Modèles de mélange et de Markov caché non-paramétriques : propriétés asymptotiques de la loi a posteriori et efficacité

Elodie, Edith Vernet

université PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2016SACLS418

Thèse de doctorat
de
l'Université Paris-Saclay
préparée à
l'Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

École Doctorale n°574
École doctorale de mathématiques Hadamard (EDMH)

Spécialité de doctorat : Mathématiques appliquées

Par

Elodie VERNET

Modèles de mélange et de Markov caché non paramétriques : propriétés asymptotiques de la loi a posteriori et efficacité

Thèse présentée et soutenue à Orsay le 15 novembre 2016

Composition du Jury :

CASTILLO Ismael	Professeur, UPMC	Examinateur
DOUC Randal	Professeur, Télécom SudParis	Examinateur
GASSIAT Élisabeth	Professeur, Université Paris-Sud	Codirecteur de thèse
MATIAS Catherine	Directeur de recherche, CNRS	Président
MOULINES Éric	Professeur, École Polytechnique	Rapporteur
NICKL Richard	(Professeur, University of Cambridge)	Rapporteur
ROUSSEAU Judith	(Professeur, Université Paris-Dauphine)	Codirecteur de thèse

Résumé

Les modèles latents sont très utilisés en pratique, comme en génomique, économétrie, reconnaissance de parole, étude de population... Comme la modélisation paramétrique des lois d'émission, c'est-à-dire les lois d'une observation sachant l'état latent, peut conduire à de mauvais résultats en pratique, un récent intérêt pour les modèles latents non paramétriques est apparu dans les applications. Or ces modèles ont peu été étudiés en théorie. Dans cette thèse je me suis intéressée aux propriétés asymptotiques des estimateurs (dans le cas fréquentiste) et de la loi a posteriori (dans le cadre Bayésien) dans deux modèles latents particuliers : les modèles de Markov cachés et les modèles de mélange. J'ai tout d'abord étudié la concentration de la loi a posteriori dans les modèles non paramétriques de Markov cachés. Plus précisément, j'ai étudié la consistance puis la vitesse de concentration de la loi a posteriori. Enfin je me suis intéressée à l'estimation efficace du paramètre de mélange dans les modèles semi-paramétriques de mélange.

Mots-clés:

Statistiques non et semi-paramétriques, statistiques Bayésiennes, statistiques asymptotiques, modèle de Markov caché, modèle de mélange

Nonparametric Mixture Models and Hidden Markov Models: Asymptotic Behaviour of the Posterior Distribution and Efficiency

Abstract

Latent models have been widely used in diverse fields such as speech recognition, genomics, econometrics. Because parametric modeling of emission distributions, that is the distributions of an observation given the latent state, may lead to poor results in practice, in particular for clustering purposes, recent interest in using nonparametric latent models appeared in applications. Yet little thoughts have been given to theory in this framework. During my PhD I have been interested in the asymptotic behaviour of estimators (in the frequentist case) and the posterior distribution (in the Bayesian case) in two particuliar nonparametric latent models: hidden Markov models and mixture models. I have first studied the concentration of the posterior distribution in nonparametric hidden Markov models. More precisely, I have considered posterior consistency and posterior concentration rates. Finally, I have been interested in efficient estimation of the mixture parameter in semiparametric mixture models.

Keywords:

Non and semiparametric statistics, Bayesian statistics, asymptotic statistics, hidden Markov model, mixture model

Remerciements

Mon doctorat fut une période intense, parcourue de moments de doutes, d'attentes, de déceptions, de soulagements et de joie. Je tiens à remercier tous ceux qui m'ont accompagnée, soutenue (et supportée ?) tout au long de cette thèse.

En particulier, un grand merci à mes deux directrices de thèse. Elisabeth et Judith, je vous remercie chaleureusement pour le temps que vous m'avez consacré, chacune malgré des contraintes prenantes. Merci pour le sujet que vous m'avez proposé, riche en questions ouvertes et intéressantes, pour vos conseils de rédaction et enfin de m'avoir permis de participer à de nombreuses conférences enrichissantes et motivantes.

I deeply thank Eric Moulines and Richard Nickl for the precious time they spent to read my thesis. I'd like to take this opportunity to thank Richard for funding my postdoc and for his welcome in Cambridge. I can already say that I am really lucky to experience the Cambridge bubbling of intellectual energy. Je remercie aussi vivement Ismael Castillo, Randal Douc et Catherine Matias, pour avoir accepté de participer à mon jury.

Je veux mentionner ici que cette thèse a été financée par le ministère de l'Enseignement Supérieur et de la Recherche via l'ENS de Cachan. Je leur en suis reconnaissante. J'en profite pour souligner que l'ENS de Cachan m'a fourni mes premiers contacts avec la recherche, source de motivation. I want to thank Bob Williamson for the supervision of a research internship, its memory enabled me to overcome my doubts.

C'est l'université Paris-Sud, plus précisément le laboratoire de mathématiques d'Orsay qui m'a accueillie pendant ma thèse. J'y ai particulièrement apprécié les séminaires et l'ambiance entre doctorants. À ce propos, je remercie mes co-bureaux Alba, Arno, Corentin, Etienne, Imène, Joseph, Luc, Martin, Raphaël, Robert, Thomas, Thomas et Vincent. Je tiens aussi à remercier les doctorants et anciens doctorants d'Orsay qui ont pimenté les déjeuners, en particulier Anthony, Céline, Clément, Cong, Emilie, Emilien, Jeanne, Lionel, Lison, Lucie, Lucile, Matthias, Mor, Olivier, Pierre-Antoine, Valérie, Vincent, Vincent, Solenne, Thibault et Tiago. Merci aussi à Catherine, Florence et Valérie qui m'ont facilité les tâches administratives.

Les conférences auxquelles j'ai participé ont toujours été une grande source de motivation. Merci à tous ceux qui ont fait de ces événements des succès, notamment à Botond, Claire, Clara, Eddie, Jade, JB, Julyan, Romain et Zacharie.

Enfin, ces dernières années, une grande partie de mon temps a été consacrée à ma thèse. Merci à tous ceux qui m'ont permis de m'évader et d'atténuer la pression que je m'impose. Merci à mes plus proches amis Paul, Perrine et Roxane. Je veux aussi remercier ma famille pour son

soutien infaillible. Notamment, un grand merci à mes parents et à ma sœur pour leur confiance indéfectible, leur joie de vivre et plus pragmatiquement leur aide pour la préparation du pot. Pour finir, un tendre merci à Pierre pour tout, en particulier pour ton soutien lors de la dernière ligne droite, il m'a été précieux.

CONTENTS

CHAPTER 1

INTRODUCTION

This introduction does not aim at being exhaustive. Here, I want to introduce the notions and objects which are important in my thesis along with the motivations of the research I have accomplished during my PhD. In Section 1.1, I present the models I have studied namely hidden Markov models (HMMs) and mixture models. Both models are latent models, in other words, the observations are driven by some hidden random variables. In mixture models the latent variables are independent and identically distributed while they are dependent in HMMs. Latent models are very popular in practice. Recently, there has been a increased use of nonparametric versions of latent models. Yet, in this framework theoretical guarantees for estimators or for the posterior distribution in the Bayesian framework are not well understood. In this thesis I have contributed to better understand the theoretical behaviour of both point estimators and posterior distributions in these models. In Section 1.2, I present the type of properties I have been studying. In Section 1.3, I recall the results, that were known at the beginning of my PhD, about nonparametric HMMs and semiparametric mixture models. I finish the introduction with Section 1.4, which gives an overview of the results I have obtained and perspectives.

1.1 HMMs and Mixture Models

First of all, let us introduce some notations.

1.1.1 Some General Notations for Statistical Models

A statistical model is a triple $(\mathcal{Y}^n, \mathcal{B}_n, \mathcal{P}_n)$ where $(\mathcal{Y}^n, \mathcal{B}_n)$ is a measurable space and $\mathcal{P}_n = \{P_n^\theta, \theta \in \Theta\}$ is a set of distributions on $(\mathcal{Y}^n, \mathcal{B}_n)$ parametrised by θ. The integer n represents the number of observations. In this thesis we are interested in asymptotic properties, that is the study of what happens when n tends to infinity. We consider $P_{+\infty}^\theta$ a probability distribution on $\mathcal{Y}^{\mathbb{N}}$ that will be denoted $P^\theta := P_{+\infty}^\theta$ in the following, then P_n^θ is the n-marginal of P^θ and $\mathcal{Y}^n$ is the set where the observations $(Y_1, \dots, Y_n)$ live in: $(Y_1, \dots, Y_n) \in \mathcal{Y}^n$. When Θ is finite dimensional we say that the model is parametric otherwise we say that the model is nonparametric. Throughout the thesis, we assume that the model is dominated, that is the distributions P_n^θ are absolutely continuous with respect to a unique measure λ_n. We denote p_n^θ the density functions of P_n^θ with respect to λ_n :

$$P_n^\theta = p_n^\theta \lambda_n.$$

We say that the model is well-specified when the observations $Y_1, \dots, Y_n$ are assumed to be distributed from a true distribution $P_n^{\theta^*}$ which belongs to the considered family of distributions, $\theta^* \in \Theta$. An aim is then to obtain some information about θ^* from the observations. For instance, we may want to estimate θ^* or a functional of θ^*. This inference is done with the help of an estimator $\hat{\theta}_n$, i.e. a measurable function of the observations: $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$.

In Bayesian statistics, the set of parameters Θ is endowed with a sigma-field and a probability

distribution Π on Θ is given, it is called the prior distribution. The prior distribution may reflect what is already known on the parameters. The prior distribution may also be "neutral" that is not giving any information on the parameters, we call such prior distribution noninformative prior distributions. One may also choose the prior because of its tractability. The choice of the prior affects the inference, so that this choice has to be done with care, particularly in the nonparametric setting.

From the prior and the observations $Y_1, \dots, Y_n$, we can define the posterior distribution $\Pi(\cdot|Y_1, \dots, Y_n)$ which is a distribution on the set of parameters Θ. The posterior distribution is the distribution of the parameters given the observations, by the Bayes' rule:

$$\Pi(\theta \in A|Y_1, \dots, Y_n) = \frac{\int_A p_n^\theta(Y_1, \dots, Y_n)\Pi(d\theta)}{\int_\Theta p_n^\theta(Y_1, \dots, Y_n)\Pi(d\theta)}.$$

The posterior distribution represents the knowledge on the parameters we have learnt thanks to the observations. When the prior distribution is absolutely continuous with respect to a measure ν, that is $\Pi = \pi\nu$, the posterior admits a density with respect to ν. We denote it $\pi(\cdot|Y_1, \dots, Y_n)$ and

$$\pi(\theta|Y_1, \dots, Y_n) = \frac{p_n^\theta(Y_1, \dots, Y_n)\pi(\theta)}{\int_\Theta p_n^\theta(Y_1, \dots, Y_n)\pi(\theta)\nu(d\theta)}.$$

From the posterior distribution, we can build frequentist estimators for instance the maximum a posteriori (MAP) estimator

$$\widehat{\theta}_{MAP} \in \arg\max_{\theta\in\Theta} \pi(\theta|Y_1 \dots Y_n)$$

or the posterior mean

$$\widehat{\theta}_{PM} = \int_\Theta \pi(\theta|Y_1, \dots Y_n)\nu(d\theta),$$

if they exist. Note that the posterior distribution potentially gives more information than an estimator, since given the observations, it gives a distribution on the parameter set and not only a value for the parameter.

The posterior distribution may not have an explicit expression, then statisticians may choose prior distributions for which the posterior is easier to compute or an approximation of the posterior distribution may be computed with MCMC for instance. A class of prior distributions leading to analytically computable posterior distributions is the class of conjugate prior distributions. A prior distribution is said to be conjugate for some likelihoods p^θ when the posterior distribution and the prior distribution belong to the same class of distributions. For instance, the Gaussian prior is conjugate for the likelihood $p_n^\theta(y_1, \dots, y_n) = \prod_{i=1}^n \left(1/\sqrt{n}\exp(-(y_i - \theta)^2/2)\right)$ with parameter $\theta \in \Theta = \mathbb{R}$. Indeed when $\pi_{\mu,\sigma}(\theta) = 1/\sqrt{n}\exp(-(\theta - \mu)^2/(2\sigma^2))$, then the posterior

distribution is the Gaussian distribution

$$\mathcal{N}\left(\frac{\mu}{1+n\sigma^2}+\frac{1}{\sigma^2/n+1}\frac{1}{n}\sum_{i=1}^{n}Y_i\ ,\ \frac{1}{1/\sigma^2+n}\right). \tag{1.1}$$

An important class of conjugate prior distributions for independent observations are the Dirichlet distributions. The Dirichlet distribution is a generalization of the beta distribution in higher dimension. The Dirichlet distribution with parameter $(\alpha_1,\dots,\alpha_j), j \in \mathbb{N}^*$, is a distribution absolutely continuous with respect to the Lebesgue measure on the $(j-1)$-dimensional simplex $\Delta_j = \{x \in \mathbb{R}_+^j : \sum_{i=1}^j x_i = 1\}$ with density function:

$$\frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} x_1^{\alpha_1-1}\dots x_{j-1}^{\alpha_{j-1}-1}(1-\sum_{i<j}x_i)^{\alpha_j-1}, x \in \Delta_j.$$

The Dirichlet distribution with parameter (α_1,α_2) is the beta distribution with parameter (α_1,α_2). The Dirichlet distribution is conjugate for the likelihood of the type product of a categorical distributions. In other words, if $\mathcal{Y}_n = \{1,\dots,k\}^n$, $\Theta = \Delta_k$, $p_n^\theta(y_1,\dots,y_n) = \theta_{y_1}\dots\theta_{y_n}$ and the prior is a Dirichlet distribution of parameter $(\alpha_1,\dots,\alpha_k)$ then by the Bayes' rule:

$$\pi(\theta|Y_1,\dots,Y_n) = \frac{\theta_{y_1}\dots\theta_{y_n}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1}}{\int_{\Delta_k}\theta_{y_1}\dots\theta_{y_n}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1}d\theta},$$

so that the posterior distribution is a Dirichlet distribution of parameter $(\alpha_1 + |\{i : Y_i = 1\}|,\dots,\alpha_k + |\{i : Y_i = k\}|)$.

In Bayesian nonparametric models, we need probability distributions on infinite dimensional sets. A popular distribution used in this framework is the Dirichlet process which is a generalization of the Dirichlet distribution.

Definition 1.1 (Dirichlet process)**.** *Let $\alpha > 0$, and let G be a probability distribution on some set Γ. The Dirichlet process $DP(\alpha G)$ is a process on probability measures $\mathcal{M}(\Gamma)$ on Γ such that for all realization $P \in \mathcal{M}(\Gamma)$ of the process, for all finite partition $(\Gamma_1,\dots,\Gamma_r)$ of Γ,*

$$(P(\Gamma_1),\dots,P(\Gamma_r)) \text{ is distributed as } Dir\left(\alpha G(\Gamma_1),\dots,\alpha G(\Gamma_r)\right).$$

The Dirichlet process is popular because it is conjugate when $\Theta = \mathcal{M}(\Gamma)$, $P_n^\theta = \otimes_{i=1}^n \theta$, i.e. in the i.i.d. case with distribution θ. Yet a drawback of Dirichlet processes is that it puts all its mass on discrete distributions. Thus statisticians often use Dirichlet process mixture of some kernel as prior on density functions. For instance in the case where the parameter $\theta = f$ is a density function, we can choose that under the prior distribution, $f = \int K_\sigma(\cdot - m)P(dm, d\sigma)$ where P is distributed as a Dirichlet process $DP(\alpha G)$ and $K_\sigma(\cdot - m)$ is a kernel with window σ centered at m, e.g. the Gaussian kernel where $K_\sigma(y-m) = 1\sqrt{2\pi\sigma^2}\exp(-(y-m)^2/(2\sigma^2))$.

More properties about the Dirichlet Process are given in Ghosh and R.V. Ramamoorthi [GR03] for instance. Of course, Dirichlet process mixtures of kernels are not the only possible prior distributions. For instance, Gaussian processes are another type of popular nonparametric prior distributions, see Ghosh and R.V. Ramamoorthi [GR03] and van der Vaart and van Zanten [VZ08] for instance and references therein. For more information on Bayesian statistics, see Robert [Rob01] or Gelman *et al.* [GCSR14] for example and on nonparametric Bayesian statistics see Ghosh and R.V. Ramamoorthi [GR03] for instance.

I have considered two models during my PhD which are mixture models and hidden Markov models (HMMs). I now define both models and compare them.

1.1.2 HMMs and Mixture Models: Definitions and Examples

During my PhD, I have been interested in latent models. In these models, a sequence of latent variables is hidden and the statistician only observes a noisy version of it. An important class of such models is when the latent variables live in a finite set, say $\{1, \dots, k\}$. In this case, the latent variables are often used to model populations the observations come from. In the case of mixture models the latent variables are i.i.d. while in hidden Markov models, the latent variables form a Markov chain. We now define these models formally. More information can be found in the following books and the references therein, MacDonald and Zucchini [MZ97], MacDonald and Zucchini [MZ09] and Cappé *et al.* [CMR05] on HMMs, Lee *et al.* [LMMR09] on mixture models.

Definition 1.2 (Mixture model)**.** *Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be two measurable spaces where $\mathcal{X}$ is finite. Let μ be a distribution on $\mathcal{X}$ and $(F_x)_{x\in\mathcal{X}}$ be a family of distributions on $\mathcal{Y}$. If X is distributed as μ and given $X = x$, Y is distributed from F_x then Y is distributed from a mixture model. In other words, Y is distributed from $\sum_{x\in\mathcal{X}} \mu(x)F_x(\cdot)$. We call the distributions F_x, the emission distributions.*

A commonly used mixture model is the mixture of Gaussian distributions. In this case, the emission distributions F_x are assumed to be Gaussian distributions.

A DAG representation of a mixture model is given in Figure 1.1.

In such a mixture model, statisticians cannot observe the hidden variables $X_1, \dots X_n$ i.i.d. from μ but observe the observations $Y_1, \dots, Y_n$.

Now imagine that the hidden variables $X_1, \dots, X_n$ are not i.i.d. any more but are distributed from a Markov chain. Then we obtain a hidden Markov model. Here is a formal definition of a hidden Markov model.

Definition 1.3 (Hidden Markov model)**.** *Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be two measurable sets, Q a transition matrix on $\mathcal{X} \times \mathcal{X}$, μ a probability distribution on $\mathcal{X}$ and $(F_x)_{x\in\mathcal{X}}$ a family of probability distribution on $\mathcal{Y}$. Assume $(X_t)_{t\in\mathbb{N}}$ is a Markov chain with transition matrix Q and initial distri-*

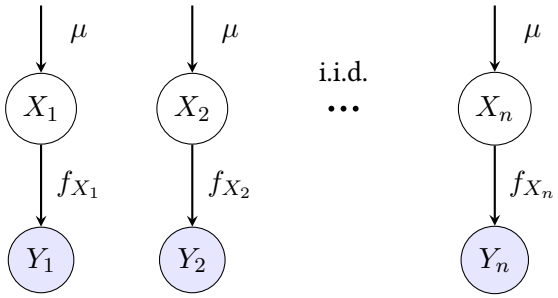


Figure 1.1 – Visualization of a mixture model.

bution μ, so that

$$X_1 \sim \mu, \quad X_{t+1}|X_1, \dots X_t \sim \sum_{x \in \mathcal{X}} Q_{X_t,x}\delta_x. \tag{1.2}$$

Assume that given the Markov chain $(X_t)_{t\in\mathbb{N}}$, the observations Y_t are independent and for all $s \in \mathbb{N}$, Y_s is distributed from F_{X_s}.
Then the sequence $(X_t, Y_t)_{t\in\mathbb{N}}$ is a hidden Markov chain.
When for all $x \in \mathcal{X}$, F_x is absolutely continuous with respect to some measure λ with density f_x, the likelihood associated to this model is

$$p_n^\theta(Y_1, \dots, Y_n) = \sum_{(x_1,\dots,x_n)\in\mathcal{X}^n} \mu_{x_1} Q_{x_1,x_2} \dots Q_{x_{n-1},x_n} f_{x_1}(Y_1) \dots f_{x_n}(Y_n),$$

where $\theta = (\mu, Q, f)$, $f = (f_x)_{x\in\mathcal{X}}$.

A DAG representation of a HMM is given in Figure 1.2.

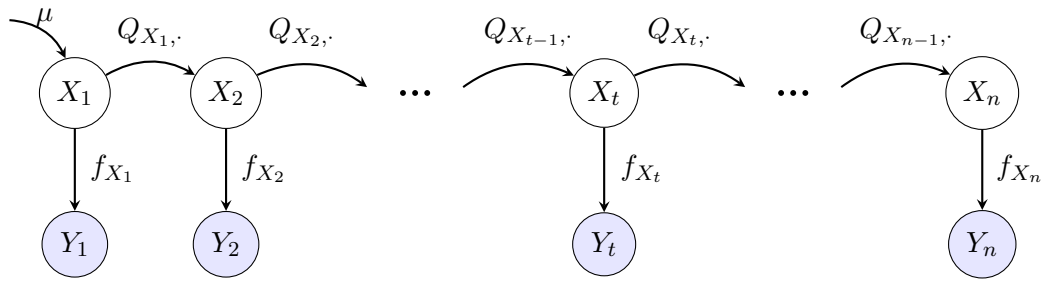


Figure 1.2 – Visualization of a HMM.

Mixture models and HMMs are very popular models. They are used in many fields of application such as speech recognition, image processing, genetics, ecology, econometrics or climate, to cite a few. Their popularity is due to their great flexibility, together with the existence of efficient algorithms, both for the Bayesian and frequentist methods.

In the following, we consider a particular case of these models that we describe now. We only consider mixture models and HMMs where the number of states for the latent variables is fi-

nite and known. In other words, we assume that there exists $k \in \mathbb{N}$, that we know, such that $\mathcal{X} = \{1, \dots, k\}$. These particular models are often used to cluster the observations into groups associated to the same latent variable. While we constraint the latent variables to live in a finite state space, we do not assume that the emission distributions have a specific parametric form. So that we consider nonparametric latent models with finite state space. Chapters 2 and 3 deal with nonparametric HMMs with finite state space. Chapter 4 deals with mixture models with finite state space where the emission distributions are a product of three distributions (for identifiability purpose, see Section 1.3.1). The latter mixture model is represented in Figure 1.3.

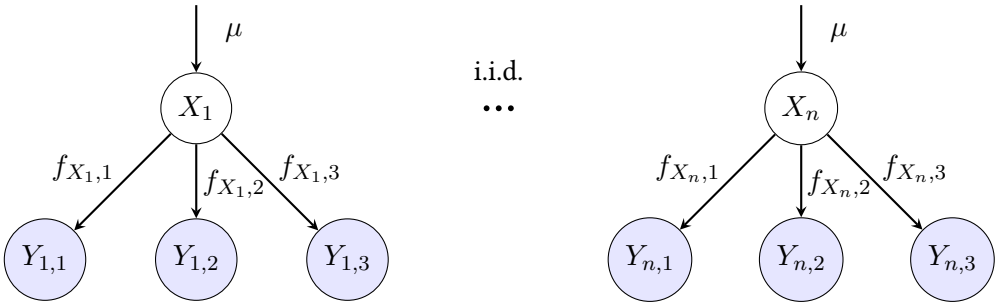


Figure 1.3 – Visualization of the multivariate mixture model.

The reason for considering a nonparametric model for the emission distributions, is that they allow for much more robust inference. In Yau *et al.* [YPRH11] for example, a nonparametric HMM with finite state space is used to model genetic data. More precisely the index t corresponds to a locus in the DNA. For each locus, the authors want to know if the fragment of DNA has been deleted twice ($X_t = 1$), once ($X_t = 2$), if nothing has happened ($X_t = 3$), if it has been replicated once ($X_t = 4$) or twice ($X_t = 5$). This variation of the number of replicates of the fragment of the DNA is called genomic copy number variation and is represented in Figure 1.4. The data consist of intensity measurements obtained after some experience on the studied DNA. The authors model the data with a HMM where the hidden states correspond to the state of deletion or duplication and the observations Y_t are the intensity of measures. The HMM is then associated with $k = 5$ states. The authors consider a location HMM, i.e.

$$Y_t = m_{X_t} + \epsilon_t,$$

with $m_1, \dots, m_5 \in \mathbb{R}$, $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{G}$ and $\mathcal{G}$ some unknown distribution. The authors use a Bayesian nonparametric approach and model the density of the noise (ϵ_t) using a DP mixture of Gaussian distributions.

Two questions then arise.

- Is this model identifiable?
- Does the Bayesian approach lead to consistency?

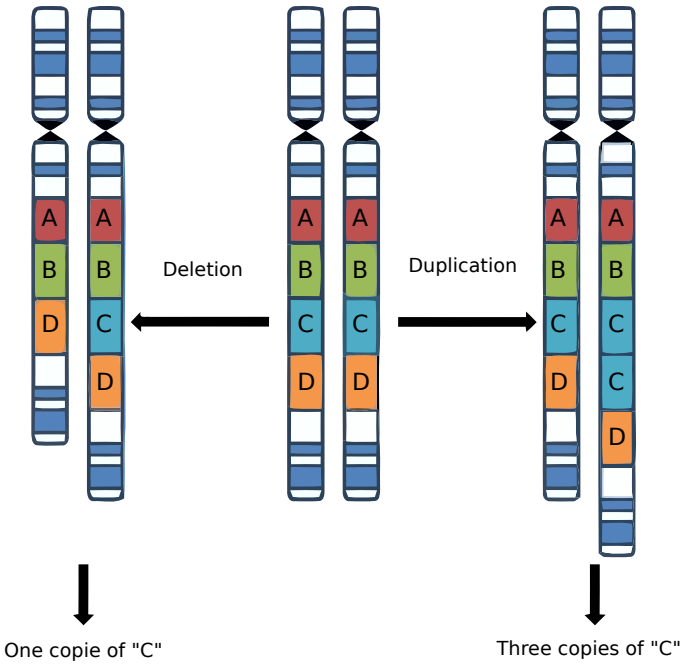


Figure 1.4 – Representation of genomic copy number variation, from http://readingroom.mindspec.org/?page_id=8221

The identifiability issue has been solved by Gassiat and Rousseau [GR16] and Gassiat *et al.* [GCR15]. In this thesis we study the asymptotic behaviour of the posterior distribution for such models. More generally, we are interested in the asymptotic properties of the posterior distribution and estimators in nonparametric HMMs and semiparametric mixture models. In Section 1.2, we present the asymptotic guarantees we examine in Chapters 2, 3 and 4.

1.2 Asymptotic Theoretical Guarantees

Before studying the asymptotic behaviour of the posterior distribution or of estimators, it is important to understand when and why the model is identifiable. **Identifiability** is the injectivity of the functional $\theta \mapsto P^\theta$: basically, it states that from the true distribution of the observations you can recover the parameters. Of course it is a very important notion in the context of the estimation of θ.

For example:

- the i.i.d. Gaussian experiment, that is $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma) =: P^{\mu,\sigma}$ is identifiable. Indeed $\mu = \mathbb{E}_{P^{\mu,\sigma}}(X)$ and $\sigma = \sqrt{Var_{P^{\mu,\sigma)}}(X)}$.
- In general, a nonparametric mixture model, with $Y_i \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{k} \mu_j f_j(\cdot)\lambda$, is not identifiable unless some extra constraint is imposed on f_j, $j \leq k$. Indeed let $\tilde{\mu} = (\mu_1/2, \mu_1/2 +$

$\mu_2, \mu_3, \ldots, \mu_k)$, $\tilde{f}_2 = 1/(\mu_1/2 + \mu_2)\ (\mu_1/2 f_1 + \mu_2 f_2)$ and $\tilde{f}_j = f_j$ for $j \in \{1, 3, \ldots, k\}$, then $\sum_j \mu_j f_j \lambda = \sum_j \tilde{\mu}_j \tilde{f}_j \lambda$. Similarly one can choose many other parameters $(\tilde{\mu}, \tilde{f})$ leading to the same distribution.

- Interestingly, if for all j, f_j can be written as $f_j = \prod_{c=1}^{3} f_{j,c}$, and under the restriction that for all $j \leq k$, $\mu_j > 0$ and $f_{j,c}\lambda$ are linearly independent for all $c \in \{1, 2, 3\}$, then the model is identifiable (up to the labelling of the hidden states). This is the model represented in Figure 1.3

More identifiability results in latent models can be found in Section 1.3.1.

Identifiability often is a prerequisite before studying more involved guarantees. Indeed, in statistics, you cannot have access to the probability P^θ but to observations $Y_1, \ldots, Y_n$ which are distributed from P^θ and an objective is to obtain information about the unknown P^θ from the observations. For instance, we may want to know from which parameter the observations come from, i.e. to estimate θ or predict the next observations (prediction). When the goal is to estimate θ then identifiability is a required property. Then we need to control that what we have built with the observations, namely an estimator or the posterior distribution, gives a " good" approximation of what we wanted to recover, where the adjective " good" has to be defined. It may be defined in an asymptotic way that is by regarding what is happening when the number of observations n increases or nonasymptotically, that is when the number of observations is fixed. I'm going to emphasise asymptotic guarantees since the results I have obtained during my PhD are asymptotic.

To study the asymptotic properties of the method of inference, we take a frequentist point of view, i.e. we assume that the observations come from a true distribution P^{θ^*}. We investigate three types of asymptotic guarantees:

- consistency results,
- rates of convergence,
- limit distributions.

In Section 1.2.1, we describe the tools used to study the asymptotic behaviour of the posterior distribution. Intuitively, when the number of observations, distributed from a true distribution P^{θ^*}, increases, the posterior should concentrate around the true parameter θ^*, i.e. the posterior distribution should converge to a Dirach measure δ_{θ^*} at θ^*. This is represented in Figure 1.5 by the plain arrow, over which there is a question mark. This is not the same as the problem of approximation of the posterior distribution for a given n by algorithms as Markov Chain Monte Carlo (MCMC), represented by the dotted line. The latter is not treated in this thesis.

Posterior consistency and posterior concentration rates are introduced in Section 1.2.1 and are studied in the framework of nonparametric HMMs in Chapters 2 and 3. In Section 1.2.2, the

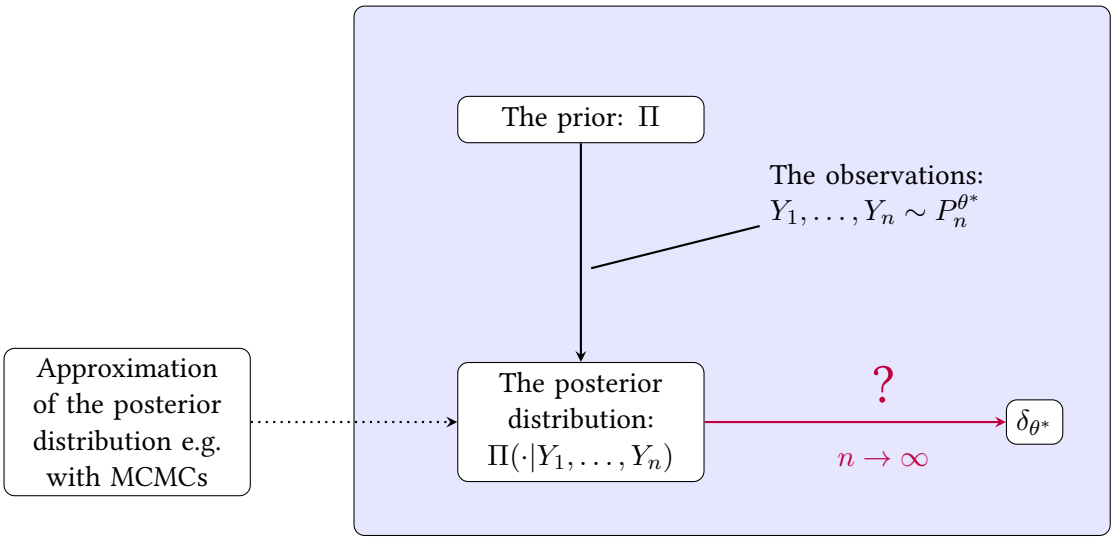


Figure 1.5 – Study of the asymptotic behaviour of the posterior distribution

notion of limit distribution linked to asymptotic efficiency is developed in the frequentist and Bayesian settings. This property is studied in the case of semiparametric mixture models in Chapter 4.

1.2.1 Posterior Consistency and Posterior Concentration Rates

In this section, we give results on the asymptotic behaviour of the posterior distribution taking a frequentist point of view, i.e. assuming that the observations come from a true distribution P^{θ^*}. The interest of studying posterior consistency and posterior concentration rates is that it sheds light on the impact of the prior distribution on the posterior distribution. This is particularly important in nonparametric models where the prior cannot be fully subjectively assessed and is difficult to apprehend, given the complexity of the parameter space. We then study the behaviour of the posterior distribution $\Pi(\cdot|Y_1, \ldots Y_n)$ when the number of observations n increases.

1.2.1.1 Definitions

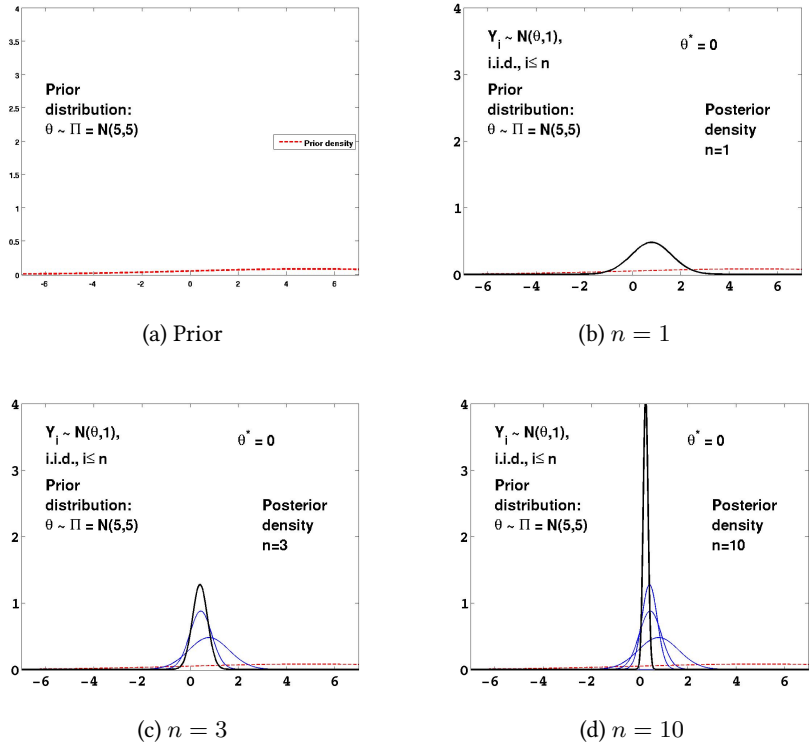
The first guarantee we may look for is posterior consistency which corresponds to answering the question: "when the observation comes from a true parameter θ^*, does the posterior distribution concentrate its mass around the true parameter θ^* when the number of observations increases?".

Definition 1.4 (posterior consistency)**.** *We say that the posterior distribution is consistent at θ^* with respect to a pseudo-metric d on Θ if*

$$\Pi\left(\{\theta : \ d(\theta, \theta^*) > \epsilon\} \mid Y_1, \ldots, Y_n\right) \to 0, \quad P^{\theta^*} - a.s., \quad \textit{for all } \epsilon > 0.$$

This notion is illustrated in Figure 1.6, with i.i.d. observations from $\mathcal{N}(\theta, 1)$ with the true pa-

rameter $\theta^* = 0$, and a Gaussian prior $\Pi = \mathcal{N}(5, 5)$ on the parameter $\theta \in \Theta := \mathbb{R}$. In Figure 1.6, the posterior seems to concentrate around the true value 0.



(a) Prior (b) $n = 1$

(c) $n = 3$ (d) $n = 10$

Figure 1.6 – Representation of the posterior density when n increases for one set of Gaussian i.i.d. observations and with a Gaussian prior–Illustration of posterior consistency

Posterior consistency is a minimal requirement, even in the Bayesian subjective point of view. Indeed if two posterior distributions are consistent everywhere then they will finally agree (merge weakly) as explained in Diaconis and Freedman [DF86].

To better understand the behaviour of the posterior, we may be interested at which rate the concentration occurs.

Definition 1.5 (posterior concentration rates)**.** *We say that the posterior distribution concentrates at rate $\epsilon_n \to 0$ at θ^*, with respect to a pseudo-metric d on Θ if there exists a constant $M > 0$ such that*

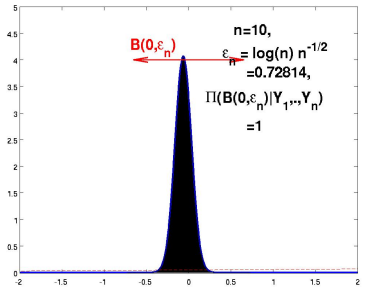
$$\Pi\left(\{\theta : \ d(\theta, \theta^*) > M\epsilon_n\} \mid Y_1, \dots Y_n\right) \to 0, \qquad \textit{in } P^{\theta^*}\textit{-probability.}$$

Posterior concentration rates are illustrated in Figure 1.7 where the framework is the same as the one of illustration 1.6. Yet in this illustration we try to know if the posterior concentrates at rate $\epsilon_n = \log(n)/\sqrt{n}$ and then at rate $\epsilon_n = \log(n)/n$ at 0. We then take a ball around 0 of radius ϵ_n, and we verify if the posterior mass of this ball tends to one in P^{θ^*} probability. In this case, it can be proved that the posterior concentrates at rate $M_n/\sqrt{n}$ at 0 for all sequence M_n increasing

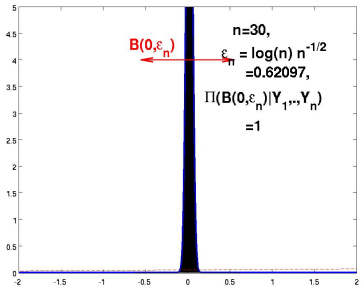
to $+\infty$ (and this is typically true in the parametric setting) and is not at a rate tending faster to zero.

Studying posterior concentration rates can give an idea of optimality of the behaviour of the posterior distribution. Indeed if the posterior concentrates at a minimax rate in the frequentist sense, then the behaviour of the posterior distribution has an optimal asymptotic behaviour. This enables us to compare two prior distributions through their associated posterior concentration rates. It also helps in comparing the Bayesian answer with frequentist estimators. Note that when the posterior concentrates at some rate, then we can build an estimator from the posterior distribution which converges to the true parameter at the same rate (see Ghosal *et al.* [GGV00]). The frequentist minimax rate often depends on the regularity of the true distribution (for instance $n^{-\beta/(2\beta+1)}$ in density estimation for a true distribution β-Hölder and the L^1-norm). If the prior distribution does not depend on the regularity parameter and the posterior distribution concentrates at the minimax rate for any regularity, so that it learns the regularity parameter from the observations; then we say that the posterior distribution concentrates at adaptive minimax rates.

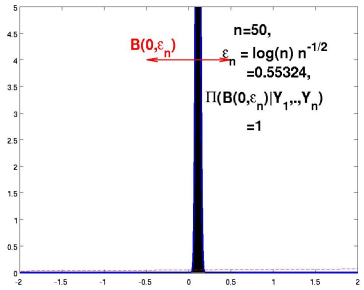
Another interest of studying posterior concentration rate is that it sheds light on how some aspects of the prior distribution influence the behaviour of the posterior distribution. This is particularly important since it is not possible to assess a prior distribution on an infinite dimensional space purely on subjective considerations.



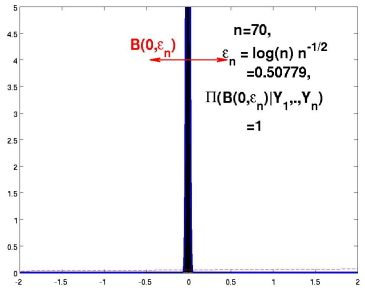
(a) $\epsilon_n = \log(n)/\sqrt{n}$, $n = 10$



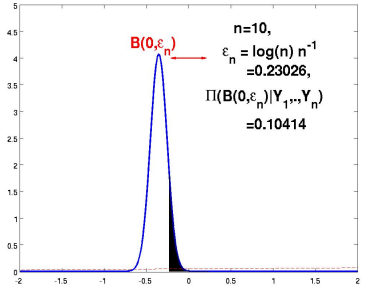
(b) $\epsilon_n = \log(n)/\sqrt{n}$, $n = 30$



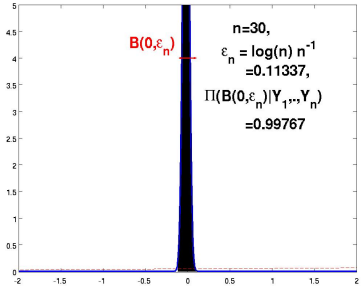
(c) $\epsilon_n = \log(n)/\sqrt{n}$, $n = 50$



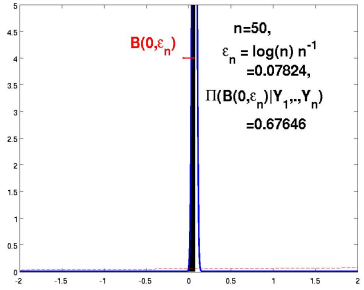
(d) $\epsilon_n = \log(n)/\sqrt{n}$, $n = 70$



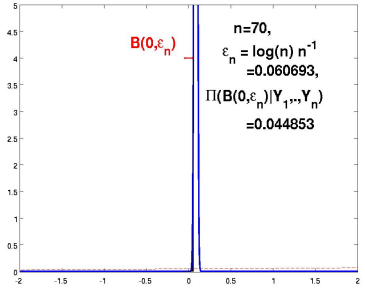
(e) $\epsilon_n = \log(n)/n$, $n = 10$



(f) $\epsilon_n = \log(n)/n$, $n = 30$



(g) $\epsilon_n = \log(n)/n$, $n = 50$



(h) $\epsilon_n = \log(n)/n$, $n = 70$

Figure 1.7 – Representation of the posterior mass of a ball $\mathcal{B}(0, \epsilon_n)$ when n increases for two decreasing rate ϵ_n–Illustration of posterior concentration rates

1.2.1.2 Technique of Proof and What is Already Known

When the observations are i.i.d. Doob's theorem ensures that the posterior distribution $\Pi(\cdot|Y_1, \ldots, Y_n)$ is consistent at Π-almost every θ^*. Yet behind this positive result, different limitations exist. First such theorem may fail with non-i.i.d. observations (see Choi and Ramamoorthi [CR08]). Moreover from Doob's theorem, we cannot know at which θ^* the posterior distribution is consistent. Finally, in the nonparametric setting, the null-set of parameters at which the posterior distribution is not consistent may be topologically huge (see Freedman [Fre65]).

These limitations underline the usefulness of a method to obtain posterior consistency at particular true parameters θ^*. A general method to prove consistency and which also leads to posterior concentrations rates when it is refined is presented now. It comes from Schwartz [Sch65], Barron [Bar88], Barron *et al.* [BSW99], Ghosal *et al.* [GGV00], Shen and Wasserman [SW01] or Ghosal and van der Vaart [GV07a] to cite but a few.

To prove posterior consistency or concentration rates, one has to look at probabilities of this type:

$$\Pi(\{\theta : \ d(\theta,\theta^*) \geq \delta_n\}|Y_1,\ldots,Y_n) = \frac{\int_{d(\theta,\theta^*)\geq\delta_n} p_n^\theta(Y_1,\ldots,Y_n)\pi(d\theta)}{\int_\Theta p_n^\theta(Y_1,\ldots,Y_n)\pi(d\theta)}$$
$$= \frac{\int_{d(\theta,\theta^*)\geq\delta_n} p_n^\theta(Y_1,\ldots,Y_n)/p_n^{\theta^*}(Y_1,\ldots,Y_n)\pi(d\theta)}{\int_\Theta \exp(\log(p_n^\theta(Y_1,\ldots,Y_n)) - \log(p_n^{\theta^*}(Y_1,\ldots,Y_n)))\pi(d\theta)} := \frac{N_n}{D_n}$$

and prove it is small in some sense. To control this quantity, it is common to

(A0.1) use some test Φ_n to test θ^* against $\theta : \ d(\theta,\theta^*) \geq \delta_n$, with small errors $\mathbb{E}^{\theta^*}(\Phi_n)$ and $\sup_{\theta:\ d(\theta,\theta^*)\geq\delta_n} \mathbb{E}^\theta(1-\Phi_n)$,

and

(A0.2) prove that the prior puts not too small probability in some neighbourhood of the true parameter θ^* (the neighbourhood is usually formed in terms of a neighbourhood of the log-likelihood associated to the true parameter). This assumption enables to obtain lower bounds for the denominator D_n.

Then, under assumptions of type (A0.1) and (A0.2),

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left[\Pi(\{\theta : \ d(\theta,\theta^*) \geq \delta_n|Y_1,\ldots Y_n)\right] \\
&\quad \leq \mathbb{E}^{\theta^*}(\Phi_n) + \mathbb{E}^{\theta^*}\left[(1-\Phi_n)\Pi(\{\theta : \ d(\theta,\theta^*) \geq \delta_n|Y_1,\ldots Y_n)\right] \\
&\quad \leq \mathbb{E}^{\theta^*}(\Phi_n) + \mathbb{E}^{\theta^*}\left[(1-\Phi_n)\frac{N_n}{D_n}\right]
\end{aligned}$$

is small because

- $\mathbb{E}^{\theta^*}(\Phi_n)$ is controlled by an assumption of type (A0.1),

- D_n is lower bounded with probability tending to 1 using an assumption of type (A0.2) and
- $\mathbb{E}^{\theta^*}((1-\Phi_n)N_n) = \int_{d(\theta,\theta^*)\geq\delta_n} \mathbb{E}^{\theta}((1-\Phi_n))\pi(d\theta)$ is small using an assumption of type (A0.1).

The needed tests of Assumption (A0.1) may not exist if the set Θ of parameters is too big (e.g. considering the set of density of functions for Θ and the L^1-norm $d(\cdot,\cdot) = \|\cdot - \cdot\|_{L^1}$) then Assumption (A0.1) can be replaced by

(A0.3) the existence of a sequence of sets $\Theta_n \subset \Theta$ (often more and more complex) such that, we can build some tests to distinguish θ^* against $\theta \in \Theta_n : \ d(\theta,\theta^*) \geq \delta_n$ and such that the prior mass of Θ_n is decreasing fast enough .

The existence of the needed tests is often implied thanks to an assumption on the complexity of the sets Θ_n. This may be measured with covering numbers for instance. The δ-covering number $N(\delta, S, d)$ of a set S with respect to a pseudo-metric d, for $\delta > 0$ is the minimum number of d-balls of radius δ needed to cover the set A. Then Assumption (A0.3) can be interpreted as "the prior has to penalize enough too complex sets".

We make precise these statements in the case of posterior consistency with respect to the L^1-norm in the framework of density estimation with real i.i.d. observations. Then Θ is a set of density functions on $\mathbb{R}$ and $p_n^\theta(y_1,\dots,y_n) = \prod_{i=1}^n \theta(y_i)$. The needed test (A0.3) can be built using Hoeffding's inequality (see Ghosh and R.V. Ramamoorthi [GR03]). Here the neighbourhood of Assumption (A0.2) is expressed via the Kullback-Leibler divergence, namely

$$\mathcal{B}_{KL}(\theta^*,\epsilon) = \left\{\theta\in\Theta : \ \int \log\left(\frac{\theta^*(y)}{\theta(y)}\right)\theta^*(y)\lambda(dy)\right\}.$$

Theorem 1.1 (Ghosh and R.V. Ramamoorthi [GR03]). *Assume that*

(B0.1) for all $\delta > 0$, there exists $\Theta_n \subset \Theta$ and $\beta > 0$,

$$\Pi(\Theta_n^c) \leq \exp(-\beta n), \quad \sum_n N(\delta/2, \Theta_n, \|\cdot - \cdot\|_{L_1(\lambda)})\exp(-n\delta^2/2) < +\infty,$$

(B0.2) for all $\epsilon > 0$, $\Pi(\mathcal{B}_{KL}(\theta^*,\epsilon)) > 0$.

Then the posterior is consistent at θ^* with respect to the L^1-norm.

A general result on posterior consistency is given in Barron [Bar88]. We state here one of its results which holds in a general setting, not necessarily in the i.i.d. case nor for density estimation.

Theorem 1.2 (Barron [Bar88]). *Assume that*

(C0.1) for all $\epsilon > 0$, there exists $\Theta_n \subset \Theta$, $S_n \subset \mathcal{Y}^n$ and positive constants β_1, β_2, C_1 and C_2 such that

$$\Pi(\Theta_n^c) \leq C_1\exp(-\beta_1 n)$$

and

$$P^{\theta^*}((Y_1,\dots,Y_n)\in S_n \text{ i.o.})=0, \qquad \sup_{\theta\in\Theta_n:\ d(\theta,\theta^*)>\epsilon} P^{\theta}((Y_1,\dots,Y_n)\in S_n^c)\leq C_2\exp(-n\beta_2)$$

(C0.2) for all $\epsilon>0$,

$$\Pi^{\theta^*}\left(\exists N,\ \forall n\geq N,\ \int_\Theta \frac{p_n^\theta(Y_1,\dots,Y_n)}{p_n^{\theta^*}(Y_1,\dots,Y_n)}\pi(d\theta)>\exp(-n\epsilon)\right)=1.$$

Then the posterior is consistent at θ^* *with respect to* d.

Note that in the setting of Theorem 1.1, Assumption (C0.1) is implied by (B0.1) and Assumption (C0.2) is implied by (B0.2). For more information on posterior consistency see Ghosh and R.V. Ramamoorthi [GR03], Rousseau [Rou15] and references therein.

To obtain posterior concentration rates, the neighbourhoods of Assumption (A0.2) considered in Ghosal *et al.* [GGV00] and Ghosal and van der Vaart [GV07a] are of the form

$$\mathcal{B}_n^2(\theta^*,\epsilon)=\left\{\theta:\ \mathbb{E}^{\theta^*}\left(\log\left(\frac{p_n^{\theta^*}(Y_1,\dots,Y_n)}{p_n^{\theta}(Y_1,\dots,Y_n)}\right)\right)\leq n\epsilon,\ Var^{\theta^*}\left(\log\left(\frac{p_n^{\theta^*}(Y_1,\dots,Y_n)}{p_n^{\theta}(Y_1,\dots,Y_n)}\right)\right)\leq n\epsilon\right\}. \tag{1.3}$$

The next theorem holds in a general setting, not necessarily in the i.i.d. case nor for density estimation.

Theorem 1.3 (Ghosal and van der Vaart [GV07a])**.** *Let* ϵ_n *be a positive sequence tending to* 0 *such that* $1/(n\epsilon_n^2)=O(1)$. *Assume that there exists positive constants* $C_0, C_1, C_2, C_3, K_0, K_1, K_2, K_3$, M *such that* $C_3-C_1-C_2>-1$ *and for all* n,

(D0.1) there exists $\Theta_n\subset\Theta$ *and a test function* ϕ_n *such that*

$$\begin{aligned}&\mathbb{E}^{\theta^*}(\Phi_n)=o(1),\\ &\mathbb{E}^{\theta}(1-\Phi_n)\leq K_1\exp(-C_1n\epsilon_n^2),\ \forall\theta\in\Theta_n\cap\{\theta:\ d(\theta,\theta^*)\geq M\epsilon_n\},\\ &\Pi(\Theta_n)\leq K_2\exp(-C_2n\epsilon_n^2),\end{aligned}$$

(D0.2) $\Pi\left(\mathcal{B}_n^2(\theta^*,\epsilon_n)\right)\geq K_3\exp(-C_3n\epsilon_n^2)$.

Then the posterior distribution concentrates at rate ϵ_n *at* θ^* *with respect to* d.

More results and references on posterior consistency and posterior concentration rates can be found in Ghosh and R.V. Ramamoorthi [GR03] and Rousseau [Rou15].

1.2.2 Limit Distributions in the Frequentist and Bayesian Framework. Asymptotic Efficiency

Here I give some tools which are useful to understand Chapter 4. In the latter chapter, we study limit distributions in the frequentist and the Bayesian frameworks. So that, here, we present some results on asymptotic distributions for both frequentist and Bayesian cases.

Once you have obtained a posterior concentration rate or a rate of convergence in the frequentist case, you may zoom in (i.e. do a change of scale) and be interested in limit distributions. Formally,

- in the frequentist case, with an estimator $\widehat{\theta}_n$ and rate ρ_n, you can be interested in the asymptotic distribution of $\rho_n^{-1}(\widehat{\theta}_n - \theta^*)$,
- in the Bayesian case, you may be interested in the asymptotic distribution of the posterior distribution for the rescaled parameter: $\rho_n^{-1}(\theta - \theta^*)$.

The rate of convergence of an estimator is lower bounded by the minimax rate. Similarly, the asymptotic distribution is also bounded in some sense. We first give a 'bound' for the asymptotic distribution in the well-known parametric case and then in the semiparametric case. In both sections, we only consider the i.i.d. setting.

Obtaining limit distribution is useful to build confidence intervals. In the frequentist case, if $\rho_n^{-1}(\widehat{\theta} - \theta)$ tends in distribution to F independent of θ, then $[\widehat{\theta} - \rho_n q_{1-\alpha/2}, \widehat{\theta} - \rho_n q_{\alpha/2}]$ gives an α-asymptotic confidence interval, where q_t is a t-quantile of F. In the Bayesian setting, obtaining the asymptotic posterior distribution of $\rho_n^{-1}(\theta - \theta^*)$ can help in proving that α-credible regions C_α, i.e. regions such that $\Pi(C_\alpha | Y_1, \dots, Y_n) \geq 1 - \alpha$ are also α-asymptotic confidence intervals.

We use the models of Examples 1 and 2, defined in the following, to illustrate the notions all along Section 1.2.2. These models are studied in Chapter 4. Namely, we are going to use mixture models, where the emission distribution consists of a product of three distributions on $[0, 1]$ as illustrated in Figure 1.3, so that the observations $Y_t = (Y_{t,1}, Y_{t,2}, Y_{t,3})$, $t \leq n$ live in $[0, 1]^3$.

Example 1 (Definition)**.** For the first model, no more restrictions are given on the emission distributions (the model is nonparametric). The emission distribution are $\otimes_{c=1}^3 f_{j,c}\lambda$, where $f_{j,c}$ are in $\mathcal{F}$, the set of density functions on $[0, 1]$. The parameters of the model are $\mathbf{f} = (f_{j,c})_{1 \leq j \leq k, 1 \leq c \leq 3} \in \mathcal{F}^{3k}$ and $\mu \in \Delta_k$ determining the distribution of the latent variables. More precisely, given the latent state X_t, the three components of the corresponding observation $Y_{t,1}$, $Y_{t,2}$ and $Y_{t,3}$ are independent with $Y_{t,c}$ distributed from $f_{X_t,c}\lambda$. Then the distribution of one observation is

$$g_{\mu,\mathbf{f}}(y)\lambda(dy) = \sum_{j=1}^{k} \mu_j \prod_{c=1}^{3} f_{j,c}(y_c)\lambda(dy_c).$$

Example 2 (Definition)**.** The second model is a parametric model where the emission distributions are piecewise constant functions with respect to a partition $\mathcal{I}_M = (I_m)_{m \leq M}$ of $[0, 1]$. More

precisely, the parameters of this model are the parameter $\mu \in \Delta_k$, determining the distribution of the latent variables, and $\omega_M \in (\Delta_M)^{3k}$ which parametrizes the emission distributions. The distribution of one observation is

$$g_{\mu,\omega;M}(y)\lambda(dy) = \sum_{j=1}^{k} \mu_j \prod_{c=1}^{3} f_{j,c;M}(y_c)\lambda(dy_c),$$

where $\mathbf{f}_M = (f_{j,c;M})_{j\leq k, 1\leq c\leq 3}$, $f_{j,c;M} = \sum_{m=1}^{M} (\omega_{j,c,m;M}/|I_m|)\mathbb{1}_{I_m}$, $j \leq k, 1 \leq c \leq 3$.

1.2.2.1 The Well Known Parametric Case with i.i.d. Observations

The research of limit distributions is well understood in the parametric case. In this section, we introduce some of the well known results of asymptotic efficiency in parametric models for i.i.d. observations. For more details, see van der Vaart [Vaa98]. We first give some restrictions on the limit distribution. We then introduce the notion of frequentist regular efficient estimator which reaches the distribution bound. We present assumptions such that the maximum likelihood is regular efficient. We end this section with an asymptotic distribution for the posterior distribution.

As we present results in the parametric framework, we assume that the parameter space Θ is a subset of $\mathbb{R}^d$, $d \in \mathbb{N}^*$. In particular, the best (in some sense, see Theorem 1.5) limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is the Gaussian distribution with variance the inverse of the Fisher information. We define the Fisher information, which depends on the model and the true parameter θ^*, in the following. This reveals a real limitation in the task of estimation.

To define the Fisher information, we need some regularity of the model. This regularity is called differentiability in quadratic mean:

Definition 1.6 (Differentiability in quadratic mean and Fisher information)**.** *A model $(p_\theta\lambda)_{\theta\in\Theta}$ is said to be differentiable in quadratic mean at θ if there exists $\dot{\ell}_\theta \in (L^1(p_\theta\lambda))^d$ such that*

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - (1/2)h^T\dot{\ell}_\theta\sqrt{p_\theta}\right)^2 \lambda = o(\|h\|^2). \tag{1.4}$$

In this case, $\int \dot{\ell}_\theta p_\theta \lambda = 0$, $\dot{\ell}_\theta$ is in $L^2(p_\theta\lambda))^d$ and the Fisher information is defined as

$$J_\theta = \int \dot{\ell}_\theta \dot{\ell}_\theta^T p_\theta \lambda.$$

Differentiability in quadratic mean is often proved using the following proposition:

Proposition 1.4 (Lemma 7.6 in van der Vaart [Vaa98])**.** *If*

- *$\theta \mapsto \sqrt{p_\theta(y)}$ is $\mathcal{C}^1$ for all y,*
- *$\theta \mapsto J_\theta = \int \dot{p}_\theta \dot{p}_\theta^T / p_\theta \lambda$ is defined and continuous at θ (with $\dot{p}_\theta = \partial p_\theta/\partial\theta$),*

then $(p_\theta \lambda)_{\theta\in\Theta}$ is differentiable in quadratic mean at θ and $\dot{\ell}_\theta = \dot{p}_\theta / p_\theta$.

