operatorname{card}\left(\mathbf{J}_{\boldsymbol{\omega}}\right)+q-R_{kd}\right)+\sum_{k=1}^{K}\sum_{g=1}^{G_k}\frac{\operatorname{card}\left(d_k^{[g]}\right)\left(\operatorname{card}\left(d_k^{[g]}\right)+1\right)}{2}.$$

Remarquons que la collection de modèles dans (5.4.48) est généralement grande et donc non traitable en pratique. Cela nous motive à restreindre le nombre de composantes K, les ordres des poids mononomiaux L et des moyennes polynomiales D parmi des ensembles finis $\mathcal{K}=[K_{\max}]$, $\mathcal{L}=[L_{\max}]$ et $\mathcal{D}=[D_{\max}]$, respectivement, où $K_{\max}\in\mathbb{N}^{\star}$, $L_{\max}\in\mathbb{N}^{\star}$ et $D_{\max}\in\mathbb{N}^{\star}$ peuvent dépendre de la taille de l'échantillon n. En outre, nous nous concentrons sur une sous-collection (potentiellement aléatoire) $\mathcal{J}$ de $\mathcal{P}\left([q]\times[p]\right)$, la taille contrôlée étant requise dans le cas de haute dimension. De plus, le nombre de vecteurs de rangs possibles considérés est réduit en travaillant sur un sous-ensemble (potentiellement aléatoire) $\mathcal{R}_{(K,\mathbf{J},D)}$ de $\left[\operatorname{card}\left(\mathbf{J}_{\boldsymbol{\omega}}\right)\wedge q\right]^{DK}$.

En particulier, rappelons que $\mathbf{B}$ est sélectionné parmi une liste de structures candidates $\left(\mathcal{B}_k\right)_{k\in[K]}\equiv\left(\mathcal{B}\right)_{k\in[K]}$, où $\mathcal{B}$ désigne l'ensemble de toutes les partitions possibles des covariables indexées par $[p]$ pour chaque groupe d'individus. Il convient de mentionner que la taille de $\mathcal{B}$ (nombre de Bell) est très grande même pour un nombre modéré de variables p. Cela nous empêche d'envisager une exploration exhaustive de l'ensemble $\mathcal{B}$. Motivés par les nouveaux travaux récents de Devijver & Gallopin (2018), pour chaque groupe $k\in[K]$, nous limitons notre attention à la sous-collection $\mathcal{B}_{k,\mathbf{\Lambda}}=\left(\mathcal{B}_{k,\lambda}\right)_{\lambda\in\mathbf{\Lambda}}$ de $\mathcal{B}_k$. Ici $\mathcal{B}_{k,\mathbf{\Lambda}}$ est la partition des variables correspondant à la structure bloc-diagonale de la matrice d'adjacence $\mathbf{E}_{k,\lambda}=\left[\mathbb{I}\left\{\left|\left[\mathbf{S}_k\right]_{z,z'}\right|>\lambda\right\}\right]_{z\in[q],z'\in[q]}$, qui est basé sur la valeur absolue seuillée de la matrice de covariance de l'échantillon $\mathbf{S}_k$ dans chaque cluster $k\in[K]$. Il est important de souligner que la classe de structures bloc-diagonales détectées par l'algorithme graphique du lasso lorsque le paramètre de régularisation varie est identique aux structures bloc-diagonales $\mathcal{B}_{k,\lambda}$ détectées par le seuillage de la covariance de l'échantillon pour chaque cluster $k\in[K]$ (Mazumder & Hastie, 2012).

Enfin, étant donné $S_{\mathbf{m}}$ défini comme dans (5.4.48), notre collection complète de modèles et notre sous-collection aléatoire de modèles SGaBloME sont définies, respectivement, comme suit:

$$\mathcal{S}=\left\{S_{\mathbf{m}}:\mathbf{m}\in\mathcal{M}\right\},\mathcal{M}=\mathcal{K}\times\mathcal{L}\times\mathcal{D}\times\left(\mathcal{B}_k\right)_{k\in[K]}\times\mathcal{P}\left([q]\times[p]\right)\times\left[\operatorname{card}\left(\mathbf{J}_{\boldsymbol{\omega}}\right)\wedge q\right]^{DK}, \tag{5.4.49}$$

$$\widetilde{\mathcal{S}}=\left\{S_{\mathbf{m}}:\mathbf{m}\in\widetilde{\mathcal{M}}\right\},\widetilde{\mathcal{M}}=\mathcal{K}\times\mathcal{L}\times\mathcal{D}\times\left(\mathcal{B}_{k,\mathbf{\Lambda}}\right)_{k\in[K]}\times\mathcal{J}\times\mathcal{R}_{(K,\mathbf{J},D)}. \tag{5.4.50}$$