Example 2 (Differentiability in quadratic mean)**.** We introduce the set $\underline{\Delta}_s = \{u \in \mathbb{R}_+^{s-1} : \sum_{i=1}^{s-1} u_i \leq 1\}$ which is in bijection with Δ_s (we use the same notation for elements in both sets, making the bijection implicit). So that we consider the parameter set $\Theta_M = \underline{\Delta}_k \times (\underline{\Delta}_M)^{3k}$.
Then the model $(g_{\mu,\omega;M}\lambda)_{(\mu,\omega)\in\Theta}$ is differentiable in quadratic mean at every $(\mu,\omega) \in \Theta$ with $\dot{\ell}_{\mu,\omega}(y_1,y_2,y_3) = (\dot{\ell}_\mu(y_1,y_2,y_3), \dot{\ell}_\omega(y_1,y_2,y_3)) \in \mathbb{R}^{k-1+3k(M-1)}$ defined as:

$$\dot{\ell}_{\mu;i}(y_1,y_2,y_3) = \frac{\prod_{c=1}^3 f_{i,c;M}(y_c) - \prod_{c=1}^3 f_{k,c;M}(y_c)}{g_{\mu,\omega;M}(y_1,y_2,y_3)}, \qquad \text{if } i < k,$$

$$\dot{\ell}_{\omega;i}(y_1,y_2,y_3) = \mu_j \frac{\left(\frac{\mathbb{1}_{I_m}(y_c)}{|I_m|} - \frac{\mathbb{1}_{I_M}(y_c)}{|I_M|}\right) \prod_{c'\neq c} f_{j,c';M}(y_{c'})}{g_{\mu,\omega;M}(y_1,y_2,y_3)}, \qquad \text{if } i = j3(M-1) + (c-1)(M-1) + m.$$

We now recall some of the the limitations inherent to any estimation procedure. First, the Cramér-Rao bound says that the variance of unbiased estimators of $\psi(\theta)$, under regularity conditions, is lower-bounded by

$$\psi'(\theta) J_\theta^{-1} \psi'(\theta).$$

In particular, the variance of an unbiased estimator of θ is lower-bounded by the inverse of the Fisher information J_θ^{-1}.

The following restriction holds for any regular estimator of $\psi(\theta)$, that is for any estimator $\widehat{\psi}_n = \widehat{\psi}_n(Y_1,\ldots,Y_n)$ such that for all h,

$$\sqrt{n}\left(\widehat{\psi}_n - \psi(\theta + h/\sqrt{n})\right)$$

has a limit distribution L_θ, under $P_{\theta+h/\sqrt{n}}$, which does not depend on h:

Theorem 1.5 (Convolution Theorem, Theorem 8.8 in van der Vaart [Vaa98])**.** *Let $(p_n^\theta \lambda)_{\theta\in\Theta}$ be a model differentiable in quadratic mean at θ with invertible Fisher information J_θ. Let $\widehat{\psi}_n$ be a regular estimator of $\psi(\theta)$ with limit distribution L_θ and ψ differentiable at θ,*
then there exists a probability distribution Q_θ such that L_θ is equal to the product convolution of the Gaussian distribution $\mathcal{N}(0, \dot{\psi}_\theta J_\theta^{-1} \dot{\psi}_\theta^T)$ and Q_θ.

Hence, the distribution of a regular estimator of θ around the truth cannot be more concentrated than the Gaussian distribution with variance the inverse information again. In this sense, the best asymptotic distribution is $\mathcal{N}(0, \dot{\psi}_\theta J_\theta^{-1} \dot{\psi}_\theta^T)$.

Regular efficient estimators are estimators achieving this lower bound. They are considered as the best asymptotic estimators. We say that a sequence of estimators $\widehat{\psi}_n$ of $\psi(\theta)$ is regular efficient when it is regular and under P^θ,

$$\sqrt{n}\left(\widehat{\psi}_n - \psi(\theta)\right) \tag{1.5}$$

converges in distribution to a Gaussian distribution $\mathcal{N}(0, \dot{\psi}_\theta J_\theta^{-1} \dot{\psi}_\theta^T)$. Note that a sequence of estimators $\widehat{\psi}_n$ of $\psi(\theta)$ is regular efficient if and only if

$$\sqrt{n}(\widehat{\psi}_n - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_\theta J_\theta^{-1} \dot{\ell}_\theta(Y_i) + o_{P^\theta}(1),$$

where $o_{P^\theta}(1)$ is a sequence tending to 0 in P^θ-probability.

For more information on efficiency, see van der Vaart [Vaa98]. Besides, we may wonder if there exist estimators which achieve this bound. In the following, we present a type of estimator which often is regular efficient.

The renowned maximum likelihood estimator (m.l.e.) is defined as

$$\widehat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n^\theta(Y_1, \dots, Y_n),$$

where $\ell_n^\theta(Y_1, \dots, Y_n) = \log(p_n^\theta(Y_1 \dots, Y_n))$ is the log-likelihood. As we only consider i.i.d. observations, $\ell_n^\theta(Y_1, \dots, Y_n) = \sum_{i=1}^n \log(p^\theta(Y_i))$. The m.l.e. is often (under regularity and identifiability assumptions) a regular efficient estimator. Before giving assumptions leading to asymptotic efficiency of the m.l.e., we give some assumptions under which the m.l.e. is consistent.

Proposition 1.6 (Consistency of the m.l.e., application of Theorem 5.7 in van der Vaart [Vaa98])**.** *Let Θ be a compact subset of $\mathbb{R}^d$. Assume that*

- *for all y, $\theta \mapsto p^\theta(y)$ is continuous,*
- *there exists a function $h \in L_1(p^{\theta^*}\lambda)$ such that $\sup_{\theta \in \Theta} |\log(p^\theta)| \leq h$,*
- *the model is identifiable at θ^* i.e. $p^\theta = p^{\theta^*}$ implies $\theta = \theta^*$*

then the m.l.e. is consistent at θ^*: $\|\theta - \theta^*\|$ tends to zero in p^{θ^*}-probability.

We now apply Proposition 1.6 to the maximum likelihood estimator in the following model.

Example 2 (Consistency of the m.l.e.)**.** To apply Proposition 1.6, we need some identifiability result. We assume that the true parameter satisfies that the μ_i^* are positive for all $i \leq k$, and that for all $c \in \{1, 2, 3\}$, $f_{1,c;M}^* \lambda, \dots, f_{k,c;M}^* \lambda$ are linearly independent distributions. Then, using Theorem 8 of Allman *et al.* [AMR09] (reproduced in Theorem 1.3.1), the equality $g_{\mu,\omega;M} = g_{\mu^*,\omega^*;M}$ implies that there exists a permutation $\sigma \in \mathcal{S}_k$ such that for all $1 \leq i \leq k$, $\mu_i = \mu_{\sigma(i)}^*$, $\omega_{i,c,m} = \omega_{\sigma(i),c,m}^*$. It means that the identifiability assumption holds up to label switching.

To avoid multiple maxima, we constrain the set of parameters in order that only one labelling (one permutation) is possible. Namely, we assume that the true parameter satisfies $\mu_1^* < \mu_2^* < \dots < \mu_k^*$ and we consider

$$\tilde{\Theta}_M = \{(\mu, \omega) \in \Theta_M : \ \mu_1 \leq \mu_2 \dots \leq \mu_k\}.$$

The constraint we have assumed on the true parameter is not necessary but just ease the calculations (if this assumption does not hold then another constraint, as $\mu_2^* < \mu_1^* < \mu_3^* < \cdots$ or $\omega_{1,c,1}^* < \omega_{2,c,1}^*, \cdots$, will work). Then all the assumptions of Proposition 1.6 hold and the maximum likelihood computed on the set $\tilde{\Theta}_M$ is then consistent.

Considering the whole space Θ and the computing the m.l.e. on this space:

$$(\widehat{\mu}_M, \widehat{\omega}_M) \in \underset{(\mu,\omega)\in\Theta}{\arg\max}\, \ell_n^\theta(Y_1, \ldots, Y_n),$$

we then obtain that there exists a sequence $(\sigma_n)_n$ of permutations in $\mathcal{S}_k$ such that

$$\|(\sigma_n(\widehat{\mu}_M), \sigma_n(\widehat{\omega}_M)) - (\mu^*, \omega^*)\|$$

tends to zero in $P^{(\mu^*,\omega^*)}$-probability.

Here are some assumptions implying that the m.l.e. is regular efficient.

Proposition 1.7 (Regular efficiency of the m.l.e., Theorem 7.6 in van der Vaart [Vaa98])**.** *Let $(p^\theta \lambda)_{\theta\in\Theta}$ be a model which is differentiable in quadratic mean at $\theta^* \in \overset{\circ}{\Theta}$ where the Fisher information J_{θ^*} is invertible. Assume there exists $\dot{\ell} \in L_2(P^{\theta^*})$ such that for all θ_1, θ_2 in a neighbourhood of θ^**

$$|\log(p^{\theta_1}(y)) - \log(p^{\theta_2}(y))| \leq \dot{\ell}(y)\|\theta_1 - \theta_2\|. \tag{1.6}$$

If the m.l.e. $\widehat{\theta}_n$ is consistent at θ^* then

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) = \frac{J_{\theta^*}^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(Y_i) + o_{P^{\theta^*}}(1),$$

where $o_{P^{\theta^*}}(1)$ represents a sequence tending to zero in P^{θ^*}-probability. So that, $\widehat{\theta}_n$ is regular efficient.

Let us apply this proposition to the model of Example 2.

Example 2 (Efficiency of the m.l.e.)**.** Consider here $\theta^* = (\mu^*, \omega^*) \in \overset{\circ}{\Theta}_M$. Assumption (1.6) is verified since $\theta \mapsto \log(p^\theta)$ is $\mathcal{C}^1$ with all the components of $\dot{\ell}_{\theta^*}$ bounded by some constant depending on the true parameter and the partition. The invertibility of the Fisher information is proved in Chapter 4. We have already proved the consistency of the m.l.e. so that, we then obtain its regular efficiency (up to label switching).

In the Bayesian framework, an interesting result related to efficiency is the Bernstein von Mises (BvM) Theorem. This theorem gives the asymptotic distribution of the posterior distribution if correctly zoomed in. When the posterior is consistent, it tends to the Dirach mass at the true parameter (see Figure 1.6). So that if we want to see a shape around the true parameter, we

need to zoom in. That is why we change the parametrization. Instead of studying the posterior distribution for θ, we study the posterior distribution for the new parameter $s = \sqrt{n}(\theta - \widehat{\theta}_n)$, doing as if we were focusing on the m.l.e. $\widehat{\theta}_n$ and zooming in with a scale $\sqrt{n}$. We can then write the posterior distribution for s as,

$$\Pi_s(s \in S|Y_1, \dots Y_n) = \frac{\int_{(1/\sqrt{n})S+\widehat{\theta}_n} \prod_{i=1}^n p^\theta(Y_i)\Pi(d\theta)}{\int_\Theta \prod_{i=1}^n p^\theta(Y_i)\Pi(d\theta)},$$

if the observations are i.i.d. from p^θ. Then, under some assumptions, the posterior for the new parameter s tends to a centered Gaussian distribution with covariance the inverse of the Fisher information:

Theorem 1.8 (BvM, Theorem 10.1 in van der Vaart [Vaa98])**.** *Let Θ be a compact subset of $\mathbb{R}^d$ and $(p_\theta \lambda)_{\theta\in\Theta}$ a model differentiable in quadratic mean at θ^* with an invertible Fisher information J_{θ^*}. Assume that*

- *for all y, $\theta \mapsto p^\theta(y)$ is continuous,*
- *the model is identifiable at θ^*, i.e. $p^\theta = p^{\theta^*}$ implies $\theta = \theta^*$,*
- *the prior Π is absolutely continuous with respect to the Lebesgue measure, with density π continuous and positive at θ^* ,*
- *$\widehat{\theta}_n$ is a regular efficient estimator .*

Then

$$\sup_{A\subset\mathbb{R}^d} |\Pi_s(A|Y_1, \dots, Y_n) - \mathcal{N}(0, J_{\theta^*})(A)|$$

tends to zero in P^{θ^*}-probability.

Example 2 (BvM)**.** Given our previous results for this model, as soon as the prior Π_M, defined on Θ_M is absolutely continuous with respect to the Lebesgue measure with a density continuous and positive at θ^*, we obtain a BvM theorem for the associated posterior.

1.2.2.2 The Semiparametric Case with i.i.d. Observations

Here, we define a semiparametric model as a model $(P^\theta)_{\theta\in\Theta}$ where the parameter θ can be decomposed into two components $\theta = (\mu, \eta)$: one of interest μ (often living in a finite-dimensional set) and one nuisance parameter η (often living in a non-finite dimensional set). As in Section 1.2.2.1, we only consider models where the observations are i.i.d.. Here are some examples of semiparametric models.

Example 1 (Semiparametric model)**.** where the parameter of interest is μ (the weights of the mixture) and the emission distributions $f_1, \dots, f_k$ are nuisance parameters is a semiparametric model.

A similar example consists in estimating the transition matrix Q in HMMs (Figure 1.2) without being interested in the emission distributions.

In the following we introduce the tools to obtain limitations in the estimation task in the semiparametric case. Note that in the semiparametric case, it is less easy than in the parametric case to build estimators reaching the distribution bound. Similarly, it is more difficult to obtain results on the posterior distribution.

In semiparametric models, we can obtain limitations in the estimation task by considering linear submodels. Indeed, the estimation task is easier in submodels than in the whole model, when less information is available on the nuisance parameter. By considering a family of linear submodels $\{P_u^t, t \in \mathbb{R}\}_{u \in U}$, passing through θ, we can build a family of score functions called a tangent set at θ and denoted $\dot{\mathcal{P}}$.

Definition 1.7 (Tangent set)**.** *Let a family, indexed by u in some set U, of linear submodels $\{P_u^t, t \in \mathbb{R}\}_{u \in U}$, where for each submodel associated to $u \in U$, $P_u^0 = P^\theta$ and $\{P_u^t, t \in \mathbb{R}\}$ is differentiable in quadratic mean at 0 with score function g_u. Then the tangent set associated with this family at θ is $\{g_u, u \in U\}$.*

In the parametric case with a model $\{P^\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ differentiable in quadratic mean at θ with score $\dot{\ell}_\theta$, the maximal tangent set is $\{u^T \dot{\ell}_\theta, u \in \mathbb{R}^d\}$.

Example 2 (Tangent set)**.** From the previous remark, the maximal tangent set at θ is

$$\{u^T \dot{\ell}_{\mu,\omega}, u \in \mathbb{R}^{k-1+3k(M-1)}\} = \{v^T \dot{\ell}_\mu + w^T \dot{\ell}_\omega, v \in \mathbb{R}^{k-1}, w \in \mathbb{R}^{3k(M-1)}\}.$$

Example 1 (Tangent set)**.** We consider the following submodel,

$$\mathcal{M}_h^{i,c} = \Big\{ P_{\mu,f^t} : f_{j,c'}^t = f_{j,c'} \text{ if } (j,c') \neq (i,c), \\ f_{i,c}^t = \tilde{f}_{i,c;h}^t := f_i \, k(t) \big(2(1+\exp(-2th))\big)^{-1}, \; t \in \mathbb{R} \Big\},$$

with $h \in L_2(f_i \lambda)$ and $\int h f_i \lambda = 0$, $k(t) = (\int f_{i,c}(y)/(2(1+\exp(-2tg(y))))dy)^{-1}$, $1 \leq i \leq k$ and $1 \leq c \leq 3$. These submodels can be seen as linear submodels where the nonparametric emission distributions are varying while the parametric component μ is fixed. Using Proposition 1.4, we obtain that this submodel is differentiable in quadratic mean at 0 with score

$$\frac{h(y_c)\mu_i \prod_{c'=1}^3 f_{i,c'}(y_{c'})}{g_{\mu,f}(y)}.$$

Then $\dot{\mathcal{P}}_f$, the set spanned by the previous scores, is the tangent set associated with the family of submodels

$$\{P^{\mu,\tilde{f}_h^t}, t \in \mathbb{R}\}_{h=(h_{j,c}),\; h_{j,c} \in L^2(f_{i,c}\lambda),\; \int h_{j,c} f_{j,c} \lambda = 0}$$

at θ. Considering the submodel $\{P^{(\mu+tv,f)}, t \in \mathbb{R}\}_{v \in \mathbb{R}^k,\; \sum_j v_j = 0}$ and using again Proposition 1.4,

we obtain the following tangent set $\dot{\mathcal{P}}_\mu = \{v^T \dot{\ell}_\mu, v \in \mathbb{R}^k,\ \sum_j v_j = 0\}$, where

$$(\dot{\ell}_\mu)_i = \frac{\prod_{c=1}^3 f_{i,c}(y_c)}{g_{\mu,f}(y_1, y_2, y_3)}.$$

Finally, we obtain the tangent set

$$\dot{\mathcal{P}} = \{v^T \dot{\ell}_\mu + s,\ v \in \mathbb{R}^k,\ \sum_j v_j = 0, s \in \dot{\mathcal{P}}_f\},$$

for the family of submodels

$$\{P^{\mu+tv, \tilde{f}_h^t}, t \in \mathbb{R}\}_{v \in \mathbb{R}^k,\ \sum_j v_j = 0,\ h=(h_{j,c}),\ h_{j,c} \in L^2(f_{i,c}\lambda),\ \int h_{j,c} f_{j,c} \lambda = 0}.$$

Using Cramér-Rao bound, the variance of an estimator of $\psi(P^\theta)$, with $\psi : \mathcal{P} \mapsto \mathbb{R}$, should be lower-bounded by the supremum of $(\partial\psi(P_t)/\partial t(0))^2 / J_\theta$ over the considered submodels. Indeed, the estimation of $\psi(P^\theta)$ should not be easier when the nuisance parameter is not known than in any submodel. Note that one can choose the family of submodels. If the family is not rich enough (in particular if it does not include submodels for which the estimation is the most difficult) then the supremum bound may not be attainable.

To formalize this idea, we introduce a notion of smoothness of ψ with respect to the tangent set $\dot{\mathcal{P}}$:

Definition 1.8 (Efficient influence function)**.** *A map $\psi : \mathcal{P} \mapsto \mathbb{R}$ is said differentiable at P^θ with respect to the tangent set $\dot{\mathcal{P}}$ if there exists a measurable function $\tilde{\psi}$ such that for all $s \in \dot{\mathcal{P}}$ and submodel P^t with score s at 0, then*

$$\frac{\Psi(P^t) - \Psi(P^0)}{t} \to \int \tilde{\psi} s dP^\theta.$$

The function $\tilde{\psi}$ is called efficient influence function.

In the semiparametric case, considering submodels $(P^{(\mu,\eta)+(tv,e(t))})_{t\in\mathbb{R}}$, the tangent set $\dot{\mathcal{P}}$ is typically constituted of functions of the form $v^T\dot{\ell}_\mu + s$ where $\dot{\ell}_\mu$ is the score function associated to the parametric model where μ is varying while the nuisance parameter is fixed and s is in $\dot{\mathcal{P}}_\eta$ some subset of $L^2(dP^\theta)$, a tangent set associated to submodels where μ is fixed while the nuisance parameter is varying. This case is verified in models of Examples 1 and 2. In this case the influence function associated to the estimation of μ (so that $\psi(P^{\mu,\eta}) = \mu$) is

$$\tilde{\psi} = \tilde{J}^{-1}\tilde{\ell}, \tag{1.7}$$

where $\tilde{\ell}$, called the efficient score function, is the orthogonal projection in $L^2(P^\theta)$ of $\dot{\ell}_\mu$ on the orthogonal (in $L^2(P^\theta)$) of $\dot{\mathcal{P}}_\eta$ and when $\tilde{J} = \int \tilde{\ell}\tilde{\ell}^T dP^\theta$, the efficient information matrix, is

invertible. Indeed, in this case,

$$\frac{\Psi(P^t) - \Psi(P^0)}{t} = \frac{\Psi(P^{(\mu,\eta)+(tu,e(t))}) - \Psi(P^{(\mu,\eta)})}{t} = u$$

with

$$u = \langle u^T \dot{\ell}_\mu + s, \tilde{J}^{-1}\tilde{\ell}\rangle_{L^2(P^\theta)} = \langle u^T \dot{\ell}_\mu + s, \tilde{\psi}\rangle_{L^2(P^\theta)}.$$

When the problem admits some influence function $\tilde{\psi}$, then the supremum of the lower bounds of the variance given by Cramér-Rao is

$$\sup_{u\in U} \frac{\partial\psi(P_u^t)}{\partial t}(0)\ I_u^{-1}\ \frac{\partial\psi(P_u^t)}{\partial t}(0) = \sup_{g\in\dot{\mathcal{P}}} \frac{(\int \tilde{\psi} g dP^\theta)^2}{\int g^2 dP^\theta} = \int \tilde{\psi}^2 dP^\theta,$$

using Cauchy-Schwarz inequality. Then the variance of the efficient influence function,

$$\tilde{I} := \int \tilde{\psi}^2 dP^\theta$$

is considered as the best asymptotic variance to estimate $\psi(\theta)$. When Equation (1.7) holds, then $\tilde{I} = \tilde{J}^{-1}$.

More formally, we can obtain a convolution theorem in semiparametric models which says that in some sense the best attainable distribution is again the Gaussian distribution with covariance $\tilde{I}$.

Theorem 1.9 (Theorem 25.20 in van der Vaart [Vaa98]). *Let $(P^\theta)_{\theta\in\Theta}$ be some model and $\dot{\mathcal{P}}$ an associated tangent set which is a convex cone. Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable with respect to $\dot{\mathcal{P}}$ at θ with influence function $\tilde{\psi}$. Let $\widehat{\psi}_n$ be a regular sequence of estimators of $\psi(\theta)$ with limit distribution L_θ,*
then there exists a probability distribution Q_θ such that L_θ is equal to the product convolution of the Gaussian distribution $\mathcal{N}(0, \tilde{I})$ and Q_θ.

We say that an estimator $\widehat{\psi}_n$ is asymptotically efficient with respect to some tangent set when it is regular and under P^θ

$$\sqrt{n}\left(\widehat{\psi}_n - \psi(P^\theta)\right)$$

tends in distribution to a Gaussian distribution $\mathcal{N}(0, \tilde{I})$. As in the parametric setting, this is equivalent to saying that:

$$\sqrt{n}\left(\widehat{\psi}_n - \psi(P^\theta)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}(Y_i) + o_{P^{\theta^*}}(1).$$

Building estimators which are asymptotically efficient is more difficult in the semiparametric case than in the parametric case. In particular, there exists settings where such estimators do not

exist, see Ritov and Bickel [RB90]. In Chapter 4, we are interesting in building such estimators in the mixture model represented in Figure 1.3.

More results on efficiency in semiparametric models can be found in van der Vaart [Vaa02], Bickel *et al.* [BKRW98] or Bickel *et al.* [BKRW05] for instance.

Semiparametric BvMs are a current active field of research. BvM have been obtained in particular models in Kim and Lee [KL04], Kim [Kim06], Boucheron and Gassiat [BG09], De Blasi and Hjort [DH09] and Bontemps [Bon11] to cite a few. While Shen [She02], Castillo [Cas12a], Bickel and Kleijn [BK12b], Rivoirard and Rousseau [RR12a] and Castillo and Rousseau [CR15] give assumptions leading to semiparametric BvM in general settings.

1.3 Theoretical Guarantees in Nonparametric HMMs and Semiparametric Mixture Models

We have presented some theoretical properties which may be studied in nonparametric and semiparametric settings in the previous section. In this section, we recall some results which have been obtained in the specific context of nonparametric HMMs and semiparametric mixture models.

1.3.1 Identifiability for Nonparametric Latent Models

Identifiability in general latent models is far from being automatic, as we have already pointed out in Section 1.2. Particularly, identifiability in nonparametric latent models is still an active research area.

Concerning the two particular latent models we consider in this thesis, recent results ensure their identifiability up to label switching.

Definition 1.9 (Label switching). *With $\mathcal{X} = \{1, \dots, k\}$ the state space, a relabelling of this state space through a permutation $\sigma \in \mathcal{S}_k$ does not change the model. The permutation of the labels of the hidden states is called label switching.*

For instance, in the case of HMMs, for all $\sigma \in \mathcal{S}_k$, if $\sigma(\mu)_i = \mu_{\sigma(i)}$, $\sigma(Q)_{i,j} = Q_{\sigma(i),\sigma(j)}$, $\sigma(F)_i = F_{\sigma(i)}$ for al $1 \leq i, j \leq k$, then the model associated to the HMM with initial distribution $\sigma(\mu)$, transition matrix $\sigma(Q)$ and emission distributions $\sigma(F)_i$, $i \leq k$ is the same as the model associated to the HMM with initial distribution μ, transition matrix Q and emission distributions F_i, $i \leq k$, i.e.

$$P_n^{(\sigma(\mu),\sigma(Q),\sigma(F))} = P_n^{(\mu,Q,F)}, \quad \text{for all } n.$$

The mixture model represented in Figure 1.3 is identifiable up to label switching using Theorem 8 in Allman *et al.* [AMR09] which says the following.

Theorem 1.10 (Allman *et al.* [AMR09]). *For a mixture model with distribution* $P^{\mu,(P_{i,c})} = \sum_{i=1}^{k} \mu_i \prod_{c=1}^{C} P_{i,c}$ *where* C *and* k *are known.*

If $C \geq 3$, $\mu_i^* > 0$ *for all* $i \leq k$ *and the distributions* $P_{1,c}, \ldots, P_{k,c}$ *are linearly independent for all* $c \leq C$, *then the model is identifiable up to label switching,*

i.e. if $P^{\mu,(P_{i,c})} = P^{\tilde{\mu},(\tilde{P}_{i,c})}$ *then there exists a permutation* $\sigma \in \mathcal{S}_k$ *such that* $\mu_i = \tilde{\mu}_{\sigma(i)}$ *and* $P_{i,c} = \tilde{P}_{\sigma(i),c}$ *for all* $i \leq k$, $c \leq C$.

Using the same type of constraint, that is when the emission distributions are a product of distributions, Bonhomme *et al.* [BJR16b] obtain an identifiability result even when k is not known. Note that other types of assumptions on mixture models are considered to obtain identifiability. For instance, identifiability results are obtained when the emission distribution are translated version of one symmetric distribution as in Bordes *et al.* [BMV06] and Hunter *et al.* [HWH07]. Other results are obtained in the case of $k = 2$ where an emission distribution belongs in a parametric family and the other is in a non-finite dimensional set with diverse constraint as in Bordes *et al.* [BDV06] or Hohmann and Holzmann [HH13] to cite a few.

Using this result, Gassiat *et al.* [GCR15] have proved identifiability up to label switching of nonparametric HMMs with finite state space, i.e. for the model represented in Figure 1.2 under very general assumptions.

Theorem 1.11 (Gassiat *et al.* [GCR15]). *Let* k *be a known integer. Assume the transition matrix* Q *has full rank with stationary distribution* $\underline{\mu}$ *and the emission distributions* $F_1, \ldots, F_k$ *are linearly independent. As soon as* $C \geq 3$,
if $P_C^{\tilde{\underline{\mu}},\tilde{Q},\tilde{F}} = P_C^{\underline{\mu},Q,F}$, *with* $\tilde{\underline{\mu}}$ *a stationary distribution associated to* $\tilde{Q}$, *then there exists a permutation* $\sigma \in \mathcal{S}_k$ *such that* $\tilde{\underline{\mu}}_i = \underline{\mu}_{\sigma(i)}$, $\tilde{Q}_{i,j} = Q_{\sigma(i),\sigma(j)}$, $\tilde{F}_i = F_{\sigma(i)}$ *for al* $1 \leq i, j \leq k$.

In the context of HMMs, Gassiat and Rousseau [GR16] have proved identifiability up to label switching of HMMs with translated emission distributions. Moreover Alexandrovich *et al.* [AHL16] have obtained identifiability of HMMs in the case where the number k of states is unknown.

1.3.2 Asymptotic results in Nonparametric HMMs

Results on identifiability in nonparametric HMMs are very recent, so few theoretical guarantees have been studied in nonparametric HMMs.

In Gassiat and Rousseau [GR16], a nonparametric HMM with translated emission distributions has been considered, following the model considered in Yau *et al.* [YPRH11]. The authors propose a consistent estimator of k along with $\sqrt{n}$ convergent, asymptotic normal estimators of the transition matrix and the translation parameters. They also propose an estimator of the marginal stationary density of an observation and deduce a minimax adaptive estimate of the translated density function in the case where $\max_j \mu_j^* > 1/2$.

For the model presented in Figure 1.2 and that I study in Chapters 2 and 3, De Castro *et al.* [DGLar] propose a frequentist penalized least squares estimator for the emission distributions which converges to the truth at an adaptive rate.

Consistency of estimators of the smoothing distribution, that is the distribution of a hidden state given the observations, is studied in De Castro *et al.* [DGC15].

Bayesian HMMs where the emission distributions are parametrised with a parameter living in a finite dimensional set but the number k of possible latent states is not known is studied in Gassiat and Rousseau [GR14] and van Havre *et al.* [HRWM16]. In particular, a test of type (A0.3) to obtain posterior concentration rates in HMMs is developed in Gassiat and Rousseau [GR14]. We use these tests in Chapters 2 and 3.

1.3.3 Asymptotic Behaviours in Semiparametric Mixture Models

Mixture models are often used to estimate density functions. In particular, results on the quality of approximation of density functions with mixture models can be found in the Bayesian literature as in Kruijer *et al.* [KRV10], Scricciolo [Scr14] to cite but a few.

In the following we focus on results concerning the estimation of some parameter in semiparametric mixture models. Such results are very recent. Indeed identifiability in such a framework has been proved only very recently under two settings, as discussed in Section 1.3.1.

- In the first setting considered in the literature the observations are univariate. Bordes and Vandekerkhove [BV10] study this framework in the particular case where

$$Y \sim \mu g(\cdot) + (1-\mu) f(\cdot - m)$$

  with unknown $\mu \in (0,1)$, unknown $m \neq 0$, unknown symmetric density function f and known density function g. The authors give the asymptotic distribution of their estimators of μ, m and the cumulative cumulative distribution function associated with f. Xiang *et al.* [XYW14] also consider this setting but with

$$Y \sim \mu g(\cdot, \xi) + (1-\mu) f(\cdot - m)$$

  with unknown $\mu \in (0,1)$, unknown $m \neq 0$, unknown $\xi \in \mathbb{R}$, unknown symmetric density function f and known parametric family of density functions $(g(\cdot,\xi))_\xi$. The authors prove the asymptotic normality of their estimator of μ, m and ξ. See for instance Hu *et al.* [HWY16], Ma and Yao [MY15] and references therein for other asymptotic results in semiparametric univariate mixture models.

- In the second setting, the emission distributions are a product of more than three distributions so that the observations are multidimensional. This is the setting we consider

in Chapter 4. Previously, up to my knowledge, only Bonhomme *et al.* [BJR16b; BJR16a] studied asymptotics in this setting. In the case where

$$Y \in \mathbb{R}^l, \quad Y \sim \sum_{j=1}^{k} \mu_j f_j(\cdot), \quad f_j(y_1, y_2, \dots, y_l) = g_j(y_1) g_j(y_2) \dots g_j(y_l)$$

with unknown density function g_j on $\mathbb{R}$, unknown $\mu \in \Delta_k$ and unknown $k \in \mathbb{N}$, Bonhomme *et al.* [BJR16b] obtain an asymptotic normal estimator of the mixture parameter μ, which is based on a constructive identification of the parameters from the mixing distribution. Bonhomme *et al.* [BJR16a] present a constructive identification of the parameters from a mixture model where emission distributions are a product of at least three distributions or from a HMM. From this result, they propose estimators, built on simultaneous diagonalization of matrices, whose the parametric component is consistent and asymptotically normal.

Note that the behaviour of the posterior distribution is studied in Rousseau and Mengersen [RM11], in the case of mixture model where the emission distributions live in a finite dimensional set but k is not known.

1.4 My Contributions

During my PhD, I have worked on three projects. All of them were aimed at understanding the asymptotic behaviour of the posterior distribution and estimators in the framework of nonparametric finite state space HMMs and multivariate finite mixtures. In this work "nonparametric" means that the emission distributions are not constrained to live in a finite-dimensional setting, but I will always assume that I know the number k of possible states taken by the latent variable.

I have first been interested in understanding when the posterior distribution is consistent in finite state space nonparametric HHMs, see Chapter 2. Then I have been interested in posterior concentration rates in the same model, see Chapter 3. The last question I have worked on concerns the estimation of only a part of the components of the parameter namely the parameters that give the distribution of the latent variables. This semiparametric problem is studied in Chapter 4. I have studied this problem in nonparametric finite mixture models with i.i.d. observations. This is a first step for studying the analogue question in HMMs. This last work, contrary to the two previous ones, is a joint work with my two PhD supervisors Elisabeth Gassiat and Judith Rousseau.

In the following, I present my contributions in words. For more (mathematical) details, see the associated chapters.

1.4.1 Contribution 1: Posterior Consistency in Nonparametric HMMs with Finite State Space, Chapter 2, Vernet [Ver15b]

My first contribution deals with posterior consistency in nonparametric HMMs with finite state space. This contribution is detailed in Chapter 2 which also corresponds to the paper Vernet [Ver15b] published in EJS.

I quickly recall here the setting. The model is parametrized by $\theta = (Q, f)$ where Q is the transition matrix and f the vector of emission distributions, see Figure 1.2 for a visualization of the model. We use a prior distribution $\Pi = \Pi_Q \otimes \Pi_f^{(k)}$ which is a product of a probability distribution Π_Q on transition matrices and a probability distribution $\Pi_f^{(k)}$ on the k emission distributions. By the Bayes' rule, we can formally write the posterior distribution as

$$\Pi(\theta \in A | Y_1, \dots, Y_n) = \frac{\int_A p_n^\theta(Y_1, \dots, Y_n)\Pi(d\theta)}{\int_\Theta p_n^\theta(Y_1, \dots, Y_n)\Pi(d\theta)},$$

where $p_n^\theta(y_1, \dots, y_n) = \sum_{1 \leq i_1, \dots, i_n \leq k} \mu_{i_1} Q_{i_1, i_2} \dots Q_{i_{n-1}, i_n} f_{i_1}(y_1) \dots f_{i_n}(y_n)$ is the likelihood.

Posterior Consistency for the Marginal Distribution P_l^θ of l Consecutive Observations

In this setting, I have studied posterior consistency for different topologies on different objects. First I was interested in knowing if the posterior distribution concentrates around parameters θ such that the corresponding distribution P_l^θ of l stationary consecutive observations is close to the true one $P_l^{\theta^*}$. It is interesting to know if the posterior concentrates with respect to this object in a prediction perspective. Indeed under this consistency, the density of the observations is consistently estimated. This study should also help in the perspective of estimating Q and f. Indeed if $l \geq 3$, P_l^θ identifies θ (see Theorem 1.11). We develop consistency for the estimation of θ in the following section.

We compare the distributions $(P_l^\theta)_\theta$ thanks to two topologies. We use the topology $\mathcal{T}_w$ associated to the weak convergence on distributions. We also consider the strongest topology $\mathcal{T}_l$ associated to the L_1-norm and which corresponds to the pseudo-distance D_l on Θ:

$$D_l(\theta, \tilde{\theta}) = \|p_l^\theta - p_l^{\tilde{\theta}}\|_{L^1(\lambda^{\otimes l})}.$$

To obtain a general consistency theorem, we used Barron [Bar88], see Theorem 1.2. We have clarified the assumptions of Theorem 1.2 in the case of HMMs. The existence of the sets Θ_n and S_n of Assumption (C0.1) was proved using the tests built in Gassiat and Rousseau [GR14], which are based on a generalisation of Hoeffding's inequality to dependent data by Rio [Rio00]. The main question was to develop an explicit set of parameters associated to log-likelihoods close to the true one, to explicit Assumption (C0.2). This is done in Lemma 2.2.

Then, we have obtained that if Π_Q puts enough mass in the neighbourhood of Q^* (see Assump-

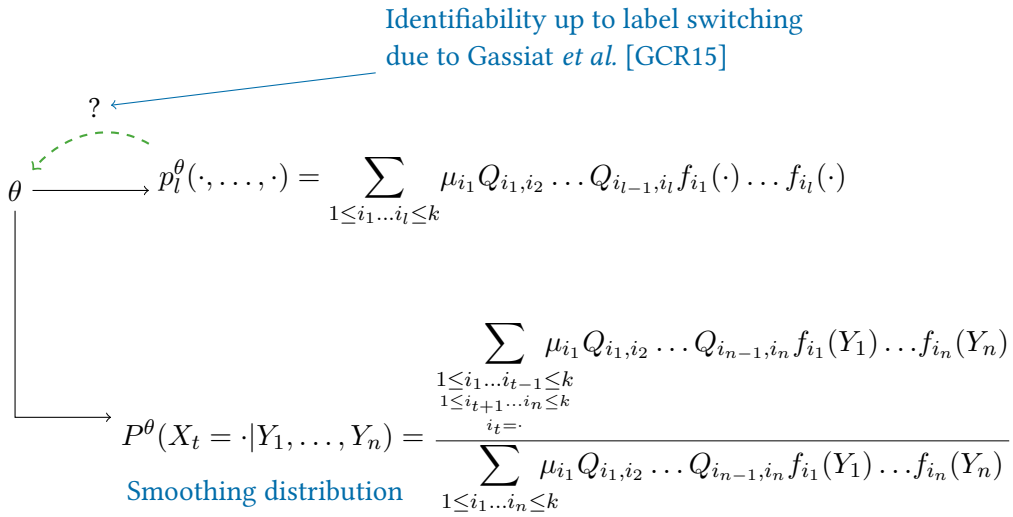


Figure 1.8 – Illustration of the issue to obtain information on θ and the smoothing distribution from p_l^θ

tion (A1.1a)) and Π_f puts enough mass in the neighbourhood of f^* (see Assumptions (A1.1b), (A1.1c) and (A1.1d)) then the posterior distribution is consistent at $\theta^* = (Q^*, f^*)$ with respect to $\mathcal{T}_w$, see Theorem 2.1. We have also proved that if moreover Π_f does not put too much mass on too big sets (see Assumption (A1.2)) then the posterior distribution is consistent at $\theta^* = (Q^*, f^*)$ with respect to $\mathcal{T}_l$, see Theorem 2.1.

The two previous topologies $\mathcal{T}_w$ and $\mathcal{T}_l$ are the easiest types of topology to prove posterior consistency results. Indeed, the topologies concerned the distributions P_l^θ for which it is easier to build tests as (C0.1). Yet one may be interested in other quantities than the distribution of the observations, like the parameter θ by itself or smoothing distributions (the distribution of a hidden state given the observations) for instance. So that we wanted to understand what it means on Q and f or on $P^{(Q,f)}(X_t = \cdot|Y_1,\dots,Y_n)$ when $p_l^{(Q,f)}$ is close to $p_l^{(Q^*,f^*)}$ in L^1, see Figure 1.8 for an illustration.

Posterior Consistency for the Parameters Q and f

If we are interested in estimating the parameter θ itself (and not P_l^θ), i.e. the transition matrix Q and the emission density functions f_j, $j \le k$, it is useful to obtain posterior consistency for the topology $\mathcal{T}_{Q,f}$ which is the product topology of the sup norm on transition matrices and the weak topology (associated to a distance d_{weak} on the emission distributions up to label switching). Thus, we want to know if the posterior distribution concentrates around parameters (Q, f) where $\|Q - \sigma(Q^*)\|$ and $\max_{1\le j\le k} d_{\text{weak}}(F_j, \sigma(F^*)_j)$ are small.

Obtaining posterior consistency with respect to $\mathcal{T}_{Q,f}$ from posterior consistency with respect to

$\mathcal{T}_w$ or $\mathcal{T}_l$ is linked to identifiability (see Section 1.3.1). Then, understanding the inverse of

$$\theta \in \Theta \mapsto p_l^\theta \tag{1.8}$$

can help. Of course, we cannot hope to recover exactly (Q, f) in generality but only the parameter up to label switching (as for the identifiability). Indeed, imagine that the prior is compatible with label switching, that is

$$\Pi(U) = \Pi(\sigma U), \quad \forall U \subset \Theta, \quad \forall \sigma \in \mathcal{S}_k,$$
$$\sigma U = \{((Q_{\sigma(i),\sigma(j)})_{i,j}, (f_{\sigma(1)}, \dots, f_{\sigma(k)})) \in \Theta : \ (Q, f) \in U\}.$$

Then the posterior mass of a set U of parameters is also equal to the posterior mass of the parameters in U for which the label of the hidden states have been switched with a permutation σ, formally

$$\Pi(U|Y_1, \dots, Y_n) = \frac{\int_U p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)}{\int_\Theta p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)} = \frac{\int_U p_n^{\sigma\theta}(Y_1, \dots Y_n)\Pi(d\sigma\theta)}{\int_\Theta p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)} = \Pi(\sigma U|Y_1, \dots, Y_n),$$

for all permutations $\sigma \in \mathcal{S}_k$. Then, the best behaviour of the posterior concerning consistency, would be that the posterior concentrates around the set $\{\theta^*\}_{\mathcal{S}_k} = \cup_{\sigma\in\mathcal{S}_k}\sigma\{\theta^*\}$ composed of the permutations of the true parameters. If the prior distribution is more general, when the number of observation increases, the prior should be 'forgotten' and we should ask the concentration of the posterior distribution at the same set $\{\theta^*\}_{\mathcal{S}_k}$.

In Theorem 2.3, we obtain that posterior consistency for D_l (with $l \geq 3$) along with identifiability for the true parameter implies posterior consistency with respect to $\mathcal{T}_{Q,f}$. The transfer of the property of posterior consistency from one topology to another is done thanks to continuity arguments of the inverse of (A.1).

Posterior Consistency for the Smoothing Distribution

Finite state space HMMs are often used to cluster the observations given the hidden states. In this context, smoothing distribution that is the distribution of a hidden state given the observations

$$P^\theta(X_t = \cdot|Y_1, \dots, Y_n)$$

are important quantities. More precisely, we consider the distribution of a finite sequence of consecutive hidden states given the observations, that is a m-joint smoothing distribution:

$$P^\theta((X_1, \dots, X_m) = (\cdot, \dots, \cdot)|Y_1, \dots, Y_n), \quad m \in \mathbb{N} \text{ fixed }, n \geq m.$$

In Chapter 2, we also study posterior consistency with respect to the m-joint smoothing distribution, i.e., we want to know if the posterior distribution concentrates around parameters for

which the associated m-joint smoothing distribution is close to the true one (up to label switching). This type of consistency should lead to a posterior distribution which clusters well the observations with respect to their hidden states.

In Theorem 2.8, we prove that identifiability of the model for the true parameter, posterior consistency for the D_l pseudo-metric and posterior consistency for $\mathcal{T}_{Q,f}$ leads to a posterior which concentrates around parameters θ for which the associated m-joint smoothing distribution are close to the true smoothing distribution in the particular context of discrete observations.

Note that consistency for the estimation of smoothing distributions have been studied in De Castro *et al.* [DGC15] in the frequentist point of view. In De Castro *et al.* [DGC15], the total variation distance between two smoothing distribution associated with two parameters $\tilde{\theta}$ and θ is controlled with the Frobenius norm $\|\tilde{Q} - Q\|_{FB}$ and the L^1-norms $\|\tilde{f}_j - f_j\|_{L^1}$. This enables to prove that consistency for the transition matrix and the emission distribution with respect to L^1 implies consistency for the smoothing distributions. To deduce a Bayesian result, a Bayesian control of $\|\tilde{f}_j - f_j\|_{L^1}$ is needed. As far as I know, such a control only exists in the case of discrete observations thanks to Chapter 2. Indeed in this case, weak topology on distributions and L_1 topology are the same and Theorem 2.1 can be used. Then we obtain Bayesian consistency for the smoothing distribution in the same framework I obtained results previously, see Theorem 2.8.

Applications to Different Prior Distributions and Settings

In Section 2.3, I propose concrete frameworks and prior distributions leading to posterior consistency for the different topologies introduced before. We consider:

- continuous observations, with emission distribution i.i.d. as a mixture of Gaussian distributions under the prior distribution, in Section 2.3.1,
- continuous observations, with translated emission distribution $f_j = g(\cdot - m_j)$ and g distributed as a mixture of Gaussian distributions under the prior distribution, in Section 2.3.2,
- discrete observations, with emission distribution i.i.d. as a Dirichlet process under the prior distribution, in Section 2.3.3.

Limitation

• The assumption on the support of the prior on the transition matrices Π_Q, which is assumed to obtain posterior consistency with respect to D_l, requires to know a lower bound on the components of the transition matrix. This assumption enables to control the mixing properties of the HMMs, to ensure the existence of some tests, namely the one built in Gassiat and Rousseau [GR14]. In Chapter 3 (on posterior concentration rates), I don't need this assumption but I make a stronger assumption on f^* and Π_f.

Perspectives

• As noticed before, we still do not know assumptions under which the posterior is consistent for the L_1-norm on the emission distributions (up to label switching, of course). It would be interesting to find some, first because it would ensure a good estimation of the emission distributions. It would also ensure a good clustering of the observations using De Castro *et al.* [DGC15].

• Another perspective is to study posterior consistency when the number k of possible states of the Markov chain is unknown. This setting has been studied in Gassiat and Rousseau [GR14] and van Havre *et al.* [HRWM16] but mostly when the emission distributions are parametrised by a finite dimensional parameter. Mixing the techniques of Gassiat and Rousseau [GR14], van Havre *et al.* [HRWM16] and Chapter 2 should be conclusive.

1.4.2 Contribution 2: Posterior Concentration Rates for Nonparametric HMMs with Finite State Space, Chapter 3, Vernet [Ver15a]

For the same setting, I have also studied posterior concentration rates, that is at which rate the posterior concentrates. This contribution is detailed in Chapter 3, and is also available on arXiv: Vernet [Ver15a].

Concerning the concentration rates, I have only used the topology associated to the pseudo-metric $D_l(\theta, \tilde{\theta}) = \|p_l^\theta - p_l^{\tilde{\theta}}\|_{L_1}$. I recall that this topology is interesting for the estimation of the density of the observations, and then for prediction. It is also a first step to obtain posterior concentration rates with respect to the L^1-norm on the emission distributions.

Posterior Concentration Rates with respect to D_l

My aim was to obtain explicit and applicable assumptions on Π_Q, Π_f, Q^* and f^* leading to rates. To do so, I have used Ghosal and van der Vaart [GV07a] (see Theorem 1.3) and I have clarified their assumptions in the HMM case. Posterior concentration rates are more difficult to obtain than posterior consistency. Indeed, to clarify Assumption (D0.2), we need to control the neighbourhood described in (1.3), which requires a better understanding of the likelihood around the true parameter. I have developed new controls of this "neighbourhood" helped by results on parametric HMMs like Douc and Matias [DM01] and Douc *et al.* [DMR04], see Lemma 3.2 and 3.3. It asked me much work to obtain assumptions which are satisfied by usual prior distributions. The existence of the set Θ_n and the test ϕ_n was again proved using the test built in Gassiat and Rousseau [GR14].

Finally, I have obtained a general theorem (Theorem 3.1) which relates the concentration rate with respect to D_l to the prior $(\Pi_Q, \Pi_f^{[k]})$ and the true parameter (Q^*, f^*). The resultant rate has the following form $\epsilon_n/\underline{q}_n$ where ϵ_n depends on the "nonparametric setting", namely $\Pi_f^{(k)}$ and f^* while $\underline{q}_n$ depends on Π_Q. So that the rate ϵ_n is deteriorated by $\underline{q}_n$, that is by the freedom

given to Π_Q concerning the mixing of the Markov chain associated to Q.

Application to Different Prior Distributions and Settings

This theorem is applied in several frameworks and leads to minimax rates up to $\log(n)$, in different settings, see Section 3.4.

Particularly, minimax rates are obtained in the case of discrete observations with emission distributions i.i.d. from a Dirichlet process under the prior distribution. More precisely, we obtain rates equal to $1/\sqrt{n}$ up to $\log(n)$. See Section 3.4.1.

Moreover adaptive concentration rates are reached in the case of continuous observations, emission density functions i.i.d. from a Dirichlet process mixture of Gaussian distributions under the prior distribution, and Hölder-type classes of functions. More precisely we obtain rates equal to $n^{-\beta/(2\beta+1)}$ up to $\log(n)$ when the emission density functions are in a β-Hölder class of functions in Section 3.4.2.

In the two previous settings, we obtain minimax rates providing that Π_Q penalizes enough the border of Δ_k^k. More generally, if $\Pi_f^{(k)} = \Pi_f^{\otimes k}$ with Π_f leading to minimax posterior concentration rates with respect to the L_1-norm on densities in the case of density estimation with i.i.d. observations, then the posterior distribution should concentrate at a minimax rate (in HMM) providing that Π_Q penalizes enough the border of Δ_k^k.

Note that, when adaptive, the obtained rates for a class of functions and $\Pi_f^{(k)} = \Pi_f^{\otimes k}$ is the same as in the i.i.d. case for the L_1-norm on densities with the same class and prior Π_f. So that, in our examples, the dependency generated by HMMs on observations does not deteriorate rates compared to the i.i.d. setting. The same remark is done in De Castro *et al.* [DGLar] and Bonhomme *et al.* [BJR16a] where rates of convergence for frequentist estimators are studied.

This contribution concerns concentration rates. Yet, if a posterior distribution concentrates at a rate decreasing to zero with respect to D_l, then it is consistent with respect to D_l. Thus posterior consistency for the topology $\mathcal{T}_{Q,f}$ (useful for the estimation of θ) is also implied by the assumptions leading to a posterior concentration rate decreasing to zero with respect to D_l, using Theorem 2.3 of Chapter 2.

Perspectives

• The assumption on Π_Q, concerning the penalization of transition matrices too close to the border of Δ_k^k, is much weaker than the one we assume to prove posterior consistency. Yet it is still strong and such prior distributions are not used in practice. It would be interesting to know if this assumption is necessary or not.

• A perspective of this work is the transfer of the rate with respect to D_l to a rate with respect to the L_1-norm on the emission distributions. This transfer is even more difficult in the case of

rates than in the case of consistency. Indeed, the transfer of consistency from one topology to another is linked to continuity while the transfer of rate from one topology to another is related to modulus of continuity. Then this problem may be solved thanks to a better understanding of the inverse of the function $\theta \mapsto p_l^\theta$ again. This transfer has been recently done in other settings, as in De Castro *et al.* [DGLar] for the L_2-norm. These works might be good approaches for the resolution of this perspective.

1.4.3 Contribution 3: Efficient Semiparametric Estimation and Model Selection for Multidimensional Mixtures (joint work with E. Gassiat and J. Rousseau), Chapter 4, Gassiat *et al.* [GRV16]

My last contribution concerns a semiparametric problem. It is a collaborative work with my two thesis supervisors Elisabeth Gassiat (Paris-Sud University) and Judith Rousseau (CEREMADE). It is also available on arXiv: Gassiat *et al.* [GRV16].

Our goal is to study asymptotic efficiency for a component of the parameter, namely the transition matrix in the case of HMMs or the mixture parameter in the case of mixture models.

For the time being, we only have results in the case of mixture models and not for HMMs. Indeed, the likelihood and score functions are easier to handle in mixture models than in HMMs.

We now present the setting used in Chapter 4. Again the hidden states X_t live in a finite states space $\{1, \dots, k\}$ where k is known. These states are i.i.d. from some distribution $\sum_{i=1}^k \mu_i \delta_i$. Moreover the observations Y_t, $t \in \mathbb{N}$, live in $[0,1]^3$. Given a hidden state X_t, the three components $Y_{t,1}$, $Y_{t,2}$ and $Y_{t,3}$ of the observation Y_t are independent with respective distribution $f_{X_t,1}d\lambda$, $f_{X_t,2}d\lambda$ and $f_{X_t,3}d\lambda$. This model can be visualized in Figure 1.3. We have seen in Section 1.3.1, that this model is identifiable up to label switching under general assumptions.

Asymptotic Efficiency

To obtain regular efficient estimators, we use approximation models. Namely we project the emission distributions on the set of histograms associated to a fixed partition $\mathcal{I}_M$ of $[0,1]$ we then consider the models of Example 2 in Section 1.2.2. Thus, the parameters of this model are the parameter $\mu \in \Delta_k$, determining the distribution of the latent variables, and $\omega_M \in (\Delta_M)^{3k}$ which parametrizes the emission distributions. The distribution of one observation is

$$g_{\mu,\omega;M}(y)\lambda(dy) = \sum_{j=1}^{k} \mu_j \prod_{c=1}^{3} f_{j,c;M}(y_c)\lambda(dy_c),$$

where $\mathbf{f}_M = (f_{j,c;M})_{j\le k, 1\le c\le 3}$, $f_{j,c;M} = \sum_{m=1}^{M} (\omega_{j,c,m;M}/|I_m|)\mathbb{1}_{I_m}$, $j \le k, 1 \le c \le 3$.

Following Section 1.2.2.1, a maximum likelihood estimator $(\widehat{\theta}_M, \widehat{\omega}_M)$, associated to the approximation model is, up to label switching, asymptotically normal around (θ^*, ω^*), where $\omega^*_{i,j,m} =$

$\int_{I_m} f^*_{j,c} d\lambda$. Moreover $\widehat{\theta}$ is regular and asymptotically Gaussian around θ^*, yet as asymptotic variance the inverse of the Fisher information associated to the approximated model, which may be different from the inverse of the efficient Fisher information for the complete semiparametric model. Yet by refining the partition slowly enough when the number of observations increases, we obtain an estimator $\widehat{\theta}_{M_n}$ regular efficient of θ^*, see Theorem 4.5.

More precisely, we first obtain that when the partition is refined the Fisher information, associated to the approximated model, increases; see 4.2. Moreover, when the partition is refined such that the sets have a size tending to zero, then the Fisher information of the approximated models is tending to the efficient Fisher information for the semiparametric model; see 4.3. Finally, we prove the existence of a refinement M_n of the partition such that the associated sequence of m.l.e. θ_{M_n} is regular efficient in the semiparametric model; see Theorem 4.5.

Similarly, if we have a family of prior distributions $(\Pi_M)_M$, one for each model associated with a partition $\mathcal{I}_M$, which are absolutely continuous with respect to the Lebesgue measure and positive on their defining sets; then by refining the partition slowly enough, we obtain a Bernstein von Mises type theorem. That is there exists a refinement L_n of the partition such that the associated sequence of posterior distributions $\Pi_{L_n}(|Y_1, \dots, Y_n)$ verifies a Bernstein von Mises theorem; see Theorem 4.5.

Model Selection

The two previous results are existence results but are not constructive, namely they don't give clue on the choice of the refinement M_n. Moreover, in Section 4.3.1, we state that if the refinement M_n is done too quickly in the case of m.l.e., then the sequence of m.l.e. $\widehat{\theta}_{M_n}$ is tending almost surely to the uniform weight and thus is not even consistent. So that the choice of the partition has a real impact on the estimation.

Then we propose a procedure to select the refinement of a collection of partitions based on cross validation. In Theorem 4.8, we obtain an oracle inequality for the risk of the selected estimator, but as if we had less ($a_n << n$) observations than we actually have (n). This choice could lead to too conservative selections. We think that this conservatism does not change the good asymptotic properties of the estimator.

Finally, we apply our selection criterion in simulations. We were there surprised to see that even in a finite horizon setting (when n is fixed), our 'conservative' procedure is doing well. See Section 4.4.

Perspectives

- We would like to obtain a rate on the refinement (on M_n) which ensures asymptotic efficiency.
- We would also like to generalize the last results in the case of HMMs. This issue is far from

being obvious using the results obtained in Chapter 4. Indeed even in parametric HMMs, the asymptotic properties of the m.l.e. are not immediate, see DoMa01, Douc *et al.* [DMR04], Douc *et al.* [DMOH11] and references in Cappé *et al.* [CMR05] for instance.

• Finally, in the Bayesian setting, it would be interesting to know if in the two latent models studied in this thesis (HMMs and mixture models), there exists a prior distribution leading to a posterior distribution with optimal concentration simultaneously for both the parameter describing the latent model (transition matrix or mixture parameter) and for the emission distributions.

1.4.4 Summary

The results I have obtained during my PhD on nonparametric HMMs and semiparametric multidimensional mixture models with finite state space is summarized in the following tabular:

estimation of the	posterior consistency, Chapter 2	posterior concentration rates, Chapter 3	asymptotic efficiency, Chapter 4
density p_l^θ	✓	✓	
parameter describing the distribution of the latent states μ or Q	✓		✓
emission density functions $f_1, \dots, f_k$	✓ (in weak topology)		
smoothing distribution $P(X_t = \cdot \vert Y_1, \dots, Y_n)$	✓ (when the observations are discrete)		
	in nonparametric HMMs		in semiparametric mixture models

The checked cells correspond to problems I have studied and for which I have obtained results. Brackets are used to precise some restrictions. As far as I know, the results of the second and third columns are the first results on the asymptotic behaviour of the posterior distribution in nonparametric HMMs with finite state space. Similarly, the result corresponding to the check cell in the fourth column is the only result we are aware on asymptotic efficiency for semiparametric mixture models with three independent observation per latent variable. The empty cells correspond to open problems I would be happy to work on.

CHAPTER 2

POSTERIOR CONSISTENCY IN NONPARAMETRIC HIDDEN MARKOV MODELS WITH FINITE STATE SPACE

In this chapter we study posterior consistency for different topologies on the parameters for hidden Markov models with finite state space. We first obtain weak and strong posterior consistency for the marginal density function of finitely many consecutive observations. We deduce posterior consistency for the different components of the parameter. We also obtain posterior consistency for marginal smoothing distributions in the discrete case. We finally apply our results to independent emission distributions, translated emission distributions and discrete HMMs, under various types of prior distributions.

2.1 Introduction

Hidden Markov models (HMMs) have been widely used in diverse fields such as speech recognition, genomics or econometrics since their introduction in Baum and Petrie [BP66]. The books MacDonald and Zucchini [MZ97], MacDonald and Zucchini [MZ09], and Cappé *et al.* [CMR05] provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. Finite state space HMMs are stochastic processes $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $(X_t)_{t\in\mathbb{N}}$ is a Markov chain taking values in a finite set, and conditionally to $(X_t)_{t\in\mathbb{N}}$, the random variables Y_t, $t \in \mathbb{N}$, are independent, the distribution of Y_t depending only on X_t. The conditional distributions of Y_t given X_t, for all possible values of X_t, are called emission distributions. The name "hidden Markov model" comes from the fact that the observations are the Y_t's only, one cannot access to the states $(X_t)_t$ of the Markov chain. Finite state space HMMs can be used to model heterogeneous variables coming from different populations, the states of the (hidden) Markov chain defining the population the observed variable comes from. HMMs are very popular dynamical models especially because of their computational tractability since there exist efficient algorithms to compute the likelihood and to recover the posterior distribution of the hidden states given the observations.

Frequentist asymptotic properties of estimators of HMMs parameters have been studied since the 1990s. Consistency and asymptotic normality of the maximum likelihood estimator have been established in the parametric case, see Douc and Matias [DM01], Douc *et al.* [DMR04], and references in Cappé *et al.* [CMR05], see also Douc *et al.* [DMOH11] for the most general consistency result up to now. As to Bayesian asymptotic results, there are only very few and recent results, see de Gunst and Shcherbakova [GS08] when the number of hidden states is known, Gassiat and Rousseau [GR14] when the number of hidden states is unknown. All these results concern parametric HMMs.

Nonparametric HMMs in the sense that the form of the emission distribution is not specified have only very recently been considered, since identifiability remained an open problem until Gassiat and Rousseau [GR16] and Gassiat *et al.* [GCR15], who prove a general identifiability result. Because parametric modelling of emission distributions may lead to poor results in practice, in particular for clustering purposes, recent interest in using nonparametric HMMs appeared in applications, see Yau *et al.* [YPRH11], Gassiat *et al.* [GCR15] and references therein. Theoretical results for estimation procedures in nonparametric HMMs have also been obtained only very recently: Dumont and Le Corff [DL14] concerns regression models with hidden (Markovian) regressors and unknown regression functions in Gaussian noise, and Gassiat and Rousseau [GR16] is about translated emission distributions.

In this chapter, we obtain posterior consistency results for Bayesian procedures in finite state space nonparametric HMMs. To our knowledge, this is the first result on posterior consistency in such models. In Section 2.2.2, we prove posterior consistency in terms of the weak topology and the L_1-norm on marginal densities of consecutive observations. Our main result is obtained

under assumptions on the emission densities and on the prior which are very similar to the ones in the i.i.d. case, see Theorem 2.1. This result relies on a new control of the Kullback-Leibler divergence for HMMs, see Lemma 2.2. Yet estimating the distribution of consecutive observations is not the main objective of a practitioner. Classifying the observations according to their corresponding hidden states or estimating the parameters of the model often are the questions of interest, see for instance Yau *et al.* [YPRH11], Whiting *et al.* [WLM03] and Couvreur and Couvreur [CC00]. In Section 2.2.3 we build upon the recent identifiability result to deduce from Theorem 2.1 posterior consistency for each component of the parameters. We obtain in general posterior consistency for the transition matrix of the Markov chain and for the emission probability distribution in the weak topology, see Theorem 2.3. Stronger results are established in particular cases, see Corollary 2.6 and Theorem 2.8. Finally, some examples of priors that fulfill the assumptions of Theorems 2.1 and 2.3 are studied in Section 2.3.

Particularly in Section 2.3.3 the discrete case is thoroughly studied with a Dirichlet process prior. Sufficient and almost necessary assumptions to apply Theorem 2.1 are given in Proposition 2.9. Moreover in this framework, posterior consistency of the marginal smoothing distributions, used in segmentation or classification, is derived in Theorem 2.8.

All proofs are given in Appendices 2.4 and 2.5.

2.2 Settings and Main Theorem

2.2.1 Notations

We now precise the model and give some notations. Recall that finite state space HMMs are stochastic processes $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $(X_t)_{t\in\mathbb{N}}$ is a Markov chain taking values in a finite set, and conditionally on $(X_t)_{t\in\mathbb{N}}$, the random variables Y_t, $t \in \mathbb{N}$, are independent. The distribution of Y_t depending only on X_t is called the emission distribution. The number k of hidden states is known, so that the state space of the Markov chain is set to $\{1, \dots, k\}$. Throughout Chapter 2, for any integer n, an n-uple $(x_1, \dots, x_n)$ is denoted $x_{1:n}$.

Let $\Delta_k = \{(x_1, \dots, x_k) : x_i \geq 0,\ i = 1, \dots, k;\ \sum_{i=1}^k x_i = 1\}$ denote the $(k-1)$-dimensional simplex. Let Q denote the $k \times k$ transition matrix of the Markov chain, so that identifying Q as the k-uple of transition distributions (the lines of the matrix), we write $Q \in \Delta_k^k$. We denote $\mu \in \Delta_k$ the initial probability measure, that is the distribution of X_1. For $\underline{q} \geq 0$, we also define

$$\Delta^k(\underline{q}) = \{Q \in \Delta_k^k \ :\ \min_{i,j\leq k} Q_{i,j} \geq \underline{q}\},$$

so that $\Delta^k(0) = \Delta_k^k$. We now recall some properties of Markov chains with transition matrix in $\Delta^k(\underline{q})$. Note that $\underline{q}$ needs to be less than $\frac{1}{k}$ for $\Delta^k(\underline{q})$ to be nonempty. Then for all Q in $\Delta^k(\underline{q})$, $\max_{i,j} Q_{i,j} \leq 1-(k-1)\underline{q}$. Also, if $Q \in \Delta^k(\underline{q})$, then for any $i \in \{1, \dots, k\}$ and $A \subset \{1, \dots, k\}$, $\sum_{j\in A} Q_{i,j} \geq k\underline{q}u(A)$, with u the uniform probability on $\{1, \dots, k\}$. Besides if $Q \in \Delta^k(\underline{q})$ with

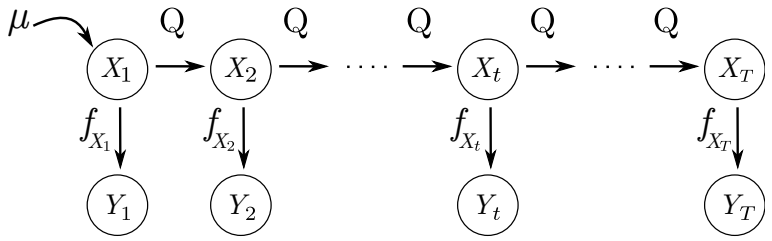


Figure 2.1 – The model.

$\underline{q} > 0$, the chain is irreducible, positive recurrent and admits a unique stationary probability measure denoted μ^Q for which $\underline{q} \leq \mu^Q(i) \leq 1 - (k-1)\underline{q}$, $1 \leq i \leq k$.

We assume that the observation space is $\mathbb{R}^d$ endowed with its Borel sigma field. Let $\mathcal{F}$ be the set of probability density functions with respect to a reference measure λ on $\mathbb{R}^d$. $\mathcal{F}^k$ is the set of possible emission densities, that is for $f = (f_1, \ldots, f_k) \in \mathcal{F}^k$, the distribution of Y_t conditionally to $X_t = i$ will be $f_i\lambda$, $i = 1, \ldots, k$. See Figure 2.1 for a visualization of the model.

Let

$$\Theta = \{\theta = (Q, f) \ : \ Q \in \Delta_k^k, f \in \mathcal{F}^k\}$$

and

$$\Theta(\underline{q}) = \{\theta = (Q, f) \ : \ Q \in \Delta^k(\underline{q}), f \in \mathcal{F}^k\}.$$

Then $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$) denotes the probability distribution of $(X_t, Y_t)_{t\in\mathbb{N}}$ under θ and initial probability measure $\mu^\theta := \mu^Q$ (respectively μ). Let p_l^θ ($p_l^{\theta,\mu}$ resp.) denote the probability density of $Y_1, \ldots, Y_l$ with respect to $\lambda^{\otimes l}$ under $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$). and P_l^θ ($P_l^{\theta,\mu}$ resp.) the marginal distribution of $Y_1, \ldots, Y_l$ under $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$). So for any $\theta \in \Theta$, initial probability measure μ, and measurable set A of $\{1, \ldots, k\}^l \times (\mathbb{R}^d)^l$:

$$\begin{aligned}
&\mathbb{P}^{\theta,\mu}((X_{1:l}, Y_{1:l}) \in A) \\
&\quad = \int \sum_{x_1,\ldots,x_l=1}^{k} \mathbb{1}_{(x_1,\ldots,x_l,y_1,\ldots,y_l)\in A}\, \mu_{x_1} Q_{x_1,x_2} \ldots Q_{x_{l-1},x_l} \\
&\qquad f_{x_1}(y_1) \ldots f_{x_l}(y_l)\lambda(dy_1) \ldots \lambda(dy_l),
\end{aligned}$$

$$p_l^{\theta,\mu}(y_1, \ldots, y_l) = \sum_{x_1,\ldots,x_l=1}^{k} \mu_{x_1} Q_{x_1,x_2} \ldots Q_{x_{l-1},x_l} f_{x_1}(y_1) \ldots f_{x_l}(y_l),$$

and $P_l^{\theta,\mu} = p_l^{\theta,\mu}\lambda^{\otimes l}$.

We denote by $\delta_\mu \otimes \pi$ the prior on $\Delta_k \times \Theta$, where $\mu \in \Delta_k$ is an initial probability measure. We assume that π is a product of probability measures on Θ, $\pi = \pi_Q \otimes \pi_f$ such that π_Q is a probability distribution on Δ_k^k and π_f is a probability distribution on $\mathcal{F}^k$.

We assume throughout Chapter 2 that the observations are distributed from $\mathbb{P}^{\theta^*}$ so that their

distribution is a stationary HMM. We are interested in posterior consistency, that is to prove that with $\mathbb{P}^{\theta^*}$-probability one, for all neighbourhood U of θ^*:

$$\lim_{n\to+\infty} \pi(U|Y_{1:n}) = 1.$$

The choice of a topology on the parameters arises here. For any distance or pseudometric D, we denote $N(\delta, A, D)$ the δ-covering number of the set A with respect to D, that is the minimum number N of elements $a_1, \dots, a_N$ such that for all $a \in A$, there exists $n \leq N$ such that $D(a, a_n) \leq \delta$.

For $k \times k$ matrices M, we use

$$\|M\| = \max_{1\leq i,j\leq k} |M_{i,j}|.$$

For probability distributions P_1 and P_2, let p_1 and p_2 be their respective densities with respect to some dominated measure ν. We use the L_1-norm:

$$\|p_1 - p_2\|_{L_1(\nu)} = \int |p_1 - p_2| d\nu$$

and the Kullback-Leibler divergence:

$$KL(P_1, P_2) = \begin{cases} \int p_1 \log(\frac{p_1}{p_2}) d\nu & \text{if } P_1 << P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

We also denote $KL(p_1, p_2)$ for $KL(p_1\nu, p_2\nu)$. On $\mathcal{F}^k$ we use the distance $d(\cdot,\cdot)$ defined for all $g = (g_1, \dots, g_k)$, $\tilde{g} = (\tilde{g_1}, \dots, \tilde{g_k})$ by

$$d(g, \tilde{g}) = \max_{1\leq j\leq k} \|g_j - \tilde{g_j}\|_{L_1(\lambda)}.$$

On $\Theta(\underline{q})$, we use the following pseudometric for $l \geq 3$, $l \in \mathbb{N}$,

$$D_l(\theta, \theta') = \int |p_l^\theta(y_1, \dots, y_l) - p_l^{\theta'}(y_1, \dots, y_l)| \lambda(dy_1) \dots \lambda(dy_l) = \|p_l^\theta - p_l^{\theta'}\|_{L_1(\lambda^{\otimes l})}.$$

Then a D_l-neighbourhood of θ is a set which contains a set $\{\theta' : D_l(\theta, \theta') < \varepsilon\}$ for some $\varepsilon > 0$. We also use the weak topology on marginal distributions $(P_l^\theta)_\theta$. We recall that in any neighbourhood of P_l^θ in the weak topology on probability measures there is a subset which is a union of sets of the form

$$\left\{ P : \left| \int h_j dP - \int h_j p_l^\theta d\lambda^{\otimes l} \right| < \varepsilon_j,\ j = 1, \dots, N \right\},$$

where for all $1 \leq j \leq N$, $\varepsilon_j > 0$ and h_j is in the set $\mathcal{C}_b((\mathbb{R}^d)^l)$ of all bounded continuous functions from $(\mathbb{R}^d)^l$ to $\mathbb{R}$. We prove posterior consistency in this general nonparametric context

using this weak topology on marginal distributions $(P_l^\theta)_\theta$ and the D_l-pseudometric in Section 2.2.2. We study the posterior consistency for the transition matrix and the emission distributions separately in Section 2.2.3.

Finally the sign $\lesssim$ is used for inequalities up to a multiplicative constant possibly depending on fixed parameters.

2.2.2 Main Theorem

In this section, we state our general theorem on posterior consistency for nonparametric hidden Markov models in the weak topology on marginal distributions $(P_l^\theta)_\theta$ and the D_l-topology. Fix $l \geq 3$. We consider the following assumptions:

(A1.0) For all $1 \leq i \leq k$, $\int f_i^*(y)|\log(f_i^*(y))|\lambda(dy) < +\infty$,

(A1.1) for all $\varepsilon > 0$ small enough there exists a set $\Theta_\varepsilon \subset \Theta(\underline{q})$ such that $\pi(\Theta_\varepsilon) > 0$ and for all $\theta = (Q, f) \in \Theta_\varepsilon$,

- (A1.1a) $\|Q - Q^*\| < \varepsilon$,
- (A1.1b) $\max_{1\leq i\leq k} \int f_i^*(y) \max_{1\leq j\leq k} \log(\frac{f_j^*(y)}{f_j(y)})\lambda(dy) < \varepsilon$,
- (A1.1c) for all $y \in \mathbb{R}^d$ such that $\sum_{i=1}^k f_i^*(y) > 0$, $\sum_{j=1}^k f_j(y) > 0$,
- (A1.1d) $\sup_{y:\sum_{i=1}^k f_i^*(y)>0} \max_{1\leq j\leq k} f_j(y) < +\infty$,

(A1.2) for all $n > 0$, for all $\delta > 0$, there exists a set $\mathcal{F}_n \subset \mathcal{F}^k$ and a real number $r_1 > 0$ such that $\pi_f\big((\mathcal{F}_n)^c\big) \lesssim e^{-nr_1}$ and such that

$$\sum_{n>0} N\left(\frac{\delta}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \exp\left(-\frac{n\delta^2 k^2 \underline{q}^2}{32l}\right) < +\infty.$$

Theorem 2.1. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leq i \leq k$, $\mu_i \geq \underline{q}$.*

a) If Assumptions (A1.0) and (A1.1) holds then, for all weak neighbourhood U of $P_l^{\theta^*}$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(U|Y_{1:n}) = 1\right) = 1.$$

b) Moreover if Assumptions (A1.0), (A1.1) and (A1.2) hold then, for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\ \{\theta :\ D_l(\theta,\theta^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

Remark 2.2.1. We assume everywhere in Chapter 2 that the support of π_Q is included in $\Delta^k(\underline{q})$. It means the results of this chapter can only be applied to priors π_Q on transition matrices which

vanish close to the border of Δ_k^k. This assumption is satisfied by a product of truncated Dirichlet distribution, i.e. if the lines $Q_{i,\cdot}$ of Q are independently distributed from a law proportional to:

$$Q_{i,1}^{\alpha_1-1}\dots Q_{i,k}^{\alpha_k-1}\mathbb{1}_{\{\underline{q}\leq Q_{i,j}\leq 1,\ \forall 1\leq j\leq k\}}dQ_{i,1}\dots dQ_{i,k}$$

where $\alpha_1,\dots,\alpha_k>0$.

The restriction on $\Delta^k(\underline{q})$ comes from the test built in Gassiat and Rousseau [GR14]. On this set, HMMs are geometrically ergodic. It is a common assumption in the literature see Douc and Matias [DM01], Douc *et al.* [DMR04], or Douc *et al.* [DMOH11] for instance. Besides Gassiat and Rousseau [GR14] explain the difficulty which appears when the Markov chain does not mix well. They are also able to obtain a less restrictive assumption on the support of the prior on transition matrices. In return they assume a more restrictive assumption on the log-likelihood, compare Equations (2.11) and (2.13) with their Assumption C1.

In the case of density estimation with i.i.d. observations, it is usual to control the Kullback-Leibler support of the prior to show weak posterior consistency and to control, in addition, a metric entropy to obtain strong consistency, see Chapter 4 of Ghosh and R.V. Ramamoorthi [GR03]. Assumptions (A1.1) and (A1.2) are similar in spirit. Assumptions (A1.0) and (A1.1) replace the assumption on the true density function being in the Kullback-Leibler support of the prior in the i.i.d. case. (A1.1a) ensures that the transition matrices of Θ_ε are in a ball of radius ε around the true transition matrix. Under (A1.1b) the emission densities are in an ε Kullback-Leibler ball around the true one. (A1.0), (A1.1b), (A1.1c) and (A1.1d) are assumptions under which the log-likelihood converges $\mathbb{P}^{\theta^*}$-a.s. and in $L_1(\mathbb{P}^{\theta^*})$. (A1.2) is very similar to the assumptions of the metric entropy of Theorem 4.4.4 in Ghosh and R.V. Ramamoorthi [GR03].

In Appendix 2.4, the proof of Theorem 2.1 relies on the method of Barron [Bar88]. It consists of controlling Kullback-Leibler neighbourhoods and building tests. The construction of tests is quite straightforward thanks to Rio's inequality [Rio00] which generalizes Hoeffding's inequality. To prove a), we use the usual strategy presented in Section 4.4.1 in Ghosh and R.V. Ramamoorthi [GR03] together with Rio's inequality [Rio00] and Gassiat and Rousseau [GR14]. To prove b), we use the tests of Gassiat and Rousseau [GR14]. To control the Kullback-Leibler neighbourhoods, we use the following lemma whose proof is given in Appendix 2.4.

Lemma 2.2. *Let θ^* be in $\Theta(\underline{q})$. If (A1.1) holds then, for all $0<\varepsilon<1$, there exists $N\in\mathbb{N}$ such that for all $n\geq N$ and for all $\theta\in\Theta_\varepsilon$:*

$$\frac{1}{n}KL(\mathbb{P}_n^{\theta^*},\mathbb{P}_n^{\theta,\mu})\leq\frac{3}{\underline{q}}\,\varepsilon.$$

2.2.3 Consistency of Each Component of the Parameter

In this Section we look at the consequences of Theorem 2.1 on posterior consistency for the transition matrix and the emission distributions separately. Estimating consistently the components of the parameter is of great importance. First one may want to know the proportion of each population or the probability of moving from one population to another, i.e. the transition matrix. Secondly, these components are important to recover the smoothing distribution, i.e. the distribution of a hidden state given the observations, and then to cluster the observations, see Cappé *et al.* [CMR05] and Theorem 2.8.

In practice, estimating the marginal density of l consecutive observations is not the first purpose. Yet estimating the parameters and the hidden states is often the goal. For instance, Whiting *et al.* [WLM03] give an algorithm to estimate the stationary probability measure of the Markov chain derived from the transition matrix. While Yau *et al.* [YPRH11] and Couvreur and Couvreur [CC00] are interested in estimating the hidden states.

The consistency for each component of the parameter, i.e. the transition matrix and the emission distributions, does not directly result from consistency of the marginal distribution of the observations, see Dumont and Le Corff [DL14]. Identifiability seems to be necessary to obtain this implication yet it is not sufficient. We obtain posterior consistency for the components of the parameter thanks to the result of identifiability of Gassiat *et al.* [GCR15] and as usually by proving the continuity of the functional

$$\begin{cases} ((p_l^\theta)_\theta, L_1) & \to & (\Theta, \text{the topology } \mathcal{T} \text{ described in the following}) \\ p_l^\theta & \mapsto & \theta \end{cases}.$$

We use a product topology on the set of parameters. In particular we study consistency in the topology associated with the sup norm on transition matrices $\|\cdot\|$ and the weak topology on probability measures for the emission distributions up to label switching. To deal with label switching, we need the following definitions. Let $\mathcal{S}_k$ denote the symmetric group on $\{1, \dots, k\}$. Let σ be a permutation in $\mathcal{S}_k$, for all matrices $Q \in \Delta_k^k$, we denote σQ the following matrix: for all $1 \leq i, j \leq k$,

$$(\sigma Q)_{i,j} = Q_{\sigma(i),\sigma(j)}.$$

If $(X_t, Y_t)_{t\in\mathbb{N}}$ is distributed from $P^{(Q,f)}$ and $\tilde{X}_t = \sigma^{-1}(X_t)$, for $\sigma \in \mathcal{S}_k$, then $(\tilde{X}_t, Y_t)_{t\in\mathbb{N}}$ is distributed from $P^{(\sigma Q,(f_{\sigma(1)},\dots,f_{\sigma(k)}))}$, i.e the labels of the Markov chain have been switched but $(Y_t)_{t\in\mathbb{N}}$ has the same distribution. Then, in generality, from the distribution of the observations one can at most recover the parameter up to label switching. Gassiat *et al.* [GCR15] proved that it is possible by knowing the joint distribution of at least three consecutive observations.

In Theorem 2.3, whose proof is given in Appendix 2.4, we prove that under the assumption of identifiability, posterior consistency in the D_l topology implies that the posterior concentrates around (Q^*, f^*) up to label switching, i.e. around $\{\sigma Q^*, (f^*_{\sigma(1)}, \dots, f^*_{\sigma(k)})\}_{\sigma\in\mathcal{S}_k}$. In other words

we obtain posterior consistency considering neighbourhoods of the form

$$\left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)} \in U_{f_i^*}, i = 1 \dots k\right\}$$

where U_{Q^*} is a neighbourhood of Q^* and for all $1 \leq i \leq k$, $U_{f_i^*}$ is a weak neighbourhood of $f_i^*\lambda$. That is to say we consider the product topology $\mathcal{T}$ of the sup norm topology on transition matrices and of the weak topology on the emission distributions up to label switching.

Theorem 2.3. *Let $\theta^* = (Q^*, f^*) \in \Theta$ such that $f_1^*\lambda, \dots, f_k^*\lambda$ are linearly independent and Q^* has full rank.*

If the posterior is consistent for the D_l pseudo-metric with $l \geq 3$, i.e. if for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\ \{\theta :\ D_l(\theta, \theta^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

then the posterior is consistent for the topology $\mathcal{T}$, i.e. for all weak neighbourhood $U_{f_i^*}$ of $f_i^*\lambda$, for all $1 \leq i \leq k$ and for all neighbourhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\ \left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)}\lambda \in U_{f_i^*},\ 1 \leq i \leq k\right\} \middle| Y_{1:n}\right) = 1\right) = 1. \tag{2.1}$$

Remark 2.2.2. In particular, Equation (2.1) implies that for all $\varepsilon > 0$

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\bigcup_{\sigma\in\mathcal{S}_k} \{Q : \|Q - \sigma Q^*\| < \varepsilon\} \,\middle|\, Y_{1:n}\right) = 1\right) = 1.$$

It means that under the assumptions of Theorem 2.3, the posterior concentrates around $\{\sigma Q^*, \sigma \in \mathcal{S}_k\}$. Equation (2.1) also implies that for all $N \in \mathbb{N}$, for all $h_i \in \mathcal{C}_b(\mathbb{R}^d)$ and for all $\varepsilon_i > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\bigcup_{\sigma\in\mathcal{S}_k} \left\{f : \left|\int h_i f_j d\lambda - \int h_i f^*_{\sigma(j)} d\lambda\right| < \varepsilon_i,\right.\right.\right.$$
$$\left.\left.\left.\text{for all } 1 \leq i, j \leq k\right\} \middle| Y_{1:n}\right) = 1\right) = 1.$$

This last result allows to consistently recover smooth functionals of the emission distributions $(f_j^*)_j$ such as $\int_K f_j^* d\lambda$ where K is compact. We obtain stronger results in Sections 2.3.2 and 2.3.3.

The uncertainty due to label switching can be removed if there is only one possible permutation σ associated to a parameter θ as in Proposition 2.4, proved in Appendix 2.4. This Proposition 2.4 may be useful if one knows some characteristics of the hidden states which order them. The function H, in Proposition 2.4, enables to order the hidden states and then to get rid of label switching.

Proposition 2.4. *Let $\theta^* \in \Theta$ such that $f_1^*\lambda, \dots, f_k^*\lambda$ are linearly independent and Q^* has full rank. Let $H : (\Theta, \mathcal{T}_1) \to \mathbb{R}^k$ be a continuous function, where $\mathcal{T}_1$ is the product topology of the sup norm topology on transition matrices and of the weak topology on the emission distributions. Assume that for all permutation $\sigma \in \mathcal{S}_k$ and for all $\theta = (Q, f) \in \Theta$,*

$$H_i((\sigma Q, f_{\sigma(1)}, \dots, f_{\sigma(k)})) = H_{\sigma(i)}(\theta), \tag{2.2}$$

$$H_1(\theta^*) < \cdots < H_k(\theta^*), \tag{2.3}$$

$$\pi\Big(\big\{\theta : \ H_1(\theta) < \cdots < H_k(\theta)\big\}\Big) = 1. \tag{2.4}$$

If the posterior is consistent for the topology $\mathcal{T}$, i.e. for all weak neighbourhood $U_{f_i^*}$ of $f_i^*\lambda$, for all $1 \le i \le k$ and for all neighbourhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\Bigg(\lim_{n\to+\infty} \pi\Bigg(\big\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)}\lambda \in U_{f_i^*},\ 1 \le i \le k\big\} \Bigg| Y_{1:n}\Bigg) = 1\Bigg) = 1 \tag{2.1}$$

then for all weak neighbourhood $U_{f_i^*}$ of $f_i^*\lambda$, for all $1 \le i \le k$ and for all neighbourhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\left\{ Q \in U_{Q^*},\ f_i\lambda \in U_{f_i^*},\ 1 \le i \le k\right\} \middle| Y_{1:n}\right) = 1\right) = 1.$$

Here we give some examples of possible functions H:

$$H_i(\theta) = Q_{i,i} \quad \text{or} \quad H_i(\theta) = \int \phi f_i d\lambda, \tag{2.5}$$

where ϕ is bounded and continuous. Even if in practice, one would often like to use $H_i(\theta) = \int y f_i(y)\lambda(dy)$, Proposition 2.4 does not allow it. Indeed, in this case H is not continuous. Yet taking a continuous truncated version of the identity for ϕ in Equation (2.5) may help.

2.3 Examples of Priors on f

In this section we apply Theorems 2.1 and 2.3 for different types of priors and emission models. In Section 2.3.1 we deal with emission distributions which are independent mixtures of Gaussian distributions. Translated emission distributions are studied in Section 2.3.2. Finally we consider the discrete case with Dirichlet process priors in Section 2.3.3.

Assumptions (A1.1b) and (A1.2) are purposely designed to resemble the types of assumptions found in density estimation for i.i.d. observations. This allows us to use existing results on consistency in the case of i.i.d. observations. This is done in Sections 2.3.1 and 2.3.2 with a prior based on a usual prior on densities, which is a mixture of Gaussian distributions such as in Tokdar [Tok06]. Two ways of using a prior on densities are considered. In Section 2.3.1, the emission distributions are independently distributed under a usual prior on densities. In Section 2.3.2,

the emission distributions are designed from a unique density, distributed from a usual prior, which is translated. Contrariwise in the discrete case we develop a new method to deal with the Dirichlet process prior in Section 2.3.3.

2.3.1 Independent Mixtures of Gaussian Distributions

We consider the well known location-scale mixture of Gaussian distributions as prior model for each f_i, namely each density under the prior is written as

$$g(y) = \int_{\mathbb{R}\times(0,+\infty)} \phi_\sigma(y-z)dP(z,\sigma) =: \phi * P, \tag{2.6}$$

where ϕ_σ is the Gaussian density with mean zero and variance σ^2, and P is a probability measure on $\mathbb{R}\times(0,+\infty)$. In this part, λ is the Lebesgue measure on $\mathbb{R}$. Let π_P be a probability measure on the set of probability measures on $\mathbb{R}\times(0,+\infty)$. Denote π_g the distribution of g expressed as (2.6) when $P\sim\pi_P$. Then we consider the prior distribution on $f=(f_1,\dots,f_k)$ defined by $\pi_f = \pi_g^{\otimes k}$. We need the following assumptions to apply Theorem 2.1 and 2.3:

(B1.1)
$$\pi_P\left(P \ : \ \int \frac{1}{\sigma} dP(z,\sigma) < \infty\right) = 1,$$

(B1.2) for all $1\le j\le k$, f_j^* is positive, continuous on $\mathbb{R}$ and bounded by $M<\infty$,

(B1.3) for all $1\le i\le k$, $1\le j\le k$,

$$\int_{\mathbb{R}} f_i^*(y)\log\left(\frac{f_j^*(y)}{\psi_j(y)}\right)\lambda(dy) < \infty$$

where $\psi_j(y) = \inf_{t\in[y-1,y+1]} f_j^*(t)$,

(B1.4) for all $1\le i\le k$, there exists $\eta>0$ such that

$$\int_{\mathbb{R}} |y|^{2(1+\eta)} f_i^*(y)\lambda(dy) < \infty,$$

(B1.5) for all $\beta>0$, $\kappa>0$, there exist a real number $\beta_0>0$, two increasing and positive sequences a_n and u_n tending to $+\infty$, and a sequence l_n decreasing to 0 such that

$$\pi_P\Big(P \ : \ P((-a_n,a_n]\times(l_n,u_n]) < 1-\kappa\Big) \le \exp(-n\beta_0),$$

$$\text{with} \qquad \frac{a_n}{l_n} \le n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) \le n\beta.$$

Proposition 2.5. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leq i \leq k$, $\mu_i \geq \underline{q}$. Assume that Q^* is in the support of π_Q and that the weak support of π_P contains all probability measures that are compactly supported.*

Then

- *(B1.1), (B1.2), (B1.3), (B1.4) imply (A1.1)*
- *and (B1.5) implies (A1.2).*

In particular in the case where π_P is the Dirichlet process $DP(\alpha G_0)$ with base measure αG_0, where G_0 is a probability measure on $\mathbb{R} \times (0, +\infty)$ and $\alpha > 0$, Assumption (B1.1) holds as soon as

$$\int_{\mathbb{R}\times(0,+\infty)} \frac{1}{\sigma} G_0(dz, d\sigma) < +\infty. \tag{2.7}$$

Indeed,

$$\begin{aligned}\int\int \frac{1}{\sigma} P(dz, d\sigma)\pi_P(dP) &= \int\int\int_{[\sigma,+\infty)} \frac{1}{t^2}\lambda(dt)P(dz, d\sigma)\pi_P(dP) \\ &= \int \frac{1}{\sigma} G_0(dz, d\sigma).\end{aligned}$$

Moreover using Remark 3.1 of Tokdar [Tok06], Assumption (B1.5) easily holds as soon as for all $\beta > 0$, there exist a real number $\beta_0 > 0$, two increasing and positive sequences a_n and u_n tending to $+\infty$ and a sequence l_n decreasing to 0 such that

$$\begin{aligned} G_0\left((-a_n, a_n] \times (l_n, u_n]\right)^c) &\leq \exp(-n\beta_0), \\ \frac{a_n}{l_n} \leq n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) &\leq n\beta. \end{aligned} \tag{2.8}$$

2.3.2 Translated Emission Distributions

In this section we consider the special case of translated emission distributions, that is to say for all $1 \leq j \leq k$,

$$f_j(\cdot) = g(\cdot - m_j),$$

where g is a density function on $\mathbb{R}$ with respect to λ and for all $1 \leq j \leq k$, m_j is in $\mathbb{R}$. In this part, λ is still the Lebesgue measure on $\mathbb{R}$ and $d = 1$. This model has been in particular considered by Yau *et al.* [YPRH11] for the analysis of genomic copy number variation. First a corollary of Theorem 2.3 is given. Then the particular case of location-scale mixture of Gaussian distributions on g is studied.

Let

$$\Xi = \{\xi = (Q, m, g), Q \in \Delta_k^k, m \in \mathbb{R}^k, m_1 = 0 < m_2 < \cdots < m_k, g \in \mathcal{F}\}$$

and

$$\Xi(\underline{q}) = \{\xi = (Q, m, g) \in \Xi, Q \in \Delta^k(\underline{q})\}.$$

To $\xi = (Q, m, g) \in \Xi$, we associate $\theta = (Q, (g(\cdot - m_1), \dots, g(\cdot - m_k))) \in \Theta$. We then denote $\mathbb{P}^\xi$ for $\mathbb{P}^\theta$. We assume that π_f is a product of probability measures,

$$\pi_f = \pi_m \otimes \pi_g,$$

where π_g is a probability measure on $\mathcal{F}$ and π_m is a probability measure on $\mathbb{R}^k$. Note that under Ξ, the model is completely identifiable, see Theorem 2.1 of Gassiat and Rousseau [GR16]. The uncertainty due to label switching is resolved here. In Corollary 2.6, additionally to posterior consistency for the transition matrices, we obtain posterior consistency for the parameters of translation m_j and for the weak convergence on the translated probability measure $g\lambda$. Under a stronger assumption, we get posterior consistency for the L_1-topology on the translated density distribution.

Fix $l \geq 3$. The following assumption replaces (A1.2) in the context of translated emission distributions:

(C1.2) for all $n > 0$, for all $\delta > 0$, there exists a set $\mathcal{F}_n \subset \mathbb{R}^k \times \mathcal{F}$ and a real number $r_1 > 0$ such that $\pi_f\big((\mathcal{F}_n)^c\big) \lesssim e^{-nr_1}$ and

$$\sum_{n>0} N\left(\frac{\delta}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \exp\left(-\frac{n\delta^2 k^2 \underline{q}^2}{32l}\right) < +\infty.$$

Corollary 2.6. *Let $\xi^* = (Q^*, m^*, g^*)$ be in $\Xi(\underline{q})$ such that $m_1^* = 0 < m_2^* < \cdots < m_k^*$ and Q^* has full rank.*

If the posterior is consistent for the D_l pseudometric with $l \geq 3$, i.e. if for all $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to\infty} \pi(\ \{\xi :\ D_l(\xi, \xi^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi(\{Q : \|Q - Q^*\| < \varepsilon\} \,\middle|\, Y_{1:n}) = 1\right) = 1,$$

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi(\{m : \forall 1 \leq j \leq k,\ |m_j - m_j^*| < \varepsilon\} \,\middle|\, Y_{1:n}) = 1\right) = 1,$$

and for all $N \in \mathbb{N}$, for all $h_i \in \mathcal{C}_b(\mathbb{R}^d)$, for all $\varepsilon_i > 0,\ 1 \leq i \leq N$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi\left(\left\{g : \left|\int h_i g d\lambda - \int h_i g^* d\lambda\right| < \varepsilon_i\right\} \,\middle|\, Y_{1:n}\right) = 1\right) = 1.$$

If moreover $\max_{1\leq j\leq k}\mu_j^* > 1/2$ *and* g^* *is uniformly continuous; then, for all* $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty}\pi\left(\{g : \|g-g^*\|_{L_1(\lambda)} < \varepsilon\}\,|Y_{1:n}\right) = 1\right) = 1.$$

The proof of Corollary 2.6, in Appendix 2.5, relies on the identifiability result of Gassiat and Rousseau [GR16] and Theorem 2.3.

In the same way as in Section 2.3.1, we propose to apply Theorem 2.1 and Corollary 2.6 to a prior based on location-scale mixtures of Gaussian distributions. In this part, we study a particular prior on the translated emission density g which is the location-scale mixture of Gaussian distributions. Then g is a sample drawn from π_g if

$$g(y) = \int_{\mathbb{R}\times(0,+\infty)}\phi_\sigma(y-z)dP(z,\sigma)$$

where P is a sample drawn from π_P and π_P is a probability measure on probability measures on $\mathbb{R}\times(0,+\infty)$. The following assumption help in proving (C1.2):

(D1.6) for all $\beta > 0$, $\kappa > 0$, there exist a real number $\beta_0 > 0$, three increasing sequences of positive numbers m_n, a_n and u_n tending to $+\infty$, and a sequence l_n decreasing to 0 such that

$$\pi_P\left(P\ :\ P((-a_n,a_n]\times(l_n,u_n]) < 1-\kappa\right) \leq \exp(-n\beta_0),$$

$$\pi_m\left(([-m_n,m_n]^k)^c\right) \leq exp(-n\beta_0),$$

$$\frac{a_n}{l_n} \leq n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) \leq n\beta, \qquad \log\left(\frac{m_n}{l_n}\right) \leq n\beta.$$

Proposition 2.7. *Let* $\underline{q} > 0$ *and* ξ^* *in* $\Xi(\underline{q})$. *Assume that the support of* π_Q *is included in* $\Delta^k(\underline{q})$ *and that for all* $1 \leq i \leq k$, $\mu_i \geq \underline{q}$. *Assume that* Q^* *is in the support of* π_Q, *that* m^* *is in the support of* π_m *and that the weak support of* π_P *contains all probability measures that are compactly supported.*

If (B1.1) is verified and (B1.2), (B1.3) and (B1.4) are verified with $f_j(\cdot) = g(\cdot - m_j)$, $1 \leq j \leq k$ *then (A1.1) holds.*

Moreover (D1.6) implies (C1.2).

The proof of Proposition 2.7 is very similar to that of Proposition 2.5 and is given in Appendix 2.5.

Corollary 2.6 and Proposition 2.7 are less general than Theorem 2.3 and Proposition 2.5 respectively. In Corollary 2.6 and Proposition 2.7, it is assumed that the true emission distributions are translated versions of a unique density g^*. In practice, we expect priors on translated emission distributions not to be as robust as priors for which the emission distributions are i.i.d. such as

priors of Section 2.3.1. Particularly if the true emission distributions have different tails, priors on translated emission distributions may lead to poor estimations.

2.3.3 Independent Discrete Emission Distributions

Discrete emission distributions, i.e. when the support of λ is included in $\mathbb{N}$, have been successfully used, for instance in genomics in Gassiat *et al.* [GCR15].

Note that for discrete emission distributions, weak and l_1 topologies are the same so that weak posterior consistency implies l_1 posterior consistency. Thus Assumption (A1.2) becomes unnecessary in Theorems 2.1 and 2.3. Moreover posterior consistency for the emission distributions in the weak topology in Theorem 2.3 implies posterior consistency for the emission distributions in l_1.

In the discrete case, we prove in Theorem 2.8 that posterior consistency for the marginal distribution of finitely many observations, for the transition matrix and for the emission distributions in l_1 together with the restriction of the prior π_Q on $\Delta^k(\underline{q})$ imply posterior consistency for the marginal smoothing:

Theorem 2.8. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leq i \leq k$, $\mu_i \geq \underline{q}$. If $f_1^*\lambda, \ldots, f_k^*\lambda$ are linearly independent, Q^* has full rank, and (A1.0) and (A1.1) hold; then, for all finite integer m,*

$$\lim_{n\to+\infty} \pi\Bigg(\Big\{\theta : \exists \sigma \in \mathcal{S}_k, \max_{1\leq a_j\leq k,\ 1\leq j\leq m} |P^\theta(X_i = \sigma(a_i),\ \forall\, 1 \leq i \leq m \,|Y_{1:n}) \\ - P^{\theta^*}(X_i = a_i,\ \forall\, 1 \leq i \leq m \mid Y_{1:n})| < \varepsilon\Big\} \Big| Y_{1:n}\Bigg) = 1 \text{ in } P^{\theta^*}\text{-probability.}$$

The proof of Theorem 2.8 is given in Appendix 2.4.

In the following we apply Theorems 2.1, 2.3 and 2.8 to a specific prior on the set of probability measures on $\mathbb{N}$ in the case of a HMM with discrete emission distributions. We consider a Dirichlet process $DP(\alpha G_0)$ with α a positive number and G_0 some probability measure on $\mathbb{N}$. We then consider a prior probability measure on Θ defined by

$$\pi = \pi_Q \otimes DP(\alpha G_0)^{\otimes k}.$$

In Proposition 2.9, we give sufficient and almost necessary conditions to obtain (A1.1). Proposition 2.9 is proved in Appendix 2.4.

Proposition 2.9. *Let $\underline{q} > 0$. Assume that the support of the prior π_Q is included in $\Delta^k(\underline{q})$, that Q^* is in the support of π_Q and that for all $1 \leq i \leq k$, $\mu_i \geq \underline{q}$.*

If

(E1.1) for all $1 \leq i \leq k$, $\sum_{l \in \mathbb{N}} \frac{f_i^*(l)}{G_0(l)} < +\infty$

then (A1.1) holds.

Moreover if

(T1.1) for all $1 \leq i \leq k$, $\sum_{l \in \mathbb{N}} f_i^*(l)(-\log f_i^*(l)) < +\infty$

then (A1.1b) implies (E1.1).

Remark 2.3.1. Therefore (E1.1) is not only sufficient to prove (A1.1b) but up to the weak Assumption (T1.1) it is also necessary. Assumption (E1.1) relies on the mutual control of the tails of the base measure G_0 and the true emission distributions f_j^*. Proposition 2.9 suggests choosing a heavy tailed probability measure G_0 with $G_0(l) > 0$, for all $l \in \mathbb{N}$.

Remark 2.3.2. We deduce from Proposition 2.9 that

$$\left\{ g^* : \mathbb{N} \to (0,1) \quad \text{such that} \quad \sum_{l \in \mathbb{N}} g^*(l) = 1, \right. \\ \left. \sum_{l \in \mathbb{N}} g^*(l)(-\log(g^*(l)) < +\infty \quad \text{and} \quad \sum_{l \in \mathbb{N}} \frac{g^*(l)}{G_0(l)} < +\infty \right\} \tag{2.9}$$

is a subset of the Kullback-Leibler support of the Dirichlet process $DP(\alpha G_0)$.

Acknowledgements

I want to thank Elisabeth Gassiat and Judith Rousseau for their valuable comments. I also want to thank the reviewer and the editor for their helpful comments. The research was partly supported by the grants ANR Banhdits and ANR Ipanema.

2.4 Proofs of Key Results

Proof of Lemma 2.2

For all $\theta,\ \theta^* \in \Delta^k(\underline{q})$ the Kullback-Leibler divergence between $p_n^{\theta^*}$ and p_n^{θ} is by definition equal to

$$\frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\frac{\sum_{i_1,\dots,i_n=1}^k \mu_{i_1}^* Q_{i_1,i_2}^* \dots Q_{i_{n-1},i_n}^* f_{i_1}^*(Y_1)\dots f_{i_n}^*(Y_n)}{\sum_{j_1,\dots,j_n=1}^k \mu_{j_1} Q_{j_1,j_2} \dots Q_{j_{n-1},j_n} f_{j_1}(Y_1)\dots f_{j_n}(Y_n)}\right)\right).$$

Multiplying and dividing each term of the sum in the numerator by

$$\mu_{i_1} Q_{i_1,i_2} \dots Q_{i_{n-1},i_n} f_{i_1}(Y_1)\dots f_{i_n}(Y_n),$$

we obtain

$$\frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\frac{\sum_{i_1,\cdots,i_n=1}^{k}\frac{\mu^*_{i_1}Q^*_{i_1,i_2}\cdots Q^*_{i_{n-1},i_n}f^*_{i_1}(Y_1)\cdots f^*_{i_n}(Y_n)}{\mu_{i_1}Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\cdots f_{i_n}(Y_n)}\mu_{i_1}Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\cdots f_{i_n}(Y_n)}{\sum_{j_1,\ldots,j_n=1}^{k}\mu_{j_1}Q_{j_1,j_2}\ldots Q_{j_{n-1},j_n}f_{j_1}(Y_1)\ldots f_{j_n}(Y_n)}\right)\right)$$
$$\leq \frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leq i_1,\ldots,i_n\leq k}\frac{\mu^*_{i_1}Q^*_{i_1,i_2}\ldots Q^*_{i_{n-1},i_n}f^*_{i_1}(Y_1)\ldots f^*_{i_n}(Y_n)}{\mu_{i_1}Q_{i_1,i_2}\ldots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\ldots f_{i_n}(Y_n)}\right)\right)$$

by bounding the quotient in each term of the sum of the numerator by its maximum. Since the maximum of a product of positive factors is bounded by the product of the maxima,

$$\begin{aligned}
&\frac{1}{n}KL(p_n^{\theta^*},p_n^{\theta,\mu})\\
&\leq \frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leq i_0\leq k}\frac{\mu^*_{i_0}}{\mu_{i_0}}\left(\max_{1\leq i,j\leq k}\frac{Q^*_{i,j}}{Q_{i,j}}\right)^{n-1}\max_{1\leq i_1\leq k}\frac{f^*_{i_1}(Y_1)}{f_{i_1}(Y_1)}\cdots\max_{1\leq i_n\leq k}\frac{f^*_{i_n}(Y_n)}{f_{i_n}(Y_n)}\right)\right)\\
&\leq \frac{1}{n\underline{q}}\max_{1\leq i_0\leq k}\left|\mu_{i_0}-\mu^*_{i_0}\right|+\frac{n-1}{n\underline{q}}\max_{1\leq i,j\leq k}\left|Q_{i,j}-Q^*_{i,j}\right|\\
&\qquad\qquad+\max_{1\leq j\leq k}\int f^*_j(y)\max_{1\leq i\leq k}\log\frac{f^*_i(y)}{f_i(y)}\lambda(dy).
\end{aligned}$$

The last inequality comes from the following inequalities

$$\begin{aligned}
&\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leq i_s\leq k}\frac{f^*_{i_s}(Y_s)}{f_{i_s}(Y_s)}\right)\right)\\
&=\sum_{j_1,\ldots,j_n=1}^{k}\mu^*_{j_1}Q^*_{j_1,j_2}\ldots Q^*_{j_{n-1},j_n}\int f^*_{j_s}(y)\log\max_{1\leq i_s\leq k}\left(\frac{f^*_{i_s}(y)}{f_{i_s}(y)}\right)\lambda(dy)\prod_{1\leq t\neq s\leq n}\int f^*_{j_t}(y)\lambda(dy)\\
&\leq\max_{1\leq j_1\leq k}\int f^*_{j_1}(y)\max_{1\leq i_1\leq k}\log\frac{f^*_{i_1}(y)}{f_{i_1}(y)}\lambda(dy),
\end{aligned}$$

$$\log\left(\max_{1\leq i_0\leq k}\frac{\mu^*_{i_0}}{\mu_{i_0}}\right)\leq\frac{1}{\underline{q}}\max_{1\leq i_0\leq k}\left|\mu_{i_0}-\mu^*_{i_0}\right|,$$

and

$$\log\left(\max_{1\leq i,j\leq k}\frac{Q^*_{i,j}}{Q_{i,j}}\right)\leq\frac{1}{\underline{q}}\max_{1\leq i,j\leq k}\left|Q_{i,j}-Q^*_{i,j}\right|$$

because $\min_{1\leq i,j\leq k}(\mu_i,\mu^*_i,Q_{i,j},Q^*_{i,j})\geq\underline{q}$.

Then for all $\varepsilon>0$, for n large enough, for all $\theta\in\Theta_\varepsilon$,

$$\frac{1}{n}KL(p_n^{\theta^*},p_n^{\theta,\mu})\leq\frac{3}{\underline{q}}\,\varepsilon.$$

Proof of Theorem 2.1

This proof relies on Theorem 5 of Barron [Bar88]. We do not assume (A1.2) in the first part of the proof. First we prove that for all $a > 0$,

$$\mathbb{P}^{\theta^*}\left(\frac{\int_\Theta p_n^\theta(Y_1,\dots,Y_n)\pi(d\theta)}{p_n^{\theta^*}(Y_1,\dots,Y_n)} \leq \exp(-an) \text{ i.o.}\right) = 0 \tag{2.10}$$

that is to say

$$p_n^{\theta^*}(y_1,\dots,y_n)\lambda(dy_1)\dots\lambda(dy_n)$$

and

$$\int_\Theta p_n^\theta(y_1,\dots,y_n)\lambda(dy_1)\dots\lambda(dy_n)\pi(d\theta)$$

merge with probability one.

Let $\varepsilon > 0$. Note that Assumption (A1.1a) implies that $Q^* \in \Delta^k(\underline{q})$. Then by Lemma 2.2, there exists a real $\tilde{\varepsilon} > 0$ such that for n large enough, for all $\theta \in \Theta_{\tilde{\varepsilon}}$,

$$\frac{1}{n}KL(p_n^{\theta^*}, p_n^{\theta,\mu}) < \varepsilon. \tag{2.11}$$

Assumptions (A1.0), (A1.1b) and (A1.1d) imply that

$$\sum_{i=1}^k \int f_i^*(y)\left|\log\left(\sum_{j=1}^k f_j(y)\right)\right|\lambda(dy) < +\infty. \tag{2.12}$$

Indeed

$$\begin{aligned}&\int f_i^*(y)\left|\log\left(\sum_{j=1}^k f_j(y)\right)\right|\lambda(dy)\\ &\leq \int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy) + \int_{\{y:\ f_i(y)\geq 1\}} f_i^*(y)\log(k\max_{1\leq j\leq k} f_j(y))\lambda(dy)\end{aligned}$$

and

$$\int_{\{y:\ f_i(y)\geq 1\}} f_i^*(y)\log(k\max_{1\leq j\leq k} f_j(y))\lambda(dy)$$

is finite under (A1.1d) and

$$\int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy)$$

is finite under (A1.0), (A1.1b) and (A1.1d) since

$$\int f_i^*(y) \max_{1\le j\le k} \log\left(\frac{f_j^*(y)}{f_j(y)}\right)\lambda(dy) \ge \int f_i^*(y)\log(f_i^*(y))\lambda(dy) + \int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy) + \int_{\{y:\ f_i(y)\ge 1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy).$$

Moreover by Proposition 1 of Douc *et al.* [DMR04], if $\theta \in \Theta(\underline{q})$ and if (A1.1c), (A1.1d) and (2.12) hold,

$$\frac{1}{n}\log\left(\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta,\mu}(Y_{1:n})}\right)$$

converges $\mathbb{P}^{\theta^*}$-almost surely and in $L^1(\mathbb{P}^{\theta^*})$. Let $\bar{L}(\theta)$ denote this limit:

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta,\mu}(Y_{1:n})}\right) =: \bar{L}(\theta),\ \mathbb{P}^{\theta^*}\text{-a.s. and in } L^1(\mathbb{P}^{\theta^*}).$$

Then using Equation (2.11), for all $\theta \in \Theta_{\tilde{\varepsilon}}$,

$$\bar{L}(\theta) \le \varepsilon. \tag{2.13}$$

So that for all $\varepsilon > 0$, there exists $\tilde{\varepsilon}$ such that

$$\pi\left(\theta : \bar{L}(\theta) < \varepsilon\right) \ \ge\ \pi(\Theta_{\tilde{\varepsilon}}) \ > 0.$$

By Lemma 10 of Barron [Bar88], for all $a > 0$, (2.10) is verified.

We now have to build the tests described in Theorem 5 in Barron [Bar88], to obtain posterior consistency first for the weak topology and secondly for the D_l-pseudometric. In the case of the weak topology, we follow the ideas of Section 4.4.1 in Ghosh and R.V. Ramamoorthi [GR03]. Using page 142 of Ghosh and R.V. Ramamoorthi [GR03], it is sufficient to consider

$$U = \left\{P\ :\ \int h dP - \int h p_l^{\theta^*} d\lambda^{\otimes l} < \varepsilon,\right\},$$

for all $\varepsilon > 0$ and $0 \le h \le 1$ in the set $\mathcal{C}_b((\mathbb{R}^d)^l)$. Choosing α and γ as in page 128 of Ghosh and R.V. Ramamoorthi [GR03], if

$$S^n = \left\{y_1,\ldots,y_n\ :\ \frac{l}{n}\sum_{j=0}^{n/l-1} h(y_{jl+1},\ldots,y_{jl+l}) > \frac{\alpha+\gamma}{2}\right\},$$

then

$$P^{\theta^*}(S^n) = P^{\theta^*}\left\{\sum_{j=0}^{n/l-1}\left(h(y_{jl+1},\dots,y_{jl+l}) - \int h p_l^{\theta^*} d\lambda^{\otimes l}\right) > \frac{n}{l}\frac{\gamma-\alpha}{2}\right\} \\ \leq \exp\left(-\frac{n(\gamma-\alpha)^2(\min_{i,j} Q^*_{i,j})^2}{2l(2-k\min_{i,j} Q^*_{i,j})^2}\right) \tag{2.14}$$

and for all $\theta \in \Theta(\underline{q})$ such that $\int h dP^\theta - \int h p_l^{\theta^*} d\lambda^{\otimes l} \geq \varepsilon$,

$$P^\theta((S^n)^c) \leq P^\theta\left\{\sum_{j=0}^{n/l-1}\left(-h(y_{jl+1},\dots,y_{jl+l}) + \int h p_l^{\theta} d\lambda^{\otimes l}\right) \geq \frac{n}{l}\frac{\gamma-\alpha}{2}\right\} \\ \leq \exp\left(-\frac{n(\gamma-\alpha)^2(\min_{i,j} Q_{i,j})^2}{2l(2-k\min_{i,j} Q_{i,j})^2}\right) \leq \exp\left(-\frac{n(\gamma-\alpha)^2\underline{q}^2}{2l}\right), \tag{2.15}$$

using the upper bound from the proof of Theorem 4 of Gassiat and Rousseau [GR14] based on Corollary 1 of Rio [Rio00].

Using Theorem 5 of Barron [Bar88] and combining Equations (2.14) and (2.15),

$$P^{\theta^*}\left(\pi\left(\left\{\theta : \int h dP^\theta - \int h p_l^{\theta^*} d\lambda^{\otimes l} < \varepsilon\right\}^c \,\middle|\, Y_{1:n}\right) \geq e^{-nr}, \text{ i.o.}\right) = 0$$

which implies that for all weak neighbourhood U of $P_l^{\theta^*}$,

$$P^{\theta^*}\left(\pi(U^c|Y_{1:n}) \geq \exp(-nr) \text{ i.o. }\right) = 0,$$

so that

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(U|Y_{1:n}) = 1\right) = 1.$$

We now assume (A1.2) and obtain consistency for the D_l-pseudometric. Let $\varepsilon > 0$ and let

$$U = \left\{\theta \;:\; D_l(\theta,\theta^*) < \frac{2\varepsilon}{k\underline{q}}\right\} \supset \left\{\theta \;:\; D_l(\theta,\theta^*) < \varepsilon\frac{2-k\min_{1\leq i,j\leq k} Q_{i,j}}{k\min_{1\leq i,j\leq k} Q_{i,j}}\right\},$$

be a D_l-neighbourhood of θ^*. Let

$$B_n^c = \Delta^k(\underline{q}) \times \mathcal{F}_n,$$

so that

$$\pi(B_n) = \pi_f(\mathcal{F}_n{}^c) \lesssim \exp(-nr_1). \tag{2.16}$$

In the proof of Theorem 4 of Gassiat and Rousseau [GR14], it is proved that for all n large enough,

there exists a test ψ_n such that

$$\begin{aligned}\mathbb{E}^{\theta^*}(\psi_n) &\leq N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \exp\left(-\frac{n\varepsilon^2}{8l}\frac{k^2(\min_{i,j} Q^*_{i,j})^2}{(2-k\min_{i,j} Q^*_{i,j})^2}\right) \\ &\leq N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \exp\left(-\frac{n\varepsilon^2 k^2 \underline{q}^2}{32l}\right)\end{aligned} \tag{2.17}$$

$$\sup_{\theta \in U^c \cap B_n^c} \mathbb{P}^{\theta,\mu}(1-\psi_n) \leq \exp\left(-\frac{n\varepsilon^2}{32l}\right). \tag{2.18}$$

Note that for all $\theta, \tilde{\theta}$ in $\Theta(\underline{q})$,

$$D_l(\theta, \tilde{\theta}) \leq \sum_{1\leq i\leq k} |\mu_i^\theta - \mu_i^{\tilde{\theta}}| + k(l-1)\|Q - \tilde{Q}\| + l \max_{1\leq j\leq k} \|f_j - \tilde{f}_j\|_{L_1(\lambda)}.$$

The function $Q \to \mu^Q$ is continuous on the compact $\Delta^k(\underline{q})$ and thus is uniformly continuous: there exists $\alpha > 0$ such that for all $\theta, \tilde{\theta}$ in $\Theta(\underline{q})$ such that $\|Q - \tilde{Q}\| < \alpha$ then $\sum_{1\leq i\leq k} |\mu_i^\theta - \mu_i^{\tilde{\theta}}| < \frac{\varepsilon}{36}$. This implies that

$$\begin{aligned}&N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \\ &\quad\leq N\left(\min\left(\frac{\varepsilon}{36k(l-1)}, \alpha\right), \Delta^k(\underline{q}), \|\cdot\|\right) N\left(\frac{\varepsilon}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \\ &\quad\leq \left(\max\left(\frac{36k(l-1)}{\varepsilon}, \frac{1}{\alpha}\right)\right)^{k(k-1)} N\left(\frac{\varepsilon}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right).\end{aligned} \tag{2.19}$$

Then combining Equations (2.16), (2.17), (2.18), (2.19) and using Theorem 5 of Barron [Bar88], there exists $r > 0$ such that

$$\mathbb{P}^{\theta^*}\left(\pi\left(U^c | Y_{1:n}\right) \geq \exp(-nr) \text{ i.o.}\right) = 0. \tag{2.20}$$

And Equation (2.20) implies that for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\{\theta : D_l(\theta, \theta^*) < \varepsilon\} \mid Y_{1:n}) = 1\right) = 1.$$

Proof of Theorem 2.3

It is sufficient to show that for all weak neighbourhood U_{f^*} of $f^*\lambda$ and neighbourhood U_{Q^*} of Q^*, there exists a D_3-neighbourhood U_{θ^*} of θ^* such that

$$U_{\theta^*} \subset \left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)} \in U_{f_i^*},\ i = 1 \dots k\right\}. \tag{2.21}$$

Following Gassiat *et al.* [GCR15], it is equivalent to show that for all sequences θ^n in $\Theta(\underline{q})$ such that $D_3(\theta^n, \theta^*) \to 0$, there exists a subsequence, that we denote again θ^n, of θ^n and $\bar{\theta} \in \Theta$ such that $\|Q^n - \bar{Q}\| \to 0$, $f_i^n \lambda$ tends to $\bar{f}_i \lambda$ in the weak topology on probability measures for all $i \leq k$ and $p_3^{(Q^*, f^*)} = p_3^{(\bar{Q}, \bar{f})}$.

Let θ^n in $\Theta(\underline{q})$ such that $D_3(\theta^n, \theta^*) \to 0$. As $\Delta^k(\underline{q})$ is a compact set, there exists a subsequence of Q^n that we denote again Q^n which tends to $\bar{Q} \in \Delta^k(\underline{q})$. Writing μ^n the (sub)sequence of the stationary distribution associated to Q_n, then $\mu^n \to \bar{\mu}$ where $\bar{\mu}$ is the stationary distribution associated to $\bar{Q}$. Moreover, using the reverse triangle inequality,

$$
\begin{aligned}
D_3(\theta^n, \theta^*) &= \| p_3^{\theta_n} - p_3^{\theta^*} \|_{L_1(\lambda^{\otimes 3})} \\
&= \int \Big| \sum_{1 \leq i_1, i_2, i_3 \leq k} \mu_{i_1}^n Q_{i_1, i_2}^n Q_{i_2, i_3}^n f_{i_1}^n(y_1) f_{i_2}^n(y_2) f_{i_3}^n(y_3) - \\
&\quad \mu_{i_1}^* Q_{i_1, i_2}^* Q_{i_2, i_3}^* f_{i_1}^*(y_1) f_{i_2}^*(y_2) f_{i_3}^*(y_3) \Big| \, \lambda(dy_1) \lambda(dy_2) \lambda(dy_3) \\
&\geq - \sum_{1 \leq i_1, i_2, i_3 \leq k} \left| \mu_{i_1}^n Q_{i_1, i_2}^n Q_{i_2, i_3}^n - \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} \right| + \\
&\quad \int \Big| \sum_{1 \leq i_1, i_2, i_3 \leq k} \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} f_{i_1}^n(y_1) f_{i_2}^n(y_2) f_{i_3}^n(y_3) - \\
&\quad \mu_{i_1}^* Q_{i_1, i_2}^* Q_{i_2, i_3}^* f_{i_1}^*(y_1) f_{i_2}^*(y_2) f_{i_3}^*(y_3) \Big| \, \lambda(dy_1) \lambda(dy_2) \lambda(dy_3),
\end{aligned}
$$

since $\sum_{1 \leq i_1, i_2, i_3 \leq k} \left| \mu_{i_1}^n Q_{i_1, i_2}^n Q_{i_2, i_3}^n - \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} \right|$ tends to zero,

$$
\begin{aligned}
\lim_n \int \Big| \sum_{1 \leq i_1, i_2, i_3 \leq k} & \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} f_{i_1}^n(y_1) f_{i_2}^n(y_2) f_{i_3}^n(y_3) - \\
& \mu_{i_1}^* Q_{i_1, i_2}^* Q_{i_2, i_3}^* f_{i_1}^*(y_1) f_{i_2}^*(y_2) f_{i_3}^*(y_3) \Big| \, \lambda(dy_1) \lambda(dy_2) \lambda(dy_3) = 0.
\end{aligned} \tag{2.22}
$$

Let $F_1^n, \ldots, F_k^n$ be the probability distribution with respective densities f_1^n, $\ldots$, f_k^n with respect to λ. Since

$$
\sum_{i_1, i_2, i_3} \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} F_{i_1}^n \otimes F_{i_2}^n \otimes F_{i_3}^n
$$

converges in total variation, it is tight and for all $1 \leq i \leq k$, $(F_i^n)_n$ is tight. By Prohorov's theorem, for all $1 \leq i \leq k$ there exists a subsequence denoted F_i^n of F_i^n which weakly converges to $\bar{F}_i$. This in turns implies that

$$
\sum_{i_1, i_2, i_3} \bar{\mu}_{i_1} \bar{Q}_{i_1, i_2} \bar{Q}_{i_2, i_3} F_{i_1}^n \otimes F_{i_2}^n \otimes F_{i_3}^n
$$

weakly converges to

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1} \bar{Q}_{i_1,i_2} \bar{Q}_{i_2,i_3} \bar{F}_{i_1} \otimes \bar{F}_{i_2} \otimes \bar{F}_{i_3},$$

which combined with (2.22), leads to

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1} \bar{Q}_{i_1,i_2} \bar{Q}_{i_2,i_3} \bar{F}_{i_1} \otimes \bar{F}_{i_2} \otimes \bar{F}_{i_3} = \sum_{i_1,i_2,i_3} \mu^*_{i_1} Q^*_{i_1,i_2} Q^*_{i_2,i_3} f^*_{i_1} \lambda \otimes f^*_{i_2} \lambda \otimes f^*_{i_3} \lambda.$$

By Gassiat *et al.* [GCR15], $\bar{Q} = Q^*$, so $\bar{\mu} = \mu^*$ and $\bar{F}_i = f^*_i \lambda$ up to a label switching, that is there exists a permutation $\sigma \in \mathcal{S}_k$ such that $\sigma \bar{Q} = Q^*$ and $\bar{F}_{\sigma(i)} = f^*_i \lambda$ so that Equation (2.21) holds.