Inégalité d'Oracle

Notez que dans cette thèse, voir Section 4.3 pour plus de détails, les structures bloc-diagonales, les variables pertinentes et les modèles à rangs épars sont conçus, par exemple, par la procédure Lasso $+l_2$-Rank dans Section 4.3.5. Néanmoins, notre inégalité d'oracle d'échantillon fini dans Theorem 5.4.18, qui est prouvée dans Section 4.3.3, tient toujours pour toute sous-collection aléatoire de $\mathcal{M}$ qui est construite par certains outils appropriés dans le cadre des modèles de régression SGaBloME.

Theorem 5.4.18 (Inégalité Oracle pour les modèles SGaBloME)**.** *Soit* $\left(\mathbf{x}_{[n]},\mathbf{y}_{[n]}\right)$ *les observations découlant de la densité conditionnelle inconnue* s_0*. Pour chaque* $\mathbf{m}\equiv\left(K,L,D,\mathbf{B},\mathbf{J},\mathbf{R}\right)\in\mathcal{M}$*, laissez* $S_{\mathbf{m}}$ *être donné par* (5.4.48)*. Supposons qu'il existe* $\tau>0$ *et* $\epsilon_{KL}>0$ *de sorte que, pour tous les* $\mathbf{m}\in\mathcal{M}$*, on puisse trouver* $\bar{s}_{\mathbf{m}}\in S_{\mathbf{m}}$ *tel que*

$$\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,\bar{s}_{\mathbf{m}}\right)\leq\inf_{t\in S_{\mathbf{m}}}\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,t\right)+\frac{\epsilon_{KL}}{n},\ \textit{et}\ \bar{s}_{\mathbf{m}}\geq e^{-\tau}s_0.$$

De plus, nous construisons une sous-collection aléatoire $\left(S_{\mathbf{m}}\right)_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *de* $\left(S_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ *comme dans* (5.4.50) *et considérer la collection* $\left(\widehat{s}_{\mathbf{m}}\right)_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *de* η*- minimiseurs de log-vraisemblance définis dans* (5.4.22)*. Alors, il existe une constante* C *telle que pour tout* $\rho\in(0,1)$*, et tout* $C_1>1$*, il existe deux constantes* κ_0 *et* C_2 *dépendant uniquement de* ρ *et de* C_1 *telles que, pour tout indice* $\mathbf{m}\in\mathcal{M}$*,* $\xi_{\mathbf{m}}\in\mathbb{R}^{+}$*,* $\Xi=\sum_{\mathbf{m}\in\mathcal{M}}e^{-\xi_{\mathbf{m}}}<\infty$*,*

$$\operatorname{pen}(\mathbf{m})\geq\kappa\left[\left(C+\ln n\right)\dim(S_{\mathbf{m}})+\left(1\vee\tau\right)\xi_{\mathbf{m}}\right],\kappa>\kappa_0,$$

l'estimateur de vraisemblance pénalisé $\eta' \, \widehat{s}_{\widehat{\mathbf{m}}}$*, défini dans* (5.4.21) *sur le sous-ensemble* $\widetilde{\mathcal{M}}$ *au lieu de* $\mathcal{M}$*, satisfait à*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, t\right) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(1\vee\tau)\frac{\Xi^2}{n} + \frac{\eta+\eta'}{\eta}.$$

Contribution du Chapitre 5

Enfin, Chapter 5 conclut le manuscrit et discute des perspectives. En particulier, nous suggérons plusieurs conjectures et problèmes ouverts comme futures directions de recherche.