In other words we have proved the continuity of the functional

$$\begin{cases} (\{p^\theta_l,\ \theta \in \Theta_I\}, L_1) & \to & (\Theta_I/\mathcal{R}_\sigma, \mathcal{T}) \\ p^\theta_l & \mapsto & \theta \end{cases}$$

where $\Theta_I = \{\theta \in \Theta : Q$ has full rank , $f_1 d\lambda \dots f_k d\lambda$ are linearly independent$\}$ and $\mathcal{R}_\sigma$ is the equivalence relation on Θ such that $\theta \mathcal{R}_\sigma \tilde{\theta}$ if there exists $\sigma \in \mathcal{S}_k$ such that for all $1 \le i, j \le k$, $Q_{i,j} = \tilde{Q}_{\sigma(i),\sigma(j)}$ and $f_i = \tilde{f}_{\sigma(i)}$; using that

$$\underset{p^\theta_l}{\left(\{p^\theta_l, \theta \in \Theta_I\}, L_1\right)} \underset{\text{continuous}}{\longrightarrow} \underset{p^\theta_l}{\left(\{p^\theta_l, \theta \in \Theta_I\}, \begin{matrix}\text{weak} \\ \text{topology}\end{matrix}\right)} \longrightarrow \underbrace{\overbrace{\left(\Theta_I/\mathcal{R}_\sigma, \mathcal{T}\right)}^{\text{compact}}}_{\theta}.$$

$$\text{continuous, bijective}$$

Proof of Proposition 2.4

To prove Proposition 2.4, using Equation (2.4), it is sufficient to prove that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\begin{aligned} &\left\{\theta \in \Theta : H_1(\theta) < \dots < H_k(\theta), \exists \sigma \in \mathcal{S}_k, \|\sigma Q - Q^*\| < \eta, \max_{1 \le i \le k} d_w(f_{\sigma(i)}, f^*_i) < \eta\right\} \\ &\subset \left\{\theta : H_1(\theta) < \dots < H_k(\theta), \|Q - Q^*\| < \varepsilon, \max_{1 \le i \le k} d_w(f_i, f^*_i) < \varepsilon\right\} \end{aligned} \tag{2.23}$$

where d_w metricizes the weak topology on $\mathcal{F}$. Using Equation (2.3),

$$\delta := \min_{1 \le i \le k-1} |H_{i+1}(\theta^*) - H_i(\theta^*)| > 0 \tag{2.24}$$

and by continuity of H for all $\varepsilon > 0$, there exists $\eta_1 > 0$ such that for all

$$\theta \in \left\{\theta \in \Theta :\ H_1(\theta) < \cdots H_k(\theta),\ \exists \sigma \in \mathcal{S}_k, \|\sigma Q - Q^*\| < \eta_1, d_w(f_{\sigma(i)}, f_i^*) < \eta_1\right\},$$

for all $1 \leq i \leq k$, $|H_i(\theta) - H_i(\theta^*)| < \delta/2$. For such θ, using Equation (2.2), we obtain for all $\sigma \in \mathcal{S}_k$,

$$|H_i((\sigma Q, f_{\sigma(1)}, \ldots, f_{\sigma(k)})) - H_{\sigma(i)}(\theta^*)| < \delta/2$$

so that using Equations (2.3), (2.24) and that $H_1(\theta) < \cdots < H_k(\theta)$, the permutation σ is equal to the identity permutation. Thus Equation (2.23) holds with $\eta = \min(\eta_1, \varepsilon)$.

Proof of Theorem 2.8

To prove Theorem 2.8 we need the following lemma:

Lemma 2.10. *Let $\varepsilon > 0$, for all $0 < \varepsilon_1 < 1$, $N > 0$, $1 \leq j < N$ and $c > 0$ such that*

$$0 < \frac{\varepsilon_1 k^N}{c(c - \varepsilon_1)} < \frac{\varepsilon}{3} \text{ and } \frac{2(1-\underline{q})^{N+1-j}}{\underline{q} + (1-\underline{q})^{N+1-j}} < \frac{\varepsilon}{3}.$$

If

$$p_N^{\theta^*}(Y_{1:N}) > c \tag{2.25}$$

then for all $n > N$,

$$\begin{aligned}
&\left\{\theta \in \Theta(\underline{q}) : \|p_N^{\theta^*} - p_N^{\theta}\|_{l_1} < \varepsilon_1,\ \exists \sigma \in \mathcal{S}_k,\ \max_{1\leq i\leq k} |\mu_{\sigma(i)}^{\theta} - \mu_i^*| < \varepsilon_1,\right. \\
&\qquad\qquad \left.\|\sigma Q - Q^*\| < \varepsilon_1,\ \max_{1\leq i\leq k} \|f_{\sigma(i)} - f_i^*\|_{l_1} < \varepsilon_1\right\} \\
&\subset \left\{\theta \in \Theta(\underline{q}) : \exists \sigma \in \mathcal{S}_k, \max_{1\leq l\leq k} \left|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:n})\right| < \varepsilon\right\}.
\end{aligned}$$

Proof of Lemma 2.10. Let $\theta \in \Theta(\underline{q})$ such that

$$\|p_N^{\theta^*} - p_N^{\theta}\|_{l_1} < \varepsilon_1$$

and there exists $\sigma \in \mathcal{S}_k$ such that

$$\max_{1\leq i\leq k} |\mu_{\sigma(i)}^{\theta} - \mu_i^*| < \varepsilon_1,\ \|\sigma Q - Q^*\| < \varepsilon_1,\ \max_{1\leq i\leq k} \|f_{\sigma(i)} - f_i^*\|_{l_1} < \varepsilon_1. \tag{2.26}$$

To bound $|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = l \mid Y_{1:n})|$, we now prove that it is sufficient to bound $|P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N})|$ with $N < n$ a well chosen fixed integer thanks to

the exponential forgetting of the HMM. Let $1 \leq a \leq k$,

$$\begin{aligned} &|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:n})| \\ &\leq A^l_{\theta^*} + |P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N})| + A^{\sigma(l)}_{\theta}, \end{aligned} \tag{2.27}$$

where for $\tilde{\theta} \in \{\theta, \theta^*\}$ and for all $1 \leq l \leq k$,

$$\begin{aligned} A^l_{\tilde{\theta}} = \Bigg| &\frac{P^{\tilde{\theta}}(Y_{1:N}, X_j = l) \sum_{1 \leq b \leq k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = l, Y_{j:N})}{\sum_{1 \leq m \leq k} P^{\tilde{\theta}}(Y_{1:N}, X_j = m) \sum_{1 \leq b \leq k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = m, Y_{j:N})} \\ &- \frac{P^{\tilde{\theta}}(Y_{1:N}, X_j = l) \sum_{1 \leq b \leq k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = a, Y_{j:N})}{\sum_{1 \leq m \leq k} P^{\tilde{\theta}}(Y_{1:N}, X_j = m) \sum_{1 \leq b \leq k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = a, Y_{j:N})} \Bigg|. \end{aligned}$$

Using Corollary 1 of Douc *et al.* [DMR04], i.e. the exponential forgetting of the HMM, we obtain for all $(b, \omega, m) \in \{1, \ldots, k\}^3$,

$$\begin{aligned} &\left| P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = m, Y_{j:N}) - P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = \omega, Y_{j:N}) \right| \\ &\leq (1 - \underline{q})^{N+1-j} \leq (1 - \underline{q})^{N+1-j} \frac{P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = \omega, Y_{j:N})}{\underline{q}} \end{aligned}$$

so that for $\tilde{\theta} \in \{\theta, \theta^*\}$ and for all $1 \leq l \leq k$

$$A^l_{\tilde{\theta}} \leq \frac{2(1 - \underline{q})^{N+1-j}}{\underline{q} + (1 - \underline{q})^{N+1-j}}. \tag{2.28}$$

Moreover, using (2.25) and (2.26), for all $1 \leq i, j \leq k$, $Y_{1:N} \in \mathbb{N}^N$,

$$\mu^{\theta}_{\sigma(i)} \geq \mu^*_i - \varepsilon_1, \quad Q_{\sigma(i),\sigma(j)} \geq Q^*_{i,j} - \varepsilon_1, \quad f_{\sigma(a_i)}(Y_i) \geq f^*_{a_i}(Y_i) - \varepsilon_1 \quad \text{and}$$

$$p^{\theta}_N(Y_{1:N}) \leq p^{\theta^*}_N(Y_{1:N})(1 + \varepsilon_1/c),$$

we obtain

$$
\begin{aligned}
&P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N}) \\
&= \frac{\sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} Q^*_{a_1,a_2} \cdots Q^*_{a_{j-1},l} Q^*_{l,a_{j+1}} \cdots Q^*_{a_{N-1},a_N} f^*_{a_1}(Y_1) \cdots f^*_{l}(Y_j) \cdots f^*_{a_N}(Y_N)}{p^{\theta^*}_N(Y_{1:N})} \\
&\quad - \frac{\sum_{a_{1:j-1}, a_{j+1:N}} \mu^{\theta}_{\sigma(a_1)} Q_{\sigma(a_1),\sigma(a_2)} \cdots Q_{\sigma(a_{N-1}),\sigma(a_N)} f_{\sigma(a_1)}(Y_1) \cdots f_{\sigma(a_N)}(Y_N)}{p^{\theta}_N(Y_{1:N})} \\
&\qquad\qquad\qquad\qquad \text{where } a_j = l \text{ as in the following,} \\
&\leq \frac{(1+\varepsilon_1/c) \sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) - \sum_{a_{1:j-1}, a_{j+1:N}} \mu^{\theta}_{\sigma(a_1)} \cdots f_{\sigma(a_N)}(Y_N)}{(1+\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})} \\
&\leq \frac{(1+\varepsilon_1/c) \sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) - \sum_{a_{1:j-1}, a_{j+1:N}} (\mu^*_{a_1} - \varepsilon_1) \cdots (f^*_{a_N}(Y_N) - \varepsilon_1)}{c + \varepsilon_1}.
\end{aligned}
$$

Expanding the product in the second sum, the numerator becomes a sum where each term is bounded by $(\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$. Indeed the first term is equal to

$$
\sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) = p^{\theta^*}_N(Y_{1:N})
$$

which gives $(\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$ when subtracted to the first sum. The other terms are a product of a positive power of ε_1 and μ^*_i, $Q^*_{i,j}$ or $f^*_{a_i}(Y_i)$ which are all bounded by 1. Thus they are bounded by $\varepsilon_1 \leq (\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$. Moreover there are k^N terms so that

$$
P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N}) \leq \frac{\varepsilon_1 k^N}{c(c+\varepsilon_1)}.
$$

Similarly

$$
P^{\theta}(X_j = \sigma(l) \mid Y_{1:N}) - P^{\theta^*}(X_j = l \mid Y_{1:N}) \leq \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)}
$$

so that

$$
\left| P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N}) \right| \leq \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)}. \tag{2.29}
$$

Combining Equations (2.27), (2.28) and (2.29), we obtain

$$
\begin{aligned}
&|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:n})| \\
&\leq 2 \frac{2(1-\underline{q})^{N+1-j}}{\underline{q} + (1-\underline{q})^{N+1-j}} + \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)} < \varepsilon.
\end{aligned}
$$

□

We prove Theorem 2.8 for $m = 1$, one may easily generalizes the proof. Let $\beta > 0$, $j > 0$ and

$\varepsilon > 0$, we fix N and $c > 0$ such that

$$\frac{2(1-\underline{q})^{N+1-j}}{\underline{q}+(1-\underline{q})^{N+1-j}} < \frac{\varepsilon}{3} \text{ and } P^{\theta^*}\left(p_N^{\theta^*}(Y_{1:N}) > c\right) > \sqrt{1-\beta} \tag{2.30}$$

then we choose ε_1 such that

$$0 < \frac{\varepsilon_1 2^{2N} k^N}{c(c-\varepsilon_1)} < \frac{\varepsilon}{3}. \tag{2.31}$$

Posterior consistency for the marginal distribution in l_1 and for all components of the parameter i.e. Theorems 2.1 and 2.3 imply that there exists M such that P^{θ^*}-a.s., for all $n \geq M$,

$$\pi\left(\{\theta \ : \ D_N(\theta, \theta^*) < \varepsilon_1\} \mid Y_{1:n}\right) > \frac{\sqrt{1-\beta}+1}{2} \tag{2.32}$$

and

$$\begin{aligned}\pi\Bigg(\{\theta \ : \ \exists \sigma \in \mathcal{S}_k, \ \max_{1\leq i\leq k} |\mu_{\sigma(i)} - \mu_i^*| < \varepsilon_1, \ \|\sigma Q - Q^*\| < \varepsilon_1, \\ \max_{1\leq i\leq k} \|f_{\sigma(i)} - f_i^*\|_{l_1}\} < \varepsilon_1 \ \Bigg| \ Y_{1:n}\Bigg) > \frac{\sqrt{1-\beta}+1}{2}.\end{aligned} \tag{2.33}$$

Using Lemma 2.10 and combining (2.30), (2.31), (2.32) and (2.33), we obtain for all $n \geq \max(N, M)$,

$$\begin{aligned}&\mathbb{E}^{\theta^*}\Bigg(\pi\Bigg(\Big\{\theta : \exists\sigma \in \mathcal{S}_k, \max_{1\leq l\leq k}\Big|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j = \sigma(l)|Y_{1:n})\Big| < \varepsilon\Big\}|Y_{1:n}\Bigg)\Bigg) \\ &\geq \mathbb{E}^{\theta^*}\Bigg(\mathbb{1}_{p_N^{\theta^*}(Y_{1:N})>c} \\ &\qquad \pi\Bigg(\Big\{\theta\colon \exists\sigma, \max_{1\leq l\leq k}\Big|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j = \sigma(l)|Y_{1:n})\Big| < \varepsilon\Big\}\Big|Y_{1:n}\Bigg)\Bigg) \\ &\geq 1-\beta.\end{aligned}$$

Then

$$\mathbb{E}^{\theta^*}\Bigg(\pi\Bigg(\Big\{\theta : \exists\sigma \in \mathcal{S}_k, \max_{1\leq l\leq k}\Big|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j = \sigma(l)|Y_{1:n})\Big| < \varepsilon\Big\}|Y_{1:n}\Bigg)\Bigg)$$

tends to 1, which concludes the proof of Theorem 2.8.

Proof of Proposition 2.9

As under $DP(\alpha G_0)^{\otimes k}$, $f_i(l)$ is distributed from Beta$(\alpha G_0(l), \alpha \sum_{m\neq l} G_0(m))$,

$$\begin{aligned}
&\int_{\mathcal{F}^k} \sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leq j\leq k} \left(-\log(f_j(l))\right) \left(DP(\alpha G_0)\right)^{\otimes k}(df) \\
&\leq \sum_{l=1}^{+\infty} f_i^*(l) \sum_{1\leq j\leq k} \int_{\mathcal{F}^k} \left(-\log(f_j(l))\right) \left(DP(\alpha G_0)\right)^{\otimes k}(df) \\
&\leq \sum_{l=1}^{+\infty} \frac{f_i^*(l)\,\Gamma(\alpha)}{\Gamma(\alpha G_0(l))\Gamma\Big(\alpha \sum\limits_{m\neq l} G_0(m)\Big)} \int_0^1 -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum\limits_{m\neq l} G_0(m)-1} \lambda(dx). \qquad (2.34)
\end{aligned}$$

On $[1/2, 1]$, $-\log(x)x^{\alpha G_0(l)-1} \leq 2\log(2)$, so that there exists a constant C_1 which does not depend on l such that

$$\int_{1/2}^1 -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum_{m\neq l} G_0(m)-1} \lambda(dx) \leq C_1. \qquad (2.35)$$

On $[0, 1/2]$, $(1-x)^{\alpha \sum_{m\neq l} G_0(m))-1} \leq 2$, so that there exists a constant C_2 which does not depend on l such that

$$\int_0^{1/2} -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum_{m\neq l} G_0(m)-1} \lambda(dx) \leq \frac{C_2}{\left(\alpha G_0(l)\right)^2}. \qquad (2.36)$$

Moreover for all $0 < \delta < 1$,

$$\frac{1}{\delta} \leq \Gamma(\delta) = \frac{\Gamma(\delta+1)}{\delta} \leq \frac{2}{\delta}. \qquad (2.37)$$

By combining Equations (2.34), (2.35), (2.36) and (2.37), for all $1 \leq i \leq k$,

$$\begin{aligned}
&\int_{\mathcal{F}^k} \sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leq j\leq k} \left(-\log(f_j(l))\right) \left(DP(\alpha G_0)\right)^{\otimes k}(df) \\
&\qquad \lesssim \sum_{l=1}^{+\infty} \frac{f_i^*(l)}{\alpha G_0(l)}
\end{aligned}$$

so that using Assumption (E1.1),

$$\begin{aligned}
\left(DP(\alpha G_0)\right)^{\otimes k} \Bigg(f_1, \ldots, f_k \ : \ &\forall 1 \leq i \leq k, \\
&\sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leq j\leq k} \left(-\log(f_j(l))\right) < +\infty \Bigg) = 1.
\end{aligned}$$

Note that for all $\varepsilon > 0$,

$$\left\{ f_1, \ldots, f_k \ : \ \forall 1 \leq i \leq k, \ \sum_{l=1}^{+\infty} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < +\infty \right\}$$
$$\subset \bigcup_{N \in \mathbb{N}} \left\{ f_1, \ldots, f_k \ : \ \forall 1 \leq i \leq k, \ \sum_{l=N}^{+\infty} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < \varepsilon \right\},$$

thus arguing by contradiction, for all $\varepsilon > 0$, there exists L_ε such that

$$\big(DP(\alpha G_0)\big)^{\otimes k} \Bigg(f_1, \ldots, f_k \ : \ \forall 1 \leq i \leq k, \\ \sum_{l > L_\varepsilon} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < \varepsilon \Bigg) > 0.$$

Using the tail free property of the Dirichlet process, for all $1 \leq j \leq k$,

$$\sum_{l > L_\varepsilon} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < \varepsilon$$

and

$$\left(\frac{f_j(1)}{\sum_{l \leq L_\varepsilon} f_j(l)}, \ldots, \frac{f_j(L_\varepsilon)}{\sum_{l \leq L_\varepsilon} f_j(l)} \right) \tag{2.38}$$

are independent given $\sum_{l > L_\varepsilon} f_j(l)$ and (2.38) given $\sum_{l > L_\varepsilon} f_j(l)$ has a Dirichlet distribution with parameter $(\alpha G_0(1), \ldots, \alpha G_0(L_\varepsilon))$. Then for all $\varepsilon > 0$, there exists L_ε such that for all $\delta \in (0, 1)$,

$$\big(DP(\alpha G_0)\big)^{\otimes k} \Bigg(f_1, \ldots, f_k \ : \ \forall 1 \leq i \leq k, \ \sum_{l > L_\varepsilon} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < \frac{\varepsilon}{2}, \\ \forall l \leq L_\varepsilon, \ |f_j(l) - f_j^*(l)| \leq c\delta \Bigg) > 0 \tag{2.39}$$

where $c = \min_{1 \leq i \leq k} \min_{l \leq L_\varepsilon, f_i^*(l) > 0} f_i^*(l)$.

For all $f_1, \ldots, f_k$ such that for all $1 \leq i, j \leq k$,

$$\sum_{l > L_\varepsilon} f_i^*(l) \max_{1 \leq j \leq k} (-\log(f_j(l))) < \frac{\varepsilon}{2}$$

and for all $l \leq L_\varepsilon$, $|f_j(l) - f_j^*(l)| \leq c\delta$,

$$\begin{aligned}
&\sum_{l\in\mathbb{N}} f_i^*(l) \max_{1\leq j\leq k} \log\left(\frac{f_j^*(l)}{f_j(l)}\right) \\
&\quad= \sum_{l\leq L_\varepsilon} f_i^*(l) \max_{1\leq j\leq k} \log\left(\frac{f_j^*(l)}{f_j(l)}\right) + \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leq j\leq k} \log(f_j^*(l)) \\
&\qquad+ \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leq j\leq k} (-\log(f_j(l))) \\
&\quad\leq \frac{\delta}{1-\delta} + 0 + \frac{\varepsilon}{2} \leq \varepsilon
\end{aligned} \tag{2.40}$$

for δ small enough. For such a δ denote

$$\begin{aligned}
\Theta_\varepsilon = \{Q \,:\, \|Q - Q^*\| \leq \varepsilon\} \times \{f_1, \ldots, f_k \,:\, &\sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leq j\leq k} (-\log(f_j(l))) < \frac{\varepsilon}{2}, \\
&\forall l \leq L_\varepsilon,\ |f_j(l) - f_j^*(l)| \leq c\delta,\ \forall 1 \leq i, j \leq k\}
\end{aligned}$$

Using Equation (2.40), (A1.1b) holds. Furthermore (A1.1d) is obviously checked. Under Assumption (E1.1), $G_0(l) > 0$ when $\sum_{i=1}^k f_i^*(l) > 0$ so that (A1.1c) holds. Using the assumption that Q^* is in the support of π_Q, (A1.1a) is checked. Then using Equation (2.39), (A1.1) holds and the first part of Proposition 2.9 follows.

We now prove the second part of Proposition 2.9. We first give a representation of a discrete Dirichlet process with independent Gamma distributed random variables.

Lemma 2.11 (Ferguson [Fer74]). *Let $(Z_l)_{l\in\mathbb{N}}$ be independent random variables such that for all $l \in \mathbb{N}$,*

$$Z_l \sim \Gamma(\alpha G_0(l), 1),$$

then $\sum_{l=1}^L Z_l$ converges almost surely and its limit has a gamma distribution $\Gamma(\alpha, 1)$.

Moreover denote

$$f : \begin{cases} \mathbb{N} & \to [0,1] \\ i & \to f(i) = Z_i / (\sum_{l=1}^{+\infty} Z_l) \end{cases},$$

then f is distributed from a Dirichlet process $DP(\alpha G_0)$.

We assume (A1.1b) i.e. for all $\varepsilon > 0$,

$$DP(\alpha G_0)^{\otimes k}\left(\left\{f \in \mathcal{F}^k, \forall i \in \{1,\ldots,k\} \sum_{l\in\mathbb{N}} f_i^*(l) \max_{1\leq j\leq k} \log\frac{f_j^*(l)}{f_j(l)} < \varepsilon\right\}\right) > 0.$$

Let $\varepsilon > 0$, define $\mathcal{F}_\varepsilon$ as the set of $f = (f_1, \ldots, f_k) \in \mathcal{F}^k$ such that for all $1 \leq i \leq k$, for all $f \in \mathcal{F}_\varepsilon$,

$$\sum_{l\in\mathbb{N}} f_i^*(l) \log\left(\frac{f_i^*(l)}{f_i(l)}\right) < \varepsilon.$$

Then $DP(\alpha G_0)^{\otimes k}(\mathcal{F}_\varepsilon) > 0$.

Since $\sum_l f_i^*(l)(-\log f_i^*(l))$ converges, then $\sum_l f_i^*(l)(-\log f_i(l))$ converges. Using Lemma 2.11, we can write f_i with independent gamma distributed random variables $(Z_l)_{l\in\mathbb{N}}$:

$$f_i(l) = \frac{Z_l}{\sum_{j\in\mathbb{N}} Z_j},$$

where $Z_l \sim \Gamma(\alpha G_0(l), 1)$. Then $\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$ converges since $\sum_{j\in\mathbb{N}} Z_j$ is finite almost surely. Since $DP(\alpha G_0)^{\otimes k}(\mathcal{F}_\varepsilon) > 0$, for all $1 \le i \le k$ with positive probability,

$$\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$$

converges. Using the Kolmogorov 0-1 law and the Three-Series Theorem (see Section 9.7.3 in Dudley [Dud02]), $\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$ converges almost surely and

$$\sum_{l\in\mathbb{N}} \mathbb{P}(|f_i^*(l)(-\log(Z_l))| > 1) < +\infty, \tag{2.41}$$

$$\sum_{l\in\mathbb{N}} \mathbb{E}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\le 1}\big) < +\infty, \tag{2.42}$$

$$\sum_{l\in\mathbb{N}} \mathrm{var}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\le 1}\big) < +\infty. \tag{2.43}$$

Equation (2.41) implies that

$$\begin{aligned}
+\infty &> \sum_{l\in\mathbb{N}} \mathbb{P}(|f_i^*(l)(-\log(Z_l))| > 1) \\
&\ge \sum_{l\in\mathbb{N}} \frac{1}{\Gamma(\alpha G_0(l))} \int_0^{\exp(-1/f_i^*(l))} x^{\alpha G_0(l)-1} e^{-x} dx \\
&\ge \sum_{l\in\mathbb{N}} \frac{1}{\alpha G_0(l)\Gamma(\alpha G_0(l))} \exp\left(-\exp\left(\frac{-1}{f_i^*(l)}\right) - \frac{\alpha G_0(l)}{f_i^*(l)}\right) \\
&\gtrsim \sum_{l\in\mathbb{N}} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right)
\end{aligned}$$

using Equation (2.37). Then

$$\lim_{l\to\infty} \frac{f_i^*(l)}{G_0(l)} = 0. \tag{2.44}$$

Moreover Equation (2.42) implies that

$$\begin{aligned}
+\infty > &\sum_l \mathbb{E}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\leq 1}\big) \\
&\geq \sum_l \left(\int_{\exp(-1/f_i^*(l))}^1 \frac{1}{\Gamma(\alpha G_0(l))} f_i^*(l)(-\log(x)) x^{\alpha G_0(l)-1} e^{-x} dx \right. \\
&\qquad \left. + \int_1^{\exp(1/f_i^*(l))} \frac{1}{\Gamma(\alpha G_0(l))} f_i^*(l)(-\log(x)) x^{\alpha G_0(l)-1} e^{-x} dx \right) \\
&\geq \sum_l \left(\frac{e^{-1} f_i^*(l)}{\Gamma(\alpha G_0(l))} \int_{\exp(-1/f_i^*(l))}^1 (-\log(x)) x^{\alpha G_0(l)-1} dx \right. \\
&\qquad \left. - \frac{1}{\Gamma(\alpha G_0(l))} \int_1^{\exp(1/f_i^*(l))} e^{-x} dx \right) \\
&\gtrsim -\alpha + \sum_l \frac{e^{-1} f_i^*(l)}{\alpha^2 G_0^2(l) \Gamma(\alpha G_0(l))} \\
&\qquad \left(1 - \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)} \right) - \frac{\alpha G_0(l)}{f_i^*(l)} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)} \right) \right) \\
&\gtrsim -\alpha + \sum_l \frac{f_i^*(l)}{G_0(l)}
\end{aligned}$$

using Equation (2.37) and that

$$\lim_{l\to\infty} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)} \right) + \frac{\alpha G_0(l)}{f_i^*(l)} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)} \right) = 0$$

using Equation (2.44). Then

$$\sum_{l\in\mathbb{N}} \frac{f_i^*(l)}{G_0(l)} < +\infty.$$

2.5 Other Proofs

Proof of Proposition 2.5

The proof uses many ideas of Tokdar [Tok06].

We now prove that Assumptions (B1.1), (B1.2), (B1.3) and (B1.4) imply (A1.1). A reproduction of the proof of Theorem 3.2. and Lemma 3.1 of Tokdar [Tok06] shows that Assumptions (B1.2), (B1.3) and (B1.4) imply that for all $\varepsilon > 0$, for all $1 \leq j \leq k$ there exists a weak neighbourhood V_j of a compactly supported probability measure $\tilde{P}_j$ such that for all $f_j = \phi * P_j$, $P_j \in V_j$,

$$\int_{\mathbb{R}} f_i^*(y) \max_{1\leq j\leq k} \log\left(\frac{f_j^*(y)}{f_j(y)} \right) \lambda(dy) < \varepsilon. \tag{2.45}$$

Let $0 < \underline{\sigma} < \bar{\sigma}$ and $\zeta > 0$ be such that for all $1 \leq j \leq k$

$$\tilde{P}_j([-\zeta, \zeta] \times [\underline{\sigma}, \bar{\sigma}]) = 1.$$

Let $\delta = \underline{\sigma}/2$. For all $1 \leq j \leq k$ define

$$U_j = \left\{ P \ : \ \left| \int_{\mathbb{R}\times(0,+\infty)} h dP - \int_{\mathbb{R}\times(0,+\infty)} h d\tilde{P}_j \right| < \varepsilon \right\},$$

where $h : \mathbb{R} \times (0, +\infty) \to [0, 1]$ is a piecewise affine continuous function such that $h(z, \sigma) = 1$ for all $z \in [-\zeta, \zeta]$ and $\sigma \in [\underline{\sigma}, \bar{\sigma}]$ and $h(z, \sigma) = 0$ for all $z \in [-\zeta - \delta, \zeta + \delta]^c$ and $\sigma \in [\underline{\sigma} - \delta, \bar{\sigma} + \delta]^c$. For all $\varepsilon > 0$, define

$$\Theta_\varepsilon = \{Q \ : \ \|Q - Q^*\| < \varepsilon\} \times (V_1 \cap U_1) \times \cdots \times (V_k \cap U_k).$$

Then for all $(Q, \phi * P_1, \ldots, \phi * P_k) \in \Theta_\varepsilon$, (A1.1b) is true according to Equation (2.45). In addition, for all $y \in \mathbb{R}$,

$$\begin{aligned} f_j(y) &\geq \int_{[-\zeta-\delta,\zeta+\delta]\times[\underline{\sigma}-\delta,\bar{\sigma}+\delta]} \phi_\sigma(y - z) P_j(dz, d\sigma) \\ &\geq \frac{1}{\bar{\sigma} + \delta} \, \phi_{\underline{\sigma}-\delta}\big(\max(|y - \zeta - \delta|, |y + \zeta + \delta|) \big) \, (1 - \varepsilon) \end{aligned}$$

which implies (A1.1c). Moreover using assumption (B1.1), Π_P-a.s. there exists $C > 0$ such that for all $1 \leq j \leq k$,

$$f_j(y) \leq \int \frac{1}{\sigma} P_j(dz, d\sigma) \leq C$$

so that (A1.1d) holds. As Θ_ε is a product of neighbourhoods of elements in the support of their respective prior, $\pi(\Theta_\varepsilon) > 0$, so (A1.1) is checked.

Now we prove that Assumption (B1.5) implies Assumption (A1.2). Let $\delta > 0$. For all $a, l, u, \kappa > 0$, such that $l < u$ denote $\mathcal{F}^\kappa_{a,l,u} = \{\phi * P : P((-a, a] \times (l, u]) > 1 - \kappa\}$. Using Section 4 of Tokdar [Tok06], there exist b_0, b_1, b_2 only depending on κ such that

$$\begin{aligned} \log(N(3\kappa, (\mathcal{F}^\kappa_{a,l,u})^k, d)) &\leq k \log(N(3\kappa, \mathcal{F}^\kappa_{a,l,u}, \|\cdot\|_{L_1(\lambda)})) \\ &\leq k b_0 \left(b_1 \frac{a}{l} + b_2 \log\left(\frac{u}{l}\right) + 1 \right). \end{aligned} \tag{2.46}$$

Choosing $\kappa = \frac{\delta}{3*36l}$ and $\beta < \frac{\delta^2 k \underline{q}^2}{32 l b_0 (b_1 + b_2)}$, Assumption (B1.5) implies that Assumption (A1.2) holds.

Proof of Corollary 2.6

To prove the first part of Corollary 2.6, we use Theorem 2.3 because $m_1^* < \cdots < m_k^*$ implies the linear independence of $g^*(\cdot - m_1^*)\lambda, \ldots, g^*(\cdot - m_k^*)\lambda$. Then it is sufficient to prove that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\begin{aligned} &\left\{\theta :\ \exists \sigma \in \mathcal{S}_k,\ \max_{1\le i\le k} d_w\big(g(\cdot - m_{\sigma(i)}), g^*(\cdot - m_i^*)\big) < \eta,\ \|\sigma Q - Q^*\| < \eta\right\} \\ &\subset \left\{\theta :\ d_w(g, g^*) < \varepsilon,\ \max_{1\le j\le k} |m_j - m_j^*| < \varepsilon,\ \|Q - Q^*\| < \varepsilon\right\}, \end{aligned} \tag{2.47}$$

where d_w metricizes the weak topology on $\mathcal{F}$. Let ξ^n be a sequence of $\Theta(\underline{q})$ such that for all n there exists $\sigma_n \in \mathcal{S}_k$ such that for all $1 \le i \le k$,

$$d_w\big(g^n(\cdot - m^n_{\sigma_n(i)}), g^*(\cdot - m_i^*)\big) \to 0 \text{ and } \|\sigma_n Q^n - Q^*\| \to 0.$$

As there exists a finite number of permutation in $\mathcal{S}_k$, there exists a subsequence, that we denote again ξ^n, of ξ^n such that there exists a permutation σ not depending on n such that for all n and for all $1 \le i \le k$,

$$d_w\big(g^n(\cdot - m^n_{\sigma(i)}), g^*(\cdot - m_i^*)\big) \to 0 \text{ and } \|\sigma Q^n - Q^*\| \to 0.$$

Particularly $g^n(\cdot)\lambda$ weakly tends to $g^*(\cdot - m^*_{\sigma^{-1}(1)})\lambda$. As weak convergence implies pointwise convergence of the characteristic functions and for all $t \in \mathbb{R}$,

$$\int e^{ity} g^n(y - m^n_{\sigma(j)})\lambda(dy) = e^{itm^n_{\sigma(j)}} \int e^{ity} g^n(y)\lambda(dy)$$

then

$$\lim_{n\to\infty} e^{itm^n_{\sigma(j)}} = e^{it(m_j^* - m^*_{\sigma^{-1}(1)})}$$

for all t such that $\int e^{ity} g^*(y)\lambda(dy) \neq 0$. As any characteristic function is uniformly continuous and equal to 1 at 0, there exists $\alpha > 0$ such that $\int e^{ity} g^*(y - m^*_{\sigma^{-1}(1)})d\lambda(y) \neq 0$ for all $|t| < \alpha$. Thus for all $1 \le j \le k$,

$$\lim_{n\to\infty} m^n_{\sigma(j)} = m_j^* - m^*_{\sigma^{-1}(1)}.$$

Since

$$0 = m_1^* < m_2^* < \cdots < m_k^* \text{ and } 0 = m_1^n < m_2^n < \cdots < m_k^n$$

then the permutation σ is equal to the identity permutation. Then Equation (2.47) holds and this implies the first part of Corollary 2.6. In fact we have proved the continuity of

$$\begin{cases} (\{p_l^\xi,\ \xi \in \Xi(\underline{q}), \text{rank}(Q) = k\}, L_1) & \to & (\Delta^k(0), |||) \times (\mathbb{R}, |)^k \times (\mathcal{F}, d_w) \\ p_l^\xi & \mapsto & \xi \end{cases}. \tag{2.48}$$

If moreover $\max_{1\leq j\leq k}\mu_j^* > \frac{1}{2}$ and g^* is uniformly continuous, if

$$\lim_{n\to\infty} D_3(\xi^n,\xi^*)=0$$

then

$$\lim_{n\to\infty} D_1(\xi^n,\xi^*)=0$$

and by continuity of the functional defined in (2.48),

$$\lim_{n\to\infty}\max_{1\leq j\leq k}|\mu_j^n-\mu_j^*|=0$$

and

$$\lim_{n\to\infty}\max_{1\leq j\leq k}|m_j^n-m_j^*|=0$$

so that

$$\lim_{n\to\infty}\max_{1\leq j\leq k}\|g^*(\cdot-m_j^n)-g^*(\cdot-m_j^*)\|_{L_1(\lambda)}=0$$

since g^* is uniformly continuous. Using the following inequality proved in the proof of Corollary 1 in Gassiat and Rousseau [GR16]

$$\begin{aligned}\|D_1(\xi^n,\xi^*)\|_{L_1} \geq &\left(2\max_{1\leq j\leq k}\mu_j^*-1\right)\|g^n-g^*\|_{L_1(\lambda)}\\ &-\max_{1\leq j\leq k}|\mu_j^n-\mu_i^*|-\max_{1\leq j\leq k}\|g^*(\cdot-m_j^n)-g^*(\cdot-m_j^*)\|_{L_1(\lambda)}\end{aligned}$$

we obtain that $\lim_{n\to\infty}\|g^n-g^*\|_{L_1(\lambda)}=0$ which implies the last part of Corollary 2.6.

Proof of Proposition 2.7

As in the proof of Proposition 2.5, many ideas come from Tokdar [Tok06]. We first prove (A1.1) assuming that (B1.1), (B1.2), (B1.3) and (B1.4) are verified with $f_j(\cdot)=g(\cdot-m_j),\ 1\leq j\leq k$. With the same ideas of the proof of Theorem 3.2 in Tokdar [Tok06], for all $\varepsilon>0$ there exists a probability measure $\tilde{P}$ on $\mathbb{R}\times(0,+\infty)$ such that there exists $0<\underline{\sigma}<\bar{\sigma}$ and $a>0$ satisfying

$$\tilde{P}((-a,a]\times(\underline{\sigma},\bar{\sigma}])=1$$

and

$$\int g^*(y-m_i^*)\max_{1\leq j\leq k}\log\frac{g^*(y-m_j^*)}{\phi*\tilde{P}(y-m_j^*)}\lambda(dy)\leq\frac{\varepsilon}{3},$$

using Assumptions (B1.2), (B1.3) and (B1.4).

Let $G=[-a,a]\times[\underline{\sigma},\bar{\sigma}]$. Using the proof of Lemma 3.1 in Tokdar [Tok06] for all $C>\max_{1\leq j\leq k}|m_j^*|+$

$a+\bar{\sigma}$, for all $m_j \in [m_j^*-a, m_j^*+a]$, and for all P such that $P(G) > \frac{\underline{\sigma}}{\bar{\sigma}}$,

$$\begin{aligned}\int_{|y|>C} g^*(y-m_i^*) \max_{1\le j\le k} \log \frac{\phi * \tilde{P}(y-m_j^*)}{\phi * P(y-m_j)} \lambda(dy) \\ \le \int_{|y|>C} g^*(y-m_i) \max_{1\le j\le k} \frac{1}{2}\left(\frac{|y|+|m_j^*|+2a}{\underline{\sigma}}\right)^2 \lambda(dy) < \infty.\end{aligned} \tag{2.49}$$

Using assumption (B1.4) and Equation (2.49), we fix C such that

$$\int_{|y|>C} g^*(y-m_i^*) \max_{1\le j\le k} \log \frac{\phi * \tilde{P}(y-m_j^*)}{\phi * P(y-m_j)} \lambda(dy) \le \frac{\varepsilon}{3}.$$

Let $G_\delta = [-a-\delta, a+\delta] \times [\underline{\sigma}-\delta, \bar{\sigma}+\delta]$, with δ chosen in $(0, \min(\frac{\underline{\sigma}}{2}, \frac{a}{2})]$. Let $h : \mathbb{R} \times (0,+\infty) \to [0,1]$ be a piecewise affine continuous function such that $h(z,\sigma) = 1$ on G and $h(z,\sigma) = 0$ on G_δ^c. Let

$$c = \inf_{\substack{\underline{\sigma}-\delta \le \sigma \le \bar{\sigma}+\delta, \\ |y| \le C, \\ |\theta| \le a + \max_j |m_j^*| + \delta}} \phi_\sigma(y-\theta).$$

By Arzelà-Ascoli theorem there exists $y_1, \ldots, y_I$ such that for all $y \in [-C, C]$ and $1 \le j \le k$, there exists $1 \le i \le I$ such that

$$\sup_{(z,\sigma)\in G_\delta} \left|\phi_\sigma\left(y-m_j^*-z\right) - \phi_\sigma\left(y_i-m_j^*-z\right)\right| < c\delta.$$

Let

$$\begin{aligned}V_\delta = \Big\{ P \ : \ \Big| \int h(z,\sigma)\phi_\sigma(y_i-m_j^*-z)dP(z,\sigma) - \\ \int h(z,\sigma)\phi_\sigma(y_i-m_j^*-z)d\tilde{P}(z,\sigma)\Big| < c\delta \Big\}.\end{aligned}$$

For all $P \in V_\delta$, for all $m_j \in [m_j^* - \frac{c\sigma\delta\sqrt{2}}{\sqrt{\pi}}, m_j^* + \frac{c\sigma\delta\sqrt{2}}{\sqrt{\pi}}]$ and for all $1 \le j \le k$, we get

$$\left| \frac{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)dP(z,\sigma)}{\int h(z,\sigma)\phi_\sigma(y-m_j-z)d\tilde{P}(z,\sigma)} - 1 \right| \le 4\delta$$

thus

$$\begin{aligned}
&\int_{|y|\leq C} g^*(y-m_i^*) \max_{1\leq j\leq k} \log \frac{\phi * \tilde{P}(y-m_j^*)}{\phi * P(y-m_j^*)} \lambda(dy) \\
&\quad \leq \int_{|y|\leq C} g^*(y-m_i^*) \max_{1\leq j\leq k} \log \frac{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)d\tilde{P}(z,\sigma)}{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)dP(z,\sigma)} \lambda(dy) \\
&\quad \leq \frac{4\delta}{1-4\delta}.
\end{aligned}$$

Then for δ small enough, for all $g = \phi * P$ such that $P \in V_\delta \cap \{P : P(G) > \frac{\underline{\sigma}}{\bar{\sigma}}\} = \tilde{V}_\delta$, for all $m_j \in [m_j^* - \frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}}, m_j^* + \frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}}] = M_j^\delta$ and for all $1 \leq i \leq k$,

$$\max_{1\leq i\leq k} \int g^*(y-m_i^*) \max_{1\leq j\leq k} \log \left(\frac{g^*(y-m_j^*)}{g(y-m_j)} \right) dy < \varepsilon, \tag{2.50}$$

moreover,

$$\begin{aligned}
g(y-m_i) &\geq \int_G \phi_\sigma(y-m_i-z)P(dz,d\sigma) \\
&\geq \frac{\underline{\sigma}}{\bar{\sigma}} \phi_{\underline{\sigma}}(\max(|y-m_i-a|, |y-m_i+a|))P(G) \\
&\geq \frac{\underline{\sigma}}{\bar{\sigma}} \phi_{\underline{\sigma}}(\max(|y-m_i-a|, |y-m_i+a|)) \frac{\underline{\sigma}}{\bar{\sigma}} > 0.
\end{aligned} \tag{2.51}$$

Assumption (B1.1) ensures that (A1.1d) holds. Finally for all $\varepsilon > 0$, there exists $\delta > 0$ such that (A1.1) holds with $\Theta_\varepsilon = \{Q : \|Q - Q^*\| < \min(\varepsilon, \underline{q}/2)\} \times M_1^\delta \times \cdots \times M_k^\delta \times \tilde{V}_\delta$ using Equations (2.50) and (2.51).

We now prove (C1.2) thanks to Assumption (D1.6). Let

$$\mathcal{F}_{a,l,u,\underline{m}} = [-\underline{m}, \underline{m}]^k \times \mathcal{F}_{a,l,u},$$

where $\mathcal{F}_{a,l,u} = \mathcal{F}_{a,l,u}^2$ is defined in the proof of Proposition 2.5. Note that for all $(m, \phi * P)$, $(\tilde{m}, \phi * \tilde{P}) \in \mathcal{F}_{a,l,u,\underline{m}}$, for all $1 \leq i \leq k$,

$$\begin{aligned}
&\|\phi * P(\cdot - m_i) - \phi * \tilde{P}(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} \\
&\leq \|\phi * P(\cdot - m_i) - \phi * P(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} + \|\phi * P(\cdot) - \phi * \tilde{P}(\cdot)\|_{L_1(\lambda)}.
\end{aligned}$$

The second term is dealt with in the proof of Proposition 2.5. As to the first part,

$$\|\phi * P(\cdot - m_i) - \phi * P(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} \leq \frac{1}{l}\sqrt{\frac{2}{\pi}}|m_i - \tilde{m}_i|$$

then for all $\kappa > 0$, $a, l, u, \underline{m} > 0$ such that $l < u$,

$$N(3\kappa, \mathcal{F}_{a,l,u,\underline{m}}, d) \leq \left(\frac{2\underline{m}}{l\kappa} + 1\right)^k N(2\kappa, \mathcal{F}_{a,l,u}, \|\cdot\|_{L_1(\lambda)}).$$

For all $\kappa > 0$, let

$$\mathcal{F}^{\kappa}_{a,l,u,\underline{m}} = [-\underline{m}, \underline{m}]^k \times \mathcal{F}^{\kappa}_{a,l,u}.$$

Following the ideas of Lemmas 4.1 and 4.2 in Tokdar [Tok06], there exist c_0, c_1, c_2, c_3 only depending on κ such that

$$\log\left(N(\kappa, \mathcal{F}^{\kappa}_{a,l,u,\underline{m}}), d\right) \leq c_0 \left(c_1 k \log\frac{m}{l} + c_2\frac{a}{l} + c_3 \log\frac{u}{l} + 1\right),$$

so that (D1.6) implies (C1.2) with suitable choices of κ and β.

CHAPTER 3

NONPARAMETRIC HIDDEN MARKOV MODELS WITH FINITE STATE SPACE: POSTERIOR CONCENTRATION RATES

The use of nonparametric hidden Markov models with finite state space is flourishing in practice while few theoretical guarantees are known in this framework. Here, we study asymptotic guarantees for these models in the Bayesian framework. We obtain posterior concentration rates with respect to the L_1-norm on joint marginal densities of consecutive observations in a general theorem. We apply this theorem to two cases and obtain minimax concentration rates up to logarithmic factor. We consider discrete observations with emission distributions distributed from a Dirichlet process and continuous observations with emission distributions distributed from Dirichlet process mixtures of Gaussian distributions.

3.1 Introduction

Hidden Markov models (HMMs) are stochastic processes much used in practice in fields as diverse as genomics, speech recognition, econometrics or climate. A hidden Markov chain is a sequence $(X_t, Y_t)_{t\in\mathbb{N}}$ where the sequence $(X_t)_{t\in\mathbb{N}}$ is a nonobserved Markov chain and the sequence of observations $(Y_t)_{t\in\mathbb{N}}$ is a noisy version of the chain $(X_t)_{t\in\mathbb{N}}$. In this chapter we consider the case where the state space of the underlying Markov chain is finite. In this situation, HMMs are often employed to classify dependent data with respect to the hidden states X_t, $t \in \mathbb{N}$. Their popularity is due to their tractability. Since their introduction in Baum and Petrie [BP66], many algorithms have been developed to infer these models. The books Cappé *et al.* [CMR05], MacDonald and Zucchini [MZ97] and MacDonald and Zucchini [MZ09] give an overview of this family of models.

Parametric HMMs suffer from a lack of robustness so that nonparametric HMMs are used more and more in applications. Indeed two constraints weaken parametric HMMs: the necessary assumption of a bound on the number of states of the Markov chain and the limitations of the parametric modeling of emission distributions (the distributions of an observation Y_t given the hidden states X_t). To deal with these issues, HMMs with an infinite countable number of states for the Markov chain are applied in Beal and Krishnamurthy [BK12a] to gene expression time course clustering, Jochmann [Joc15] to U.S. inflation dynamics and in Fox *et al.* [FJSW09] to segmentation of visual motion capture data. To handle speaker diarization, Fox *et al.* [FSJW11] proposes a model where the number of states of the Markov chain is not bounded and the emission distributions are not restricted to live in a parametric family. HMMs, where the number of states is known but the emission distributions set is not assumed to be parametric, are used in Langrock *et al.* [LKSD15] for whales dive modeling, Yau *et al.* [YPRH11] for genetic copy number variants, Whiting *et al.* [WLM03] for climate state identification, Lefèvre [Lef03] for speech recognition, Gassiat *et al.* [GCR15] for gene expression identification, see also the references herein. This last framework, namely HMMs where the number of states of the Markov chain is known and emission distributions may live in infinite-dimensional sets is the one we consider in this chapter.

The use of nonparametric HMMs is flourishing in practice while few theoretical properties are known. Many theoretical results exist for parametric HMMs particularly for the maximum likelihood estimator, see Cappé *et al.* [CMR05] and references herein for instance, see also de Gunst and Shcherbakova [GS08] for a Bernstein von Mises property of the posterior. In the nonparametric framework, there exist few theoretical guarantees of the asymptotic behavior of estimators or posterior since identifiability for general HMMs with finite state space was still an issue until recently. General identifiability is proved in Gassiat *et al.* [GCR15] when the number of states of the Markov chain is known and in Alexandrovich *et al.* [AHL16] when this number is unknown. Gassiat *et al.* [GCR15] prove that under mild assumptions, the knowledge of the marginal joint density of at least three consecutive observations (Y_t, Y_{t+1}, Y_{t+2}) gives the parameters of the

HMM up to label switching (i.e. the transition matrix of the Markov chain and the emission distributions). Here, we are interested in obtaining asymptotics in the Bayesian framework for the marginal joint density of consecutive observations.

In the Bayesian nonparametric setting, asymptotic analysis typically takes the following two forms: posterior consistency and posterior concentration rates. The posterior is said to be consistent at a parameter θ^* if it concentrates its mass around θ^*, when the observations come from θ^* and the number of observations increases. Posterior consistency is related to the merging of posterior distributions associated to two prior distributions, see Diaconis and Freedman [DF86]. In a nonparametric setup, where it is not feasible to construct a fully subjective prior (on an infinite dimensional space), it is a minimal requirement, see Ghosh and R.V. Ramamoorthi [GR03]. To go further on, one can study the rate at which this concentration occurs, see Ghosal and van der Vaart [GV07a]. Obtaining a minimax posterior concentration rates is a criterion of optimality. In particular, minimax concentration rates lead to minimax Bayesian estimators Ghosal *et al.* [GGV00] and to minimax size of credible regions Hoffmann *et al.* [HRS+15]. The concentration rate analysis also allows a better understanding of the impact of the prior, see Rousseau [Rou15] for a discussion.

In Bayesian HMMs where the number of states of the Markov chain is known, Vernet [Ver15b] provides assumptions leading to posterior consistency for the L_1-norm of the marginal density of consecutive observations. Here, we pursue the study of the asymptotic behavior of the posterior distribution in this framework and with the same topology. Namely, we study posterior concentration rates for nonparametric HMMs with respect to the L_1-norm of the marginal joint density of consecutive observations. We first give a general theorem relating the posterior concentration rate to the prior and the true model (Theorem 3.1). Then we apply the theorem to different setups, where we obtain minimax rates (Section 3.4). To the best of our knowledge, these are the first results on posterior concentration rates in nonparametric HMMs.

Let us mention the few other asymptotic results we know in the framework of nonparametric HMMs. In the nonparametric frequentist framework with a finite and known number of states, De Castro *et al.* [DGLar] offers an oracle inequality for a penalized least-squares estimator of the emission distributions. In the framework of HMMs with an unknown number of states and emission distributions living in a finite-dimensional set, posterior concentration rates are studied in Gassiat and Rousseau [GR14]. Gassiat and Rousseau [GR16] proposes asymptotics for the particular case of translated HMMs with finite state space. Finally, convergence with respect to smoothing distributions is studied in De Castro *et al.* [DGC15].

Chapter 3 is organized as follows. In Section 3.2, we precise the studied model and the notations. In Section 3.3, we state general assumptions under which the posterior concentration rate is derived (Theorem 3.1). We have chosen to define a set of assumptions as close as possible to those typically obtained in density estimation for i.i.d. models, see Ghosal *et al.* [GGV00]. The proof of this theorem is given in Section 3.3.2. All the other proofs are postponed to the appendices.

Theorem 3.1 is applied in Section 3.4 in two cases. In Section 3.4.1, the observations are assumed to be discrete and the prior on emission distributions is based on a Dirichlet process. We obtain a minimax rate which is $1/\sqrt{n}$ up to a power of $\log n$ in Corollary 3.5. In Section 3.4.2, the observations are assumed to be continuous and the emission distributions follow independently Dirichlet process mixtures of Gaussian distribution. Minimax rates of concentration are obtained for Hölder-type functional classes, see Corollary 3.7.

3.2 Bayesian Hidden Markov Models and Notations

We consider observations coming from homogeneous hidden Markov models with finite state space. Hidden Markov chains are discrete time stochastic processes $(X_t, Y_t)_{t\in\mathbb{N}}$ satisfying the following properties. The sequence $(X_t)_{t\in\mathbb{N}}$ is a Markov chain. Conditionally on the hidden chain $(X_t)_{t\in\mathbb{N}}$, the observations Y_t are independent with Y_t only depending on X_t. The states $(X_t)_{t\in\mathbb{N}}$ are latent, they are called the hidden states. The statistician observes the sequence $(Y_t)_{t\leq n}$ where n is an integer. Throughout Chapter 3, for any integer n, an n-uple $(y_1, \dots, y_n)$ is denoted $y_{1:n}$.

We first introduce the notations concerning the Markov chain $(X_t)_{t\in\mathbb{N}}$. For all $t \in \mathbb{N}$, X_t belongs to $\{1, \dots, k\}$, where k is assumed to be known in this chapter. A transition matrix Q and an initial probability distribution μ describe the distribution of the underlying Markov chain

$$X_1 \sim \sum_{1\leq i\leq k} \mu_i \delta_i, \quad X_t|(X_{t-1} = i) \sim \sum_{1\leq j\leq k} Q_{i,j}\delta_j,$$

where δ_i denotes the Dirac measure at i. The set of all initial probability distributions is the $k-1$-simplex $\Delta_k = \{x \in [0,1]^k : \sum_{1\leq i\leq k} x_i = 1\}$. We denote Δ_k^k the set of all transition matrices such that each row of the matrix is an element of Δ_k. In the following we need $\Delta_k(\underline{q}) = \{\mu \in \Delta_k : \mu_i \geq \underline{q}\ \forall i\}$ and $\Delta_k^k(\underline{q}) = \{Q \in \Delta_k^k : Q_{i,j} \geq \underline{q}\ \forall i, j\}$, for $\underline{q} \in (0, 1)$. When Q is in $\Delta_k^k(\underline{q})$, with $\underline{q} \in (0, 1)$, then the uniform mixing coefficients, defined in Rio [Rio00], associated to the corresponding Markov chain are bounded by $\phi(m) \leq (1 - \underline{q})^m$, moreover the corresponding Markov chain is irreducible and positive recurrent.

The observations Y_t are assumed to live in $\mathbb{R}^d$ which is endowed with its Borel sigma field. The distribution of Y_t is assumed to be absolutely continuous with respect to some measure λ on $\mathbb{R}^d$. Conditionally on $(X_t)_{t\in\mathbb{N}}$, Y_t is distributed from a distribution $f_{X_t}\lambda$ depending on the state X_t:

$$Y_t|(X_s)_{s\in\mathbb{N}} \sim Y_t|X_t \sim f_{X_t}\lambda.$$

The distributions $f_i\lambda$, $1 \leq i \leq k$ are called the emission distributions. The set of probability density functions with respect to λ is denoted $\mathcal{F}$. The vector $f = (f_1, \dots, f_k) \in \mathcal{F}^k$ is formed with the k emission density functions.

Then the model is completely described by the parameters μ and $\theta = (Q, f)$ where $\mu \in \Delta^k$ and $\theta = (Q, f) \in \Delta_k^k \times \mathcal{F}^k =: \Theta$. The model can be visualized in Figure 3.1.

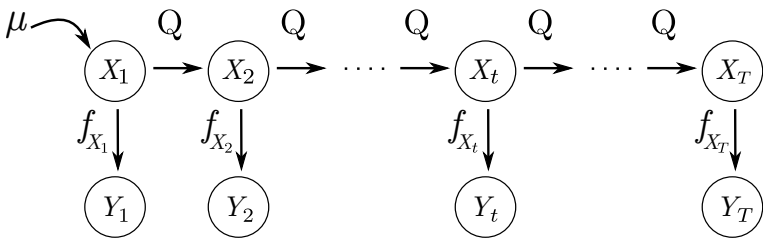


Figure 3.1 – The model

Let $P^{\mu,\theta}$ be the probability distribution of the process $(X_t, Y_t)_{t\in\mathbb{N}}$ under (μ, θ). Then for any $l \in \mathbb{N}$, $\theta \in \Theta$, initial probability μ, and measurable set A of $\{1, \dots, k\}^l \times (\mathbb{R}^d)^l$, note that:

$$P^{\mu,\theta}((X_{1:l}, Y_{1:l}) \in A) = \int \sum_{x_1,\dots,x_l=1}^{k} \mathbb{1}_{(x_1,\dots,x_l,y_1,\dots,y_l)\in A}\, \mu_{x_1} Q_{x_1,x_2} \dots Q_{x_{l-1},x_l} \\ f_{x_1}(y_1) \dots f_{x_l}(y_l) \lambda(dy_1) \dots \lambda(dy_l).$$

Note that when Q is in $\Delta_k^k(\underline{q})$, with $\underline{q}$ positive, there exists a unique stationary initial distribution μ^Q associated with Q. When μ is not specified, the stationary distribution associated with the transition matrix Q is considered in the place of μ. In other words, we define $P^{(Q,f)} := P^{\mu^Q,(Q,f)}$. The joint distribution of l consecutive observations ($(Y_1, \dots Y_l)$ for instance) under the stationary process associated with θ is denoted P_l^θ. Let p_l^θ denote the density of P_l^θ with respect to $\lambda^{\otimes l}$. Then,

$$p_l^\theta(y_1, \dots, y_l) = \sum_{x_1,\dots,x_l=1}^{k} \mu_{x_1}^Q Q_{x_1,x_2} \dots Q_{x_{l-1},x_l} f_{x_1}(y_1) \dots f_{x_l}(y_l), \quad \lambda^{\otimes l} \text{ a.s..}$$

The log-likelihood for a sequence of observations $Y_{1:l}$ under a parameter θ is denoted

$$L_l^\theta := \log\left(p_l^\theta(Y_1, \dots, Y_l)\right).$$

The dependency of L_l^θ with $Y_{1:l}$ is implicit and can be deduced from the context.

Working in the Bayesian framework, we put a prior Π on the set of parameters Θ. We choose a product probability measure $\Pi = \Pi_Q \otimes \Pi_f^{(k)}$ where Π_Q is a probability distribution on Δ_k^k and $\Pi_f^{(k)}$ is a probability distribution on $\mathcal{F}^k$. To a realization θ from Π, we implicitly associate a stationary initial distribution μ^Q. In other words, we generalize Π to a distribution $\tilde{\Pi}$ on $\Delta_k \times \Theta$ such that under Π and conditionally on $\theta = (Q, f)$, $\mu = \mu^Q$. Then using the Bayes' theorem,

the posterior is expressed by

$$\Pi(\theta \in A|Y_{1:n}) = \frac{\int_A p_n^\theta(Y_{1:n})\Pi(d\theta)}{\int_\Theta p_n^\theta(Y_{1:n})\Pi(d\theta)}.$$

We are interested in the asymptotic behaviour of the posterior that is to say when the number n of observations $Y_{1:n}$ tends to infinity. For this purpose, we take a frequentist point of view, assuming that the observations come from the true parameters μ^* and $\theta^* = (Q^*, f^*)$. We suppose that the true initial distribution μ^* is stationary. We also assume that there exists $\underline{q}^* > 0$ such that

$$Q^* \in \Delta_k^k(\underline{q}^*). \tag{3.1}$$

Vernet [Ver15b] shows posterior consistency at θ^* under general assumptions. In this chapter, we consider posterior concentration rates at θ^*. Recall that the posterior is said to concentrate at rate ϵ_n, a sequence decreasing to 0, for the loss $D(\cdot,\cdot)$ if there exists a constant $M > 0$ such that

$$\Pi(\theta : D(\theta, \theta^*) \geq M\epsilon_n|Y_{1:n}) = o_{P^{\theta^*}}(1),$$

where $Z = o_{P^{\theta^*}}(1)$ means that Z converges in probability to 0. We choose to study the concentration of the posterior from the density estimation point of view. We compare two parameters θ and $\tilde{\theta}$ by computing the L_1-distance between the joint densities p_l^θ and $p_l^{\tilde{\theta}}$. For two distributions P_1 and P_2, let p_1 and p_2 be their respective densities with respect to a dominated measure ν. The L_1-metric is defined by

$$\|p_1 - p_2\|_{L_1(\nu)} = \int |p_1 - p_2|\nu$$

and let

$$KL(p_1, p_2) = \int p_1 \log\left(\frac{p_1}{p_2}\right)\nu$$

be the Kullback-Leibler divergence between p_1 and p_2. For an integer $l \geq 1$, we use the pseudo-distance D_l on Θ defined by

$$D_l(\theta, \tilde{\theta}) = \|p_l^\theta - p_l^{\tilde{\theta}}\|_{L_1(\lambda^{\otimes l})}.$$

We study the posterior rate of concentration with respect to this pseudo-distance D_l. On $\mathcal{F}^k$, we use the distance $d(\cdot,\cdot)$ such that for all $(f, \tilde{f}) \in (\mathcal{F}^k)^2$

$$d(f, \tilde{f}) = \max_{1\leq i\leq k} \|f_i - \tilde{f}_i\|_{L_1(\lambda)},$$

on $\mathbb{R}^d$, $d \geq 2$, we use the supremum norm $\|\cdot\|$. For a positive real ϵ, a pseudo distance D defined on a set A, let $N(\epsilon, A, D)$ be the covering number that is to say the minimum number of balls of radius ϵ (in the pseudo-distance D) needed to cover A. Throughout Chapter 3 the notation $\lesssim$ means less or equal up to a multiplicative constant which is not important in the context.

3.3 General Theorem

3.3.1 Assumptions and Main Theorem

In this section, we state the general Theorem 3.1 which gives posterior concentration rates with respect to the D_l pseudo-metric. As in Ghosal *et al.* [GGV00] for instance, we propose a set of conditions which relates the rate $\epsilon_n/\underline{q}_n$ to the prior and the true model. We apply this theorem to the case of discrete observations in Section 3.4.1 and to the case of continuous observations in Section 3.4.2 where minimax rates are achieved. Now, we enumerate the assumptions of Theorem 3.1. Assumptions (A2) and (B2) concern the prior on the emission distributions $\Pi_f^{(k)}$ and the vector of the true emission distributions f^*. Assumptions (C2) and (D2) involve the prior on transition matrices Π_Q and the true transition matrix Q^*.

For two given sequences $\epsilon_n > 0$ and $\tilde{\epsilon}_n$ tending to 0, such that $\tilde{\epsilon}_n \leq \epsilon_n$ for all n, we introduce the sequence u_n of positive numbers such that

$$\begin{aligned} &\text{(i)} \quad u_n = 1, \text{ for all } n \in \mathbb{N}; \text{ if } \tilde{\epsilon}_n \gtrsim n^{-s}, \text{ for some } s < 1/2, \\ &\text{(ii)} \quad u_n = (\log(n))^{3/2}, \text{ for all } n \in \mathbb{N}; \text{ otherwise.} \end{aligned} \tag{3.2}$$

We consider the following assumptions

(A2) there exist a positive constant C_f and a sequence B_n of subsets of $\mathcal{F}^k$ such that

$$\Pi_f^{(k)}(B_n) \gtrsim \exp(-C_f n \tilde{\epsilon}_n^2)$$

and such that for all $f \in B_n$,

$$\int f_i^*(y) \log^2\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy) \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \text{ for all } 1 \leq i, j \leq k, \tag{A2.1}$$

there exist a set $S \subset \mathcal{Y}$ and functions $\tilde{f}_1, \ldots \tilde{f}_k$, which may depend on f, satisfying

$$\int_S \frac{|f_j^*(y) - f_j(y)|^2}{f_j^*(y)} \lambda(dy) \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \text{ for all } 1 \leq j \leq k, \tag{A2.2}$$

$$\int_{S^c} \tilde{f}_j(y) \lambda(dy) \leq \tilde{\epsilon}_n^2, \text{ for all } 1 \leq j \leq k, \tag{A2.3}$$

$$\int_{S^c} f_j^*(y) \lambda(dy) \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \text{ for all } 1 \leq j \leq k, \tag{A2.4}$$

$$\int_S f_i^*(y) \max_{1 \leq j \leq k} \log\left(\frac{\tilde{f}_j(y)}{f_j(y)}\right) \lambda(dy) \leq \tilde{\epsilon}_n^2, \text{ for all } 1 \leq i \leq k, \tag{A2.5}$$

$$\int_S \frac{|f_j^*(y) - \tilde{f}_j(y)|^2}{\tilde{f}_j(y)} \lambda(dy) \leq \tilde{\epsilon}_n^2. \text{ for all } 1 \leq j \leq k, \tag{A2.6}$$

(B2) there exist positive constants C and C' and a sequence $(\mathcal{F}_n)_{n\geq 1}$ of subsets of $\mathcal{F}^k$ such that

$$\Pi_f^{(k)}(\mathcal{F}_n^c) = o(\exp(-Cn\tilde{\epsilon}_n^2))$$

and

$$N\left(\frac{\epsilon_n}{12}, \mathcal{F}_n, d\right) \lesssim \exp\left(C' n \epsilon_n^2\right),$$

(C2) there exists a positive constant C_Q such that $C_Q + C_f + 2C_K < C$ with $C_K = 4 + \log\left(2/\underline{q}^*\right) + 10^4 k^2/\underline{q}^{*5}$,

$$\Pi_Q\left(\left\{Q : \|Q - Q^*\| \leq \frac{\tilde{\epsilon}_n}{\sqrt{u_n}}\right\}\right) \gtrsim \exp(-C_Q n \tilde{\epsilon}_n^2),$$

(D2) there exists a sequence $\underline{q}_n$ of $(0, 1/k]$ such that

$$\Pi_Q\Big(\left(\Delta_k^k(\underline{q}_n)\right)^c\Big) = o(\exp(-Cn\tilde{\epsilon}_n^2)).$$

Under the above assumptions, we prove that the posterior distribution concentrates at the rate $\epsilon_n/\underline{q}_n$.

Theorem 3.1. *Let $\epsilon_n \geq \tilde{\epsilon}_n > 0$ be two sequences tending to 0 such that $n\tilde{\epsilon}_n^2 \to +\infty$. Assume (A2), (B2), (C2) and (D2).*

Then, for all $l \in \mathbb{N}$, there exists a positive constant M such that

$$\Pi\left(\theta : D_l(\theta, \theta^*) \geq M \frac{\epsilon_n}{\underline{q}_n} \middle| Y_{1:n}\right) = o_{\mathbb{P}^{\theta^*}}(1).$$

We now discuss Assumptions (A2) to (D2). We have purposely considered assumptions which are as similar as possible to those considered in the set up of density estimation with i.i.d. observations see Ghosal *et al.* [GGV00]. In particular Assumption (B2) is the typical entropy assumption on a sieve which captures most of the prior mass (see Assumptions (2.2) and (2.3) of Ghosal *et al.* [GGV00]). Assumption (A2) is slightly more involved than the Kullback-Leibler condition (2.4) of Ghosal *et al.* [GGV00] in the case of i.i.d. observations.

This paragraph explains the differences between Assumption (A2) and condition (2.4) of Ghosal *et al.* [GGV00] and can be omitted at first reading. Following Ghosal and van der Vaart [GV07a], posterior rates of convergence are obtained by controlling "Kullback-Leibler" neighbourhoods and constructing some tests. Under (A2) and condition (2.4) of Ghosal *et al.* [GGV00], the Kullback-Leibler neighbourhoods are controlled. Recall that in Ghosal *et al.* [GGV00], the control of

$\mathbb{E}^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$ and $Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$ is obtained by controlling

$$\int f_i^*(y) \log\left(\frac{f_i^*(y)}{f_i(y)}\right) \lambda(dy) \qquad \text{and} \qquad \int f_i^*(y) \log^2\left(\frac{f_i^*(y)}{f_i(y)}\right) \lambda(dy).$$

Here $\mathbb{E}^{\theta^*}(L_n^{\theta^*} - L_n^{\theta}) \lesssim n\tilde{\epsilon}_n^2$ and $Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta}) \lesssim n\tilde{\epsilon}_n^2 \log n$ if f and f^* satisfy Assumptions (A2.1)–(A2.6) and $\|Q - Q^*\| \leq \tilde{\epsilon}_n/\sqrt{u_n}$. The unintuitive part of (A2) comes from the introduction of $\tilde{f}_j$ as an approximation of f_j^*, which may be different from f_j in (A2.3), (A2.5) and (A2.6). Indeed, whithout the introduction of $(\tilde{f}_{j,1\leq j\leq k}, S)$, the HMM structure of the likelihood would lead to a crude upper bound of $\mathbb{E}^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$ of the form

$$n \int f_i^*(y) \max_j \log\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy),$$

see Equations (3.19) and (3.20) in the proof of Lemma 3.2. Since i may be different from j we would loose the local quadratic approximation of the Kullback-Leibler divergence and we would only obtain $n\tilde{\epsilon}_n$ instead of $n\tilde{\epsilon}_n^2$ as an upper bound of $\mathbb{E}^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$. The details of the control of the Kullback-Leibler divergence are given in Appendix 3.5.1.

Not withstanding the technical aspects discussed above, Assumptions (A2.1)–(A2.6) are verified using techniques similar to those used in the case of density estimation with i.i.d. observations to control

$$\int f_i^*(y) \log\left(\frac{f_i^*(y)}{f_i(y)}\right) \lambda(dy) \qquad \text{and} \qquad \int f_i^*(y) \log^2\left(\frac{f_i^*(y)}{f_i(y)}\right) \lambda(dy).$$

For $\Pi_f^{(k)} = (\Pi_f)^{\otimes k}$, and many families of individual prior models Π_f on the f_j's, the rate obtained by bounding $\max_j KL(f_j^*, f_j)$ in the i.i.d. set up will be the same as in our setup. For instance, if $\mathcal{Y} = [0,1]$ and f^* is bounded from below and above, a control of $\|f_j - f_j^*\|_\infty^2 \leq \tilde{\epsilon}_n^2$ or $\|f_j - f_j^*\|_2^2 \leq \tilde{\epsilon}_n^2$ and $f_j > c$ imply (A2.1)–(A2.6). This kind of controls have been derived under (hierarchical) Gaussian process priors or log linear priors as in van der Vaart and van Zanten [VZ09], Arbel *et al.* [AGR13] and Rivoirard and Rousseau [RR12b]. Condition (A2) becomes more involved when $\mathcal{Y}$ is not compact. This case is treated under nonparametric Gaussian mixtures in Section 3.4.2 and in the case where $\mathcal{Y} = \mathbb{N}$ in Section 3.4.1.

Assumption (C2) is checked as soon as Π_Q admits a positive density with respect to the Lebesgue measure which is continuous at Q^* and $\tilde{\epsilon}_n \geq \sqrt{\log(n)/n}$. The rate ϵ_n is often equal to $\tilde{\epsilon}_n$ up to $\log n$, the use of these two different rates is usual and allows more flexibility. Then the rate ϵ_n is only determined by the nonparametric part of the model, i.e. $\Pi_f^{(k)}$ and f^*, as described above.

Following the previous explanation, when Π_Q, $\Pi_f^{(k)}$, f^* and Q^* are fixed, ϵ_n is specified by Assumption (A2) and (B2). This rate ϵ_n is deteriorated via $\underline{q}_n$ which is set through Assumption (D2). The larger $\tilde{\epsilon}_n$ is, that is to say the more difficult the estimation of the nonparametric part (f^* with $\Pi_f^{(k)}$) is, the more stringent Assumption (D2) is. To avoid too small $\underline{q}_n$ which leads to

deteriorated posterior convergence rate $\epsilon_n/\underline{q}_n$, one may choose a prior Π_Q which is supported on $\Delta_k^k(\underline{q})$ for some $0 < \underline{q} \leq \underline{q}^*$. More examples of distribution Π_Q are given in Section 3.4.

In the following section, we give the proof of Theorem 3.1.

3.3.2 Proof of Theorem 3.1

To obtain posterior concentration rates in the framework of HMMs with finite state space, we use the technique of proof of Ghosal and van der Vaart [GV07a]. The key tools of this technique are a control of the prior mass on log-likelihood neighbourhoods of θ^* and the existence of certain tests. We use the tests built in Gassiat and Rousseau [GR14]. The main difficulty of the proof arises from the control of log-likelihood neighbourhoods. These neighbourhoods are controlled thanks to Lemmas 3.2 and 3.3. The proof of these lemmas are based on refinements of results of Douc and Matias [DM01] and Douc *et al.* [DMR04].

In Lemma 3.2, we control $KL(p_n^{\theta^*}, p_n^{\theta})$. Its proof is given in Section 3.5.1.

Lemma 3.2. *Let $0 < \tilde{\epsilon}_n$ be small enough. Assume that $\theta = (Q, f) \in \Theta$, is such that Assumptions (A2.1)–(A2.6) hold with $u_n = 1$ for all $n \in \mathbb{N}$ and*

$$\|Q - Q^*\| \leq \tilde{\epsilon}_n. \tag{3.3}$$

Then there exists $N > 0$ such that for all $n \geq N$,

$$\mathbb{E}^{\theta^*}(L_n^{\theta^*} - L_n^{\theta}) = KL(p_n^{\theta^*}, p_n^{\theta}) \leq C_K n \tilde{\epsilon}_n^2,$$

where C_K is defined in Assumption (C2).

As can be seen in the proof, Assumption (A2.1) can be replaced in Lemma 3.2 by the following weaker assumption:

$$\int_{S^c} f_i^*(y) \max_{1 \leq j \leq k} \log\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy) \leq 2\tilde{\epsilon}_n^2, \text{ for all } 1 \leq i \leq k, \tag{3.4}$$

which is implied by Assumptions (A2.1) and (A2.4). Note, however, that Assumption (A2.1) is used in the control of the variance $Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$ in Lemma 3.3.

Lemma 3.3 gives a control of $Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$. It is proved in Appendix 3.5.2.

Lemma 3.3. *Let $0 < \epsilon_n$ be small enough and u_n be defined by Equation (3.2) Assume that $\theta = (Q, f) \in \Theta$, is such that Assumptions (A2.1), (A2.2) and (A2.4) hold and*

$$\|Q - Q^*\| \leq \frac{\tilde{\epsilon}_n}{\sqrt{u_n}}. \tag{3.5}$$

Then there exists a positive constant C_{KL^2} such that for all $\alpha \in (0,1)$ and $n \in \mathbb{N}$

$$Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta}) = \mathbb{E}^{\theta^*}\left[\left(\log\frac{p^{\theta^*}(Y_{1:n})}{p^{\theta}(Y_{1:n})} - \mathbb{E}^{\theta^*}\left(\log\frac{p^{\theta^*}(Y_{1:n})}{p^{\theta}(Y_{1:n})}\right)\right)^2\right] \leq C_{KL^2}\frac{n}{\alpha}\left(\frac{\tilde{\epsilon}_n}{\sqrt{u_n}}\right)^{2-\alpha}.$$

We now give the proof of Theorem 3.1.

Proof of Theorem 3.1. This proof follows the lines of the proof of Theorem 1 of Ghosal and van der Vaart [GV07a] with two variants. These differences come from the tests (see Equations (3.6) and (3.7)) and the control of the Kullback-Leibler neighbourhoods (Equation (3.10)).

Using the tests built in the proof of Theorem 4 in Gassiat and Rousseau [GR14], for all $M > 0$, there exists $\psi_n \in [0,1]$ such that

$$\mathbb{E}^{\theta^*}(\psi_n) \leq N\left(\frac{\epsilon_n}{12}, \Delta_k^k(\underline{q}_n) \times \mathcal{F}_n, D_l\right)\exp\left(\frac{-n\epsilon_n^2 \underline{q}^{*2} k^4 M^2}{128 l}\right) \tag{3.6}$$

and

$$\sup_{\substack{\theta \in \Delta_k^k(\underline{q}_n)\times\mathcal{F}_n \\ D_l(\theta,\theta^*) \geq M\epsilon_n/\underline{q}_n}} \mathbb{E}^{\theta}(1-\psi_n) \leq \exp\left(-\frac{n\epsilon_n^2 k^2 M^2}{128 l}\right). \tag{3.7}$$

Since

$$D_l(\theta,\tilde{\theta}) \leq \sum_{1\leq i\leq k} |\mu_i^Q - \mu_i^{\tilde{Q}}| + k(l-1)\max_{1\leq i,j\leq k}|Q_{i,j} - \tilde{Q}_{i,j}| + l d(f,\tilde{f}),$$

we obtain

$$N\left(\frac{\epsilon_n}{12}, \Delta_k^k(\underline{q}_n) \times \mathcal{F}_n, D_l\right) \leq \left(\frac{24lk(k-1)}{\epsilon_n}\right)^{k(l-1)} N\left(\frac{\epsilon_n}{24l}, \mathcal{F}_n, d\right) \tag{3.8}$$

which leads to

$$\mathbb{E}^{\theta^*}(\psi_n) \lesssim \exp(-n\epsilon_n^2(M^2\tilde{C} - C')) \tag{3.9}$$

for some constant $\tilde{C}$, using Assumption (B2). We replace Equation (8.4) of Ghosal and van der Vaart [GV07a] by Equation (3.9). Equation (8.5) of Ghosal and van der Vaart [GV07a] is replaced by Equation (3.7).

Let α_n be a sequence tending to 0, to be specified later. We define

$$\mathcal{B}_n(\theta^*) := \left\{\theta : KL(p_n^{\theta^*}, p_n^{\theta}) \leq C_K n\tilde{\epsilon}_n^2,\ Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta}) \leq \frac{C_{KL^2} n}{\alpha_n}\left(\frac{\tilde{\epsilon}_n}{\sqrt{u_n}}\right)^{2-\alpha_n}\right\} \tag{3.10}$$

in the place of $B_n(\theta^*, \bar{\epsilon}_n, 2)$ in the notation of Ghosal and van der Vaart [GV07a], setting $\bar{\epsilon}_n = \sqrt{C_K}\tilde{\epsilon}_n$. Using Assumptions (A2) and (C2), and Lemmas 3.2 and 3.3

$$\Pi\left(\mathcal{B}_n\left(\theta^*\right)\right) \gtrsim \exp(-(C_Q + C_f)n\tilde{\epsilon}_n^2) \gtrsim \exp(-Cn\tilde{\epsilon}_n^2).$$

By choosing

(i) $\alpha_n = 1/\log(n)$, for all $n \in \mathbb{N}$; if $\tilde{\epsilon}_n \gtrsim n^{-s}$ for some $s < 1/2$,

(ii) $\alpha_n = \log(\log(n))/\log(n)$, for all $n \in \mathbb{N}$; otherwise,

and following the lines of the proof of Lemma 10 of Ghosal and van der Vaart [GV07a],

$$P^{\theta^*}\left(\mathcal{D}_n < \Pi(B_n(\theta^*))\exp(-2C_K n\tilde{\epsilon}_n^2)\right) = O\left(\frac{n(\tilde{\epsilon}_n/\sqrt{u_n})^{2-\alpha_n}}{\alpha_n(n\tilde{\epsilon}_n^2)^2}\right) = o(1), \tag{3.11}$$

with $\mathcal{D}_n = \int_{\mathcal{B}_n(\theta^*)} p_n^{\theta}(Y_{1:n})/p_n^{\theta^*}(Y_{1:n})\Pi(d\theta)$. Following the lines of the proof of Theorem 1 of Ghosal and van der Vaart [GV07a] with the above modifications, we obtain,

$$\mathbb{E}^{\theta^*}\left(\Pi\left(\theta \in \Delta_k^k(\underline{q}_n) \times \mathcal{F}_n : D_l(\theta, \theta^*) \geq M\frac{\epsilon_n}{\underline{q}_n}\middle| Y_{1:n}\right)\right) = o(1) \tag{3.12}$$

for M large enough.

This concludes the proof since

$$\mathbb{E}^{\theta^*}\left(\Pi((\Delta_k^k(\underline{q}_n) \times \mathcal{F}_n)^c | Y_{1:n})\right) = o(1)$$

using Equation (3.11) and Lemma 1 of Ghosal and van der Vaart [GV07a] with

$$\Pi((\Delta_k^k(\underline{q}_n) \times \mathcal{F}_n)^c) = o(\exp(-2n\bar{\epsilon}_n^2)\Pi\left(\mathcal{B}_n\left(\theta^*\right)\right)$$

as soon as $C > 2C_K + C_Q + C_f$ (using Assumptions (B2) and (D2)). □

3.4 Applications of the main theorem to different models and prior distributions

In this section, we apply Theorem 3.1 to different priors and different classes of emission density functions. In all examples treated in Section 3.4, the prior on emission distributions is chosen to be a product of a distribution Π_f on $\mathcal{F}$:

$$\Pi_f^{(k)} = (\Pi_f)^{\otimes k}. \tag{3.13}$$

However, Theorem 3.1 can also be applied to other priors such as priors restricted to translated emission density functions, the translation HMM is described in Equation (3.15).

In Section 3.4.1, we consider discrete observations, i.e. $\mathcal{Y} = \mathbb{N}$. We assume that the prior Π_f on each emission distributions is a Dirichlet process. We compute the rate ϵ_n obtained with this prior when the true emission distributions have an exponential decay. In Section 3.4.2, the observations are assumed to live in $\mathbb{R}$ and the emission distributions are supposed to be absolutely continuous with respect to the Lebesgue measure. We consider a Dirichlet process mixture of Gaussian distributions as a prior Π_f on each emission density functions. We compute the rate ϵ_n obtained with this prior when the emission density functions belong to functional classes of β-Hölder types.

We always assume that

(Q2.0) Π_Q is absolutely continuous with respect to the Lebesgue measure on Δ_k^k with density π_Q, $\pi_Q(Q^*) > 0$ and $\pi_Q(Q) = \pi_q(Q_{1,\cdot}) \dots \pi_q(Q_{k,\cdot})$, for all $Q \in \Delta_k^k$ and where $Q_{i,\cdot}$ denotes the i-th row of Q.

In Sections 3.4.1 and 3.4.2, we consider three different priors Π_Q on transition matrices which corresponds to three different decays of π_Q near the boundary of Δ_k^k:

(Q2.1) exponential tail:

$$\pi_q(u_1, \dots, u_k) \lesssim \exp(-\alpha_1/u_1) \dots \exp(-\alpha_k/u_k),$$

for all $u \in \Delta_k$, for some positive constants α_i, $1 \leq i \leq k$,

(Q2.2) exponential of exponential tail:

$$\pi_q(u_1, \dots, u_k) \lesssim \exp(-\beta_1 \exp(u_1^{-\alpha_1})) \dots \exp(-\beta_k \exp(u_k^{-\alpha_k})),$$

for all $u \in \Delta_k$, for some positive constants α_i and β_i, $1 \leq i \leq k$,

(Q2.3) truncated distribution:

$$\Pi_Q(\Delta_k^k(\underline{q})) = 1,$$

for some positive $\underline{q}$.

Note that Assumption (Q2.3) implies Assumption (Q2.2) which implies (Q2.1). In Gassiat and Rousseau [GR14], together with priors of type (Q2.1), more general priors are also considered, since they assume

$$\pi_q(u_1, \dots, u_k) \lesssim u_1^{\alpha_1 - 1} \dots u_k^{\alpha_k - 1},$$

for all $u \in \Delta_k$, for some positive constants α_i, $1 \leq i \leq k$. Gassiat and Rousseau [GR14] show that $\underline{q}_n$, in Assumption (D2) and Theorem 3.1, is equal to a power of $1/n$ when the emission

distributions belong to a parametric family. We do not consider this type of priors since they lead to deteriorated rates $\epsilon_n/\underline{q}_n$. Under (Q2.1), Gassiat and Rousseau [GR14] obtain $\underline{q}_n$ equal to a power of $1/(\log n)$ when the emission distributions belong to a parametric family. We obtain the same rate $\underline{q}_n$ in the case of discrete observations and emission distributions with exponential decay (more generally, it would be the case as soon as $\epsilon_n = n^{-1/2}\log(n)^t$ for some positive t). However, in the case of emission distributions absolutely continuous with respect to Lebesgue measure (Section 3.4.2), Assumption (D2) leads to a rate $\underline{q}_n$ at least polynomial in $1/n$ with priors satisfying (Q2.1). While priors verifying (Q2.2) or (Q2.3) lead to a rate $\underline{q}_n$ equal to a power of $1/\log n$ as soon as ϵ_n is a power of n; and thus do not deteriorate the posterior concentration rate (up to $\log(n)$). Note that Assumptions (Q2.0) and (Q2.3) are compatible if and only if $\underline{q} \leq \underline{q}^*$. Thus, the use of a prior verifying (Q2.3) requires a knowledge of a lower bound $\underline{q}^*$ of $\min_{1\leq i,j\leq k} Q^*_{i,j}$.

In Vernet [Ver15b], posterior consistency is derived under Assumption (Q2.3) while weaker conditions on f^* and $\Pi_f^{(k)}$ (see Assumptions (A0), (A1) and (A2) in Vernet [Ver15b]) compared to (A2), (B2) and (C2). Here, we manage to obtain posterior concentration rates under Assumptions (Q2.1) and (Q2.2) which are weaker than Assumption (Q2.3) because stronger conditions on $\Pi_f^{(k)}$ and f^* are assumed.

3.4.1 Discrete Observations

In this section, we apply Theorem 3.1 to the case of discrete observations so that λ is the count measure on $\mathbb{N}$. HMMs with discrete observations are used in different applications, as in Borchers *et al.* [BZH+13] for animal abundance estimation, in Gassiat *et al.* [GCR15] for gene expression identification, or in Linderman *et al.* [LJWC14] for neural representation of spatial navigation, to cite a few.

In the framework of discrete distribution estimation with i.i.d. observations, Han *et al.* [HJW14] have proved that no rates can be obtained with the L_1 loss without constraint on the considered distributions. Moreover, they obtain a minimax rate proportional to $1/\log n$ over the set $\{f \in \mathcal{F} : \sum_{i\in\mathbb{N}} -f(i)\log(f(i)) \leq C\}$. Rates of convergence are more widely studied in the case of the L_2 norm, for instance with monotony constraint in Jankowski and Wellner [JW09], with log-concave constraint in Balabdaoui *et al.* [BJRP13], with convex constraint in Durot *et al.* [DHKR13] and with envelope constraint in Boucheron and Gassiat [BG09].

In the nonparametric Bayesian framework, the Dirichlet process is a very popular prior. Here, we consider a Dirichlet process $DP(G)$ on the emission distributions, with G some finite positive measure on $\mathbb{N}$:

$$\Pi_f = DP(G).$$

Canale and Dunson [CD11] propose other priors for discrete observations based on discretization of continuous mixtures of kernels and gives an overview of the priors used in the case of discrete

observations.

We first verify Assumptions (A2), (B2) and (C2). As it can be seen from the proof (see Lemma 3.14 in Appendix 3.5.3), we need a heavy tail condition on G:

(P2) there exists positive constants $a \leq A$ and $\alpha \geq 2$ such that for all $1 \leq j \leq k$ and for all $l \in \mathbb{N}$

$$al^{-\alpha} \leq G(l) \leq Al^{-\alpha}.$$

Here, we consider the following class of discrete distributions which is based on an envelope constraint:

$$\mathcal{D}(m, c, K) = \Big\{ f \in \mathcal{F} : f(l) \leq d\exp(-cl^m),\ \sum_{l \leq N} \frac{-\log(f(l))}{l} \lesssim N^K, \text{ for all } N \text{ large enough} \Big\}, \quad (3.14)$$

where c, K and m are positive constants. We also consider the following assumption linking the tails of the true emission distributions:

(I.2) there exists $\delta > 0$ such that for all N large enough and all $1 \leq i, j \leq k$,

$$\sum_{l \geq N} f_i^*(l) \log^2\left(f_j^*(l)\right) \lesssim \exp(-N^m(c-\delta)).$$

Under these assumptions, we obtain the following rates $\tilde{\epsilon}_n$ and ϵ_n:

Theorem 3.4. *Assume there exist positive constants c, K and m such that for all $1 \leq j \leq k$, $f_j^* \in \mathcal{D}(m, c, K)$ and that Assumptions (Q2.0), (I.2) and (P2) hold.*

Then Assumptions (A2), (B2) and (C2) hold with

$$\tilde{\epsilon}_n = \frac{1}{\sqrt{n}}(\log n)^{t_0} \quad \text{and} \quad \epsilon_n = \frac{1}{\sqrt{n}}(\log n)^{t},$$

where $t > 4t_0$ and $t_0 \geq \max(1/m + 1, K/m)/2$.

Theorem 3.4 leads to the following posterior concentration rates $(\epsilon_n/\underline{q}_n)$ which are minimax (up to $\log n$):

Corollary 3.5. *Assume there exist positive constants c, K and m such that for all $1 \leq j \leq k$, $f_j^* \in \mathcal{D}(m, c, K)$ and that Assumptions (Q2.0), (I.2) and (P2) hold. Moreover suppose that Π_Q satisfies*

- *(Q2.1), then the posterior concentrates with rate $\frac{1}{\sqrt{n}}(\log n)^{t+2t_0}$;*
- *(Q2.2), then the posterior concentrates with rate $\frac{1}{\sqrt{n}}(\log n)^{t}$;*

- *(Q2.3), then the posterior concentrates with rate* $\frac{1}{\sqrt{n}}(\log n)^t$*;*

with $t > 4t_0$ *and* $t_0 \geq 1/2 \max(1/m + 1, K/m)$.

3.4.2 Dirichlet Process Mixtures of Gaussian Distributions–Adaptivity to Hölder Function Classes

Dirichlet process mixtures of Gaussian distributions are commonly used to model densities on $\mathbb{R}$ or $\mathbb{R}^d$. In particular, there exist efficient algorithms to sample from the posterior distribution in the i.i.d. framework. In the translation HMM:

$$Y_t = m_{X_t} + \epsilon_t, \tag{3.15}$$

where $\epsilon_t \overset{\text{i.i.d.}}{\sim} g\lambda$, $m_j \in \mathbb{R}$ and X_t is a Markov chain with translation matrix Q; Yau *et al.* [YPRH11] use a Dirichlet process mixtures of Gaussian distributions on g. In the context of i.i.d. observations posterior concentration rates have been derived with such prior models in Ghosal and van der Vaart [GV07b], Kruijer *et al.* [KRV10] and Shen *et al.* [STG13]. In the framework of HMMs, we propose to apply Theorem 3.1 when Π_f is a Dirichlet process mixture of Gaussian distributions.

We assume that the reference measure λ is the Lebesgue measure on $\mathbb{R}$. We also assume that the prior on $\mathcal{F}^k$ is a product of Dirichlet process mixture of Gaussian distributions:

$$\begin{aligned} &Y_t | X_t = j \sim f_j, \quad f_j = \int \phi_\sigma(\cdot - \mu) dP_j(\mu), \\ &P_j \overset{\text{i.i.d.}}{\sim} DP(G), \text{ for all } 1 \leq j \leq k, \quad \sigma \sim \pi_\sigma \lambda, \end{aligned}$$

where ϕ_σ is the Gaussian density function with variance σ^2 and mean zero, $DP(G)$ is the Dirichlet process with finite positive base measure G and π_σ is a distribution on $\mathbb{R}$.

We define the same functional classes as in Kruijer *et al.* [KRV10]:

$$\begin{aligned} \mathcal{P}(\beta, \bar{g}, \gamma) := \Big\{ f \in \mathcal{F} : \ &\log f \text{ is locally } \beta\text{-Hölder with derivatives } g^m = (\log f)^{(m)} \\ &\text{and } |g^{\lfloor \beta \rfloor}(y) - g^{\lfloor \beta \rfloor}(x)| \leq r! \bar{g}(y) |y - x|^{\beta - \lfloor \beta \rfloor}, \text{ as soon as } |x - y| \leq \gamma \Big\} \end{aligned} \tag{3.16}$$

where $\beta > 0$, $\bar{g}$ is a polynomial function, $\gamma > 0$ and $\lfloor \beta \rfloor$ is the largest integer smaller than β. We also consider the following tail conditions:

(T2.1) there exist positive constants M_0, τ_0, γ_0 such that for all $1 \leq i \leq k$ and all $y \in \mathbb{R}$

$$f_i^*(y) \leq M_0 \exp(-\tau_0 |y|^{\gamma_0}),$$

(T2.2) for all $1 \leq i, j \leq k$ there exist constants $T_{i,j}$, $M_{i,j}$, $\tau_{i,j}$, $\gamma_{i,j} < \gamma_0$ such that

$$f_i^*(y) \leq f_j^*(y) M_{i,j} \exp(\tau_{i,j}|y|^{\gamma_{i,j}}), \quad |y| \geq T_{i,j},$$

(T2.3) for all $1 \leq i \leq k$, f_i^* is positive and there exist $c_i > 0$, $y_i^m < y_i^M$ such that f_i^* is nondecreasing on $(-\infty, y_i^m)$, nonincreasing on $(y_i^M, +\infty)$ and $f_i^*(y) \geq c_i$ for $y \in (y_i^m, y_i^M)$.

Assumptions (T2.1) and (T2.3) are the same tail assumptions as those used in Kruijer *et al.* [KRV10]. The new Assumption (T2.2) links the tail of each emission distributions.

We now describe the assumptions concerning the prior on the emission distributions:

(G2.1) $G([-y, y]^c) \lesssim \exp(-C_1 y^{a_1})$ for all sufficiently large $y > 0$, for some positive constant a_1,

(S2.1) $\Pi_\sigma(\sigma \leq x) \lesssim \exp(-C_2 x^{-a_2})$ for all sufficiently small $x > 0$, for some positive constant a_2,

(S2.2) $\Pi_\sigma(\sigma > x) \lesssim x^{-a_3}$ for all sufficiently large $x > 0$, for some positive constant a_3,

(S2.3) there exists $a_6 \leq 1$ such that

$$\Pi_\sigma(x \leq \sigma \leq x(1+s)) \gtrsim x^{-a_4} s^{a_5} \exp(-C_3 x^{-a_6})$$

for all $s \in (0, 1)$, sufficiently small $x > 0$, for some positive constants a_4, a_5 and a_6.

The gamma and Gaussian distributions satisfy Assumption (G2.1). The inverse gamma distribution verifies (S2.1), (S2.2) and (S2.3).

Theorem 3.6. *Assume that there exist β, $\bar{g}$ and γ such that for all $1 \leq j \leq k$, $f_j^* \in \mathcal{P}(\beta, \bar{g}, \gamma)$ and Assumptions (Q2.0), (T2.1)–(T2.3), (G2.1) and (S2.1)–(S2.3) hold.*

Then Assumptions (A2), (B2) and (C2) hold with

$$\tilde{\epsilon}_n = n^{-\frac{\beta}{2\beta+1}} \log(n)^{t_0} \quad \text{and} \quad \epsilon_n = n^{-\frac{\beta}{2\beta+1}} \log(n)^{t}, \tag{3.17}$$

where $t > t_0 \geq (2 + 2/\gamma_0 + 1/\beta)/(1/\beta + 2)$.

The proof of Theorem 3.6 is given in Appendix 3.5.4. Using Theorem 3.1 and 3.6, we directly deduce posterior rate of convergence under the Assumptions of Theorem 3.6 and the different types of priors Π_Q.

Corollary 3.7. *Assume that there exist β, $\bar{g}$ and γ such that for all $1 \leq j \leq k$, $f_j^* \in \mathcal{P}(\beta, \bar{g}, \gamma)$ and that Assumptions (Q2.0), (T2.1)–(T2.3), (G2.1) and (S2.1)–(S2.3) hold. Moreover suppose that Π_Q satisfies*

- *(Q2.1), then the posterior concentrates with rate $n^{\frac{-\beta+1}{2\beta+1}} (\log n)^{3t}$;*

- *(Q2.2), then the posterior concentrates with rate* $n^{-\frac{\beta}{2\beta+1}}(\log n)^{t+1/(\max_{1\leq i\leq k}\alpha_i)}$;
- *(Q2.3), then the posterior concentrates with rate* $n^{-\frac{\beta}{2\beta+1}}(\log n)^{t}$;

with $t > (2 + 2/\gamma_0 + 1/\beta)/(1/\beta + 2)$.

The minimax rate, with respect to D_l in the HMM framework for emission density functions belonging to functional classes of β-Hölder type, is larger than $n^{-\beta/(2\beta+1)}$. Indeed with a hidden Markov chain (X_t, Y_t) distributed from a parameter $\theta = (Q, f)$ such that $Q_{i,j} = 1/k$ and $f_i = f_1$ for all $1 \leq i, j \leq k$, the observations (Y_t) are i.i.d. from $f_1\lambda$. Thus, priors satisfying (Q2.0), (T2.1)–(T2.3), (G2.1), (S2.1)–(S2.3) and (Q2.2) lead to minimax rates (up to $\log n$). As these priors do not depend on the regularity of the functional class considered, they ensure adaptive Bayesian density estimation in the framework of HMMs.

Acknowledgements

I want to thank Elisabeth Gassiat and Judith Rousseau for their valuable comments. This work was partly supported by the grants ANR Banhdits and Calibration.

3.5 Proofs

3.5.1 Proof of Lemma 3.2 : control of the Kullback Leibler divergence between θ^* and θ

First denote $\epsilon = \tilde{\epsilon}_n$. Using Assumptions (3.1) and (3.3), there exists $\underline{q} > 0$ such that for all $1 \leq i, j \leq k$, $Q^*_{i,j} \geq \underline{q}$ and $Q_{i,j} \geq \underline{q}$, more precisely, $\underline{q}$ can be chosen equal to $\underline{q}^*/2$ as soon as n is large enough. Let $q_t^{\theta,Y_{1:t-1}}$ be the conditional density function of Y_t given $Y_{1:t-1}$ with respect to λ:

$$q_t^{\theta,Y_{1:t-1}} = \sum_{i=1}^{k} f_i(\cdot) Q_{t,i}^{\theta,Y_{1:t-1}},$$

where $Q_{t,i}^{\mu,\theta,Y_{1:t-1}} = \mathbb{P}^\theta(X_t = i | Y_{1:t-1}, X_1 \sim \mu)$, where $t \geq 1$, $1 \leq i \leq k$, $\mu \in \Delta_k$. When μ is not specified ($Q_{t,i}^{\theta,Y_{1:t-1}}$), the stationary initial probability distribution is considered:

$$Q_{t,i}^{\theta,Y_{1:t-1}} = Q_{t,i}^{\mu^Q,\theta,Y_{1:t-1}}.$$

Note that, when $\mu \in \Delta_k(\underline{q})$, $\tilde{\theta} \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$,

$$Q_{t,i}^{\mu,\tilde{\theta},Y_{1:t-1}} = \frac{\sum_{j=1}^k Q_{t-1,j}^{\mu,\tilde{\theta},Y_{1:t-2}} Q_{j,i} f_j(Y_{t-1})}{\sum_{\iota=1}^k Q_{t-1,\iota}^{\mu,\tilde{\theta},Y_{1:t-2}} f_\iota(Y_{t-1})} \geq \underline{q}, \tag{3.18}$$

for all $1 \leq i \leq k, t \geq 1$.

$$\begin{aligned}
&KL(p_n^{\theta^*}, p_n^{\theta}) \\
&= \mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right) \\
&= \mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int_S q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right) \\
&\quad + \mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int_{S^c} q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right).
\end{aligned} \tag{3.19}$$

Using Equation (3.18), for all $1 \leq r \leq k$,

$$\begin{aligned}
\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)} = \frac{\sum_{i=1}^k f_i^*(y) Q_{t,i}^{\theta^*, Y_{1:t-1}}}{\sum_{i=1}^k f_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}} &\leq \frac{k \max_{1 \leq j \leq k} f_j^*(y)}{\underline{q} f_r(y)} \\
&\leq \max_{1 \leq j \leq k} \frac{k f_j^*(y)}{\underline{q} f_j(y)}
\end{aligned} \tag{3.20}$$

then Assumptions (3.4) and (A2.4) lead to

$$\mathbb{E}^{\theta^*}\left(\int_{S^c} q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right) \leq \left(\log \frac{1}{\underline{q}} + 2\right) \epsilon^2. \tag{3.21}$$

We now control the expectation of the third line of Equation (3.19)

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int_S q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right) \\
&\leq \mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int_S q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{q_t^{\theta^*, Y_{1:t-1}}(y)}{\sum_{i=1}^k \tilde{f}_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}}\right) \lambda(dy)\right) \\
&\quad + \mathbb{E}^{\theta^*}\left(\sum_{t=1}^{n} \int_S q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{\sum_{i=1}^k \tilde{f}_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right).
\end{aligned} \tag{3.22}$$

We control the expectation of the third line of Equation (3.22), using

$$\frac{\sum_{i=1}^k \tilde{f}_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}}{q_t^{\theta, Y_{1:t-1}}(y)} = \frac{\sum_{i=1}^k \tilde{f}_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}}{\sum_{i=1}^k f_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}} \leq \max_{1 \leq i \leq k} \frac{\tilde{f}_i(y)}{f_i(y)};$$

by Lemma 3.8, and Assumption (A2.5), to obtain

$$\mathbb{E}^{\theta^*}\left(\int_S q_t^{\theta^*, Y_{1:t-1}}(y) \log\left(\frac{\sum_{i=1}^k \tilde{f}_i(y) Q_{t,i}^{\theta, Y_{1:t-1}}}{q_t^{\theta, Y_{1:t-1}}(y)}\right) \lambda(dy)\right) \leq \epsilon^2. \tag{3.23}$$

We bound the expectation of the second line of Equation (3.22), using the inequality recalled at the top of page 1234 of Kruijer *et al.* [KRV10],

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\int_S q_t^{\theta^*,Y_{1:t-1}}(y)\log\left(\frac{q_t^{\theta^*,Y_{1:t-1}}(y)}{\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}}\right)\lambda(dy)\right)\\
&\leq \mathbb{E}^{\theta^*}\left(\int_S \frac{\left(q_t^{\theta^*,Y_{1:t-1}}(y)-\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}\right)^2}{\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}}\lambda(dy)\right)\\
&\quad+\mathbb{E}^{\theta^*}\left(\int_{S^c}\left(\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}\right)-q_t^{\theta^*,Y_{1:t-1}}(y)\lambda(dy)\right).
\end{aligned} \tag{3.24}$$

The expectation of the second line of Equation (3.24) is controlled as follows

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\int_S \frac{\left(q_t^{\theta^*,Y_{1:t-1}}(y)-\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}\right)^2}{\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}}\lambda(dy)\right)\\
&\leq 2\mathbb{E}^{\theta^*}\left(\frac{k\max_{1\leq i\leq k}\left(Q_{t,i}^{\theta^*,Y_{1:t-1}}-Q_{t,i}^{\theta,Y_{1:t-1}}\right)^2}{\underline{q}}\right)+2\mathbb{E}^{\theta^*}\left(\int_S \max_{1\leq i\leq k}\frac{\left(f_i^*(y)-\tilde{f}_i(y)\right)^2}{\underline{q}\tilde{f}_i(y)}\lambda(dy)\right)\\
&\leq \left(\frac{16k(1+2k)}{\underline{q}^4}+1\right)\frac{2}{\underline{q}}\epsilon^2+\frac{16k\rho^{2(t-1)}}{\underline{q}},
\end{aligned} \tag{3.25}$$

where $\rho=(1-k\underline{q})/(1-(k-1)\underline{q})\leq 1-\underline{q}$, using Lemma 3.10 and then Assumption (A2.6) and Lemma 3.10. The expectation of the third line of Equation (3.24) is controlled thanks to Assumption (A2.3):

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\int_{S^c}\left(\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}\right)-q_t^{\theta^*,Y_{1:t-1}}(y)\lambda(dy)\right)\\
&\leq \mathbb{E}^{\theta^*}\left(\int_{S^c}\sum_{i=1}^k \tilde{f}_i(y)Q_{t,i}^{\theta,Y_{1:t-1}}\lambda(dy)\right)\\
&\leq \max_{1\leq i\leq k}\int_{S^c}\tilde{f}_i(y)\lambda(dy)\leq\epsilon^2.
\end{aligned} \tag{3.26}$$

We conclude the proof by combining Equations (3.19), (3.21), (3.22), (3.23), (3.24), (3.25) and (3.26). □

Lemma 3.8. *For all $a_i,b_i,c_i,d_i\geq 0$, $1\leq i\leq k$,*

$$\frac{\sum_{1\leq i\leq k}a_ib_i}{\sum_{1\leq j\leq k}c_jd_j}=\frac{\sum_{1\leq i\leq k}a_i/c_i*b_i/d_i*c_id_i}{\sum_{1\leq j\leq k}c_jd_j}\leq\max_{1\leq i\leq k}\frac{a_i}{c_i}\max_{1\leq j\leq k}\frac{b_j}{d_j}.$$

3.5.2 Proof of Lemma 3.3 with technical lemmas: control of $Var^{\theta^*}(L_n^{\theta^*} - L_n^{\theta})$

3.5.2.1 Proof of Lemma 3.3

First denote $\epsilon = \tilde{\epsilon}_n/\sqrt{u_n}$. Using Assumptions (3.1) and (3.5), there exists $\underline{q} > 0$ such that for all $1 \leq i, j \leq k$, $Q^*_{i,j} \geq \underline{q}$ and $Q_{i,j} \geq \underline{q}$, more precisely, $\underline{q}$ can be chosen equal to $\underline{q}^*/2$ as soon as n is large enough. Let Var be the variance of $L_n^{\theta^*} - L_n^{\theta}$:

$$Var := \mathbb{E}^{\theta^*}\left[\left(\log\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta}(Y_{1:n})} - \mathbb{E}^{\theta^*}\left(\log\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta}(Y_{1:n})}\right)\right)^2\right].$$

Denoting $Z_t = \log\left(\frac{P^{\theta^*}(Y_t|Y_{1:t-1})}{P^{\theta}(Y_t|Y_{1:t-1})}\right)$, then

$$Var = \mathbb{E}^{\theta^*}\left(\left(\sum_{t=1}^n Z_t - \mathbb{E}^{\theta^*}\left(\sum_{t=1}^n Z_t\right)\right)^2\right).$$

We want to bound Var by $Cn\epsilon^{(2-\alpha)/2}$, for any $\alpha > 0$. In this purpose, we split the sum in two parts:

$$Var \leq 2\underbrace{\mathbb{E}^{\theta^*}\left[\left(\sum_{t=1}^n\left(Z_t - \mathbb{E}^{\theta^*}\left(Z_t|Y_{1:t-1}\right)\right)\right)^2\right]}_{=S_1} + 2\underbrace{\mathbb{E}^{\theta^*}\left[\left(\sum_{t=1}^n\left(\mathbb{E}^{\theta^*}\left(Z_t|Y_{1:t-1}\right) - \mathbb{E}^{\theta^*}(Z_t)\right)\right)^2\right]}_{=S_2}, \tag{3.27}$$

S_1 is the expectation of the square of a sum of martingale increments, for which the covariances are zero so that only n terms remain. S_2 is further controlled using the exponential forgetting of Markov chain. First, we control S_1:

$$\begin{aligned} S_1 &= \sum_{t=1}^n \mathbb{E}^{\theta^*}\left(\left(Z_t - \mathbb{E}^{\theta^*}\left(Z_t|Y_{1:t-1}\right)\right)^2\right) \\ &\quad + 2\sum_{1\leq r<t\leq n}\mathbb{E}^{\theta^*}\left(\left(Z_r - \mathbb{E}^{\theta^*}(Z_r|Y_{1:r-1})\right)\mathbb{E}^{\theta^*}\left(Z_t - \mathbb{E}^{\theta^*}(Z_t|Y_{1:t-1})|Y_{1:t-1}\right)\right) \\ &\leq \sum_{t=1}^n \mathbb{E}^{\theta^*}\left(Z_t^2\right) \end{aligned} \tag{3.28}$$

using that

$$\mathbb{E}^{\theta^*}(\mathbb{E}^{\theta^*}(Z_t|Y_{1:t-1})^2) = \mathbb{E}^{\theta^*}\left[\mathbb{E}^{\theta^*}\left[(\mathbb{E}^{\theta^*}(Z_t|Y_{1:t-1})^2|Y_{1:t-1}\right]\right] \leq \mathbb{E}^{\theta^*}(Z_t^2).$$

As to S_2:

$$\begin{aligned} S_2 &= \sum_{t=1}^{n} \mathbb{E}^{\theta^*}\left(\left(\mathbb{E}^{\theta^*}\left(Z_t|Y_{1:t-1}\right) - \mathbb{E}^{\theta^*}\left(Z_t\right)\right)^2\right) \\ &\quad + 2\sum_{1\le r<t\le n} \mathbb{E}^{\theta^*}\left(\left(\mathbb{E}^{\theta^*}(Z_r|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_r)\right)\mathbb{E}^{\theta^*}\left(\mathbb{E}^{\theta^*}(Z_t|Y_{1:t-1}) - \mathbb{E}^{\theta^*}(Z_t)|Y_{1:r-1}\right)\right) \\ &\le \sum_{t=1}^{n} \mathbb{E}^{\theta^*}(Z_t^2) + 2\sum_{1\le r<t\le n} \sqrt{\mathbb{E}^{\theta^*}(Z_r^2)}\sqrt{\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^2\right)}, \end{aligned} \tag{3.29}$$

using that

$$\mathbb{E}^{\theta^*}(\mathbb{E}^{\theta^*}(Z_t|Y_{1:t-1})^2) \le \mathbb{E}^{\theta^*}(Z_t^2). \tag{3.30}$$

and Cauchy-Schwarz inequality to bound the second term.

Combining (3.27), (3.28) et (3.29), we obtain

$$Var \le 4\sum_{t=1}^{n} \mathbb{E}^{\theta^*}\left(Z_t^2\right) + 4\sum_{1\le r<t\le n} \sqrt{\mathbb{E}^{\theta^*}(Z_r^2)}\sqrt{\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^2\right)}. \tag{3.31}$$

Then using Lemmas 3.9 and 3.11,

$$\begin{aligned} Var \le 4\left(\frac{16}{\underline{q}(1-\rho^2)} + Cn\epsilon^2\right) + 16\rho^{-\frac{5\alpha}{4}}\left(2\epsilon + \frac{10}{\underline{q}}\right)^{\alpha/2} \\ \sum_{1\le r<t\le n}\left(\frac{16\rho^{2(r-1)}}{\underline{q}^2} + \tilde{C}\epsilon^2\right)^{1-\alpha/4}\rho^{\frac{\alpha}{4}(t-r)}. \end{aligned} \tag{3.32}$$

Since,

$$\sum_{1\le r<t\le n}\left(\frac{16\rho^{2(r-1)}}{\underline{q}^2} + \tilde{C}\epsilon^2\right)^{1-\alpha/4}\rho^{\frac{\alpha}{4}(t-r)} \le 2\frac{\rho^{\alpha/4}}{1-\rho^{\alpha/4}}\left(\frac{16}{\underline{q}^2(1-\rho)} + n\tilde{C}\epsilon^{2-\alpha/2}\right) \tag{3.33}$$

therefore there exists a constant $C_{KL^2} > 0$ only depending on k and $\underline{q}^*(= 2\underline{q})$ such that

$$Var \le C_{KL^2}\frac{n}{\alpha}\left(\frac{\tilde{\epsilon}_n}{\sqrt{u_n}}\right)^{2-\alpha}. \tag{3.34}$$

3.5.2.2 Lemma 3.9: Control of $\mathbb{E}^{\theta^*}\left(Z_t^2\right)$

Lemma 3.9. *For all* $\theta, \theta^* \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$

$$\begin{aligned}
&\mathbb{E}^{\theta^*}(Z_t^2) \\
&\leq \frac{16\rho^{2(t-1)}}{\underline{q}^2} + 2 \max_{1\leq i\leq k} \int f_i^*(y) \max_{1\leq j\leq k} \log^2\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy) \\
&\quad + \frac{32\,\|Q-Q^*\|^2}{\underline{q}^4(1-\rho)^2} + \frac{32}{\underline{q}^4(1-\rho)^2} \int \min\left(\sum_{1\leq i\leq k} \frac{|f_i^*(y)-f_i(y)|^2}{f_i^*(y)}, \underline{q}^2k^2 \sum_{1\leq j\leq k} f_j^*(y)\right) \lambda(dy),
\end{aligned} \tag{3.35}$$

with $Z_t = \log\left(\frac{P^{\theta^*}(Y_t|Y_{1:t-1})}{P^{\theta}(Y_t|Y_{1:t-1})}\right)$ *and* $\rho = \frac{1-k\underline{q}}{1-(k-1)\underline{q}} \leq 1-\underline{q}$.
If moreover Assumptions (A2.1), (A2.4), (A2.6) *and* (3.5) *hold, then*

$$\mathbb{E}^{\theta^*}(Z_t^2) \leq \frac{16\rho^{2(t-1)}}{\underline{q}^2} + \tilde{C}\frac{\tilde{\epsilon}_n^2}{u_n} \tag{3.36}$$

where $\tilde{C} \leq 33(1+2k)/\underline{q}^6$.

Proof of Lemma 3.9. Let $Q_{t,i}^{\mu,\theta,Y_{1:t-1}} = \mathbb{P}^{\theta}(X_t = i|Y_{1:t-1}, X_1 \sim \mu)$, where $t \geq 1$, $1 \leq i \leq k$, $\mu \in \Delta_k$ and when μ is not specified, the stationary initial probability distribution is considered. Then the conditional density function of Y_t given $Y_{1:t-1}$ with respect to λ is:

$$\sum_{i=1}^{k} f_i(\cdot) Q_{t,i}^{\theta,Y_{1:t-1}};$$

so that

$$\begin{aligned}
\mathbb{E}^{\theta^*}\left(Z_t^2\right) &= \mathbb{E}^{\theta^*}\left(\left(\log \frac{\sum_{i=1}^k f_i^*(Y_t) Q_{t,i}^{\theta^*,Y_{1:t-1}}}{\sum_{j=1}^k f_j(Y_t) Q_{t,j}^{\theta,Y_{1:t-1}}}\right)^2\right) \\
&\leq 2\mathbb{E}^{\theta^*}\left(\max_{1\leq j\leq k} \log^2\left(\frac{f_j^*(Y_t)}{f_j(Y_t)}\right)\right) + \frac{2}{\underline{q}^2}\mathbb{E}^{\theta^*}\left(\left(\sum_{j=1}^k |Q_{t,j}^{\theta^*,Y_{1:t-1}} - Q_{t,j}^{\theta,Y_{1:t-1}}|\right)^2\right)
\end{aligned} \tag{3.37}$$

using Equation (3.18) and Lemma 3.8.

Combining Equation (3.37) and Lemma 3.10 (Equation (3.39)), we obtain Equation (3.35). Moreover, using Assumption (A2.1),

$$\mathbb{E}^{\theta^*}\left(\max_{1\leq j\leq k} \log^2\left(\frac{f_j^*(Y_t)}{f_j(Y_t)}\right)\right) \leq \epsilon^2. \tag{3.38}$$

Finally, combining Equations (3.37), (3.38) and Lemma 3.10 (Equation (3.40)), we obtain Equation (3.36). □

Lemma 3.10. *For all $\theta, \theta^* \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$ and $\mu, \mu^* \in \Delta_k(\underline{q})$,*

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\left(\sum_{j=1}^{k}|Q_{t,j}^{\theta^*,Y_{1:t-1}} - Q_{t,j}^{\theta,Y_{1:t-1}}|\right)^2\right) \\
&\leq 8\rho^{2(t-1)} + \frac{16\,\|Q-Q^*\|^2}{\underline{q}^2(1-\rho)^2} \\
&\quad + \frac{16}{\underline{q}^2(1-\rho)^2}\int \min\left(\sum_{1\leq i\leq k}\frac{|f_i^*(y)-f_i(y)|^2}{f_i^*(y)}, \underline{q}^2k^2\sum_{1\leq j\leq k}f_j^*(y)\right)\lambda(dy),
\end{aligned} \tag{3.39}$$

with $\rho = \frac{1-k\underline{q}}{1-(k-1)\underline{q}} \leq 1-\underline{q}$.

If moreover Assumptions (A2.1), (A2.4), (A2.6) and (3.5) hold, then

$$\mathbb{E}^{\theta^*}\left(\left(\sum_{j=1}^{k}|Q_{t,j}^{\theta^*,Y_{1:t-1}} - Q_{t,j}^{\theta,Y_{1:t-1}}|\right)^2\right) \leq 8\rho^{2(t-1)} + C'\frac{\tilde{\epsilon}_n^2}{u_n} \tag{3.40}$$

where $C' \leq 16(1+2k)/\underline{q}^4$.