Model Selection and Approximation in High-dimensional Mixtures of Experts Models: From Theory to Practice

Abstract: In this thesis, we study the approximation capabilities, model estimation and selection properties, of a rich family of mixtures of experts (MoE) models in a high-dimensional setting, including MoE with Gaussian experts and softmax (SGaME) or Gaussian gating functions (GLoME). Firstly, we improve upon universal approximation results in the context of unconditional mixture distributions, and study such capabilities for MoE models in a variety of contexts, including conditional probability density functions (PDF) approximation and approximate Bayesian computation. More precisely, we prove that to an arbitrary degree of accuracy, location-scale mixtures of a continuous PDF can approximate any continuous PDF, uniformly, on a compact set; location-scale mixtures of an essentially bounded PDF, resp. of conditional PDF, can approximate any PDF, resp. any continuous conditional PDF whenever the input and output variables are both compactly supported, in Lebesgue spaces. Next, we establish non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds on the penalties hold true, in high-dimensional regression scenarios for a variety of MoE regression models, including GLoME and SGaME, based on an inverse regression strategy or a Lasso penalization, respectively. We show that the performance in Jensen–Kullback–Leibler type loss of our penalized maximum likelihood estimator is roughly comparable to that of oracle model, given large enough the constant in front of the penalty. This penalty is only known up to a multiplicative constant, proportional to the dimension of model and is calibrated by slope heuristic criterion. Finally, to support our theoretical results and the statistical study of non-asymptotic model selection in a variety of MoE models, we perform numerical studies by considering simulated and real data, which highlight the performance of our finite-sample oracle inequality results.
Keywords: Mixture of experts; mixture models; universal approximation; penalized maximum likelihood; non-asymptotic model selection; high-dimensional statistics; Lasso; EM algorithm; ABC.

***

Sélection et approximation de modèle dans les modèles de mélange d'experts de grande dimension: de la théorie à la pratique

Résumé: Dans cette thèse, nous étudions les capacités d'approximation et les propriétés d'estimation et de sélection de modèles, d'une riche famille de mélanges d'experts (MoE) dans un cadre de grande dimension. Cela inclut des MoE avec experts gaussiens et fonctions softmax (SGaME) ou gaussiennes normalisées (GLoME) pour modéliser la distribution de la variable latente conditionnellement aux experts. Tout d'abord, nous améliorons les résultats d'approximation universelle dans les mélanges inconditionnels, et étudions ces capacités d'approximation pour les MoE dans une variété de contextes, y compris en approximation de fonctions de densités de probabilité (FDP) conditionnelles et en calcul bayésien approximatif. Plus précisément, nous prouvons qu'à un degré de précision arbitraire, les mélanges de translatées dilatées d'une FDP continue peuvent approximer toute FDP continue, uniformément, sur un ensemble compact; les mélanges de translatées dilatées d'une FDP essentiellement bornée, resp. d'une FDP conditionnelle, peuvent approcher toute FDP, resp. toute FDP conditionnelle continue lorsque les variables d'entrée et de sortie sont toutes deux à support compact, dans les espaces de Lebesgue. Par la suite, nous établissons des limites de risque non-asymptotiques qui prennent la forme d'inégalités d'oracle faibles, à condition que les limites inférieures des pénalités soient vraies, dans des scénarios de régression à grande dimension, pour une variété de modèles de régression MoE, y compris GLoME et SGaME, en s'appuyant sur une stratégie de régression inverse ou une pénalisation Lasso, respectivement. Nous montrons que la performance en perte de type Jensen-Kullback-Leibler de notre estimateur de maximum de vraisemblance pénalisé est à peu près comparable à celle du modèle oracle, si la constante devant la pénalité est suffisamment grande. Cette pénalité n'est connue que jusqu'à une constante multiplicative, proportionnelle à la dimension du modèle et est calibrée par un critère heuristique de pente. Enfin, pour appuyer nos résultats théoriques et l'étude statistique de la sélection non-asymptotique de modèles dans une variété de modèles de MoE, nous réalisons des études numériques en considérant des données simulées et réelles, qui mettent en évidence la performance de nos résultats notamment ceux d'inégalités d'oracle en échantillon fini.
Mots-clés: Mélange d'experts; mélange de lois; approximation universelle; maximum de vraisemblance pénalisé; sélection non-asymptotique de modèle; grande dimension; Lasso; algorithme EM.