Proof of Lemma 3.10. We first control $\sum_{i=1}^{k}|Q_{t,i}^{\theta^*,Y_{1:t-1}} - Q_{t,i}^{\mu^*,\theta,Y_{1:t-1}}|$. For this purpose, we are going to use a modified version of Proposition 1 of Douc and Matias [DM01]. By Proposition 1 of Douc and Matias [DM01] and for all θ, θ^* in $\Delta_k^k(\underline{q}) \times \mathcal{F}^k$ we can control the L_1-norm between two conditional probabilities of the state t when the initial probabilities are equal.

$$\begin{aligned}
&\sum_{i=1}^{k}|Q_{t,i}^{\theta^*,Y_{1:t-1}} - Q_{t,i}^{\mu^*,\theta,Y_{1:t-1}}| \\
&\quad\leq \sum_{j=1}^{k}\left|Q_{t-1,j}^{(Q_{2,\cdot}^{\theta^*,Y_1},Q,f),Y_{2:t-1}} - Q_{t-1,j}^{(Q_{2,\cdot}^{\theta^*,Y_1},Q^*,f^*),Y_{2:t-1}}\right| + \frac{1}{2}\left(\frac{1-k\underline{q}}{1-(k-1)\underline{q}}\right)^{t-1} * \\
&\qquad\sum_{u=1}^{k}\left|\frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta}\mu^*)_v} - \frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta^*}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta^*}\mu^*)_v}\right|
\end{aligned} \tag{3.41}$$

with $\left(A^{Y,\theta}_{i,j}\right)_{1\leq i,j\leq k} = (Q_{j,i}f_j(Y))_{1\leq i,j\leq k}$. And

$$\begin{aligned}
&\sum_{u=1}^{k}\left|\frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta}\mu^*)_v} - \frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta^*}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta^*}\mu^*)_v}\right| \\
&\quad= \sum_{u=1}^{k}\left|\frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta^*}\mu^* - A^{Y_1,\theta}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta^*}\mu^*)_v}\right. \\
&\qquad - \frac{\sum_{w=1}^{k}\sum_{i=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{i,w}(A^{Y_1,\theta^*}\mu^* - A^{Y_1,\theta}\mu^*)_w}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta^*}\mu^*)_v} \\
&\qquad \left.\frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta}\mu^*)_v}\right| \\
&\quad\leq 2\sum_{u=1}^{k}\min\left(\left|\frac{\sum_{j=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta})_{j,u}(A^{Y_1,\theta^*}\mu^* - A^{Y_1,\theta}\mu^*)_u}{\sum_{v=1}^{k}(A^{Y_{t-1},\theta}\dots A^{Y_2,\theta}A^{Y_1,\theta^*}\mu^*)_v}\right|, 1\right) \\
&\quad\leq 2\min\left(\max_{1\leq u\leq k}\frac{\sum_{i=1}^{k}\mu_i^*|Q_{i,u}f_i(Y_1) - Q^*_{i,u}f_i^*(Y_1)|}{\sum_{i=1}^{k}\mu_i^*Q^*_{i,u}f_i^*(Y_1)}, k\right) \\
&\quad\leq 2\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{(1-(k-1)\underline{q})}{\underline{q}}\frac{\sum_{i=1}^{k}|f_i^*(Y_1) - f_i(Y_1)|\mu_i^*}{\sum_{j=1}^{k}f_j^*(Y_1)\mu_j^*}, k\right)\right),
\end{aligned} \tag{3.42}$$

using Lemma 3.8. Combining Equations (3.41) and (3.42),

$$\begin{aligned}
&\sum_{i=1}^{k}|Q_{t,i}^{\theta^*,Y_{1:t-1}} - Q_{t,i}^{\mu^*,\theta,Y_{1:t-1}}| \\
&\quad\leq \sum_{j=1}^{k}\left|Q_{t-1,j}^{(Q_{2,\cdot}^{\theta^*,Y_1},Q,f),Y_{2:t-1}} - Q_{t-1,j}^{(Q_{2,\cdot}^{\theta^*,Y_1},Q^*,f^*),Y_{2:t-1}}\right| + \left(\frac{1-k\underline{q}}{1-(k-1)\underline{q}}\right)^{t-1} * \\
&\qquad\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{(1-(k-1)\underline{q})}{\underline{q}}\frac{\sum_{i=1}^{k}|f_i^*(Y_1) - f_i(Y_1)|\mu_i^*}{\sum_{j=1}^{k}f_j^*(Y_1)\mu_j^*}, k\right)\right).
\end{aligned} \tag{3.43}$$

By repeating the arguments of Equation (3.42), we show that

$$\begin{aligned}
&\sum_{i=1}^{k}|Q_{2,i}^{(Q_{t-1}^{\theta^*,Y_{1:t-2}},Q^*,f^*),Y_{1:t-1}} - Q_{2,i}^{(Q_{t-1}^{\theta^*,Y_{1:t-2}},Q,f),Y_{1:t-1}}| \\
&\quad= \sum_{i=1}^{k}\left|\frac{\sum_{i=1}^{k}A_{j,i}^{Y_{t-1},\theta^*}Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}{\sum_{i,u=1}^{k}A_{u,i}^{Y_{t-1},\theta^*}Q_{t-1,i}^{\theta^*,Y_{1:t-2}}} - \frac{\sum_{i=1}^{k}A_{j,i}^{Y_{t-1},\theta}Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}{\sum_{i,u=1}^{k}A_{u,i}^{Y_{t-1},\theta}Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}\right| \\
&\quad\leq 2\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{(1-(k-1)\underline{q})}{\underline{q}}\frac{\sum_{i=1}^{k}|f_i^*(Y_{t-1}) - f_i(Y_{t-1})|Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}{\sum_{j=1}^{k}f_j^*(Y_{t-1})Q_{t-1,j}^{\theta^*,Y_{1:t-2}}}, k\right)\right).
\end{aligned}$$

By induction on (3.43),

$$\begin{aligned}
&\sum_{i=1}^{k}|Q_{t,i}^{\theta^*,Y_{1:t-1}} - Q_{t,i}^{\mu^*,\theta,Y_{1:t-1}}| \\
&\quad\leq 2\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{(1-(k-1)\underline{q})}{\underline{q}}\frac{\sum_{i=1}^{k}|f_i^*(Y_{t-1}) - f_i(Y_{t-1})|Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}{\sum_{j=1}^{k} f_j^*(Y_{t-1})Q_{t-1,j}^{\theta^*,Y_{1:t-2}}}, k\right)\right) \\
&\qquad+\sum_{u=3}^{t}\left(\frac{1-k\underline{q}}{1-(k-1)\underline{q}}\right)^{u-1} \\
&\qquad\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{(1-(k-1)\underline{q})}{\underline{q}}\frac{\sum_{i}|f_i^*(Y_{t-u+1}) - f_i(Y_{t-u+1})|Q_{t-u+1,i}^{\theta^*,Y_{1:t-u}}}{\sum_{j} f_j^*(Y_{t-u+1})Q_{t-u+1,j}^{\theta^*,Y_{1:t-u}}}, k\right)\right).
\end{aligned} \tag{3.44}$$

Using Corollary 1 of Douc *et al.* [DMR04], we can control the ℓ_1-norm between two conditional probabilities of the state t for the same parameter θ but different initial probabilities: for all $\theta \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$, $\mu, \tilde{\mu} \in \Delta_k$ and for all $y_{1:l-1} \in \{y \,:\, \exists i,\ f_i^*(y) > 0\}^{l-1}$

$$\sum_{i=1}^{k}\left|Q_{i,l}^{\mu,\theta,y_{1:l-1}} - Q_{i,l}^{\tilde{\mu},\theta,y_{1:l-1}}\right| \leq 2\rho^{l-1}. \tag{3.45}$$

Combining Equations (3.44) and (3.45), we obtain

$$\begin{aligned}
&\sum_{i=1}^{k}|Q_{t,i}^{\mu^*,\theta^*,Y_{1:t-1}} - Q_{t,i}^{\mu,\theta,Y_{1:t-1}}| \\
&\quad\leq 2\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{1}{\underline{q}}\frac{\sum_{i=1}^{k}|f_i^*(Y_{t-1}) - f_i(Y_{t-1})|Q_{t-1,i}^{\theta^*,Y_{1:t-2}}}{\sum_{j=1}^{k} f_j^*(Y_{t-1})Q_{t-1,j}^{\theta^*,Y_{1:t-2}}}, k\right)\right) \\
&\qquad+2\sum_{u=3}^{t}\left(\frac{1-k\underline{q}}{1-(k-1)\underline{q}}\right)^{u-1} \\
&\qquad\underbrace{\left(\frac{\|Q-Q^*\|}{\underline{q}} + \min\left(\frac{1}{\underline{q}}\frac{\sum_{i}|f_i^*(Y_{t-u+1}) - f_i(Y_{t-u+1})|Q_{t-u+1,i}^{\theta^*,Y_{1:t-u}}}{\sum_{j} f_j^*(Y_{t-u+1})Q_{t-u+1,j}^{\theta^*,Y_{1:t-u}}}, k\right)\right)}_{=\Delta_{t-u+1}} \\
&\qquad+4\left(\underbrace{\frac{1-k\underline{q}}{1-(k-1)\underline{q}}}_{=\rho}\right)^{t-1} \\
&\quad\leq \frac{2}{\rho}\sum_{u=1}^{t-1}\rho^u\Delta_{t-u} + 4\rho^{t-1}.
\end{aligned} \tag{3.46}$$

Then

$$\left(\sum_{i=1}^{k}|Q_{t,i}^{\mu^*,\theta^*,Y_{1:t-1}}-Q_{t,i}^{\mu,\theta,Y_{1:t-1}}|\right)^2\leq\frac{8}{\rho^2}\left(\sum_{u=1}^{t-1}\rho^u\right)\left(\sum_{u=1}^{t-1}\rho^u\Delta_{t-u}^2\right)+8\rho^{2(t-1)},\tag{3.47}$$

using Cauchy-Schwarz inequality. Moreover, using Lemma 3.8,

$$\begin{aligned}
&\mathbb{E}^{\theta^*}(\Delta_{t-u}^2)\\
&\leq 2\frac{\|Q-Q^*\|^2}{\underline{q}^2}+\frac{2}{\underline{q}^2}\mathbb{E}^{\theta^*}\left(\min\left(\left(\frac{\sum_i|f_i^*(Y_{t-u})-f_i(Y_{t-u})|Q_{t-u,i}^{\theta^*,Y_{1:t-u-1}}}{\sum_j f_j^*(Y_{t-u})Q_{t-u,j}^{\theta^*,Y_{1:t-u-1}}}\right)^2,(\underline{q}k)^2\right)\right)\\
&\leq 2\frac{\|Q-Q^*\|^2}{\underline{q}^2}\\
&\qquad+\frac{2}{\underline{q}^2}E\left(\int\sum_{1\leq\iota\leq k}Q_{t-u,\iota}^{\theta^*,Y_{1:t-u-1}}f_\iota(y)\min\left(\left(\frac{\sum_i|f_i^*(y)-f_i(y)|Q_{t-u,i}^{\theta^*,Y_{1:t-u-1}}}{\sum_j f_j^*(y)Q_{t-u,j}^{\theta^*,Y_{1:t-u-1}}}\right)^2,(\underline{q}k)^2\right)\lambda(dy)\right)\\
&\leq 2\frac{\|Q-Q^*\|^2}{\underline{q}^2}+\frac{2}{\underline{q}^2}\int\min\left(\sum_{1\leq i\leq k}\frac{|f_i^*(y)-f_i(y)|^2}{f_i^*(y)},\underline{q}^2k^2\sum_{1\leq j\leq k}f_j^*(y)\right)\lambda(dy).
\end{aligned}\tag{3.48}$$

Combining Equations (3.47) and (3.48), we obtain Equation (3.39) which implies Equation (3.40) under Assumptions (A2.1), (A2.4), (A2.6) and (3.5). This concludes the proof of Lemma 3.10.

□

3.5.2.3 Lemma 3.11: Control of $\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1})-\mathbb{E}^{\theta^*}(Z_t)|^2\right)$

In the following lemma we show that $\mathbb{E}^{\theta^*}\left(\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1})-\mathbb{E}^{\theta^*}(Z_t)|\right)^2\right)$ geometrically decreases to 0 when t tends to $+\infty$, using the exponential forgetting of the Markov chain.

Lemma 3.11. *For all $\theta,\theta^*\in\Delta_k^k(\underline{q})\times\mathcal{F}^k$ and $\alpha\in(0,2)$,*

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1})-\mathbb{E}^{\theta^*}(Z_t)|^2\right)\\
&\leq 8\mathbb{E}^{\theta^*}(Z_t^2)^{\frac{2-\alpha}{2}}\rho^{-\frac{5\alpha}{2}}\rho^{\frac{\alpha}{2}(t-r)}\left(2\max_{1\leq j\leq k}\int f_j^*(y)\max_{1\leq i\leq k}\left|\log\frac{f_i^*(y)}{f_i(y)}\right|\lambda(dy)+\frac{10}{\underline{q}}\right)^\alpha,
\end{aligned}\tag{3.49}$$

where $Z_t=\log\left(\frac{P^{\theta^*}(Y_t|Y_{1:t-1})}{P^{\theta}(Y_t|Y_{1:t-1})}\right)$. If moreover Assumption (A2.1) holds then

$$\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1})-\mathbb{E}^{\theta^*}(Z_t)|^2\right)\leq 8\mathbb{E}^{\theta^*}(Z_t^2)^{\frac{2-\alpha}{2}}\rho^{-\frac{5\alpha}{2}}\rho^{\frac{\alpha}{2}(t-r)}\left(2\frac{\tilde{\epsilon}_n}{\sqrt{u_n}}+\frac{10}{\underline{q}}\right)^\alpha.\tag{3.50}$$

Proof of Lemma 3.11. Denote

$$L_t = (X_t, Y_t, Q_{t,\cdot}^{\theta, Y_{1:t-1}}, Q_{t,\cdot}^{\theta^*, Y_{1:t-1}})$$

for all $t \in \mathbb{N}$, then $(L_t)_{t\in\mathbb{N}}$ is the extended Markov chain with transition kernel Π_θ more precisely described in Douc and Matias [DM01] at page 384. Let

$$h\colon \begin{cases} \{1,\dots,k\} \times \{y : \exists 1 \le j \le k, f_j^*(y) > 0\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\}^2 \longrightarrow \mathbb{R} \\ l = (x, y, \mu, \mu^*) \longmapsto h(l) = \log\left(\frac{\sum_{i=1}^k \mu_i^* f_i^*(y)}{\sum_{i=1}^k \mu_i f_i(y)}\right) \end{cases}$$

then

$$h(L_t) = Z_t = \log\left(\frac{p^{\theta^*}(Y_t|Y_{1:t-1})}{p^{\theta}(Y_t|Y_{1:t-1})}\right)$$

and for all $r \le t$ and $0 < \alpha < 2$,

$$\begin{aligned} &\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^2\right) \\ &\quad = \mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^{2-\alpha}|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^{\alpha}\right) \\ &\quad \le 2^{2-\alpha}\mathbb{E}^{\theta^*}\Big(\Big(\max(\mathbb{E}^{\theta^*}(|Z_t|), \mathbb{E}^{\theta^*}(|Z_t||Y_{1:r-1}))\Big)^{2-\alpha} |\mathbb{E}^{\theta^*}(h(L_t)|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(h(L_t))|^{\alpha}\Big). \end{aligned} \tag{3.51}$$

The following term is geometrically decreasing, using Lemma 3.12

$$\begin{aligned} &|\mathbb{E}^{\theta^*}(h(L_t)|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(h(L_t))| \\ &\le \int\int \left|\int h(l_t)\Pi_\theta^{t-r}(l_r, dl_t) - \int h(l_t)\Pi_\theta^{t-r}(\tilde{l}_r, dl_t)\right| P^{\theta}(dl_r|Y_{1:r-1})P^{\theta}(d\tilde{l}_r). \end{aligned} \tag{3.52}$$

More precisely, using Equation (3.52) and Lemma 3.12 with $m = \lfloor \frac{t-r+1}{2} \rfloor$ and $u = t - r$, we obtain

$$|\mathbb{E}^{\theta^*}(h(L_t)|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(h(L_t))| \le \rho^{\frac{t-r}{2}-\frac{5}{2}}\left(2\max_{1\le j\le k}\int f_j^*(y)\max_{1\le i\le k}\left|\log\frac{f_i^*(y)}{f_i(y)}\right|\lambda(dy) + \frac{10}{\underline{q}}\right). \tag{3.53}$$

Therefore using Equations (3.51) and (3.53),

$$\begin{aligned} &\mathbb{E}^{\theta^*}\left(|\mathbb{E}^{\theta^*}(Z_t|Y_{1:r-1}) - \mathbb{E}^{\theta^*}(Z_t)|^2\right) \\ &\quad \le 2^{2-\alpha}\mathbb{E}^{\theta^*}\left(\max(\mathbb{E}^{\theta^*}(\left(|Z_t|\right), \mathbb{E}^{\theta^*}(|Z_t||Y_{1:r-1}))\right)^{2-\alpha}\right) \\ &\qquad \rho^{-\frac{5\alpha}{2}}\rho^{\frac{\alpha}{2}(t-r)}\left(2\max_{1\le j\le k}\int f_j^*(y)\max_{1\le i\le k}\left|\log\frac{f_i^*(y)}{f_i(y)}\right|\lambda(dy) + \frac{10}{\underline{q}}\right)^{\alpha} \end{aligned} \tag{3.54}$$

By convexity of the square function and concavity of $x \mapsto x^{\frac{2-\alpha}{2}}$, with $0 < \alpha < 2$,

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\max(\mathbb{E}^{\theta^*}(|Z_t|), \mathbb{E}^{\theta^*}(|Z_t||Y_{1:r-1}))^{2-\alpha}\right) \\
&\leq \mathbb{E}^{\theta^*}\left(\max(\mathbb{E}^{\theta^*}(Z_t^2)^{1/2}, \mathbb{E}^{\theta^*}(Z_t^2|Y_{1:r-1})^{1/2})^{2-\alpha}\right) \\
&\leq \mathbb{E}^{\theta^*}(Z_t^2)^{\frac{2-\alpha}{2}} + \mathbb{E}^{\theta^*}(\mathbb{E}^{\theta^*}(Z_t^2|Y_{1:r-1})^{\frac{2-\alpha}{2}}) \\
&\leq 2\mathbb{E}^{\theta^*}(Z_t^2)^{\frac{2-\alpha}{2}}.
\end{aligned} \tag{3.55}$$

Combining Equations (3.54) and (3.55), we get Equation (3.49). Besides, using Assumption (A2.1) and Cauchy–Schwarz inequality ,

$$\max_{1\leq j\leq k} \int f_j^*(y) \max_{1\leq i\leq k} \left|\log \frac{f_i^*(y)}{f_i(y)}\right| \lambda(dy) \leq \epsilon \tag{3.56}$$

so that Equation (3.50) holds.

□

Lemma 3.12 is an improved version of Proposition 2 of Douc and Matias [DM01].

Lemma 3.12. *For all integers $u > 0, m < u$, for all $z, \tilde{z} \in \{1, \ldots, k\} \times \{y : \exists 1 \leq j \leq k, f_j^*(y) > 0\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\}$ and for all $\theta, \theta^* \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$,*

$$\begin{aligned}
&\left|\int h(l)\Pi_\theta^u(z, dl) - \int h(l)\Pi_\theta^u(\tilde{z}, dl)\right| \\
&\leq \frac{4}{\underline{q}}\rho^{u-1} + \frac{4}{\underline{q}}\rho^m + 2\rho^{m-2}\left(\max_{1\leq j\leq k} \int f_j^*(y) \max_{1\leq i\leq k} \left|\log \frac{f_i^*(y)}{f_i(y)}\right| \lambda(dy) + \log\left(\frac{1}{\underline{q}}\right)\right)
\end{aligned} \tag{3.57}$$

where Π_θ is the transition kernel of the extended Markov chain $L_t = (X_t, Y_t, Q_{t,\cdot}^{\theta, Y_{1:t-1}}, Q_{t,\cdot}^{\theta^*, Y_{1:t-1}})$ and

$$h\colon \begin{cases} \{1, \ldots, k\} \times \{y : \exists 1 \leq j \leq k, f_j^*(y) > 0\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\}^2 \longrightarrow \mathbb{R} \\ l = (x, y, \mu, \mu^*) \longmapsto h(l) = \log\left(\frac{\sum_{i=1}^k \mu_i^* f_i^*(y)}{\sum_{i=1}^k \mu_i f_i(y)}\right). \end{cases}$$

Proof of Lemma 3.12. We improve the result of Proposition 2 of Douc and Matias [DM01] by defining h on

$$\mathcal{Z} = \{1, \ldots, k\} \times \{y : \exists 1 \leq j \leq k, f_j^*(y) > 0\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\} \times \{\mu \in \Delta_k : \mu_i > \underline{q}\ \forall i\}$$

and using that if $z \in \mathcal{Z}$ then $\Pi_{\theta^*}(z, \mathcal{Z}) = 1$. Then we obtain

$$\operatorname{lip}(h, x, y) = \frac{1}{\underline{q}} \tag{3.58}$$

since for all $(x, y, \mu, \mu^*), (x, y, \tilde{\mu}, \tilde{\mu}^*) \in \mathcal{Z}$

$$\begin{aligned}
|h(x,y,\mu,\mu^*) - h(x,y,\tilde{\mu},\tilde{\mu}^*)| &= \left|\log\left(\frac{\sum_{i=1}^k \mu_i^* f_i^*(y)}{\sum_{i=1}^k \mu_i f_i(y)}\right) - \log\left(\frac{\sum_{i=1}^k \tilde{\mu}_i^* f_i^*(y)}{\sum_{i=1}^k \tilde{\mu}_i f_i(y)}\right)\right| \\
&= \left|\log\left(\frac{\sum_{i=1}^k \mu_i^* f_i^*(y)}{\sum_{i=1}^k \tilde{\mu}_i^* f_i^*(y)}\right) - \log\left(\frac{\sum_{i=1}^k \mu_i f_i(y)}{\sum_{i=1}^k \tilde{\mu}_i f_i(y)}\right)\right| \\
&\leq \max_{1\leq i\leq k}\left|\log\left(\frac{\mu_i^*}{\tilde{\mu}_i^*}\right)\right| + \max_{1\leq i\leq k}\left|\log\left(\frac{\mu_i}{\tilde{\mu}_i}\right)\right| \\
&\leq \frac{1}{\underline{q}}\left(\sum_{i=1}^k |\mu_i^* - \tilde{\mu}_i^*| + \sum_{i=1}^k |\mu_i - \tilde{\mu}_i|\right)
\end{aligned}$$

using . Moreover

$$k(h,x,y) = \log\frac{1}{\underline{q}} + \max_{1\leq i\leq k}\left|\log\left(\frac{f_i^*(y)}{f_i(y)}\right)\right| \tag{3.59}$$

because, using Lemma 3.8,

$$\begin{aligned}
|h(x,y,\mu,\mu^*)| &= \left|\log\left(\frac{\sum_{i=1}^k \mu_i^* f_i^*(y)}{\sum_{i=1}^k \mu_i f_i(y)}\right)\right| \\
&\leq \max_{1\leq i\leq k}\left|\log\left(\frac{\mu_i^* f_i^*(y)}{\mu_i f_i(y)}\right)\right| \\
&\leq \max_{1\leq i\leq k}\left|\log\left(\frac{\mu_i^*}{\mu_i}\right)\right| + \max_{1\leq i\leq k}\left|\log\left(\frac{f_i^*(y)}{f_i(y)}\right)\right| \\
&\leq \log\frac{1}{\underline{q}} + \max_{1\leq i\leq k}\left|\log\left(\frac{f_i^*(y)}{f_i(y)}\right)\right|.
\end{aligned}$$

Moreover instead of using Proposition 1 of Douc and Matias [DM01] we use Corollary 1 of Douc *et al.* [DMR04] so that for all $\theta \in \Delta_k^k(\underline{q}) \times \mathcal{F}^k$, $\mu, \tilde{\mu} \in \Delta_k(\underline{q})$ and for all $y_{1:l-1} \in \{y : \exists i,\ f_i^*(y) > 0\}^{l-1}$

$$\sum_{i=1}^k \left|Q_{i,l}^{\mu,\theta,y_{1:l-1}} - Q_{i,l}^{\tilde{\mu},\theta,y_{1:l-1}}\right| \leq 2\rho^{l-1}. \tag{3.60}$$

Then, let $z = (x, y, \mu, \mu^*) \in \mathcal{Z}$ and $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\mu}, \tilde{\mu}^*) \in \mathcal{Z}$ using the proof of Proposition 1 of Douc and Matias [DM01]:

$$
\begin{aligned}
&|\int h(l)\Pi^u(z,dl) - \int h(l)\Pi^u(\tilde{z},dl)| \\
&\leq \underbrace{\left|\int h(l)\Pi^u((x,y,\mu,\mu^*),dl) - \int h(l)\Pi^u((x,\tilde{y},\tilde{\mu},\tilde{\mu}^*),dl)\right|}_{=A} \\
&\quad + \underbrace{\left|\int h(l)\Pi^u(x,\tilde{y},\tilde{\mu},\tilde{\mu}^*),dl) - \int h(l)\Pi^u((\tilde{x},\tilde{y},\tilde{\mu},\tilde{\mu}^*),dl)\right|}_{=B}
\end{aligned} \tag{3.61}
$$

where

$$
\begin{aligned}
|A| \leq \Bigg| &\sum_{x_{2:u+1}=1}^{k} \int \mathrm{lip}(h, x_{u+1}, y_{u+1}) \\
&\left(\sum_{i=1}^{k} |Q_{u,i}^{Q_1^{\mu,\theta,y},\theta,y_{2:u}} - Q_{u,i}^{Q_1^{\tilde{\mu},\theta,\tilde{y}},\theta,y_{2:u}}| + \sum_{i=1}^{k} |Q_{u,i}^{Q_1^{\mu^*,\theta^*,y},\theta^*,y_{2:u}} - Q_{u,i}^{Q_1^{\tilde{\mu}^*,\theta^*,\tilde{y}},\theta^*,y_{2:u}}|\right) \\
&Q^*_{x,x_2} \dots Q^*_{x_u,x_{u+1}} f^*_{x_2}(y_2) \dots f^*_{x_{u+1}}(y_{u+1}) \lambda(dy_2) \dots \lambda(dy_{u+1})\Bigg|
\end{aligned} \tag{3.62}
$$

and for any $1 \leq m \leq u$,

$$
\begin{aligned}
|B| \leq &\sum_{x_{2:u+1}=1}^{k} \int \mathrm{lip}(h, x_{u+1}, y_{u+1}) \Bigg(\sum_{i=1}^{k} |Q_{m+1,i}^{Q_{u+1-m}^{\tilde{\mu},\theta,\tilde{y},y_{2:u-m}},\theta,y_{u-m+1:u}} - Q_{m+1,i}^{\nu,\theta,y_{u-m+1:u}}| \\
&+ \sum_{i=1}^{k} |Q_{m+1,i}^{Q_{u+1-m}^{\tilde{\mu},\theta^*,\tilde{y},y_{2:u-m}},\theta^*,y_{u-m+1:u}} - Q_{m+1,i}^{\nu^*,\theta^*,y_{u-m+1:u}}|\Bigg) \\
&\left|Q^*_{x,x_2} - Q^*_{\tilde{x},x_2}\right| Q^*_{x_2,x_3} \dots Q^*_{x_u,x_{u+1}} f^*_{x_2}(y_2) \dots f^*_{x_{u+1}}(y_{u+1}) \lambda(dy_{2:u+1}) \\
&+ \sum_{x_{m:u+1}=1}^{k} \int k(h, x_{u+1}, y_{u+1}) \\
&|Q^{*m-1}_{x,x_m} - Q^{*m-1}_{\tilde{x},x_m}| Q^*_{x_m,x_{m+1}} \dots Q^*_{x_u,x_{u+1}} f^*_{x_m}(y_m) \dots f^*_{x_{u+1}}(y_{u+1}) \lambda(dy_{2:u+1}).
\end{aligned} \tag{3.63}
$$

Combining Equations (3.58), (3.60) and (3.62),

$$
|A| \leq \frac{4}{\underline{q}} \rho^{u-1} \tag{3.64}
$$

and using Equations (3.58), (3.59), (3.60) and (3.63)

$$
|B| \leq \frac{4}{\underline{q}} \rho^m + 2\rho^{m-2} \left(\max_{1 \leq j \leq k} \int f_j^*(y) \max_{1 \leq i \leq k} \left| \log \frac{f_i^*(y)}{f_i(y)} \right| \lambda(dy) + \log\left(\frac{1}{\underline{q}}\right) \right) \tag{3.65}
$$

therefore using Equations (3.61), (3.64) and (3.65) we obtain

$$\begin{aligned}&\left|\int h(l)\Pi^u(z,dl) - \int h(l)\Pi^u(\tilde{z},dl)\right| \\ &\leq \frac{4}{\underline{q}}\rho^{u-1} + \frac{4}{\underline{q}}\rho^m + 2\rho^{m-2}\left(\max_{1\leq j\leq k}\int f_j^*(y)\max_{1\leq i\leq k}\left|\log\frac{f_i^*(y)}{f_i(y)}\right|\lambda(dy) + \log\left(\frac{1}{\underline{q}}\right)\right).\end{aligned} \tag{3.66}$$

□

3.5.3 Proof of Theorem 3.4 (discrete observations)

Assumption (B2) will be checked using Proposition 2 of Shen *et al.* [STG13] that we recall here.

Lemma 3.13. *[Proposition 2 of Shen* et al. *[STG13]] Let H be a positive integer, $\bar{z}$ and ϵ be positive, denote*

$$\mathcal{H}_{H,\bar{z},\epsilon} = \left\{ f = \sum_{h=1}^{+\infty}\pi_h\delta_{z_h} \ : \ \sum_{h>H}\pi_h < \epsilon, z_h \in [0,\bar{z}], h \leq H \right\}^k.$$

Then

$$(DP(G))^{\otimes k}\left(\mathcal{H}^c_{H,\bar{z},\epsilon}\right) \leq \frac{kH}{G(\mathbb{N})}G((\bar{z},+\infty)) + k\left(\frac{eG(\mathbb{N})}{H}\log\frac{1}{\epsilon}\right)^H \tag{3.67}$$

$$N(4\epsilon, \mathcal{H}_{H,\bar{z},\epsilon}, d) \lesssim (\bar{z}+1)^{kH}\epsilon^{-kH}. \tag{3.68}$$

We now give the proof of Theorem 3.4.

Proof of Theorem 3.4. We first prove Assumption (A2) with $\tilde{f}_j = f_j$, for all $1 \leq j \leq k$ using Lemma 3.14 with $\epsilon = \tilde{\epsilon}_n^2/u_n$, $L = L_n = \left(-\log\left(\tilde{\epsilon}_n^2/(u_n\log\log n)\right)/(c-\delta)\right)^{1/m}$ and $S_L = \{1,\dots,L\}$. Using that $f_i^* \in \mathcal{D}(m,c,K)$ for all $1 \leq i \leq k$ and Assumption (P2), we get

$$\begin{aligned}\sum_{l>L_n}\frac{f_i^*(l)}{(G(l))^2} &\lesssim \sum_{l>L_n}\exp(-cl^m)l^{2\alpha} \lesssim \sum_{l>L_n}\exp(-(c-\delta)l^m)l^{m-1} \\ &\lesssim \int_{L_n}^{\infty}\exp(-(c-\delta)x^m)x^{m-1}\lambda(dx) \lesssim \exp(-(c-\delta)L_n^m) \lesssim \tilde{\epsilon}_n^2\end{aligned}$$

which proves Equation (3.76). Equation (3.77) is proved similarly. Equation (3.75) follows using Assumption (I.2). Then, we can apply Lemma 3.14 so that

$$\begin{aligned}&DP(G)^{\otimes k}\Bigg(f \ : \forall 1\leq i,j\leq k \ \sum_{l=1}^{+\infty}f_i^*(l)\log^2\left(\frac{f_j^*(l)}{f_j(l)}\right) \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \\ &\sum_{l=1}^{L_n}\frac{\left(f_j^*(l)-f_j(l)\right)^2}{f_j^*(l)} \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \sum_{l>L_n}f_j(l) \leq \frac{\tilde{\epsilon}_n^2}{u_n}, \sum_{l=1}^{L_n}\frac{\left(f_j^*(l)-f_j(l)\right)^2}{f_j(l)} \leq \frac{\tilde{\epsilon}_n^2}{u_n}\Bigg) \\ &\gtrsim \prod_{j=1}^{k}\left(\left(\frac{\tilde{\epsilon}_n^2}{4u_n}\right)^{(L_n-1+G(\mathbb{N}))/2} f_j^*(l^*)^{L_n-2}\left(\frac{1}{3}\right)^{L_n}\prod_{l=1}^{L_n}G(l)f_j^*(l)^{G(l)}\right).\end{aligned} \tag{3.69}$$

Moreover using that $f_i^* \in \mathcal{D}(m, c, K)$ for all $1 \leq i \leq k$ and Equation (P2), we obtain

$$\log \prod_{l=1}^{L_n} G(l) \gtrsim -L_n \log(L_n), \tag{3.70}$$

$$\log \left(\prod_{l=1}^{L_n} f_j^*(l)^{G(l)} \right) \gtrsim \sum_{l=1}^{L_n} G(l) \log f_j^*(l) \gtrsim -L_n^K. \tag{3.71}$$

Combining Equations (3.69) with $l^* = \text{argmax}_l(\min_{1 \leq j \leq k} f_j^*(l))$, (3.70) and (3.71), Assumption (A2) of Theorem 3.1 is true if

$$(-\log(\tilde{\epsilon}_n))^{\max(1/m+1, K/m)} \lesssim n\tilde{\epsilon}_n^2.$$

Then we choose

$$\tilde{\epsilon}_n = \frac{1}{\sqrt{n}} (\log n)^{t_0}$$

with $2t_0 > \max(1/m + 1, K/m)$ and Assumption (A2) holds.

Using Assumption (Q2.0), for $\tilde{\epsilon}_n$ small enough,

$$\Pi_Q \left(\left\{ Q : \|Q - Q^*\| \leq \frac{\tilde{\epsilon}_n}{\sqrt{u_n}} \right\} \right) \geq \frac{\pi(Q^*)}{2} \lambda \left(\left\{ Q : \|Q - Q^*\| \leq \frac{\tilde{\epsilon}_n}{\sqrt{u_n}} \right\} \right) \gtrsim \left(\frac{\tilde{\epsilon}_n}{\sqrt{u_n}} \right)^{k(k-1)} \tag{3.72}$$

so that Assumption (C2) holds.

Using Lemma 3.13 with $\bar{z} = \exp((\log n)^{2t_0 + t/2})$, $H = (n\epsilon_n^2)/((\log n)^{2t_0+t})$ and $\epsilon_n = (\log n)^t/\sqrt{n}$,

$$\begin{aligned} &\Pi(\mathcal{F}_n^c) \\ &\lesssim (\log n)^{t-2t_0} \exp\left(-(\alpha - 1)(\log n)^{2t_0 + t/2}\right) + \exp\left(-(t - 2t_0 - 1)(\log n)^{t-2t_0} (\log\log n)^2\right) \\ &= o(\exp(-C'(\log n)^{2t_0})) = o(\exp(-C' n\tilde{\epsilon}_n^2)), \end{aligned} \tag{3.73}$$

if $t > 4t_0$. Moreover

$$\log\left(N\left(\frac{\epsilon_n}{12}, \mathcal{F}_n, D_l\right)\right) \lesssim (\log n)^{3t/2} + (\log n)^{t-2t_0+1} \lesssim n\epsilon_n^2 \tag{3.74}$$

so that Assumption (B2) holds. This concludes the proof of Theorem 3.4.

□

Lemma 3.14. *Let S_L be a subset of $\{1, \ldots, L\}$. If*

$$\max_{1 \leq i,j \leq k} \sum_{S_L^c} f_i^*(l) \log^2(f_j^*(l)) \leq \frac{\epsilon}{8} \tag{3.75}$$

and

$$\max_{1\leq i\leq k}\sum_{S_L^c}\frac{f_i^*(l)}{(G(l))^2}\leq\frac{\epsilon}{384\log^2(2)k(G(\mathbb{N}))^2}, \tag{3.76}$$

and if there exists $\delta>0$ *such that*

$$\max_{1\leq i\leq k}\sum_{l\in S_L^c}f_i^*(l)\leq\epsilon^{1+\delta}, \tag{3.77}$$

then, for all $l^*\in S_L$ *and all* $\epsilon>0$ *small enough,*

$$\begin{aligned}P_G:=DP(G)^{\otimes k}\Bigg(f\ :\forall 1\leq i,j\leq k\ \sum_{l=1}^{+\infty}f_i^*(l)\log^2\left(\frac{f_j^*(l)}{f_j(l)}\right)\leq\epsilon,\\ \sum_{l\in S_L}\frac{\left(f_j^*(l)-f_j(l)\right)^2}{f_j^*(l)}\leq\epsilon,\sum_{l\in S_L^c}f_j(l)\leq\epsilon,\sum_{l\in S_L}\frac{\left(f_j^*(l)-f_j(l)\right)^2}{f_j(l)}\leq\epsilon\Bigg)\\ \gtrsim\prod_{j=1}^k\left(\left(\sqrt{\frac{\epsilon}{4}}\right)^{L-1+G(\mathbb{N})}\left(\frac{1}{3}\right)^L f_j^*(l^*)^{L-2}\left[\prod_{l\in S_L}G(l)f_j^*(l)^{G(l)}\right]\right).\end{aligned}$$

Proof of Lemma 3.14. Note that if for all $l\in S_L$ and for all $1\leq j\leq k$,

$$\left(1-\sqrt{\frac{\epsilon}{4}}\right)f_j^*(l)\leq f_j(l)\leq\left(1+\sqrt{\frac{\epsilon}{4}}\right)f_j^*(l)$$

then for all $l\in S_L$,

$$\log^2\left(\frac{f_j^*(l)}{f_j(l)}\right)\leq\frac{\epsilon}{2},\quad\frac{|f_j^*(l)-f_j(l)|^2}{f_j^*(l)^2}\leq\frac{\epsilon}{4}\quad\text{and}\quad\frac{|f_j^*(l)-f_j(l)|^2}{f_j(l)^2}\leq\epsilon$$

so that

$$\sum_{l\in S_L}f_i^*(l)\log^2\left(\frac{f_j^*(l)}{f_j(l)}\right)\leq\frac{\epsilon}{2},\sum_{l\in S_L}\frac{\left|f_j^*(l)-f_j(l)\right|^2}{f_j^*(l)}\leq\epsilon\text{ and }\sum_{l\in S_L}\frac{\left|f_j^*(l)-f_j(l)\right|^2}{f_j(l)}\leq\epsilon.$$

Moreover using Assumptions (3.75) and (3.77) if for all $1\leq i,j\leq k$,

$$\sum_{l\in S_L^c}f_i^*(l)\log^2\left(f_j(l)\right)\leq\frac{\epsilon}{8},$$

then

$$\sum_{l\in S_L^c}f_i^*(l)\log^2\left(\frac{f_j^*(l)}{f_j(l)}\right)\leq 2\sum_{l\in S_L^c}f_i^*(l)\log^2(f_j^*(l))+2\sum_{l\in S_L^c}f_i^*(l)\log^2(f_j(l))\leq\frac{\epsilon}{2}.$$

Combining the two last remarks, we obtain

$$
\begin{aligned}
P_G \geq \prod_{j=1}^{k} \Bigg(DP(G) \Bigg(f_j \ : \ & \left(1-\sqrt{\frac{\epsilon}{4}}\right) f_j^*(l) \leq f_j(l) \leq \left(1+\sqrt{\frac{\epsilon}{4}}\right) f_j^*(l),\ \forall l \in S_L, \\
& \sum_{l \in S_L^c} f_i^*(l) \log^2\left(f_j(l)\right) \leq \frac{\epsilon}{8}, \forall 1 \leq i \leq k, \text{ and } \sum_{l \in S_L^c} f_j(l) \leq \epsilon \Bigg) \Bigg) \\
\geq \prod_{j=1}^{k} \Bigg(DP(G) \Bigg(f_j \ : \ & \exp\left(-\sqrt{\frac{\epsilon}{16 \max_{1 \leq i \leq k} \sum_{l \in S_L^c} f_i^*(l)}}\right) \leq \sum_{m \in S_L^c} f_j(m) \leq \epsilon, \\
& \sum_{l \in S_L^c} f_i^*(l) \log^2\left(\frac{f_j(l)}{\sum_{m \in S_L^c} f_j(m)}\right) \leq \frac{\epsilon}{32},\ \forall 1 \leq i \leq k \text{ and} \\
& \left(1-\sqrt{\frac{\epsilon}{4}}\right) \frac{f_j^*(l)}{\sum_{m \in S_L} f_j(m)} \leq \frac{f_j(l)}{\sum_{m \in S_L} f_j(m)} \leq \left(1+\sqrt{\frac{\epsilon}{4}}\right) \frac{f_j^*(l)}{\sum_{m \in S_L} f_j(m)},\ \forall l \in S_L \Bigg) \Bigg)
\end{aligned}
$$

indeed if

$$
\exp\left(-\sqrt{\frac{\epsilon}{16 \max_{1 \leq i \leq k} \sum_{l \in S_L^c} f_i^*(l)}}\right) \leq \sum_{m \in S_L^c} f_j(m)
$$

and

$$
\sum_{l \in S_L^c} f_i^*(l) \log^2\left(\frac{f(l)}{\sum_{m \in S_L^c} f(m)}\right) \leq \frac{\epsilon}{32},\ \forall 1 \leq i \leq k
$$

then

$$
\begin{aligned}
& \sum_{l \in S_L^c} f_i^*(l) \log^2\left(f_j(l)\right) \\
& \quad \leq 2 \sum_{l \in S_L^c} f_i^*(l) \log^2\left(\frac{f_j(l)}{\sum_{m \in S_L^c} f_j(m)}\right) + 2 \sum_{l \in S_L^c} f_i^*(l) \log^2\left(\sum_{m \in S_L^c} f_j(m)\right) \leq \frac{\epsilon}{8}.
\end{aligned}
$$

Using the tail free property of the Dirichlet process, $(f_j(l)/\sum_{m \in S_L^c} f_j(m))_{l \in S_L^c}, \sum_{m \in S_L^c} f_j(m)$

and $(f_j(1)/\sum_{m\in S_L} f_j(m), \ldots, f_j(L)/\sum_{m\in S_L} f_j(m))$ are independent. So that we obtain

$$P_G \geq \int_{\exp\left(-\sqrt{\frac{\epsilon}{16\max_{1\leq i\leq k}\sum_{l\in S_L^c} f_i^*(l)}}\right)} \frac{\Gamma(G(\mathbb{N}))}{\Gamma(G(S_L))\Gamma(G(S_L^c))} a^{G(S_L^c)-1}(1-a)^{G(S_L)-1}$$
$$\underbrace{Dir(G|_{S_L})\left(x \,:\, \left(1-\sqrt{\frac{\epsilon}{4}}\right)\frac{f_j^*(l)}{1-a} \leq x_l \leq \left(1+\sqrt{\frac{\epsilon}{4}}\right)\frac{f_j^*(l)}{1-a},\ \forall l\in S_L\right)}_{D(a)} \lambda(da)$$
$$DP(G|_{S_L^c})\left(\sum_{l\in S_L^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right) < \frac{\epsilon}{32},\ \forall 1\leq i\leq k\right). \tag{3.78}$$

We first control the integral of Equation (3.78). Note that if

$$x \in V_{SL} := \left\{x\in\Delta_{|S_L|} :\ x_l \in V_l,\ \forall l\in S_L\setminus\{l^*\}\right\}$$

where for all $l\in S_L\setminus\{l^*\}$

$$V_l := \left\{x_l :\ \left(1-\sqrt{\frac{\epsilon}{16}}f_j^*(l^*)\right)\frac{f_j^*(l)}{1-a} \leq x_l \leq \left(1+\sqrt{\frac{\epsilon}{16}}f_j^*(l^*)\right)\frac{f_j^*(l)}{1-a}\right\}$$

and if

$$a\in V_A := \left\{\sum_{l\in S_L} f_j^*(l) - \sqrt{\frac{\epsilon}{16}}f_j^*(l^*) \leq 1-a \leq \sum_{l\in S_L} f_j^*(l) + \sqrt{\frac{\epsilon}{16}}f_j^*(l^*)\right\} \tag{3.79}$$

then for all $l\in S_L$,

$$\left(1-\sqrt{\frac{\epsilon}{4}}\right)\frac{f_j^*(l)}{1-a} \leq x_l \leq \left(1+\sqrt{\frac{\epsilon}{4}}\right)\frac{f_j^*(l)}{1-a}, \tag{3.80}$$

where $x_{l^*} = 1-\sum_{l\in S_L, l\neq l^*} x_l$. So that

$$D(a) \geq \frac{\Gamma(G(S_L))}{\prod_{m\in S_L}\Gamma(G(m))}\mathbb{1}_{V_A}(a)\int_{V_{SL}}\left(1-\sum_{m\in S_L\setminus\{l^*\}} x_m\right)^{G(l^*)-1}\prod_{l\in S_L\setminus\{l^*\}} x_l^{G(l)-1}\lambda(dx) \tag{3.81}$$

where

$$\left(1-\sum_{m\in S_L\setminus\{l^*\}} x_m\right)^{G(l^*)-1} \geq \left(\frac{f_j^*(l^*)}{1-a}\right)^{G(l^*)-1}\min\left((1/2)^{G(l^*)-1}, (3/2)^{G(l^*)-1}\right) \tag{3.82}$$

and

$$\int_{V_l} x_l^{G(l)-1}\lambda(dx_l) \geq \left(\frac{f_j^*(l)}{1-a}\right)^{G(l)} f_j^*(l^*)\sqrt{\frac{\epsilon}{4}}\min\left((1/2)^{G(l)-1},(3/2)^{G(l)-1}\right), \tag{3.83}$$

using Equation (3.80). Then combining Equations (3.81), (3.82) and (3.83); $D(a)$ is bounded from below by

$$\mathbb{1}_{V_A}(a)\frac{\Gamma(G(S_L))}{\prod_{m\in S_L}\Gamma(G(m))}\left(\frac{2}{3}\right)^L\left(\frac{1}{2}\right)^{G(\mathbb{N})}\left(\frac{\epsilon}{4}\right)^{(L-1)/2}\left(\frac{1}{1-a}\right)^{G(S_L)-1} f_j^*(l^*)^{L-2}\prod_{l\in S_L} f_j^*(l)^{G(l)}.$$

So that

$$\begin{aligned}
&\int_{\exp\left(-\sqrt{\frac{\epsilon}{16\max_{1\leq i\leq k}\sum_{l\in S_L^c} f_i^*(l)}}\right)}^{\epsilon} \frac{\Gamma(G(\mathbb{N}))}{\Gamma(G(S_L))\Gamma(G(S_L^c))} a^{G(S_L^c)-1}(1-a)^{G(S_L)-1}D(a)\lambda(da)\\
&\geq \frac{\Gamma(G(\mathbb{N}))}{G(S_L^c)\Gamma(G(S_L^c))\prod_{l\in S_L}\Gamma(G(l))}\left(\frac{\epsilon}{4}\right)^{(L-1)/2}\left(\frac{2}{3}\right)^L\left(\frac{1}{2}\right)^{G(\mathbb{N})} f_j^*(l^*)^{L-2}\prod_{l\in S_L} f_j^*(l)^{G(l)}\\
&\quad\left(\left(\min\left(\sum_{m\in S_L^c} f_j^*(m)+\sqrt{\frac{\epsilon}{16}}f_j^*(l^*),\epsilon\right)\right)^{G(S_L^c)}\right.\\
&\quad\left.-\left(\max\left(\exp\left(-\sqrt{\epsilon/\left(16\max_{1\leq i\leq k}\sum_{l\in S_L^c} f_i^*(l)\right)}\right),\sum_{m\in S_L^c} f_j^*(m)-\sqrt{\frac{\epsilon}{16}}f_j^*(l^*)\right)\right)^{G(S_L^c)}\right)\\
&\gtrsim \left(\sqrt{\frac{\epsilon}{4}}\right)^{L-1+G(\mathbb{N})}\left(\frac{1}{3}\right)^L f_j^*(l^*)^{L-2}\prod_{l\in S_L} G(l)f_j^*(l)^{G(l)}.
\end{aligned} \tag{3.84}$$

using that for all $0<a<1$,

$$\frac{1}{a}\leq\Gamma(a)\leq\frac{2}{a}. \tag{3.85}$$

and that under Assumption (3.77), for ϵ small enough

$$\exp\left(-\frac{\epsilon^{-\delta/2}}{4}\right)\geq\exp\left(-\sqrt{\frac{\epsilon}{16\max_{1\leq i\leq k}\sum_{l\in S_L^c} f_i^*(l)}}\right)\geq 0\geq\sum_{m\in S_L^c} f_j^*(m)-\sqrt{\frac{\epsilon}{16}}f_j^*(l^*).$$

We now control the last term of Equation (3.78). Using Markov's inequality,

$$\begin{aligned} & DP(G|_{S_L^c})\left(\sum_{l\in S_l^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right) < \frac{\epsilon}{32},\ \forall 1\leq i\leq k\right) \\ & \geq 1 - DP(G|_{S_L^c})\left(\sum_{i=1}^{k}\sum_{l\in S_L^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right) > \frac{\epsilon}{32}\right) \\ & \geq 1 - \frac{\mathbb{E}^{\theta^*}\left(\sum_{i=1}^{k}\sum_{l\in S_L^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right)\right)}{\frac{\epsilon}{32}}. \end{aligned} \tag{3.86}$$

As $f_j(l)/\sum_{m\in S_L^c} f_j(m)$ is distributed from $\beta\big(G(l), G(S_L^c\setminus\{l\})\big)$,

$$\begin{aligned} & \mathbb{E}^{\theta^*}\left[\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right)\right] \\ & = \frac{\Gamma(G(S_L^c))}{\Gamma(G(l))\Gamma(G(S_L^c\setminus\{l\}))}\ \Bigg(\underbrace{\int_0^{1/2}\log^2(x)x^{G(l)-1}(1-x)^{G(S_L^c\setminus\{l\})-1}\lambda(dx)}_{I_1} \\ & \qquad + \underbrace{\int_{1/2}^{1}\log^2(x)x^{G(l)-1}(1-x)^{G(S_L^c\setminus\{l\})-1}\lambda(dx)}_{I_2}\Bigg), \end{aligned} \tag{3.87}$$

with

$$\begin{aligned} I_1 & \leq 2\int_0^{1/2}\log^2(x)x^{G(l)-1}\lambda(dx) = \frac{4\log^2(2)(1/2)^{G(l)}}{G(l)^3}\left(\frac{G(l)^2}{2} + \frac{G(l)}{\log 2} + \frac{1}{\log^2 2}\right) \\ & \leq \frac{12\log^2(2)(G(\mathbb{N}))^2}{G(l)^3}, \end{aligned} \tag{3.88}$$

and

$$I_2 \leq 2\log^2(2)\int_{1/2}^{1}(1-x)^{G(S_L^c\setminus\{l\})-1}\lambda(dx) \leq \frac{2\log^2(2)}{G(S_L^c\setminus\{l\})}. \tag{3.89}$$

Combining Equations (3.85) (3.87), (3.88) and (3.89), we obtain

$$\mathbb{E}^{\theta^*}\left(\sum_{l\in S_L^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right)\right) \leq 24G(\mathbb{N})^2\log^2(2)\sum_{l\in S_L^c}\frac{f_i^*(l)}{(G(l))^2}. \tag{3.90}$$

Then using Assumption (3.76) and Equations (3.86) and (3.90)

$$DP(G|_{S_L^c})\left(\sum_{l\in S_L^c} f_i^*(l)\log^2\left(\frac{f_j(l)}{\sum_{m\in S_L^c} f_j(m)}\right) < \frac{\epsilon}{16k},\ \forall 1\leq i\leq k\right) \geq 1/2 \tag{3.91}$$

Lemma 3.14 follows combining Equations (3.78), (3.84) and (3.91). □

3.5.4 Proof of Theorem 3.6 (Dirichlet process mixtures of Gaussian distributions)

Proof of Theorem 3.6 . Let $\sigma_n = \tilde{\epsilon}_n/(\log(1/\tilde{\epsilon}_n))$, $\tilde{\epsilon}_n = n^{-\beta/(2\beta+1)}(\log n)^{t_0}$. Following the computations of the proof of Theorem 4 of Shen *et al.* [STG13] and using Assumption (S2.3), Lemma 3.15 ensures that Assumption (A2) holds with $t_0 \geq (2+2/\gamma_0+1)/(1/\beta+2)$. Using Assumption (Q2.0), Assumption (C2) holds. Using Theorem 5 of Shen *et al.* [STG13], Assumptions (G2.1), (S2.1) and (S2.2); Assumption (B2) holds with $\epsilon_n = n^{-\beta/(2\beta+1)}(\log(n))^t$, $t > t_0$. This concludes the proof of Theorem 3.6. □

The following lemma is a generalization of Lemma 4 of Kruijer *et al.* [KRV10] in the HMM context. In other words, we give a set of density functions $(f_j)_{1\leq j\leq k}$ satisfying Assumptions (A2.1)–(A2.6) in Lemma 3.15.

Lemma 3.15. *Assume that there exist β, L and γ such that for all $1 \leq j \leq k$, $f_j^* \in \mathcal{P}(\beta, L, \gamma)$ and Assumptions (T2.1)–(T2.3) hold. Let σ be a positive real small enough.*

Then for all $1 \leq j \leq k$, there exists a discrete measure $m_j = \sum_{i=1}^{N_j} \mu_j^i \delta_{z_j^i}$ supported on $\{x : f_j^*(x) \geq K_j\sigma^{2\beta+H_1}\}$ with $H_1 > 2\beta$, K_j a constant small enough and $N_j = O(\sigma^{-1}|\log\sigma|^{2/\gamma_0})$ such that Assumptions (A2.1)–(A2.6) hold with $f_j = \phi_\sigma * m_j$ for all $1 \leq j \leq k$ and $\sigma^{2\beta} \leq \tilde{\epsilon}_n^2/u_n$.

Assumptions (A2.1)–(A2.6) also hold with $f_j = \phi_{\tilde{\sigma}} * \tilde{m}_j$, for all $\tilde{\sigma} \in [\sigma, \sigma + \sigma^{\delta' H_1+2}]$ and for all $\tilde{m}_j = \sum_{i=1}^{+\infty} \tilde{\mu}_j^i \delta_{\tilde{z}_j^i}$ such that $\tilde{\mu}_j^{1:N_j} \in \mathcal{B}(\mu_j, \sigma^{\delta' H_1+2})$ and $\tilde{z}_j^i \in \mathcal{B}(z_j^i, \sigma^{\delta' H_1+2})$, for all $1 \leq i \leq N_j$, where $\delta' \geq 1 + \beta/H_1$.

Proof of Lemma 3.15. The proof of Lemma 3.15 is based on Kruijer *et al.* [KRV10]. First notice that $f_j^* \in \mathcal{P}(\beta, \bar{g}, \gamma)$ implies that for all integer $m \leq \beta$, $|g_j^m|$, where $g_j^m = (\log f_j)^{(m)}$, is bounded by a polynomial. Then Assumption (T2.1) implies that Assumption (C2) of Kruijer *et al.* [KRV10] holds and stronglier implies that there exists $\delta > 0$ such that for all $1 \leq i, j, \iota \leq k$ and all integer $m \leq \beta$,

$$\begin{aligned}
&\int |g_j^m(x)|^{(2\beta+\delta)/m} f_i^*(x)\lambda(dx) < \infty,\\
&\int |\bar{g}(x)|^{2+\delta/\beta} f_i^*(x)\lambda(dx) < \infty,\\
&\int |g_j^m(x)|^{(2\beta+\delta)/m} f_i^*(x)\log f_\iota^*(x)\lambda(dx) < \infty,\\
&\int |\bar{g}(x)|^{2+\delta/\beta} f_i^*(x)\log f_\iota^*(x)\lambda(dx) < \infty.
\end{aligned} \tag{3.92}$$

Let $\sigma > 0$, we consider

$$S = \bigcap_{j=1}^{k} \left(A_\sigma^j \cap E_\sigma^j \right),$$

where

$$A_\sigma^j = \left\{ x \ : \ |l_m^j(x)| \leq B\sigma^{-m}|\log(\sigma)|^{-m/2}, \forall 1 \leq m \leq \lfloor \beta \rfloor, \ |\bar{g}(x)| \leq B\sigma^{-\beta}|\log(\sigma)|^{-\beta/2} \right\}$$

and

$$E_\sigma^j = \left\{ x \ : \ f_j^*(x) \geq \sigma^{H_1} \right\}.$$

Using Assumptions 3.16, (T2.1), (T2.3) and Lemma 2 of Kruijer *et al.* [KRV10], there exists k density functions h_β^j such that for all $1 \leq j \leq k$ and all $x \in S$,

$$h_\beta^j * \phi_\sigma(x) = f_j^*(x) \left(1 + O(R^j(x)\sigma^\beta) \right) + O((1 + R^j(x))\sigma^H) \tag{3.93}$$

where R^j is defined as in Equation (16) page 1232 of Kruijer *et al.* [KRV10] and H is as large as we want. Using Assumptions 3.16, (T2.1) and (T2.3), the proof of Lemma 2 of Kruijer *et al.* [KRV10] is easily generalizable in this context so that

$$\int_{S^c} h_\beta^j * \phi_\sigma(y)\lambda(dy) \lesssim \sigma^{2\beta}. \tag{3.94}$$

The generalization can be proved using Equation (3.92) and by replacing Equation (56) of Kruijer *et al.* [KRV10] by Equation (3.101).

As at page 1251 in Kruijer *et al.* [KRV10], we denote

$$\tilde{h}_j^\beta = \frac{\mathbb{1}_{x:h_j^\beta(x) \geq \sigma^{H_2}} h_j^\beta}{\int_{x:h_j^\beta(x) \geq \sigma^{H_2}} h_j^\beta d\lambda}$$

and using Lemma 12 of Kruijer *et al.* [KRV10], for all $1 \leq j \leq k$, there exist k discrete distributions $m_j = \sum_{i=1}^{N_j} \mu_j^i \delta_{z_j^i}$ supported in $\{x \ : \ f_j^*(x) \geq K_j\sigma^{2\beta+H_1}\}$ for $H_1 > 2\beta$, a constant K_j small enough and with $N_j = O(\sigma^{-1}|\log\sigma|^{2/\gamma_0})$ such that

$$\begin{aligned} &\|\tilde{h}_j^\beta * \phi_\sigma - m_j * \phi_\sigma\|_\infty \leq \sigma^{-1} e^{-C|\log\sigma|^{2/\gamma_0}}, \\ &\|\tilde{h}_j^\beta * \phi_\sigma - m_j * \phi_\sigma\|_1 \leq \sigma^{-1} e^{-C'|\log\sigma|^{2/\gamma_0}} \end{aligned} \tag{3.95}$$

for any C', C large enough.

It is now sufficient to prove that Assumptions (A2.1) to (A2.6) hold with $\epsilon_n^{4/(2-\alpha)} = O(\sigma^{2\beta})$, $f_j = \tilde{m}_j * \phi_{\tilde{\sigma}}$ and $\tilde{f}_j = h_j^\beta * \phi_{\tilde{\sigma}}$ for all $\tilde{\sigma} \in [\sigma, \sigma + \sigma^{\delta' H_1 + 2}]$ and all discrete distributions $\tilde{m}_j = \sum_{i=1}^{N_j} \tilde{\mu}_j^i \delta_{\tilde{z}_j^i}$ such that $\tilde{\mu}_j \in \mathcal{B}(\mu_j, \sigma^{\delta' H_1 + 2}) \cap \Delta_{N_j}$ and $\tilde{z}_j \in \mathcal{B}(z_j, \sigma^{\delta' H_1 + 2})$ where $\delta' \geq 1 + \beta/H_1$.

By Lemma 2 of Kruijer *et al.* [KRV10],

$$\|f_j - m_j * \phi_\sigma\|_1 \lesssim \sigma^{H_1+1+\beta}, \quad \|f_j - m_j * \phi_\sigma\|_\infty \lesssim \sigma^{H_1+\beta}, \tag{3.96}$$

- Proof of (A2.1). We cut the integral as in the following,

$$\begin{aligned}\int f_i^*(y) \log^2 \frac{f_j^*(y)}{f_j(y)} \lambda(dy) \leq & \int_S f_i^*(y) \log^2 \frac{f_j^*(y)}{\tilde{f}_j(y)} \lambda(dy) \\ & + 2 \int_{S^c} f_i^*(y) \log^2 \frac{f_j^*(y)}{\tilde{f}_j(y)} \lambda(dy) \\ & + 2 \int f_i^*(y) \log^2 \frac{\tilde{f}_j(y)}{f_j(y)} \lambda(dy).\end{aligned} \tag{3.97}$$

  The last integral can be controlled by $O(\sigma^{2\beta})$ as in the proof of Lemma 4 of Kruijer *et al.* [KRV10].

  Using Equation (3.93) we control the first integral of the bound of Equation (3.97):

$$\int_S f_i^*(y) \log^2 \frac{f_j^*(y)}{\tilde{f}_j(y)} \lambda(dy) \leq \int_S f_i^*(y) \frac{|f_j^*(y) - \tilde{f}_j(y)|^2}{\min(\tilde{f}_j(y), f_j^*(y))} \lambda(dy) \lesssim \sigma^{2\beta}$$

  as soon as H is large enough, using Equation (3.93).

  Using that $f_j^* \lesssim \tilde{f}_j \lesssim \phi_\sigma * f_j^* \lesssim 1$ (see Remark 1 and the bottom of page 1252 of Kruijer *et al.* [KRV10]), we control the second integral in the bound of Equation (3.97),

$$\int_{S^c} f_i^*(y) \log^2 \frac{f_j^*(y)}{\tilde{f}_j(y)} \lambda(dy) \lesssim \int_{S^c} f_i^*(y) \lambda(dy) + \int_{S^c} f_i^*(y) \log^2 \left(f_j^*(y)\right) \lambda(dy)$$

  which is bounded by $O(\sigma^{2\beta})$ following the proof of (A2.4).

- Proof of (A2.2). We cut the integral into three parts:

$$\begin{aligned}&\int_S \frac{|f_i^*(y) - f_i(y)|^2}{f_i^*(y)} \lambda(dy) \\ &\lesssim \underbrace{\int_S \frac{|f_i^*(y) - h_\beta^i * \phi_\sigma(y)|^2}{f_i^*(y)} \lambda(dy)}_{I_1} + \underbrace{\int_S \frac{|h_\beta^i * \phi_\sigma(y) - \tilde{h}_\beta^i * \phi_\sigma(y)|^2}{f_i^*(y)} \lambda(dy)}_{I_2} \\ &\quad + \underbrace{\int_S \frac{|\tilde{h}_\beta^i * \phi_\sigma(y) - f_i(y)|^2}{f_i^*(y)} \lambda(dy)}_{I_3}\end{aligned}$$

  Using Equation (3.93),

$$I_1 \lesssim \sigma^{2\beta}. \tag{3.98}$$

We now control I_2 using the bound

$$\left| \frac{h^i_\beta * \phi_\sigma(y)}{\tilde{h}^i_\beta * \phi_\sigma(y)} - 1 \right| = O(\sigma^{2\beta}), \text{ for all } y \in S$$

of page 1252 of Kruijer *et al.* [KRV10] and Equation (3.93). Then

$$\begin{aligned} I_2 &= \int_S \left(\frac{|h^i_\beta * \phi_\sigma(y) - \tilde{h}^i_\beta * \phi_\sigma(y)|}{h^i_\beta * \phi_\sigma(y)} \right)^2 h^i_\beta * \phi_\sigma(y) \frac{h^i_\beta * \phi_\sigma(y)}{f^*_i(y)} \lambda(dy) \\ &\leq \int_S (O(\sigma^{2\beta}))^2 \, h^i_\beta * \phi_\sigma(y) \; 2 \; \lambda(dy) \lesssim \sigma^{2\beta}, \end{aligned} \tag{3.99}$$

using Equation (3.93). As to I_3, using Equations (3.95) and (3.96), it is upper-bounded by

$$2 \frac{\|\tilde{h}^\beta_i * \phi_\sigma - m_i * \phi_\sigma\|_\infty \|\tilde{h}^\beta_i * \phi_\sigma - m_i * \phi_\sigma\|_1}{\sigma^{H_1}} + 2 \frac{\|f_i - m_i * \phi_\sigma\|_\infty \|f_i - m_i * \phi_\sigma\|_1}{\sigma^{H_1}} \lesssim \sigma^{2\beta} \tag{3.100}$$

when $2 > \gamma_0$, such a γ_0 can always be chosen (see the first line of page 1253 of Kruijer *et al.* [KRV10]).

- Proof of (A2.3). Assumption (A2.3) is proved in Equation (3.94).
- Proof of (A2.4). It is sufficient to bound

$$\int_{(E^i_\sigma)^c} f^*_j(y) \lambda(dy)$$

and

$$\int_{(A^i_\sigma)^c} f^*_j(y) \lambda(dy).$$

Using Assumption (T2.2), for all $0 < \delta < 1$

$$\begin{aligned} &\int_{(E^i_\sigma)^c} f^*_j(y) \lambda(dy) \\ &\leq \int_{\{y : f^*_j(y) < \sigma^{H_1} M_{j,i} \exp(\tau_{j,i} |y|^{\gamma_{j,i}})\}} f^*_j(y) \lambda(dy) \\ &\leq \int_{\{y : f^*_j(y) < \sigma^{H_1} M_{j,i} \exp(\tau_{j,i} |y|^{\gamma_{j,i}})\}} (f^*_j(y))^{1/\delta} (f^*_j(y))^{1-1/\delta} \lambda(dy) \\ &\leq \sigma^{H_1/\delta} \int (M_{j,i})^{1/\delta} \exp(\tau_{j,i} |y|^{\gamma_{j,i}}/\delta) (f^*_j(y))^{1-1/\delta} \lambda(dy) \lesssim \sigma^{2\beta} \end{aligned} \tag{3.101}$$

as soon as $H_1 > 2\beta$, using Assumption (T2.1). Moreover using (3.92) and Markov inequal-

ity, as in the proof of Lemma 2 of Kruijer *et al.* [KRV10],

$$\int_{(A_\sigma^i)^c} f_j^*(y)\lambda(dy) \lesssim \sigma^{2\beta}. \tag{3.102}$$

- Proof of (A2.5). Using the same argument as in the bottom of the page 1252 of Kruijer *et al.* [KRV10], Equations (3.95) and (3.96),

$$\begin{aligned} &\int_S f_i^*(y) \max_{1\le j\le k} \log\left(\frac{\tilde{f}_j(y)}{f_j(y)}\right)\lambda(dy) \\ &\le \int_S f_i^*(y) \max_{1\le j\le k} \frac{|\tilde{f}_j(y)-f_j(y)|}{f_j(y)}\lambda(dy) \\ &\le \int_S f_i^*(y) \max_{1\le j\le k} \frac{\|\tilde{f}_j-f_j\|_\infty}{\sigma^{H_2}-\|\tilde{f}_j-f_j\|_\infty}\lambda(dy) \lesssim \sigma^{2\beta} \end{aligned} \tag{3.103}$$

- Proof of (A2.6). Using that $f_j^* \lesssim \tilde{f}_j$ (see Assumption (C3) of Kruijer *et al.* [KRV10]) Equation (3.98) implies (A2.6).

□

CHAPTER 4

EFFICIENT SEMIPARAMETRIC ESTIMATION AND MODEL SELECTION FOR MULTIDIMENSIONAL MIXTURES

This is a joint work with E. Gassiat and J. Rousseau.

In this chapter, we consider nonparametric multidimensional finite mixture models and we are interested in the semiparametric estimation of the population weights. Here, the i.i.d. observations are assumed to have at least three components which are independent given the population. We approximate the semiparametric model by projecting the conditional distributions on step functions associated to some partition. Our first main result is that if we refine the partition slowly enough, the associated sequence of maximum likelihood estimators of the weights is asymptotically efficient, and the posterior distribution of the weights, when using a Bayesian procedure, satisfies a semiparametric Bernstein von Mises theorem. We then propose a cross-validation like procedure to select the partition in a finite horizon. Our second main result is that the proposed procedure satisfies an oracle inequality. Numerical experiments on simulated data illustrate our theoretical results.

4.1 Introduction

We consider in this chapter multidimensional mixture models that describe the probability distribution of a random vector Y with at least three coordinates. The model is a probability mixture of k populations such that, given the population, the coordinates of the random vector are independently distributed. We call emission distributions the conditional distributions of the coordinates and θ the parameter that contains the probability weights of each population. It has been known for some time that such a model is identifiable. An algebraic result by Kruskal Kruskal [Kru77] in 1977 (see also Rhodes [Rho10]) proved it when the coordinates of Y take finitely many values. Kruskal's result was recently used by Allman *et al.* [AMR09] to obtain identifiability under almost no assumption on the possible emission distributions: only the fact that, for each coordinate, the k emission distributions are linearly independent. Spectral methods were proposed by Anandkumar *et al.* [AGH+14], which allowed Bonhomme *et al.* [BJR16a] to derive estimators of the emission densities having the minimax rate of convergence when the smoothness of the emission densities is known. Moreover, Bonhomme *et al.* [BJR16a] proposes an estimation procedure in the case of repeated measurements (where the emission distributions of each coordinate given a population are the same).

Chapter 4 focusses on the semiparametric estimation of the population weights when nothing is known about the emission distributions. This is a semiparametric model, where the finite dimensional parameter of interest is θ and the infinite dimensional nuisance parameters are the emission distributions.

We are in particular interested in constructing optimal procedures for the estimation of θ. Optimal may be understood as efficient, in Le Cam's theory point of view which is about asymptotic distribution and asymptotic (quadratic) loss. See [LY00], Bickel *et al.* [BKRW98], van der Vaart [Vaa98], van der Vaart [Vaa02]. The first question is: is the parametric rate attainable in the semiparametric setting? We know here, for instance using spectral estimates, that the parametric rate is indeed attainable. Then, the loss due to the nuisance parameter may be seen in the efficient Fisher information and efficient estimators are asymptotically equivalent to the empirical process on efficient influence functions. The next question is thus: how can we construct asymptotically efficient estimators? In the parametric setting, maximum likelihood estimators (m.l.e.'s) do the job, but the semiparametric situation is more difficult, because one has to deal with the unknown nuisance parameter, see Theorems in chapter 24 of van der Vaart [Vaa98] where it is necessary to control various bias/approximation terms.

From a Bayesian perspective, the issue is the validity of the Bernstein-Von Mises property of the marginal posterior distribution of the parameter of interet θ. In other words: is the marginal posterior distribution of θ asymptotically Gaussian? Is it asymptotically centered around an efficient estimator? Is the asymptotic variance of the posterior distribution the inverse of the efficient Fisher information matrix? Semiparametric Bernstein-Von Mises theorems have been

the subject of recent research, see Shen [She02], Boucheron and Gassiat [BG09], Rivoirard and Rousseau [RR12a], Castillo [Cas12a], Castillo [Cas12b], Bickel and Kleijn [BK12b], De Blasi and Hjort [DH09] and Rivoirard and Rousseau [RR12a].

The results of Chapter 4 are twofold: first we obtain asymptotically efficient semiparametric estimators using a likelihood strategy, then we propose a data driven method to perform the strategy in a finite horizon with an oracle inequality as theoretical guarantee.

Let us describe our ideas. For the multidimensional mixture model we consider, we will take advantage of the fact that, for some finite approximations of the nuisance parameter, the model is still valid for the observation process. This may be seen as a *no bias* situation. Indeed, when approximating the emission densities by step functions, the density of the observation is the multinomial distribution of the indicator function of the sets in the partition. Hence, this is a common and fairly crude modelling of densities by histograms. The no bias property of this modelling implies that, for each of these finite dimensional models, the parameter of interest, i.e. the weights of the mixture, may be efficiently estimated within the finite dimensional model. Then, under weak assumptions, and using the fact that one can approximate any density on $[0, 1]$ by such histograms based on partitions with radius (i.e. the size of the largest bin) going to zero, it is possible to prove that asymptotically efficient semiparametric estimators may be built using the sequence of m.l.es in a growing (with sample size) sequence of approximation models. In the same way, using Bayesian posteriors in the growing sequence of approximation models, one gets a Bernstein-Von Mises result. One of the important implications of the Bernstein von Mises property is that credible regions, such as highest posterior density regions or credible ellipses are also confidence regions. In the particular case of the semiparametric mixtures, this is of great interest, since the construction of a confidence region is not necessarily trivial. This is our first main result which is stated in Theorem 4.5: by considering partitions refined slowly enough when the number of observations increases, we can derive efficient estimation procedures for the parameter of interest θ and in the Bayesian approach for a marginal posterior distribution on θ which satisfies the renown Bernstein von Mises property.

We still need however in practice to choose a good partition, for a finite sample size. This can be viewed as a model selection problem. There is now a huge literature on model selection, both in the frequentist and in the Bayesian literature. Roughly speaking the methods can be split into two categories: penalized likelihood types of approaches, which include in particular AIC, BIC, MDL and marginal likelihood (Bayesian) criteria or approaches which consist in estimating the risk of the estimator in each model using for instance bootstrap or cross validation methods. In all these cases theory and practice are nowadays well grounded, see for instance Hansen and Yu [HY01], Robert [Rob01], Barbe and Bertail [BB95], Massart [Mas07], Baudry *et al.* [BMM12], Arlot and Celisse [AC10], Claeskens and Hjort [CH08], Ando [And10]. Most of the existing results above cover parametric or nonparametric models. Penalized likelihoods in particular

target models wich are best in terms of Kullback-Leibler divergences typically and therefore aim at estimating the whole nonparametric parameter. Risk estimation via bootstrap or cross - validation methods are more naturally defined in semiparametric (or more generally setups with nuisance parameters) models, however the theory remains quite limited in cases where the estimation strategy is strongly nonlinear as encountered here.

In our context, the natural risk for θ is the quadratic risk, which can not be written as some risk of the distribution of the observations, which is the basic stone in the theory of model selection based on risk estimation. To propose specific procedures, one has thus to find some way to estimate the risk of the estimator in each approximation model, and then select the model with the smallest estimated risk. We propose to use a cross-validation method similar to the one proposed in Brookhart and van der Laan [BL06]. To get theoretical results on such a strategy, the usual basic tool is to write the cross-validation criterion as a function of the empirical distribution which is not possible in our semiparametric setting. We thus divide the sample in nonoverlapping blocks of size a_n (n being the the sample size) to define the cross validation criterion. This enables us to prove our second main result: Theorem 4.8 which states an oracle inequality on the quadratic risk associated with a sample of size a_n observations, and which also leads to criterion to select a_n. Simulations indicate moreover that the approach behaves well in practice.

In Section 4.2, we first describe the model, set the notations and our basic assumptions. We recall the semiparametric tools in Section 4.2.2, where we define the score functions and the efficient information matrices. Using the fact that spectral estimators are smooth functions of the empirical distribution of the observations, we obtain that, for large enough approximation model, the efficient Fisher information matrix is full rank, see Proposition 4.1. Intuition says that with better approximation spaces, more is known about all parameters of the distribution, in particular about θ. We prove in Proposition 4.2 that indeed the efficient Fisher information matrix increases when the partition is refined. We are finally able to prove our main general result in Section 4.2.3. In Lemma 4.3, we first prove that semiparametric score functions and semiparametric efficient Fisher information matrix are the limits of the parametric ones obtained in the approximation parametric models. Thus, the fact that the semiparametric efficient Fisher information matrix is full rank is a consequence of previous results and stated in Proposition 4.4. In Theorem 4.5, we prove that it is possible to let the approximation parametric models grow with the sample size so that the sequence of maximum likelihood estimators are asymptotically efficient in the semiparametric model and so that a semiparametric Bernstein - von Mises Theorem holds. In Section 4.3, we first discuss in Section 4.3.1 the reasons to perform model selection and the fact that choosing a too large approximation space does not work, see Proposition 4.6 and Corollary 4.7. Then we propose in Section 4.3.2 our cross-validation criterion, for which we prove an oracle inequality in Theorem 4.8 and Proposition 4.9. Results of simulations are described in Section 4.4, we investigate several choices of the number and length of blocks for performing cross validation, and investigate practically also V-fold strategies. We discuss possible extensions,

open questions and further work in Section 4.5. Finally Section 4.6 is dedicated to proofs of intermediate propositions and lemmas.

4.2 Asymptotic Efficiency

4.2.1 Model and Notations

Let $(Y_n)_{n\geq 1}$ be a sequence of independent and identically distributed randoms variables taking values in $[0,1]^3$. We assume the possible marginal distribution of an observation Y_n, $n \geq 1$ is a population mixture of k distributions such that, given the population, the coordinates are independent and have some density with respect to the Lebesgue measure on $[0,1]$. The possible densities of Y_n, $n \geq 1$, are, if $\mathbf{y} = (y_1, y_2, y_3) \in [0,1]^3$:

$$g_{\theta,\mathbf{f}}(\mathbf{y}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} f_{j,c}(y_c), \quad \sum_{j=1}^{k} \theta_j = 1, \quad \theta_j \geq 0, \, \forall j \tag{4.1}$$

Here, k is the number of populations, θ_j is the probability to belong to population j for $j \leq k$ and we set $\theta = (\theta_1, \ldots, \theta_{k-1})$. For each $j = 1, \ldots, k$, $f_{j,c}$, $c = 1, 2, 3$, is the density of the c-th coordinate of the observation, given the observation coming from population j and we set $\mathbf{f} = ((f_{j,c})_{1\leq c\leq 3})_{1\leq j\leq k}$. We denote by $\mathbb{P}^\star$ the true (unknown) distribution of the sequence $(Y_n)_{n\geq 1}$, such that $\mathbb{P}^\star = P_{\theta^\star,\mathbf{f}^\star}^{\otimes\mathbb{N}}$, $dP_{\theta^\star,\mathbf{f}^\star}(\mathbf{y}) = g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y})d\mathbf{y}$, for some $\theta^\star \in \Theta$ and $\mathbf{f}^\star \in \mathcal{F}^{3k}$, where Θ is the set of possible parameters θ and $\mathcal{F}$ the set of probability densities on $[0,1]$.

We approximate the densities by step functions on some partitions of $[0,1]$. We assume that we have a collection of partitions $\mathcal{I}_M$, $M \in \mathcal{M}$, $\mathcal{M} \subset \mathbb{N}$, so that for each $M \in \mathcal{M}$, $\mathcal{I}_M = (I_m)_{1\leq m\leq M}$ is a partition of $[0,1]$ by borelian sets. It is clear that I_m changes when M changes. For each $M \in \mathcal{M}$, we now consider the model of possible densities

$$g_{\theta,\omega;M}(\mathbf{y}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \left(\sum_{m=1}^{M} \frac{\omega_{j,c,m}}{|I_m|} \mathbb{1}_{I_m}(y_c) \right). \tag{4.2}$$

Here, $\omega = (((\omega_{j,c,m})_{1\leq m\leq M-1})_{1\leq c\leq 3})_{1\leq j\leq k}$, and for each $j = 1, \ldots, k$, each $c = 1, 2, 3$, each $m = 1, \ldots, M-1$, $\omega_{j,c,m} \geq 0$, $\sum_{m=1}^{M-1} \omega_{j,c,m} \leq 1$, and we denote $\omega_{j,c,M} = 1 - \sum_{m=1}^{M-1} \omega_{j,c,m}$. Thus, $\omega_{j,c,m}$ may be thought of as

$$\omega_{j,c,m} = \int_0^1 f_{j,c} \mathbb{1}_{I_m}(u)du.$$

We denote Ω_M the set of possible parameters ω when using model (4.2) with the partition $\mathcal{I}_M$.

Let $\ell_n(\theta, \omega; M)$ be the log-likelihood using model (4.2), that is

$$\ell_n(\theta, \omega; M) = \sum_{i=1}^{n} \log g_{\theta,\omega;M}(Y_i).$$

It appears as the model of population mixture of multinomial distributions for the observations $U_i := ((\mathbb{1}_{I_m}(Y_{i,c}))_{1\le m\le M})_{1\le c\le 3}$, for which the true (unknown) parameter is given by

$$\theta = \theta^\star,\ \omega = \omega_M^\star := \left(\left(\left(\int_0^1 f_{j,c}^\star \mathbb{1}_{I_m}(u)du \right)_{1\le m\le M-1} \right)_{1\le c\le 3} \right)_{1\le j\le k}.$$

We denote, for each $M \in \mathcal{M}$, $(\widehat{\theta}_M, \widehat{\omega}_M)$ the m.l.e., that is a maximizer of $\ell_n(\theta, \omega; M)$ over $\Theta \times \Omega_M$.

Let Π_M denote a prior distribution, that is a probability distribution on the parameter space $\Theta \times \Omega_M$. The posterior distribution $\Pi_M(\cdot|Y_1, \dots, Y_n)$ is defined as follows. For any borelian subset A of $\Theta \times \Omega_M$,

$$\Pi_M(A|Y_1, \dots, Y_n) = \frac{\int_A \prod_{i=1}^n g_{\theta,\omega;M}(Y_i) d\Pi_M(\theta, \omega)}{\int_{\Theta\times\Omega_M} \prod_{i=1}^n g_{\theta,\omega;M}(Y_i) d\Pi_M(\theta, \omega)}.$$

The first requirement to get consistency of estimators or posterior distributions is the identifiability of the model. We use the following assumption.

(A3.1)
- For all $j = 1, \dots, k$, $\theta_j^\star > 0$.
- For all $c = 1, 2, 3$, the measures $f_{1,c}^\star dy, \dots, f_{k,c}^\star dy$ are linearly independent.

It is proved in Theorem 8 of Allman *et al.* [AMR09] that under (A3.1) identifiability holds up to label switching, that is, if $\mathcal{S}_k$ is the set of permutations of $\{1, \dots, k\}$,

$$\forall \theta \in \Theta,\ \forall \mathbf{f} \in \mathcal{F}^{3k},\ g_{\theta,\mathbf{f}} = g_{\theta^\star,\mathbf{f}^\star} \Longrightarrow \exists \sigma \in \mathcal{S}_k \text{ such that } {}^\sigma\theta = \theta^\star,\ {}^\sigma f = f^\star,$$

where ${}^\sigma\theta \in \Theta$, ${}^\sigma f \in \mathcal{F}^{3k}$ and ${}^\sigma\theta_j = \theta_{\sigma(j)}$, ${}^\sigma f_{j,c} = f_{\sigma(j),c}$, for all $c \in \{1, 2, 3\}$, $j \in \{1, , \dots, k\}$. We need that identifiability holds for model (4.2) also. It is straightforward that this is the case if the partition is refined enough. For any partition M, any $\omega = (\omega_m)_{1\le m\le M-1}$ such that $\omega_m \ge 0$, $m = 1, \dots, M$, with $\omega_m = 1 - \sum_{m=1}^{M-1} \omega_m$, denote f_ω the step function given by

$$f_\omega(y) = \sum_{m=1}^{M} \frac{\omega_m}{|I_m|} \mathbb{1}_{I_m}(y). \tag{4.3}$$

Introduce the following assumption on the sequence of partitions $\mathcal{I}_M$, $M \in \mathcal{M}$.

(A3.2) • For all M, the sets I_m in $\mathcal{I}_M$ are intervals with nonempty interior.
• As M tends to infinity, $\max_{1\leq m\leq M}|I_m|$ tends to 0.

Assumption (A3.2) is used to get that all functions $f_{\omega^\star_{j,c;M}}$ tend to $f^\star_{j,c}$ Lebesgue almost everywhere. To extend the results when the coordinates y_c may be multivariate, the first point of (A3.2) has to be replaced by:
• There exists $a>0$ such that for all M, for all I_m in $\mathcal{I}_M$, there exists an open ball I such that $I_m\subset I$ and $|I_m|\geq a|I|$. Here $|I|$ is the Lebesgue measure of the set I.

Then, if (A3.1) and (A3.2) hold, for M large enough, we have that for all $c=1,2,3$, the measures $f_{\omega^\star_{1,c;M}}dy,\dots,f_{\omega^\star_{k,c;M}}dy$ are linearly independent, where

$$\omega^\star_{j,c;M}:=\left(\int_0^1 f^\star_{j,c;M}\mathbb{1}_{I_m}(u)du\right)_{1\leq m\leq M-1},\ c=1,2,3,\ j=1,\dots,k.$$

We give a formal proof of this fact in Section 4.6.1. Thus, using again the identifiability result in Allman *et al.* [AMR09], under (A3.1) and (A3.2), for M large enough,

$$\forall\theta\in\Theta,\ \forall\omega\in\Omega_M,\ g_{\theta,\omega;M}=g_{\theta^\star,\omega^\star_M;M}\Longrightarrow\exists\sigma\in\mathcal{S}_k\text{ such that }{}^\sigma\theta=\theta^\star,\ {}^\sigma\omega=\omega^*_M,$$

where ${}^\sigma\omega\in\Omega_M$ and ${}^\sigma\omega_{j,c,m}=\omega_{\sigma(j),c,m;M}$, for all $m\in\{1,\dots,M\}$, $c\in\{1,2,3\}$, $j\in\{1,\dots,k\}$.

4.2.2 Efficient Influence Functions and Information

We now study the estimation of θ in model (4.1) and (4.2) from the semiparametric point of view, following Le-Cam's theory. We start with model (4.2) which is easier to analyze since it is a parametric model. For any M, $g_{\theta,\omega;M}(\mathbf{y})$ is a polynomial function of the parameter (θ,ω) and the model is differentiable in quadratic mean. Denote by $S^\star_M=(S^\star_{\theta,M},S^\star_{\omega,M})$ the score function for parameter (θ,ω) at point $(\theta^\star,\omega^\star_M)$ in model (4.2). We have for $j=1,\dots,k-1$

$$\left(S^\star_{\theta,M}\right)_j=\frac{\prod_{c=1}^3 f_{\omega^\star_{j,c;M}}-\prod_{c=1}^3 f_{\omega^\star_{k,c;M}}}{g_{\theta^\star,\omega^\star_M;M}}\tag{4.4}$$

and for $j=1,\dots,k$, $c=1,2,3$, $m=1,\dots,M-1$

$$\left(S^\star_{\omega,M}\right)_{j,c,m}=\frac{\theta^\star_j\left(\frac{\mathbb{1}_{I_m}(y_c)}{|I_m|}-\frac{\mathbb{1}_{I_M}(y_c)}{|I_M|}\right)\prod_{c'\neq c}f_{\omega^\star_{j,c';M}}}{g_{\theta^\star,\omega^\star_M;M}}\tag{4.5}$$

Denote by J_M the Fisher information, that is the variance of $S^\star_M(Y)$:

$$J_M=\mathbb{E}^\star\left[S^\star_M(Y)S^\star_M(Y)^T\right]$$

Here, $\mathbb{E}^\star$ denotes expectation under $\mathbb{P}^\star$, and $S_M^\star(Y)^T$ is the transpose vector of $S_M^\star(Y)$.

When considering the question of efficient estimation of θ in the presence of a nuisance parameter, the relevant mathematical objects are the efficient influence function and the efficient Fisher information. Let us recall well known facts, see van der Vaart [Vaa98] or van der Vaart [Vaa02] for details. The efficient score function is the projection of the score function with respect to parameter θ on the orthogonal subspace of the closure of the linear subspace spanned by the tangent set with respect to the nuisance parameter (that is the set of scores in parametric models regarding the nuisance parameter). The efficient Fisher information is the variance matrix of the efficient score function. For parametric models, direct computation gives the result. If we partition the Fisher information J_M according to the parameters θ and ω, that is

$$[J_M]_{\theta,\theta} = \mathbb{E}^\star\left[S_{\theta,M}^\star(Y)S_{\theta,M}^\star(Y)^T\right],\ [J_M]_{\omega,\omega} = \mathbb{E}^\star\left[S_{\omega,M}^\star(Y)S_{\omega,M}^\star(Y)^T\right],$$

$$[J_M]_{\theta,\omega} = \mathbb{E}^\star\left[S_{\theta,M}^\star(Y)S_{\omega,M}^\star(Y)^T\right],\ [J_M]_{\omega,\theta} = [J_M]_{\theta,\omega}^T,$$

we get that, in model (4.2), if we denote $\tilde{\psi}_M$ the efficient score function for the estimation of θ,

$$\tilde{\psi}_M = S_{\theta,M}^\star - [J_M]_{\theta,\omega}([J_M]_{\omega,\omega})^{-1}S_{\omega,M}^\star,$$

and the efficient Fisher information $\tilde{J}_M$ is

$$\tilde{J}_M = [J_M]_{\theta,\theta} - [J_M]_{\theta,\omega}([J_M]_{\omega,\omega})^{-1}[J_M]_{\theta,\omega}^T.$$

To discuss efficiency of estimators, invertibility of the efficient Fisher information is needed. Spectral methods have been proposed recently to get estimators in model (4.2), see Anandkumar *et al.* [AGH+14]. It is possible to obtain upper bounds of their local maximum quadratic risk with rate $n^{-1/2}$, which as a consequence excludes the possibility that the efficient Fisher information be singular. This is stated in Proposition 4.1 below and proved in Section 4.6.1.

Proposition 4.1. *Assume (A3.1) and (A3.2). Then, for large enough M, $\tilde{J}_M$ is nonsingular.*

In the context of mixture models, all asymptotic results are given up to label switching. We define here formally what we mean by 'up to label switching' for frequentist efficiency results with Equation (4.7) and Bayesian efficiency results with Equation (4.9).

Then, if (A3.1) and (A3.2) hold, for large enough M $\tilde{J}_M$ is nonsingular, and an estimator $\widehat{\theta}$ is asymptotically a regular efficient estimator of $\theta^\star$ if and only if

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}_M^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_M\left(Y_i\right) + o_{\mathbb{P}^\star}(1), \quad \text{up to label switching,} \tag{4.6}$$

which formally means that there exists a sequence $(\sigma_n)_n$ of $\mathcal{S}_k$ such that

$$\sqrt{n}\left({}^{\sigma_n}\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}_M^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_M\left(Y_i\right) + o_{\mathbb{P}^\star}(1). \tag{4.7}$$

To get an asymptotically regular efficient estimator, one may for instance apply a one step improvement (see Section 5.7 in van der Vaart [Vaa98]) of a preliminary spectral estimator, described in Anandkumar *et al.* [AGH+14]. Also, using the trick given in van der Vaart [Vaa98] p. 63 to get consistency of the maximum likelihood estimator (m.l.e.), one sees also that the m.l.e. $\widehat{\theta}_M$ is asymptotically a regular efficient estimator of $\theta^\star$.

In the Bayesian context, Bernstein-von Mises Theorem holds for large enough M if the prior has a positive density in the neighbourhood of $(\theta^\star, \omega_M^\star)$, see Theorem 10.1 in van der Vaart [Vaa98]. That is, if $\|\cdot\|_{TV}$ denotes the total variation distance, with $\Pi_{M,\theta}$ the marginal distribution on the parameter θ,

$$\left\| \Pi_{M,\theta}\left(\cdot | Y_1, \ldots, Y_n\right) - \mathcal{N}\left(\widehat{\theta}; \frac{\tilde{J}_M^{-1}}{n}\right) \right\|_{TV} = o_{\mathbb{P}^\star}(1), \quad \text{up to label switching,} \tag{4.8}$$

where $\widehat{\theta}$ verifies Equation (4.6),
which formally means that

$$\sup_{A \subset \Theta} \left| \Pi_{M,\theta}\big(\exists \sigma \in \mathcal{S}_k : {}^{\sigma}\theta \in A \big| Y_1, \ldots, Y_n\big) - \mathcal{N}\left({}^{\sigma_n}\widehat{\theta}; \frac{\tilde{J}_M^{-1}}{n}\right)(A) \right| = o_{\mathbb{P}^\star}(1), \tag{4.9}$$

where (σ_n) and $\widehat{\theta}$ satisfie Equation (4.7).

A naive heuristic idea is that, when using the U_i's as summaries of the Y_i's, one has less information, but more and more if the partition $\mathcal{I}_M$ is refined. Thus, efficient Fisher information should grow when partitions $\mathcal{I}_M$ are refined. The following proposition is proved in Section 4.6.2.

Proposition 4.2. *Let $\mathcal{I}_{M_1}$ be a coarser partition than $\mathcal{I}_{M_2}$, that is such that for any $I \in \mathcal{I}_{M_1}$, there exists $A \subset \mathcal{I}_{M_2}$ such that $I = \cup_{I' \in A} I'$. Then*

$$\tilde{J}_{M_2} \geq \tilde{J}_{M_1}$$

in which "$\geq$" denotes the partial order between symmetric matrices.

Thus, it is of interest to let the partitions grow so that one reaches the largest efficient Fisher information.

Let us now come back to model (4.1). Let, for $j = 1, \ldots, k$, $c = 1, 2, 3$, $\mathcal{H}_{j,c}$ be the subset of functions h in $L^2(f_{j,c}^\star dy)$ such that $\int h f_{j,c}^\star dy = 0$. Then the tangent set for $\mathbf{f}$ at point $(\theta^\star, \mathbf{f}^\star)$ is

the subspace $\dot{\mathcal{P}}$ of $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y})d\mathbf{y})$ spanned by the functions

$$\mathbf{y} \mapsto \frac{h(y_c)\prod_{c'=1}^{3} f^\star_{j,c'}(y_{c'})}{g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y})},\ h \in \mathcal{H}_{j,c},\ j = 1,\dots,k,\ c = 1,2,3,$$

and the efficient score function $\tilde{\psi}$ for the estimation of θ in the semiparametric model (4.1) is given, for $j = 1,\dots,k-1$, by

$$\tilde{\psi}_j = \left(S^\star_\theta\right)_j - \mathbb{A}\left(S^\star_\theta\right)_j,\ \left(S^\star_\theta\right)_j = \frac{\prod_{c=1}^{3} f^\star_{j,c} - \prod_{c=1}^{3} f^\star_{k,c}}{g_{\theta^\star,\mathbf{f}^\star}}, \tag{4.10}$$

with $\mathbb{A}$ the orthogonal projection onto the closure of $\dot{\mathcal{P}}$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y})d\mathbf{y})$. Then, the efficient Fisher information $\tilde{J}$ is the variance matrix of $\tilde{\psi}$.
If $\tilde{J}$ is nonsingular, an estimator $\widehat{\theta}$ is asymptotically a regular efficient estimator of $\theta^\star$ if and only if

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right) + o_{\mathbb{P}^\star}(1), \quad \text{up to label switching} \tag{4.11}$$

and a Bayesian method using a nonparametric prior Π satisfies a semiparametric Bernstein-von Mises Theorem if, with Π_θ the marginal distribution on the parameter θ,

$$\left\| \Pi_\theta\left(\cdot|Y_1,\dots,Y_n\right) - \mathcal{N}\left(\widehat{\theta}; \frac{\tilde{J}^{-1}}{n}\right)\right\|_{TV} = o_{\mathbb{P}^\star}(1), \quad \text{up to label switching} \tag{4.12}$$

for a $\widehat{\theta}$ satisfying (4.11).

4.2.3 General Result

When the sequence of models is a good approximation of model (4.1) by model (4.2), we expect that efficient score functions in (4.2) are good approximations of efficient score functions in (4.1) so that asymptotically efficient estimators in model (4.2) become efficient estimators in model (4.1). This is what Theorem 4.5 below states. The approximation assumption we shall use is the following.

(A3.3) There exists $\delta > 0$ such that for all $\mathbf{y}$ in $[0,1]^3$, $\delta \leq g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y}) \leq 1/\delta$, and

$$\lim_{M\to+\infty} \|g_{\theta^\star,\omega^\star_M;M} - g_{\theta^\star,\mathbf{f}^\star}\|_\infty = 0.$$

Note that when (A3.2) is satisfied, (A3.3) holds true as soon as the functions $f^\star_{j,c}$, $j = 1,\dots,k$, $c = 1,2,3$, are positive continuous functions.
We first obtain:

Lemma 4.3. *Under Assumptions (A3.1), (A3.2) and (A3.3), the sequence of score functions* $(\tilde{\psi}_M)_M$

converges in $L^2(g_{\theta^\star,\mathbf{f}^*}d\mathbf{y})$ *to the score function* $\tilde{\psi}$*, and the sequence of efficient Fisher informations* $(\tilde{J}_M)_M$ *converges to the efficient Fisher information matrix* $\tilde{J}$*.*

Lemma 4.3 is proved in Section 4.6.3.
To get that $\tilde{J}$ is invertible, it is enough that subsequences of approximation spaces are embedded. Introduce the following assumption.

(A3.4) There exists a sequence $(M_p)_{p\geq 1}$ such that for all p, $\mathcal{I}_{M_p}$ is a coarser partition than $\mathcal{I}_{M_{p+1}}$.

The proof of the following proposition is straightforward using Lemma 4.3, Proposition 4.1 and Proposition 4.2.

Proposition 4.4. *Under Assumptions (A3.1), (A3.2), (A3.3) and (A3.4),* $\tilde{J}$ *is nonsingular.*

We are now ready to state Theorem 4.5.

Theorem 4.5. *Under Assumptions (A3.1), (A3.2), (A3.3) and (A3.4), there exists a sequence* M_n *tending to infinity sufficiently slowly such that the m.l.e.* $\widehat{\theta}_{M_n}$ *is asymptotically a regular efficient estimator of* $\theta^\star$ *and satisfies*

$$\sqrt{n}\left(\widehat{\theta}_{M_n}-\theta^\star\right)=\frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right)+o_{\mathbb{P}^\star}(1),\quad \textit{up to label switching.}$$

Under the same assumptions and if for all M*, the prior* Π_M *has a positive density in the neighbourhood of* $(\theta^\star,\omega_M^\star)$*, then there exists a sequence* L_n *tending to infinity sufficiently slowly such that moreover*

$$\left\|\Pi_{L_n,\theta}\left(\cdot|Y_1,\ldots,Y_n\right)-\mathcal{N}\left(\theta^\star+\frac{\tilde{J}^{-1}}{n}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right);\frac{\tilde{J}^{-1}}{n}\right)\right\|_{TV}=o_{\mathbb{P}^\star}(1),\ \textit{up to label switching.}$$

Proof. If $\widehat{\theta}_M$ is the m.l.e. when using model (4.2) with partition $\mathcal{I}_M$ one has

$$\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M-\theta^\star\right)=\frac{\tilde{J}_M^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_M\left(Y_i\right)+R_n(M)$$

where for each M, $(R_n(M))_{n\geq 1}$ is a sequence of random vectors converging to 0 in $\mathbb{P}^\star$-probability as n tends to infinity. But then, there exists a sequence M_n tending to infinity sufficiently slowly so that, as n tends to infinity, $R_n(M_n)$ tends to 0 in $\mathbb{P}^\star$-probability. Now,

$$\begin{aligned}\frac{\tilde{J}_{M_n}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_{M_n}\left(Y_i\right) &= \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right)+\frac{\tilde{J}_{M_n}^{-1}-\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right)+\frac{\tilde{J}_{M_n}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{\psi}_{M_n}-\tilde{\psi})\left(Y_i\right)\\ &= \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(Y_i\right)+o_{\mathbb{P}^\star}(1)\end{aligned}$$

since, by Lemma 4.3, $\mathbb{E}^\star \| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\tilde{\psi}_{M_n} - \tilde{\psi})(Y_i) \|^2 = \| \tilde{\psi}_{M_n} - \tilde{\psi} \|^2_{L^2(g_{\theta^\star, \mathbf{f}^\star}(\mathbf{y}) d\mathbf{y})}$ tends to 0 as n tends to infinity and $(\tilde{J}_{M_n})^{-1}$ converges to $(\tilde{J})^{-1}$ as n tends to infinity, so that the first part of the theorem is proved.
On the Bayesian side, for all M, there exists a sequence $V_n(M)$ of random vectors converging to 0 in $\mathbb{P}^\star$-probability as n tends to infinity such that

$$\sup_{A \subset \Theta} \left| \Pi_{M,\theta}\left(\exists \sigma \in \mathcal{S}_k : {}^\sigma\theta \in A \middle| Y_1, \ldots, Y_n \right) - \mathcal{N}\left({}^{\sigma_{n,M}}\widehat{\theta}_M ; \frac{\tilde{J}_M^{-1}}{n} \right) \right| = V_n(M).$$

Arguing as previously, there exists a sequence L_n tending to infinity sufficiently slowly so that, as n tends to infinity, both $V_n(L_n)$ and $R_n(L_n)$ tend to 0 in $\mathbb{P}^\star$-probability. Using the fact that the total variation distance is invariant through one-to-one transformations we get

$$\begin{aligned}
&\left\| \mathcal{N}\left({}^{\sigma_{n,M}}\widehat{\theta}_M ; \frac{\tilde{J}_M^{-1}}{n} \right) - \mathcal{N}\left(\theta^\star + \frac{\tilde{J}^{-1}}{n} \sum_{i=1}^n \tilde{\psi}(Y_i) ; \frac{\tilde{J}^{-1}}{n} \right) \right\|_{TV} \\
&\quad = \left\| \mathcal{N}\left(\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M - \theta^\star \right) - \frac{\tilde{J}^{-1}}{\sqrt{n}} \sum_{i=1}^n \tilde{\psi}(Y_i) ; \tilde{J}_M^{-1} \right) - \mathcal{N}\left(0 ; \tilde{J}^{-1} \right) \right\|_{TV} \\
&\quad = \left\| \mathcal{N}\left(\tilde{J}_M^{1/2} [\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M - \theta^\star \right) - \frac{\tilde{J}^{-1}}{\sqrt{n}} \sum_{i=1}^n \tilde{\psi}(Y_i)] ; Id \right) - \mathcal{N}\left(0 ; \tilde{J}_M \tilde{J}^{-1} \right) \right\|_{TV} \\
&\quad \leq \left\| \mathcal{N}\left(\tilde{J}_M^{1/2} [\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M - \theta^\star \right) - \frac{\tilde{J}^{-1}}{\sqrt{n}} \sum_{i=1}^n \tilde{\psi}(Y_i)] ; Id \right) - \mathcal{N}(0 ; Id) \right\|_{TV} \\
&\qquad + \left\| \mathcal{N}(0, Id) - \mathcal{N}\left(0 ; \tilde{J}_M \tilde{J}^{-1} \right) \right\|_{TV}.
\end{aligned}$$

But for vectors in $m \in \mathbb{R}^{k-1}$ and symmetric positive $(k-1) \times (k-1)$ matrices Σ we have

$$\| \mathcal{N}(m, Id) - \mathcal{N}(0 ; Id) \|_{TV} \leq \| m \|$$

and

$$\begin{aligned}
\| \mathcal{N}(0, Id) - \mathcal{N}(0 ; \Sigma) \|_{TV} \leq{}& \mathbb{P}\left(\| \Sigma^{1/2} U \|^2 - \| U \|^2 \geq \log[det(\Sigma)] \right) \\
&- \mathbb{P}\left(\| U \|^2 - \| \Sigma^{-1/2} U \|^2 \geq \log[det(\Sigma)] \right)
\end{aligned}$$

where $U \sim \mathcal{N}(0, Id)$. Thus the last part of the theorem follows from the triangular inequality and the fact that using Lemma 4.3, as n tends to infinity, $\tilde{J}_{L_n} \tilde{J}^{-1}$ tends to Id and $V_n(L_n)$ and $R_n(L_n)$ tend to 0 in $\mathbb{P}^\star$-probability. □

4.3 Model Selection

In Theorem 4.5, we prove the existence of some increasing partition leading to efficiency. In this section, we propose a method to choose a partition when the number of observations n is fixed.

4.3.1 Reasons to Do Model Selection

We first explain why the choice of the model is important. We have seen in Proposition 4.2 that for a sequence of increasing partitions, the efficient matrix is nondecreasing. This suggests to choose the coarsest partition and thus M_n increasing as fast as possible. Yet, one needs to pay attention to the bias in a finite horizon (i.e. when the number of observations n is fixed). Note that in this model, we don't know any unbiased estimator of θ. Besides, typically the bias of an estimator of θ may increase when M increases. This prevents us to choose a sequence M_n tending to $+\infty$ too quickly (see Corollary 4.7).

We now illustrate this issue using the m.l.e. If the m.l.e. is unbiased asymptotically, it is biased for a finite sample. In Proposition 4.6, we give the limit of the m.l.e. when the number n of observations is fixed but M tends to infinity.

Proposition 4.6. *For almost all observations $Y_1, \dots, Y_n$, $\hat{\theta}_M(Y_1, \dots, Y_n)$ tends to*

$$\underline{\theta}_n = (\underbrace{\lfloor n/k \rfloor/n, \dots, \lfloor n/k \rfloor/n}_{r:=n-k\lfloor n/k \rfloor}, \underbrace{\lceil n/k \rceil/n, \dots, \lceil n/k \rceil/n}_{k-r})$$

up to label switching, when M tends to infinity.

Proposition 4.6 is proved in Section 4.6.4.

Using Proposition 4.6, we can deduce a constraint (leading to an upper bound in some cases), depending on the considered sequence of partitions $(\mathcal{I}_M)_{M\in\mathcal{M}}$, on sequences M_n leading to efficiency. We believe that this constraint is very conservative and leads to very conservative bounds. Corollary 4.7 below is proved in Section 4.6.5.

Corollary 4.7. *Suppose Assumption (A3.3), if $\widehat{\theta}_{M_n}$ tends to θ^* in probability, and θ^* is different from $(1/k, \dots, 1/k)$,*

then there exists $N > 0$ and a constant $C > 0$ such that for all $n \geq N$,

$$n^2 \left(\max_{m \leq M_n} |I_m| \right)^2 M_n \geq C.$$

Moreover, in the particular case where there exists $0 < C_1 \leq C_2$ such that for all $n \in \mathbb{N}$ and $1 \leq m \leq M_n$,

$$\frac{C_1}{M_n} \leq |I_m| \leq \frac{C_2}{M_n} \tag{4.13}$$

then there exists a constant $C > 0$ *such that,*

$$M_n \leq Cn^2.$$

Note that Assumption (4.13) holds as soon as the partition is regular, so that in the two following cases:

- for the uniform partition, when $\mathcal{M} = \mathbb{N}$ and for all $M \in \mathcal{M}$ $I_m = [(m-1)/M, m/M)$ for all $m < M$, $I_M = [(M-1)/M, 1]$,
- or for the dyadic regular partitions, when $\mathcal{M} = \{2^p, p \in \mathbb{N}^*\}$ and for all $M \in \mathcal{M}$ $I_m = [(m-1)/M, m/M)$ for all $m < M$, $I_M = [(M-1)/M, 1]$, which form an embedded sequence of partition.

4.3.2 Criterion for Model Selection

In this section, we propose a criterion to choose the partition when n is fixed. This criterion can be used to choose the size M of a family of partitions but also to choose between two families of partition. With a dataset, we can compute the m.l.e. (with the EM algorithm) when using model (4.2) with partition $\mathcal{I}$, or we can get an estimator of θ using its posterior distribution (the posterior mean or the posterior median for instance). We thus shall index all our estimators by $\mathcal{I}$. Note that the results of this section are valid for any family of estimators $(\tilde{\theta}_{\mathcal{I}})$ and not only for the m.l.e.

Proposition 4.6 and Corollary 4.7 show the necessity to choose an appropriate partition among a collection of partitions $\mathcal{I}_M$, $M \in \mathcal{M}$. To choose the partition we need a criterion. Since the aim is to get efficient estimators, we choose the quadratic risk as the criterion to minimize. We thus want to minimize over all possible partitions

$$R_n(\mathcal{I}) = \mathbb{E}^\star \left[\|\tilde{\theta}_{\mathcal{I}}(Y_{1:n}) - \theta^\star\|^2_{\mathcal{S}_k} \right], \tag{4.14}$$

where $Y_{1:n} = (Y_i)_{i \leq n}$ and for all $\theta, \tilde{\theta} \in \Theta$,

$$\|\theta - \tilde{\theta}\|_{\mathcal{S}_k} = \min_{\sigma \in \mathcal{S}_k} \|{}^\sigma\theta - \tilde{\theta}\|_2 = \|{}^o\theta - {}^o\tilde{\theta}\|_2, \tag{4.15}$$

with ${}^o\theta = {}^\sigma\theta$ for a permutation $\sigma \in \mathcal{S}_k$ which orders the components of ${}^\sigma\theta$, i.e. such that ${}^\sigma\theta_1 \leq \ldots \leq {}^\sigma\theta_k$. As usual, this criterion cannot be computed in practice (since we do not know θ^*). To do this on data we need for each partition $\mathcal{I}$ some estimator $C(\mathcal{I})$ of $R_n(\mathcal{I})$.

We want to emphasize here that the choice of the criterion for this problem is not easy. Indeed, the quadratic risk $R_n(\mathcal{I})$ cannot be written as the expectation of an excess loss expressed thanks to a contrast function, i.e. in the form $\mathbb{E}^\star \left[\mathbb{E}^\star \left[\gamma(\tilde{\theta}(Y_{1:n}), Y) - \gamma(\theta^*, Y) | Y_{1:n} \right] \right]$, where $\gamma : \Theta \times \mathcal{Y} \to [0, +\infty)$. Yet, the last framework is the framework of most theoretical results in model selection,

see Arlot and Celisse [AC10] or Massart [Mas07] for instance. Moreover the quadratic risk has not a usual behaviour. Indeed if we decompose it as an approximation error plus an estimation error as explained in Arlot and Celisse [AC10]:

$$R_n(\mathcal{I}) = \underbrace{\inf_{\theta\in\Theta_{\mathcal{I}}} \|\theta-\theta^*\|^2_{\mathcal{S}_k}}_{\text{approximation error}} + \underbrace{R_n(\mathcal{I}) - \inf_{\theta\in\Theta_{\mathcal{I}}} \|\theta-\theta^*\|^2_{\mathcal{S}_k}}_{\text{estimation error}}, \quad \text{where } \Theta_{\mathcal{I}} = \Theta,$$

we see that the approximation error is always zero in our model (and not decreasing as often). For these reasons, we cannot apply the usual methods and we use instead a variant of usual cross validation technique.

Consider a partition of $\{1,\cdots,n\}$ in the form $(B_b, B_{-b}, b \leq b_n)$, in other words the partition is made of $2\times b_n$ subsets of $\{1,\cdots,n\}$. By definition $B_{b_1}\cap B_{-b_2}=\emptyset$ for all $b_1, b_2 \leq b_n$. Because the maximum likelihood estimator based on any finite sample size is not unbiased, the following naive estimator of the risk is not appropriate:

$$C_{CV1}(\mathcal{I}) = \frac{1}{2b_n}\sum_{b=1}^{b_n} \|\tilde{\theta}_{\mathcal{I}}(Y_{B_b}) - \tilde{\theta}_{\mathcal{I}}(Y_{B_{-b}})\|^2_{\mathcal{S}_k}.$$

Indeed, using Proposition 4.6, $C_{CV1}(\mathcal{I})$ is tending to 0 when $\max_m |I_m|$ tends to 0. So that minimizing this criterion leads to choosing a partition $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}} C_{CV1}(\mathcal{I})$ which has a large number of sets and so $\widehat{\theta}_{\widehat{\mathcal{I}}_n}(Y_{1:n})$ may be close to $(1/k,\ldots,1/k)$ and then may not even be consistent. This can be seen when decomposing the risk $R_n(\mathcal{I})$ as:

$$R_n(\mathcal{I}) = \underbrace{Var^\star\left[{}^o\tilde{\theta}_{\mathcal{I}}(Y_{1:n})\right]}_{\text{variance}} + \underbrace{\left\|\mathbb{E}^\star\left[\tilde{\theta}_{\mathcal{I}}(Y_{1:n})\right] - \theta^\star\right\|^2_{\mathcal{S}_k}}_{\text{bias}} \tag{4.16}$$

and computing the expectation of $C_{CV1}(\mathcal{I})$ in the case where the sizes of B_b, B_{-b}, $b\leq b_n$, are all equal,

$$\mathbb{E}^\star\left[C_{CV1}(\mathcal{I})\right] = Var^\star\left[{}^o\tilde{\theta}_{\mathcal{I}}(Y_{B_b})\right]$$

suggests that $C_{CV1}(\mathcal{I})$ does not estimate the bias of Equation (4.16). As an illustration, see Figure 4.2 where the trends of $R_n(\mathcal{I})$, $Var^\star\left[{}^o\widehat{\theta}_{\mathcal{I}}(Y_{1:n})\right]$ and $\left\|\mathbb{E}^\star\left[\widehat{\theta}_{\mathcal{I}}(Y_{1:n})\right]-\theta^\star\right\|^2_{\mathcal{S}_k}$ respectively are plotted.

To address the bad behaviour of $C_{CV1}(\mathcal{I})$, we use an idea of Brookhart and van der Laan [BL06]. Choose a (fixed) base partition $\mathcal{I}_0$ (for which the criterion may also be computed) which is believed to be (almost) unbiased. And set

$$C_{CV}(\mathcal{I}) = \frac{1}{b_n}\sum_{b=1}^{b_n} \|\tilde{\theta}_{\mathcal{I}}(Y_{B_b}) - \tilde{\theta}_{\mathcal{I}_0}(Y_{B_{-b}})\|^2_{\mathcal{S}_k}.$$

Equivalently, we could choose any unbiased estimator $\tilde{\theta}$ instead of using an estimator $\theta_{\mathcal{I}_0}$ of the considered family of estimator. Figure 4.3 gives an idea of the behaviour of C_{CV} and C_{CV1} using the m.l.e.. It shows in particular that in our simulation study C_{CV} follows the same behaviour as $R_n(\mathcal{I})$, contrarywise to C_{CV1}. More details are provided in Section 4.4.

We now provide some theoretical results on the behaviour of the minimizer of $C_{CV}(\cdot)$ over a finite family of candidate partitions $\mathcal{M}_n$ compared to the minimizer of $R_{a_n}(\cdot)$ over the same family.

Let $m_n = \#\mathcal{M}_n$ be number of candidate partitions.

To do so we consider the following set of assumptions:

(A3.5) (A3.5.1) B_b, B_{-b}, $b \leq b_n$ are disjoint sets of equal size

$$\#B_b = \#B_{-b} = a_n, \quad \text{for all } b \leq b_n$$

(A3.5.2) $\tilde{\theta}_{\mathcal{I}_0,b,2}$ is not biased i.e. $\mathbb{E}^*[\tilde{\theta}_{\mathcal{I}_0,b,2}] = \theta^*$,

we obtain the following oracle inequality.

Theorem 4.8. *Suppose Assumption (A3.5). For any sequences $0 < \epsilon_n, \delta_n < 1$, with probability greater than*

$$1 - 2m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right),$$

we have

$$R_{a_n}(\widehat{\mathcal{I}}_n) \leq \frac{1+\epsilon_n}{1-\epsilon_n} \inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \frac{2\delta_n}{1-\epsilon_n}, \tag{4.17}$$

where $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}\in\mathcal{M}_n} C_{CV}(\mathcal{I})$.

As a consequence of Theorem 4.8, the following Proposition holds. Recall that $n = 2b_n a_n$.

Proposition 4.9. *Assume (A3.5). If $b_n \gtrsim n^{2/3}\log^2(n)$, $a_n \lesssim n^{1/3}/(\log^2(n))$, and $m_n \leq C_\alpha n^\alpha$, for some $C_\alpha > 0$ and $\alpha \geq 0$, then*

$$\mathbb{E}^\star\left[a_n R_{a_n}(\widehat{\mathcal{I}}_n)\right] \leq \inf_{\mathcal{I}\in\mathcal{M}_n} a_n R_{a_n}(\mathcal{I}) + o(1),$$

where $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}\in\mathcal{M}_n} C_{CV}(\mathcal{I})$.

Note that for each $\mathcal{I}$, $R_{a_n}(\mathcal{I})$ is of order of magnitude $1/a_n$ so that the remaining term is indeed small regarding the main term. Note also that this is an exact oracle inequality (with constant 1).

In Theorem 4.8 and Proposition 4.9, $\widehat{\mathcal{I}}_n$ is built with n observations while the risk is associated with $a_n < n$ observations. This leads to a conservative choice of $\widehat{\mathcal{I}}_n$, i.e. we may choose a sequence $\widehat{\mathcal{I}}_n$ (optimal with a_n observations) increasing more slowly than the optimal one (with n

observation). We think however that this conservative choice should not change the good behaviour of $\widehat{\theta}_{\widehat{\mathcal{I}}_n}$, since Theorem 4.5 implies that any sequence of partitions which grows slowly enough to infinity leads to an efficient estimator. Hence, once the sequence M_n growing to infinity is chosen, then any other sequence growing to infinity more slowly also leads to an efficient estimator.

In Proposition 4.9 and Theorem 4.8, the reference point estimate $\tilde{\theta}_{\mathcal{I}_0}(Y_{B_{-b}})$ is assumed to be unbiased. This is a strong assumption, which is not exactly satisfied in our simulation study. To consider a reasonable approximation of it, $\tilde{\theta}_{\mathcal{I}_0}(Y_{B_{-b}})$ is chosen as the m.l.e. associated to a partition with a small number of bins. The heuristic behind this choice is that the maximum likelihood is asymptotically unbiased and a small number of bins implies a smaller number of parameters to estimate, so that the asymptotic regime is attained faster. Our simulations confirm this heuristic, see Section 4.4.

4.4 Simulations

In this section, we illustrate the results obtained in Sections 4.3.1 and 4.3.2 with simulations. We compare six criteria for the model selection based on C_{CV} with different choices of size of training and testing sets. We choose the regular embedded dyadic partitions, i.e. when $\mathcal{M} = \{2^p, p \in \mathbb{N}^*\}$ and for all $M \in \mathcal{M}$, $I_m = [(m-1)/M, m/M)$ for all $m < M$, $I_M = [(M-1)/M, 1]$. Following Corollary 4.7, when n is fixed, we only consider $M = 2^P \leq M_n = n^3$ (i.e. $P \leq P_n := \lfloor 3/2 \log(n) \rfloor$). In this part, we only consider m.l.e. estimators with ordered components and approximated thanks to the EM algorithm.

For n fixed, the choice of the model, through P, is done thanks to the criterion C_{CV} with two types of choice for $(B_b), (B_{-b})$. First, we use the framework under which we were able to prove something, i.e. Assumption (A3.5.1) where all the training and testing sets are disjoints. We use different sizes a_n and b_n:

- $b_n = \lceil n^{2/3} log(n)/(20) \rceil$ and $a_n = \lfloor n/(2b_n) \rfloor$ (Assumption of Proposition 4.9, up to $\log(n)$), leading to the criterion $C_{CV}^{D,1}$ and the choice of P noted $\widehat{P}_n^{D,1} \in \arg\min_{P \leq P_n} C_{CV}^{D,1}(\mathcal{I}_{2^P})$,
- $b_n = \lceil n^{1/3} \rceil$, $a_n = \lfloor n/(2b_n) \rfloor$, leading to the criterion $C_{CV}^{D,2}$ and the choice of P noted $\widehat{P}_n^{D,2} \in \arg\min_{P \leq P_n} C_{CV}^{D,2}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n/10 \rfloor$, $b_n = \lfloor n/(2a_n) \rfloor$, leading to the criterion $C_{CV}^{D,3}$ and the choice of P noted $\widehat{P}_n^{D,3} \in \arg\min_{P \leq P_n} C_{CV}^{D,3}(\mathcal{I}_{2^P})$

We also consider the famous V-fold, where the dataset is cut into b_n disjoint sets $\tilde{B}_b$ of size a_n, leading to training sets $B_b = \tilde{B}_b$ and testing sets $B_{-b} = \{1, \dots n\} \setminus \tilde{B}_b$. We also use different sizes a_n and b_n:

- $a_n = \lfloor n^{1/3} \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,1}$ and the choice of P noted $\widehat{P}_n^{V,1} \in \arg\min_{P \leq P_n} C_{CV}^{V,1}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n^{2/3}/2 \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,2}$ and the choice of P noted $\widehat{P}_n^{V,2} \in \arg\min_{P \leq P_n} C_{CV}^{V,2}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n/10 \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,3}$ and the choice of P noted $\widehat{P}_n^{V,3} \in \arg\min_{P \leq P_n} C_{CV}^{V,3}(\mathcal{I}_{2^P})$.

Note that for criteria

- $C_{CV}^{j,1}$, $j \in \{D, V\}$, a_n is proportional to $n^{1/3}$ up to a logarithm term,
- $C_{CV}^{j,2}$, $j \in \{D, V\}$, a_n is proportional to $n^{2/3}$,
- $C_{CV}^{j,3}$, $j \in \{D, V\}$, a_n is proportional to n.

We know explain how we choose $\mathcal{I}_0$. We do not know any unbiased estimate of θ, which would match the Assumption (A3.5.2). Particularly the m.l.e. $\widehat{\theta}_M$ is unbiased asymptotically but biased with finite n. We propose to choose a m.l.e. $\widehat{\theta}_{M_0}$ with a small M_0 with the idea that when M is small the asymptotic is attained more quickly. Yet, M_0 should not be taken too small neither since otherwise the model would not be identifiable. We propose to choose the smallest $M_0 = 2^{P_0}$ such that $M_0 \geq k + 2$ (equivalently $P_0 \geq \log(k+2)/\log(2)$). This lower bound ensures that generically on $\mathcal{I}_0$ the model (4.2) is identifiable.

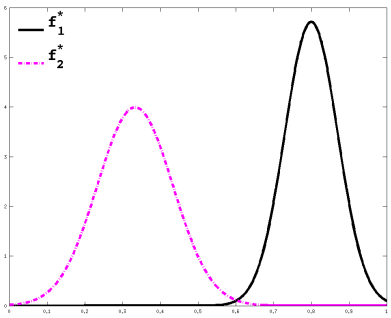
In the simulation part, we work in the repeated setting, that is $f_{j,1}^* = f_{j,2}^* = f_{j,3}^*$ and we assume that we know it, i.e. when we search for the m.l.e. in the model (4.2) associated to $M \in \mathcal{M}$, we only search for $\theta \in \Delta_k$, $\omega \in \Delta_M^k$ (and not $\omega \in \Delta_M^{3k}$) assuming that $\omega_{j,1,m} = \omega_{j,2,m} = \omega_{j,3,m} = \omega_{j,m}$. We first use three different true parameters for the simulations, in easy situations. In the three cases, $k = 2$ and the other parameters are given in Table 4.1. So that, we work with $P_0 = 2$ and $M_0 = 2^2 = 4$.

Simu.	k	p^*	$f_{1,1}^* d\lambda = f_{1,2}^* d\lambda = f_{1,3}^* d\lambda$	$f_{2,1}^* d\lambda = f_{2,2}^* d\lambda = f_{2,3}^* d\lambda$
1	2	$(0.3, 0.7)$	$\mathcal{N}(4/5, 0.07^2)$ truncated to $[0,1]$	$\mathcal{N}(1/3, 0.1^2)$ truncated to $[0,1]$
2	2	$(0.2, 0.8)$	$\mathcal{U}((0,1))$	$\mathcal{N}(2/3, 0.05^2)$ truncated to $[0,1]$
3	2	$(0.3, 0.7)$	$\beta(1,2)$	$\beta(5,3)$

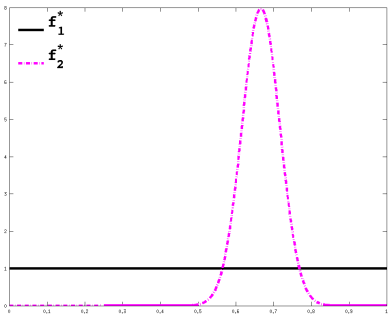
Table 4.1 – Values of the true parameters for simulation 1 to 3

The different emission distributions are represented in Figure 4.1.

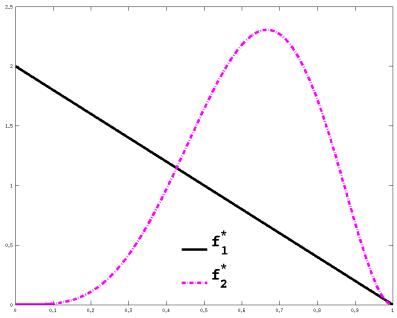
Figure 4.2 gives a taste of the trend of the risk $R_n(\mathcal{I}_{2^P})$, along with the variance $Var^*\left[{}^o\widehat{\theta}_{2^P}(Y_{1:n})\right]$ and the squared bias $\left\|\mathbb{E}^*\left[\widehat{\theta}_{2^P}(Y_{1:n})\right] - \theta^*\right\|_{\mathcal{S}_k}^2$ defined in Equation (4.16) when P increases. We illustrates these trends thanks to different true parameters and numbers of observations n. The



(a) Simulation 1



(b) Simulation 2



(c) Simulation 3

Figure 4.1 – Representation of the true emission distributions for simulations 1, 2 and 3.

different risks, bias and variances are estimated by Monte Carlo by repeating 1000 times the estimation of θ with the m.l.e. (approximated with the EM algorithm). A typical behaviour of the bias is being constant or decreasing, with small increasing values of P, then increasing a lot when P increases, and finally stabilizing to the value $\|\underline{\theta}_n - {}^o\theta^*\|$, which is a consequence of Proposition 4.6. Typically, the variance is constant or decreasing for small increasing values of P, sometimes it then increases, before decreasing to zero (which also is a consequence of Proposition 4.6) when P gets large. Then, the risk, which is the sum of the squared bias and the variance, is usually constant or decreasing for small increasing values of P and then increasing to $\|\underline{\theta}_n - {}^o\theta^*\|^2$ when P gets large.

Now we have an idea of the behaviour of the risk $R_n(\mathcal{I}_{2^P})$, we can check the behaviour of the different criteria C_{CV} and C_{CV1}. Figure 4.3 gives an idea of the pattern of some criteria for one sequence of observations $Y_{1:n}$, distributed from two different true parameters, with respect to P. We do not show all the criteria since they all look alike. Moreover the purpose of figure 4.3 is to illustrate the 'bad' behaviour of C_{CV1} compared to C_{CV} and not comparing the six criteria (which would anyway be impossible with one sequence of observations $Y_{1:n}$). Note that we do not compare the values but the behaviour. Indeed, the criteria are used to choose the best P by taking the minimum of the criterion so that the values are not important by themselves.

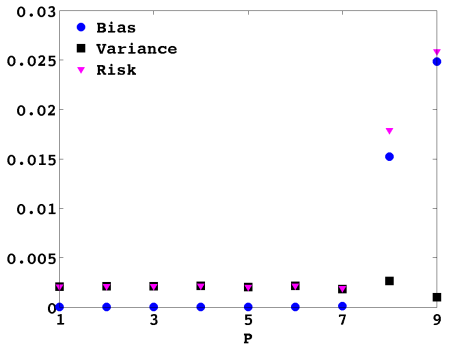
Besides, we know that the criterion C_{CV} is biased by a constant depending on $\mathcal{I}_0$. As theoretically explained in Section 4.3 and as a consequence of Proposition 4.6, we can see that the criteria C_{CV1} are tending to 0 when P increases while it is not the case for the criteria C_{CV}. Looking at Figure 4.3, the behaviour of C_{CV} seems to be correct, we precise this impression with table 4.2.

Finally we compare the six criteria $C_{CV}^{j,c}$, $j \in \{D, V\}$, $c \in \{1, 2, 3\}$, by estimating the squared risk of the associated estimator $\widehat{\theta}_{2^{\widehat{P}_n^{j,c}}}$, presented in Table 4.2. Different sizes n of samples and different true parameters are used to simulate the data. We can compare the six squared risk to $\sqrt{\min_{P \leq P_n} R_n(2^P)}$ and $\sqrt{R_n(2^{P_0})}$. The different risks are estimated by Monte Carlo by repeating 100 times the estimation. The differences of performance between the different criteria are not obvious. Besides, the performances of all the criteria are satisfactory, compared to $\sqrt{\min_{P \leq P_n} R_n(2^P)}$. Yet, we suggest not to use criterion $C_{CV}^{V,1}$ because it is longer than the others, particularly when n is large (because of large b_n). Furthermore, there is a little advantage to criteria $C_{CV}^{D,1}$ and $C_{CV}^{V,2}$.

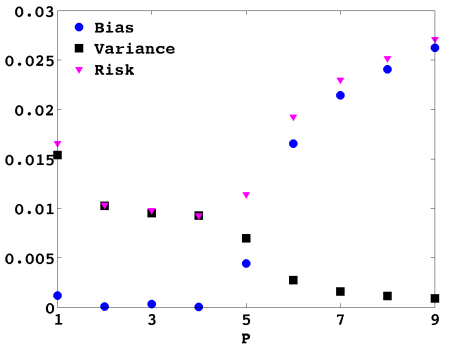
These results confirm that by using M_0 small, the criterion behaves correctly. Moreover, the fact that the choice of $\widehat{I}_n$ corresponds to a risk associated with $a_n < n$ observations does not seem to be a conservative choice even in a finite horizon (i.e. when n is fixed). We were expecting this behaviour asymptotically but not in a finite horizon.

4.5 Discussion

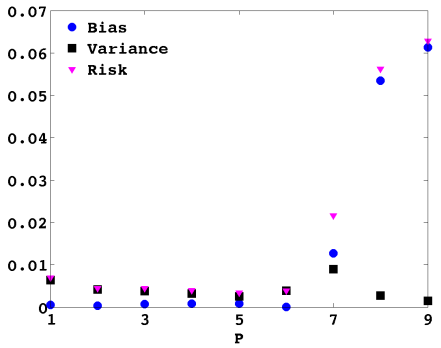
Finite mixture models all have the property that, when the approximation space for the emission distributions is that of step functions (histograms), then the model stays true for observation process. Thus there is no approximation bias regarding the parameter that describes the probability distribution of the latent variables. Extension of the results we obtain in this chapter should be possible to other nonparametric finite mixture models. This should also be the case for nonparametric hidden Markov models with translated emission distributions studied in Gassiat and Rousseau [GR16] or for general nonparametric finite state space hidden Markov models studied in De Castro *et al.* [DGLar], Vernet [Ver15b] and De Castro *et al.* [DGC15]. Here, the parameter describing the probability distribution of the latent variable is the transition matrix of the hidden Markov chain. However, semiparametric asymptotic theory for dependent observations is much more involved, see McNeney and Wellner [MW00] for the ground principles. It seems difficult to identify the score functions and the efficient Fisher information matrices for hidden Markov models even in the parametric approximation model, so that to get results such as Theorem 4.5 could be quite challenging.



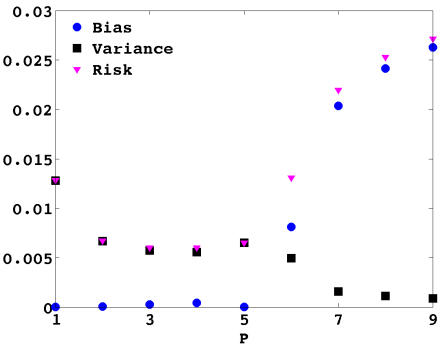
(a) Simulation 1, $n = 100$



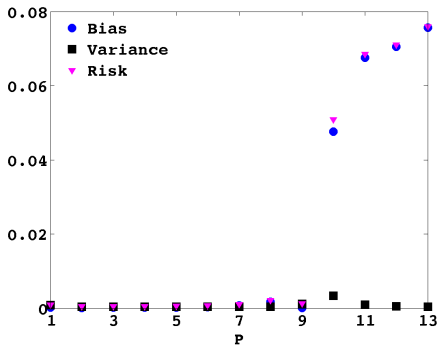
(b) Simulation 3, $n = 50$



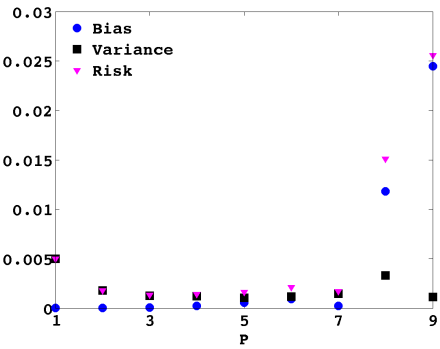
(c) Simulation 2, $n = 50$



(d) Simulation 3, $n = 100$

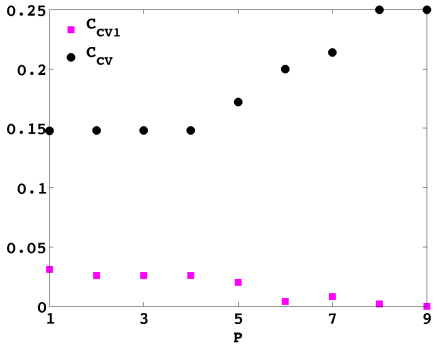


(e) Simulation 2, $n = 500$

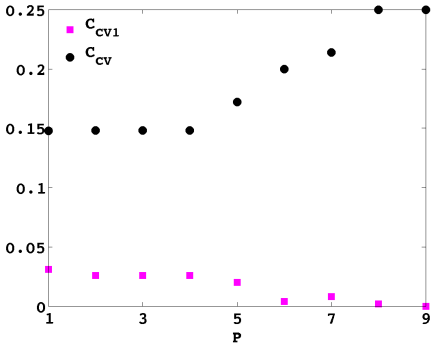


(f) Simulation 3, $n = 500$

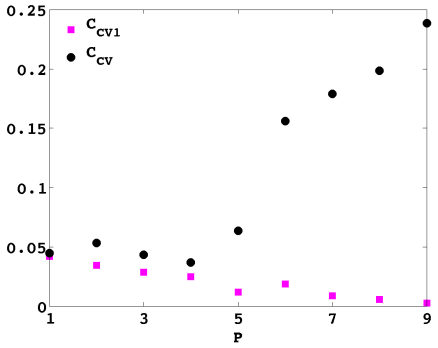
Figure 4.2 – Patterns of the risk (with black squares), the squared bias (with blue dots) and variance (with magenta triangles) with respect to $P = \log(M)/\log(2)$ for simulations 1, 2 and 3 and different values of n.



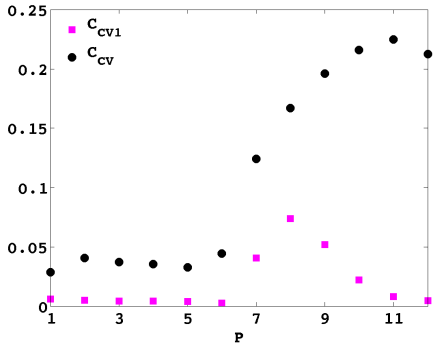
(a) Simulation 1, $n = 100$, $C^{D,1}_{CV(1)}$



(b) Simulation 1, $n = 100$, $C^{V,3}_{CV(1)}$



(c) Simulation 2, $n = 500$, $C^{D,1}_{CV(1)}$



(d) Simulation 2, $n = 500$, $C^{V,3}_{CV(1)}$

Figure 4.3 – Taste of the behaviour of C_{CV1} vs C_{CV}.

Simulation	1	1	1	1	1	2	2	2	3	3	3
n	50	100	500	1000	2000	50	100	500	50	100	500
$\sqrt{\min_{P \leq P_n} R_n(2^P)}$	0.062	0.043	0.020	0.014	0.010	0.058	0.046	0.020	0.096	0.078	0.036
$\sqrt{R_n(2^{P_0})}$	0.063	0.046	0.021	0.015	0.010	0.067	0.046	0.022	0.10	0.082	0.042
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{D,1}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.069	0.047	0.019	0.014	0.011	0.075	0.056	0.019	0.12	0.087	0.037
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{D,2}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.073	0.046	0.022	0.015	0.010	0.065	0.056	0.025	0.10	0.087	0.046
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{D,3}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.086	0.047	0.021	0.014	0.010	0.087	0.056	0.026	0.11	0.087	0.041
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{V,1}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.091	0.046	0.021	0.013	0.009	0.104	0.055	0.022	0.11	0.087	0.053
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{V,2}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.069	0.046	0.019	0.013	0.010	0.070	0.049	0.022	0.12	0.084	0.036
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\widehat{P}_n^{V,3}}}(Y_{1:n}) - \theta^\star\rVert^2\right]}$	0.103	0.046	0.019	0.014	0.009	0.10	0.049	0.022	0.14	0.083	0.035

Table 4.2 – Comparison of the squared risk of estimators associated to different criteria

4.6 Proofs

4.6.1 Proof of Proposition 4.1

Let us first prove that for large enough M, the measures $f_{\omega^\star_{1,c;M}}dy, \dots, f_{\omega^\star_{k,c;M}}dy$ are linearly independent. Indeed, if it is not the case, there exists a subsequence M_p tending to infinity as p tends to infinity and a sequence $(\alpha^{(p)})_{p\geq 1}$ in the unit ball of $\mathbb{R}^k$ such that for all $p \geq 1$,

$$\sum_{j=1}^{k} \alpha_j^{(p)} f_{\omega^\star_{j,c;M_p}}(y) = 0$$

Lebesgue a.e. Let $\alpha = (\alpha_1, \dots, \alpha_k)$ be a limit point of $(\alpha^{(p)})_{p\geq 1}$ in the unit ball of $\mathbb{R}^k$. Using Assumption (A.2) and Corollary 1.7 in Chapter 3 of Stein and Shakarchi [SS05], we have that as p tends to infinity, $f_{\omega^\star_{j,c;M_p}}(y)$ converges to $f^\star_{j,c}(y)$ Lebesgue a.e. so that we obtain $\sum_{j=1}^k \alpha_j f^\star_{j,c}(y) = 0$ Lebesgue a.e., contradicting Assumption (A3.1).
Fix now M large enough so that the measures $f_{\omega^\star_{1,c;M}}dy, \dots, f_{\omega^\star_{k,c;M}}dy$ are linearly independent. Then, one may use the spectral method described in Anandkumar *et al.* [AGH+14] to get estimators $\widehat{\theta}_{sp}$ and $\widehat{\omega}_{M;sp}$ of the parameters θ and ω_M from a sample of the multinomial distribution associated to density $g_{\theta,\omega;M}$. The estimator uses eigenvalues and eigenvectors computed from the empirical estimator of the multinomial distribution. But in a neighbourhood of $\theta^\star$ and $\omega^\star_M$, this is a continuously derivative procedure, and since on this neighbourhood, classical deviation probabilities on empirical means hold uniformy, we get easily that for any vector $V \in \mathbb{R}^k$, there exists $K > 0$ such that for all $c > 0$, for large enough n (the size of the sample):

$$\sup_{\|\theta-\theta^\star\|\leq \frac{c}{\sqrt{n}}} \mathbb{E}^\theta\left[\left(\sqrt{n}\langle\widehat{\theta}_{sp} - \theta, V\rangle\right)^2\right] \leq K.$$

Now, the multinomial model is differentiable in quadratic mean, and following the proof of Theorem 4 in Gassiat *et al.* [GPS13] one gets that, if $V^T\tilde{J}_M V = 0$, then

$$\lim_{c\to+\infty}\lim_{n\to+\infty}\sup_{\|\theta-\theta^\star\|\leq \frac{c}{\sqrt{n}}} \mathbb{E}^\theta\left[\left(\sqrt{n}\langle\widehat{\theta}_{sp} - \theta, V\rangle\right)^2\right] = +\infty.$$

Thus for all $V \in \mathbb{R}^k$, $V^T\tilde{J}_M V \neq 0$, so that $\tilde{J}_M$ is not singular.

4.6.2 Proof of Proposition 4.2

We prove the proposition when $M_1 = M$, $M_2 = M+1$, $\mathcal{I}_M = \{I_1, \dots, I_M\}$ and $\mathcal{I}_{M+1} = \{I_1, \dots, I_{M,0}, I_{M,1}\}$ with $I_M = I_{M,0} \cup I_{M,1}$, which is sufficient by induction. We denote $(\omega^{(M)}_{j,c,m})_{j,c,1\leq m\leq M}$ the parameter ω in the model with partition $\mathcal{I}_M$ and $(\omega^{(M+1)}_{j,c,m})_{j,c,1\leq m\leq M+1}$ the parameter ω in the model with partition $\mathcal{I}_{M+1}$. Define $b \in (0,1)$, $\alpha_{j,c} \in (0,1)$, $j = 1, \dots, k$,

$c = 1, 2, 3$ so that

$$|I_{M,0}| = (1-b)|I_M|,\ |I_{M,1}| = b|I_M|,\ \omega^{(M+1)}_{j,c,M} = (1-\alpha_{j,c})\omega^{(M)}_{j,c,M},\ \omega^{(M+1)}_{j,c,M+1} = \alpha_{j,c}\omega^{(M)}_{j,c,M}.$$

Then, we may write

$$g_{\theta,\omega;M}(\mathbf{y}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M} \left(\frac{\omega^{(M)}_{j,c,m}}{|I_m|} \right)^{\mathbb{1}_{I_m}(y_c)}$$

and

$$\begin{aligned} g_{\theta,\omega;M+1}(\mathbf{y}) &= \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M-1} \left(\frac{\omega^{(M+1)}_{j,c,m}}{|I_m|} \right)^{\mathbb{1}_{I_m}(y_c)} \left[\left(\frac{\omega^{(M+1)}_{j,c,M}}{|I_{M,0}|} \right)^{\mathbb{1}_{I_{M,0}}(y_c)} \left(\frac{\omega^{(M+1)}_{j,c,M+1}}{|I_{M,1}|} \right)^{\mathbb{1}_{I_{M,1}}(y_c)} \right] \\ &= \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M} \left(\frac{\omega^{(M)}_{j,c,m}}{|I_m|} \right)^{\mathbb{1}_{I_m}(y_c)} \left[\left(\frac{\alpha_{j,c}}{b} \right)^{\mathbb{1}_{I_{M,1}}(y_c)} \left(\frac{1-\alpha_{j,c}}{1-b} \right)^{\mathbb{1}_{I_{M,0}}(y_c)} \right]. \end{aligned}$$

Thus, when $y_c \notin I_M$ for $c = 1, 2, 3$, $g_{\theta,\omega;M+1}(\mathbf{y}) = g_{\theta,\omega;M}(\mathbf{y})$ and computations have to take care of $\mathbf{y}$'s such that for some c, $y_c \in I_M$. If we parametrize the model with partition $\mathcal{I}_{M+1}$ using the parameter $\left(\theta, (\omega^{(M)}_{j,c,m}), (\alpha_{j,c})\right)$ we get the same efficient Fisher information for θ as when parametrizing with $\left(\theta, (\omega^{(M+1)}_{j,c,m})\right)$. Define the function D as the difference between the gradient of $\log g_{\theta,\omega;M+1}$ and that of $\log g_{\theta,\omega;M}(\mathbf{y})$ with respect to the parameter $\left(\theta, (\omega^{(M)}_{j,c,m}), (\alpha_{j,c})\right)$:

$$D(\mathbf{y}) := \nabla \log g_{\theta,\omega;M+1}(\mathbf{y}) - \nabla \log g_{\theta,\omega;M}(\mathbf{y}),$$

in particular the last coordinates of $\nabla \log g_{\theta,\omega;M}(\mathbf{y})$ corresponding to the derivatives with respect to $(\alpha_{j,c})$ are zero. Let us denote $K^{(M+1)}$ the Fisher information obtained for this new parametrization, that is $K^{(M+1)} = \mathbb{E}^\star[(\nabla \log g_{\theta,\omega;M+1}(Y))(\nabla \log g_{\theta,\omega;M+1}(Y))^T]$. Easy but tedious computations give

$$\mathbb{E}^\star[(\nabla \log g_{\theta,\omega;M}(Y))(D(Y))^T] = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

so that

$$K^{(M+1)} = \begin{pmatrix} J_M & 0 \\ 0 & 0 \end{pmatrix} + \Delta$$

where $\Delta = \mathbb{E}^\star[D(Y)D(Y)^T]$ is positive semi-definite. As said before, $\tilde{J}_{M+1}$ is obtained from $K^{(M+1)}$ using the similar formula as from J_{M+1}. Then usual algebra gives that $\tilde{J}_{M+1} \geq \tilde{J}_M$ since Δ is positive semi-definite.

4.6.3 Proof of Lemma 4.3

Proof. Notice first that under (A3.3), $g_{\theta^\star,\mathbf{f}^*}/g_{\theta^\star,\omega_M^\star,M}$ is positively lower and upper bounded, so that the set of functions which are in $L^2(g_{\theta^\star,\mathbf{f}^*}d\mathbf{y})$ is the same as the set of functions which are in $L^2(g_{\theta^\star,\omega_M^\star,M}d\mathbf{y})$. Also, any step function which is constant over $I_{m_1}\times I_{m_2}\times I_{m_3}$, $m_1,m_2,m_3 = 1,\ldots,M$, has the same hilbertian product with $g_{\theta^\star,\mathbf{f}^*}$ and with $g_{\theta^\star,\omega_M^\star,M}$. Thus, if for any M, $\mathbb{A}_M$ is the orthogonal projection in $L^2(g_{\theta^\star,\mathbf{f}^*}d\mathbf{y})$ onto $\dot{\mathcal{P}}_M$, the set of step functions spanned by the functions $\left(S^\star_{\omega,M}\right)_{j,c,m}$, $j=1,\ldots,k$, $c=1,2,3$, $m=1,\ldots,M-1$, then for all $j=1,\ldots,k-1$,

$$(\tilde{\psi}_M)_j = \left(S^\star_{\theta,M}\right)_j - \mathbb{A}_M\left(S^\star_{\theta,M}\right)_j,$$

so that

$$(\tilde{\psi})_j - (\tilde{\psi}_M)_j = (S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j - \mathbb{A}_M\left[(S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j\right] + (\mathbb{A}_M - \mathbb{A})(S^\star_\theta)_j. \tag{4.18}$$

Notice that using (A3.3),

$$\dot{\mathcal{P}}_M \subset \dot{\mathcal{P}} \tag{4.19}$$

so that $\mathbb{A}_M\mathbb{A} = \mathbb{A}_M$. We then obtain

$$\left\|(\tilde{\psi})_j - (\tilde{\psi}_M)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})} \le \left\|(S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})} + \left\|(\mathbb{A}_M\mathbb{A} - \mathbb{A})(S^\star_\theta)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})}.$$

Using Assumption (A3.2) and Corollary 1.7 in Chapter 3 of Stein and Shakarchi [SS05], we have that as M tends to infinity, $\left(S^\star_{\theta,M}\right)_j$ converges to $(S^\star_\theta)_j$ Lebesgue a.e. Both functions are uniformly upper bounded by the finite constant $1/\theta_j^\star$ using Assumption (A.1), so that $\left(S^\star_{\theta,M}\right)_j$ converges to $(S^\star_\theta)_j$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{y})d\mathbf{y})$ as M tends to $+\infty$ and $\left\|(S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})}$ converges to 0 as M tends to $+\infty$. Using the same argument, for any function $S\in\dot{\mathcal{P}}$ there exists a sequence of functions $S_M\in\dot{\mathcal{P}}_M$ that converges to S in $L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})$. Let $(S_M)_M$ be the sequence of functions converging to $\mathbb{A}(S^\star_\theta)_j$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})$. Since for all M, $S_M\in\dot{\mathcal{P}}_M$, we have that

$$\left\|\mathbb{A}_M\left[\mathbb{A}(S^\star_\theta)_j\right] - \mathbb{A}(S^\star_\theta)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})} \le \left\|S_M - \mathbb{A}(S^\star_\theta)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})}$$

so that also $\left\|(\mathbb{A}_M\mathbb{A} - \mathbb{A})(S^\star_\theta)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})}$ converges to 0 as M tends to $+\infty$. We thus obtain that $(\tilde{\psi})_j$ converges to $(\tilde{\psi}_M)_j$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{y})$. As a consequence, $\tilde{J}_M$ converges to $\tilde{J}$ as M tends to $+\infty$. □

4.6.4 Proof of Proposition 4.6

Proposition 4.6 is easily implied by Lemma 4.10 which formalizes the following. When the sequence of observations $Y_1, \dots, Y_n$ and n are fixed, then almost surely there exists a sufficiently fine partition $\mathcal{I}_M$ such that there exists at most one component of an observation in each set I_m, $m \leq M$. Then we can reorder the sets I_m so that $Y_{i,c} \in I_{i+n(c-1)}$, for all $c \in \{1,2,3\}$ and $i \leq n$. In this case, the likelihood $\ell_n(\cdot,\cdot;M)$ is maximised at each parameter (θ,ω) belonging to the set $\mathcal{S}_M \subset \Delta_k \times (\Delta_M)^{3k}$ that we explain now (and formalise in Lemma 4.10). Each element of $\mathcal{S}_M$ corresponds to one clustering of the observations in k sets (represented by the $(A_j^*)_{j\leq k}$ in Lemma 4.10) of size as equal as possible. For each clustering, for all $j \leq k$,

• $\theta_j = \#A_j^*/n$ is the proportion of observations associated to A_j^* (then the θ_j are almost equal to $1/k$),

• for all $c \in \{1,2,3\}$ and for all $l \leq M$,

$$\omega_{j,c,l} = \begin{cases} 1/\#A_j^* & \text{if } l - n(c-1) \in A_j^* \text{ (i.e. } Y_{l-n(c-1)} \in I_l \text{ is associated to the hidden state } j), \\ 0 & \text{if } l - n(c-1) \in \{1,\dots n\} \setminus A_j^* \text{ (i.e. } Y_{l-n(c-1)} \in I_l \text{ is not associated to } j), \\ 0 & \text{otherwise (i.e. there is no observation in } I_l). \end{cases}$$

Lemma 4.10. *Let $Y_1, \dots, Y_n$ be fixed observations, as soon as for all $i \leq n$ and $c \in \{1,2,3\}$, $Y_{i,c} \in I_{i+n(c-1)}$ then the likelihood $\ell_n(\cdot,\cdot;M)$ is maximised at $(\hat{\theta}_M, \hat{\omega}_M)$ if and only if $(\hat{\theta}_M, \hat{\omega}_M) \in \mathcal{S}_M$ where*

$$\begin{aligned} \mathcal{S}_M = \big\{(\theta,\omega) : \theta_j = \#A_j^*/n,\ \omega_{j,c,l} = \mathbb{1}_{l-n(c-1)\in A_j^*}/\#A_j^*, \\ (J_1,J_2) \text{ partition of } \{1,\dots,k\},\ \#J_2 = n - k\lfloor n/k \rfloor =: r \\ (A_j^*)_{j\leq k} \text{ partition of } \{1,\dots,n\}, \\ \#A_{j_1}^* = \lfloor n/k \rfloor =: q, \text{ for } j_1 \in J_1,\ \#A_{j_2}^* = \lfloor n/k \rfloor + 1 =: q+1, \text{ for } j_2 \in J_2\big\}, \end{aligned}$$

and $n = kq + r$, $0 \leq r \leq k-1$.

Proof. Since the set of parameters is compact and the likelihood is a continuous function of the parameters then the maximum is attained.

If (θ,ω) maximises the likelihood $\ell_n(\cdot,\cdot;M)$,

(P1) then, for all $1 \leq i \leq n$, there exists $1 \leq j \leq k$ such that $\omega_{j,c,i+n(c-1)} > 0$ for all $c \in \{1,2,3\}$.
Indeed, if there exists $1 \leq i \leq n$ such that for all $1 \leq j \leq k$, $\omega_{j,c,i+n(c-1)} = 0$ for some $c \in \{1,2,3\}$, then

$$\ell_n(\theta,\omega;M) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \omega_{j,c,i+n(c-1)}\right) + \underbrace{\sum_{i=1}^{n} \log\left(1/(|I_i||I_{i+n}||I_{i+2n}|)\right)}_{\text{constant}} = -\infty.$$

(P2) and if there exists j, c, i such that $\omega_{j,c,i+n(c-1)} = 0$ and $\theta_j > 0$ then $\omega_{j,d,i+n(d-1)} = 0$ for all d.
Indeed otherwise you can give the weight $\omega_{j,d,i+n(d-1)}$, to one of the other $\omega_{j,d,s+n(d-1)}$ for which $\omega_{j,e,s+n(e-1)} > 0$, for all $e \neq d$ (which exist otherwise take $\theta_j = 0$ which would increase the likelihood) and this increases the likelihood.

(P3) and if $\theta_j > 0$, then $\omega_{j,c,l} = 0$ if $l - n(c-1) \notin \{1, \dots, n\}$.
Indeed, in this case, there is no observation in I_l so that $\omega_{j,c,l}$ does not appear in the likelihood and we conclude similarly as the previous point.

Combining all the previous remarks, we know that the maximum can only be attained (and is at least once) in one of the following sets, indexed by $J \subset \{1, \dots, k\}$ which determines the zeros of θ and $A_j \subset \{1, \dots, n\}$, $j \leq k$, which determine the zeros of ω:

$$\begin{aligned}\mathcal{S}_{J,A_1,\dots,A_k} =& \{\theta \in \Delta_k : \ \theta_j > 0, \ j \in J, \theta_j = 0, \ j \in J^c\} \\ & \times \prod_{j \leq k} \Big\{ (\omega_{j,1,\cdot}, \omega_{j,2,\cdot}, \omega_{j,3,\cdot}) \in (\Delta_M)^3 : \\ & \qquad \text{if } j \in J, \quad \underset{\text{using (P2)}}{\omega_{j,c,i+n(c-1)}} > 0 \quad \underset{\text{using (P3)}}{}, \text{ if } \ i \in A_j, \ c \in \{1,2,3\} \\ & \qquad \text{and } \omega_{j,c,l} = 0, \ \text{if } l \in \{1, \dots, M\} \setminus \{i + n(c-1), \ i \in A_j\} \Big\}.\end{aligned}$$

Note that we do not assume that $(A_j)_{j \in J}$ is a partition of $\{1, \dots, n\}$.

We fix $J \subset \{1, \dots, k\}$ and $A_j \subset \{1, \dots, n\}$, $j \in J$. Now we search for parameters $(\bar{\theta}, \bar{\omega})$ in $\mathcal{S}_{J,A_1,\dots,A_k}$ which maximize the likelihood. They are zeros of the derivative of

$$(\theta, \omega, \lambda, \mu) \mapsto \ell_n(\theta, \omega; M) + \lambda \left(\sum_{j=1}^{k} \theta_j - 1 \right) + \sum_{c=1}^{3} \mu_{j,c} \left(\sum_i \omega_{j,c,i} - 1 \right), \tag{4.20}$$

with respect to nonzero components (θ_j, $\omega_{j,c,i+n(c-1)}$, λ and $\mu_{j,c}$, for $j \in J$, $i \in A_j$, $1 \leq c \leq 3$). Annulling the partial derivatives give

$$\sum_{i \in A_j} \frac{\bar{\omega}_{j,1,i} \bar{\omega}_{j,2,i+n} \bar{\omega}_{j,3,i+2n}}{\sum_{s \in J(i)} \bar{\theta}_s \bar{\omega}_{s,1,i} \bar{\omega}_{s,2,i+n} \bar{\omega}_{s,3,i+2n}} = -\lambda, \qquad \forall j \in J \tag{4.21}$$

$$\frac{\bar{\theta}_j \prod_{d \neq c} \bar{\omega}_{j,d,i+n(d-1)}}{\sum_{s \in J(i)} \bar{\theta}_s \bar{\omega}_{s,1,i} \bar{\omega}_{s,2,i+n} \bar{\omega}_{s,3,i+2n}} = -\mu_{j,c}, \qquad \forall j \in J, \ i \in A_j, \ c \in \{1,2,3\} \tag{4.22}$$

$$\sum_{j \in J} \bar{\theta}_j = 1, \tag{4.23}$$

$$\sum_{i \in A_j} \bar{\omega}_{j,c,i+n(c-1)} = 1, \qquad \forall j \in J, \ c \in \{1,2,3\}, \tag{4.24}$$

where $J(i) = \{s \in J : \ i \in A_s\}$.

Multiplying Equation (4.22) by $\bar{\omega}_{j,c,i+n(c-1)}$ and then summing the result over $i \in A_j$ and using Equation (4.24), we obtain that $\mu_{j,c}$ does not depend on c. Then using Equations (4.22) for $c = 1$, $c = 2$ and $c = 3$, we obtain

$$\bar{\theta}_j \bar{\omega}_{j,1,i} \bar{\omega}_{j,2,i+n} = \bar{\theta}_j \bar{\omega}_{j,1,i} \bar{\omega}_{j,3,i+2n} = \bar{\theta}_j \bar{\omega}_{j,2,i+n} \bar{\omega}_{j,3,i+2n},$$

so that

$$\bar{\omega}_{j,1,i} = \bar{\omega}_{j,2,i+n} = \bar{\omega}_{j,3,i+2n}. \tag{4.25}$$

Furthermore, multiplying Equation (4.21) by $\bar{\theta}_j$ and summing the result over $j \in J$ and using Equation (4.23), we obtain $\lambda = -n$. Moreover by multiplying Equation (4.22) by $\bar{\omega}_{j,c,i+n(c-1)}$, and then summing the result over $i \in A_j$ and finally subtracting (4.21) multiplied by $\bar{\theta}_j$ to the result (ie making $\sum_{i \in A_j} \bar{\omega}_{j,c,i+n(c-1)}(4.22) - \bar{\theta}_j(4.21)$), we get

$$0 = -\mu_{j,c} - n\bar{\theta}_j. \tag{4.26}$$

Then using again Equations (4.22), (4.25) and (4.26), we get

$$\bar{\omega}^2_{j,c,i+n(c-1)} = n \sum_{s \in J(i)} \bar{\theta}_s \bar{\omega}^3_{s,1,i}, \quad \forall j \in J(i),\ \forall c \in \{1,2,3\},$$

so that $\bar{\omega}_{j,c,i+n(c-1)}$ does not depend on $j \in J(i)$ and

$$\bar{\omega}_{j,c,i+n(c-1)} = \mathbb{1}_{i \in A_j} \Big/ \left(n \sum_{s \in J(i)} \bar{\theta}_s \right), \quad \forall j \in J(i). \tag{4.27}$$

For each $\mathcal{S}_{J,A_1,\dots,A_k} =: \mathcal{S}$, we have obtained the zeros of the derivative of the log-likelihood, that we now denote $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega})$, to emphasize the dependence with the considered set $\mathcal{S}$. We now want to know which of these zeros $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega})$ are local maxima thanks to the second partial derivatives. We consider sets $\mathcal{S}_{J,A_1,\dots,A_k}$ for which there exists $i \leq n$ such that there exist j and l are in $J(i)$ and $j \neq i$. We consider a second partial derivative of

$$\tilde{\ell}_n(\theta, \tilde{\omega}; M) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \theta_j (\tilde{\omega}_{j,1,i})^3 \right)$$

that is the log-likelihood (up to an additive constant) associated to the model where for all $1 \leq m \leq k$, $1 \leq s \leq n$, $\omega_{m,1,s} = \omega_{m,2,s+n} = \omega_{m,3,s+2n}$. Assume without loss of generality that $\theta_l \geq \theta_j$, then (using that $\theta_k = 1 - \sum_{m<k} \theta_m$ and $\omega_{j,1,n} = 1 - \sum_{s<n} \omega_{j,1,s}$),

$$\frac{\partial^2 \tilde{\ell}_n}{\partial \tilde{\omega}^2_{j,1,i}}({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega}; M) = C \left(6\, {}^{\mathcal{S}}\bar{\theta}_j \sum_{m \in J(i) \setminus \{j\}} {}^{\mathcal{S}}\bar{\theta}_m - 3\, {}^{\mathcal{S}}\bar{\theta}^2_j \right) \geq C \left(6\, {}^{\mathcal{S}}\bar{\theta}_j\, {}^{\mathcal{S}}\bar{\theta}_l - 3\, {}^{\mathcal{S}}\bar{\theta}^2_j \right) > 0,$$

where $C > 0$. This implies that for all sets $\mathcal{S}_{J,A_1,\ldots,A_k} := \mathcal{S}$ where there exists $i \leq n$ such that $\#J(i) > 1$, every zeros $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega})$ is not a local maximum. So that the only possible local maxima of $\ell_n(\theta, \omega; M)$ are the zeros $({}^{\mathcal{S}_{J,A_1,\ldots,A_k}}\bar{\theta}, {}^{\mathcal{S}_{J,A_1,\ldots,A_k}}\bar{\omega})$ where $\#J(i) = 1$ for all $i \leq n$, i.e. when $(A_j)_{j\in J}$ forms a partition of $\{1, \ldots, n\}$.

So we now only consider sets A_j, $j \in J$ which form a partition of $\{1, \ldots, n\}$ and $\bar{\omega}_{j,c,i+n(c-1)} = \mathbb{1}_{i\in A_j}/(n\bar{\theta}_j)$ for $i \in A_j$, using Equation (4.27). As $\sum_{i\in A_j} \bar{\omega}_{j,1,i} = 1$, we then obtain that $\bar{\theta}_j = \#A_j/n = 1/(n\bar{\omega}_{j,1,i})$, for all $i \in A_j$. So that we now only have to choose the best partition $(A_j)_{j\in J}$ of $\{1, \ldots, n\}$ and J. Let $N_j = \#A_j$, we know that $\sum_j N_j = n$ and the log-likelihood at the local maximum $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega})$ associated to $\mathcal{S}_{J,A_1,\ldots,A_k} =: \mathcal{S}$ is

$$\ell_n({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega}; M) = \sum_{s\in J} N_s \log(N_s^{-2}) + constant.$$

So that we want to minimize

$$\sum_{s\in J} N_s \log(N_s) \text{ under the constraint } \sum_{s\in J} N_s = n \tag{4.28}$$

over $J \subset \{1, \ldots k\}$ and $N_j \in \mathbb{N}$, $j \in J$. This minimization is equivalent to the minimization of

$$\sum_{s\leq k} N_s \log(N_s) \text{ under the constraint } \sum_{s\leq k} N_s = n \tag{4.29}$$

over $N_j \in \mathbb{N}$, $j \leq k$ (since then the problem (4.29) is less constrained than for the minimization of (4.28) when J is fixed).

And, when k divides n, the minimum of (4.29) is attained at $N_s = n/k$. Otherwise, when k does not divide n, consider only two indices s_1, s_2 in $\{1, \ldots, k\}$ and assume that N_s, $s \notin \{s_1, s_2\}$ are fixed such that $N_{s_1} + N_{s_2} = S_N$ is also fixed. Then we want to minimise $-N_{s_1}\log(N_{s_1}) - (S_N - N_{s_1})\log(S_N - N_{s_1})$. Studying the function $x \in (0, S_N) \mapsto -x\log(S_N) - (S_N - x)\log(S_N - x)$, we obtain that the minimum is attained when N_{s_1} and $N_{s_2} = S_N - N_{s_1}$ are the closest of $N_S/2$. Then in both cases, the m.l.e. is attained at every $(\theta, \omega) \in \mathcal{S}_M$.

□

4.6.5 Proof of Corollary 4.7

Suppose that for all $N > 0$ and all $C > 0$, there exists $n \geq N$ such that

$$n^2 \left(\max_{m\leq M_n} |I_m| \right)^2 M_n \leq C.$$

So that there exists a subsequence $(\phi(n))_{n\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$(\phi(n))^2 \left(\max_{m\leq M_{\phi(n)}} |I_m|\right)^2 M_{\phi(n)} \underset{n\to\infty}{\longrightarrow} 0. \tag{4.30}$$

Set $\epsilon > 0$, by Proposition 4.6, there exists $N_1 > 0$ such that for all $n \geq N_1$,

$$\begin{aligned}
&P\left(\left|\widehat{\theta}_{M_n}(Y_{1:\phi(n)}) - (1/k,\dots,1/k)\right| \leq \epsilon\right) \\
&\quad\geq P\left(\left\{\exists\, 1\leq i_1,i_2\leq \phi(n),\ 1\leq c,d\leq 3,\ m\leq M_{\phi(n)} :\ Y_{i_1,c}\in I_m, Y_{i_2,d}\in I_m\right\}^c\right) \\
&\quad\geq 1 - \sum_{i_1=1}^{\phi(n)}\sum_{i_2=1}^{\phi(n)}\sum_{m=1}^{M_{\phi(n)}} P\left(Y_{i_1,c}\in I_m, Y_{i_2,d}\in I_m\right) \\
&\quad\geq 1 - (\phi(n))^2 M_{\phi(n)} \max\left(\sup g, (\sup g)^2\right)\left(\max_{m\leq M_{\phi(n)}} |I_m|\right)^2.
\end{aligned} \tag{4.31}$$

Using Equations (4.30) and (4.31) and Assumption (A3.3), then $\widehat{\theta}_{M_n}(Y_{1:\phi(n)})$ tends in probability to $(1/k,\dots,1/k)$ which contradicts the convergence in law of $\widehat{\theta}_{M_n}$ to θ^*. This concludes the proof.

4.6.6 Proof of Theorem 4.8

We first recall Lemma 2.1 in Arlot [Arl14]:

Lemma 4.11 (Sylvain Arlot). *Let $A, B, C, R : \mathcal{M} \to \mathbb{R}$. If for all $m, m' \in \mathcal{M}$,*

$$(C(m) - R(m)) - (C(m') - R(m')) \leq A(m) + B(m'),$$

then for all $\widehat{m} \in \mathcal{M}$ such that $C(\widehat{m}) \leq \inf_{m\in\mathcal{M}} C(m) + \rho$, $\rho > 0$,

$$R(\widehat{m}) - B(\widehat{m}) \leq \inf_{m\in\mathcal{M}}\{R(m) + A(m)\} + \rho.$$

We are going to use this lemma with $R(\mathcal{I}) = R_{a_n}(\mathcal{I})$, $C(\mathcal{I}) = C_{CV}(\mathcal{I})$ and

$$A(\mathcal{I}) = B(\mathcal{I}) = \epsilon_n R(\mathcal{I}) + \delta_n.$$

Using Hoeffding's inequality,

$$P(\{-B(\mathcal{I}) \leq C_{CV}(\mathcal{I}) - R_{a_n}(\mathcal{I}) \leq A(\mathcal{I})\}^c) \leq 2\exp\left(-2b_n A(\mathcal{I})^2\right),$$

since $\|\widehat{\theta}_{\mathcal{I}}(Y_{B_b}) - \widehat{\theta}_{\mathcal{I}_0}(Y_{B_{-b}})\|^2 \leq 1$, for all b. We introduce the sets

$$\mathcal{S}_{\mathcal{I}} = \{-B(\mathcal{I}) \leq C_{CV}(\mathcal{I}) - R_{a_n}(\mathcal{I}) \leq A(\mathcal{I})\} \tag{4.32}$$

for all $\mathcal{I} \in \mathcal{M}_n$. Using Lemma 4.11, on the set $\cap_{\mathcal{I}\in\mathcal{M}_n}\mathcal{S}_{\mathcal{I}}$, Equation (4.17) holds and using Equation (4.32), we obtain

$$P(\cap_{\mathcal{I}\in\mathcal{M}_n}\mathcal{S}_{\mathcal{I}}) \geq 1 - 2m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right).$$

4.6.7 Proof of Proposition 4.9

Using Theorem 4.8,

$$\begin{aligned}
&\mathbb{E}^\star\left[a_n R_{a_n}(\widehat{\mathcal{I}}_n)\right] \\
&\leq a_n\left(\frac{1+\epsilon_n}{1-\epsilon_n}\inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \frac{2\delta_n}{1-\epsilon_n}\right) + 2a_n m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right)
\end{aligned}$$

we can conclude by taking $\epsilon_n = \delta_n = 1/(\log(n)a_n)$.

APPENDIX A

RÉSUMÉ LONG

1.1 Introduction

Dans cette thèse nous considérons deux modèles latents, c'est-à-dire des modèles où une observation dépend d'un état caché (ou latent). Dans ces deux modèles sachant les états cachés $(X_t)_{t\in\mathbb{N}}$, les observations $(Y_t)_{t\in\mathbb{N}}$ sont indépendantes avec Y_t qui ne dépend que de X_t via la loi d'émission F_{X_t}. Ainsi les observations sont une version bruitée des états cachés. Dans le premier modèle étudié, les modèles de Markov cachés, les états cachés X_t, $t \in \mathbb{N}$ forment une chaîne de Markov alors que dans le second cas, les modèles de mélange, les états cachés X_t, $t \in \mathbb{N}$ sont i.i.d.. On pourra trouver une représentation de ces modèles dits de Markov cachés et de mélange dans les Figures A.1 et A.2 respectivement. Le terme modèle de Markov caché pourra être abrégé grâce à son acronyme anglais HMM.

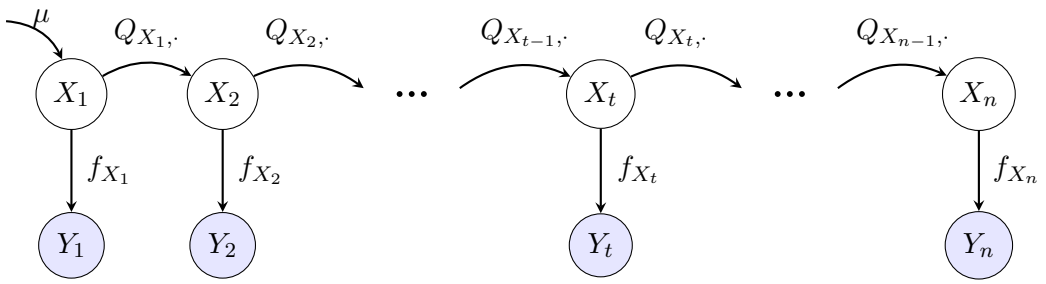


Figure A.1 – Visualisation d'un HMM.

Ces deux modèles sont très utilisés en pratique, par exemple en génomique, reconnaissance de parole, économétrie, climatologie, étude de populations. Récemment, un intérêt croissant a été donné aux modèles de mélange et de Markov cachés non paramétriques en pratique. En effet, leurs homologues paramétriques souffrent de problèmes de robustesse. La généralisation de ces modèles, dans le cas où le nombre d'état pris par les états cachés est fini, s'est faite de deux façons

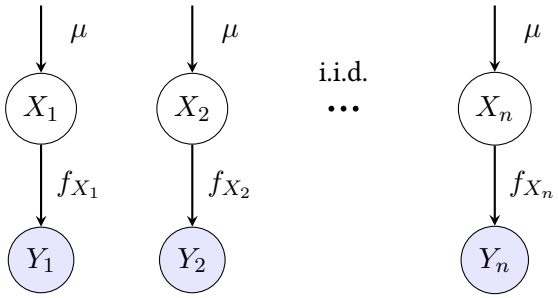


Figure A.2 – Visualisation d'un modèle de mélange.

:

- en n'imposant plus de borne sur le nombre d'états pris par les états latents,
- en ne supposant plus que les lois des observations sachant les états cachés (soit les lois d'émission) étaient paramétrisées de façon paramétrique (i.e. avec un paramètre vivant dans un espace de dimension finie).

Dans cette thèse nous ne considérons que le deuxième cas, c'est-à-dire les modèles de mélange et de Markov caché où le nombre d'états possible pour X_t est fini et connu mais les lois d'émission ne sont pas contraintes à vivre dans un espace de dimension finie.

L'utilisation de ces modèles non paramétriques induit de nombreuses questions théoriques sans réponses. Dans cette thèse nous nous attachons à obtenir des garanties théoriques sur des estimateurs ou la loi a posteriori dans ces modèles.

Le Chapitre 1 propose une introduction plus poussée (en anglais) aux problèmes étudiés, on y trouvera notamment la description des modèles considérés, une description des propriétés asymptotiques analysées dans cette thèse, et enfin la version anglaise des contributions apportées par cette thèse. Voici mes contributions en français.

1.2 Contributions

Durant ma thèse, j'ai travaillé sur trois projets permettant de mieux comprendre certaines propriétés théoriques d'estimateurs ou de la loi a posteriori dans le cadre des modèles présentées dans la partie précédente. Je me suis tout d'abord intéressée au problème de consistance de la loi a posteriori dans les modèles de Markov cachés, voir le Chapitre 2. J'ai ensuite étudié la vitesse de concentration de la loi a posteriori dans ces mêmes modèles, voir le Chapitre 3. Pour finir, j'ai considéré un problème d'estimation semi-paramétrique dans les modèles de mélange. Ce dernier projet de recherche s'est fait en collaboration avec mes deux directrices de thèse Elisabeth Gassiat et Judith Rousseau.

Dans la suite, je présente mes contributions de manière informelle, pour plus de détails (mathématiques), voir les chapitres concernés.

1.2.1 Contribution 1 : Consistance de la loi a posteriori dans les modèles de Markov cachés non paramétriques à espace d'états finis, Chapitre 2, Vernet [Ver15b]

Ma première contribution concerne la consistance de la loi a posteriori dans les modèles de Markov cachés non paramétriques à espace d'états finis. Ce sujet est développé dans le Chapitre 2 qui correspond aussi à l'article Vernet [Ver15b] publié dans EJS.

Je précise ici le cadre de cette contribution. On se place dans le cas où la chaîne de Markov (cachée) $(X_t)_{t\in\mathbb{N}}$ prend ses valeurs dans un espace d'état fini $\{1,\dots,k\}$ et on connaît le nombre d'états k . Quant aux observations Y_t, $t\in\mathbb{N}$, on suppose qu'elles vivent dans $\mathbb{R}^d$. Le modèle est paramétré par μ et $\theta=(Q,f)$ où μ est la loi initiale de la chaîne de Markov, Q est la matrice de transition, enfin $f=(f_1,\dots,f_k)$ est le vecteur constitué des k densités d'émission par rapport à une mesure λ. μ et Q décrivent le comportement de la chaîne de Markov X_t sous-jacente et f décrit la loi des observations sachant $(X_t)_{t\in\mathbb{N}}$. Donc

$$
\begin{aligned}
&X_t\in\{1,\dots,k\},\quad Y_t\in\mathbb{R}^d\\
&X_1\sim\sum_{i=1}^k\mu_i\delta_i,\quad X_{t+1}|X_1,\ \dots,\ X_t\sim X_{t+1}|X_t\sim\sum_{i=1}^k Q_{X_t,i}\delta_i\\
&Y_1,\dots,Y_s,\dots|(X_t)_{t\in\mathbb{N}}\text{ sont indépendantes},\quad Y_t|(X_t)_{t\in\mathbb{N}}\sim Y_t|X_t\sim f_{X_t}(\cdot)d\lambda,
\end{aligned}
$$

où δ_x est la mesure de Dirac en x. Ce modèle est représenté dans la Figure A.1.

On se place ici dans le cadre Bayésien, on a donc besoin d'une loi sur l'espace Θ des paramètres (la loi a priori). On utilise $\Pi=\Pi_Q\otimes\Pi_f^{(k)}$ qui est un produit d'une loi de probabilité Π_Q sur les matrices de transition et une loi de probabilité $\Pi_f^{(k)}$ sur les k densités d'émission et δ_μ pour la loi a priori sur la loi initiale avec μ une loi initiale donnée. Par la loi de Bayes, on peut formellement écrire la loi a posteriori comme suit :

$$
\Pi(\theta\in A|Y_1,\dots,Y_n)=\frac{\int_A p_n^{\mu,\theta}(Y_1,\dots,Y_n)\Pi(d\theta)}{\int_\Theta p_n^{\mu,\theta}(Y_1,\dots,Y_n)\Pi(d\theta)},
$$

où $p_n^{\mu,\theta}(y_1,\dots,y_n)=\sum_{1\le i_1,\dots,i_n\le k}\mu_{i_1}Q_{i_1,i_2}\dots Q_{i_{n-1},i_n}f_{i_1}(y_1)\dots f_{i_n}(y_n)$ est la vraisemblance.

La loi a posteriori remplace le rôle de l'estimateur dans le cadre fréquentiste. Elle permet de donner une idée du paramètre duquel proviennent les observations. Remarquez que contrairement à un estimateur fréquentiste usuel, la loi a posteriori est une loi de probabilité sur l'ensemble des paramètres.

Dans la suite, je m'intéresse aux garanties théoriques qu'on peut obtenir sur la loi a posteriori.

En particulier, je m'intéresse au comportement asymptotique de cette loi, c'est-à-dire lorsque le nombre d'observations tend vers l'infini. Dans ce but, je prends un point de vue fréquentiste en supposant que les observations proviennent d'un vrai paramètre θ^*. Dans ce cas, il paraît naturel que la loi a posteriori concentre sa masse sur le vrai paramètre θ^*. On appelle ce comportement la consistance de la loi a posteriori. Formellement, on dit que la loi a posteriori est consistante en θ^* par rapport à la pseudo-métrique d lorsque,

$$\Pi\left(\{\theta:\ d(\theta,\theta^*)>\epsilon\}\mid Y_1,\dots,Y_n\right)\to 0,\quad P^{\theta^*}-a.s.,\quad \text{pour tout } \epsilon>0.$$

La consistance est une exigence minimale sur la loi a posteriori. L'étude de la consistance de la loi a posteriori dans le cadre des HMMs à espace d'état fini est l'objet de ma première contribution. En particulier j'ai étudié cette garantie en considérant différentes pseudo-métriques d, c'est-à-dire différentes topologies sur différents objets.

Consistance de la loi a posteriori pour l'estimation de la loi marginale jointe P_l^θ de l observations consécutives

J'ai tout d'abord cherché à savoir si la loi a posteriori concentrait sa masse autour des paramètres θ tels que les lois P_l^θ, de l observations stationnaires consécutives associées (c'est-à-dire la loi de $Y_1, Y_2, \dots, Y_l$ sous une loi stationnaire associée à θ), étaient proches de $P_l^{\theta^*}$. Cette étude est intéressante dans le cadre de la prédiction. En effet, si la loi a posteriori est consistante pour cet objet P_l^θ alors la loi des observations peut être estimée de façon consistante. Cette étude se révélera aussi utile dans le cadre de l'estimation de Q et f, voir la partie suivante.

Deux topologies sont utilisées pour comparer les lois $(P_l^\theta)_\theta$. On considère la topologie $\mathcal{T}_w$ associée à la convergence en loi ainsi que la topologie plus fine $\mathcal{T}_l$ associé à la norme L_1, qui correspond à l'utilisation de la pseudo-métrique D_l sur Θ:

$$D_l(\theta,\tilde\theta)=\|p_l^\theta-p_l^{\tilde\theta}\|_{L^1(\lambda^{\otimes l})},$$

où p_l^θ est la densité associée à P_l^θ par rapport à $\lambda^{\otimes l}$.

Pour obtenir un théorème général portant sur la consistance de la loi a posteriori pour les deux topologies précédentes, j'ai utilisé la "méthode usuelle", plus précisément Barron [Bar88]. Cette méthode consiste à montrer que la loi a priori met suffisamment de poids dans le voisinage de Kullback du vrai paramètre et à prouver l'existence de certains tests (qui est souvent démontrée par le fait que la loi a priori ne met pas trop de poids sur des espaces trop grands, i.e. pénalise suffisamment les espaces complexes). Voir le Théorème 1.11 pour plus de précisions.

J'ai explicité les hypothèses provenant de Barron [Bar88] dans le cadre des HMMs. L'existence des tests était déjà démontrée dans Gassiat and Rousseau [GR14], elle s'appuie sur une généralisation de l'inégalité d'Hoeffding pour des données dépendantes par Rio [Rio00]. Il me restait donc à

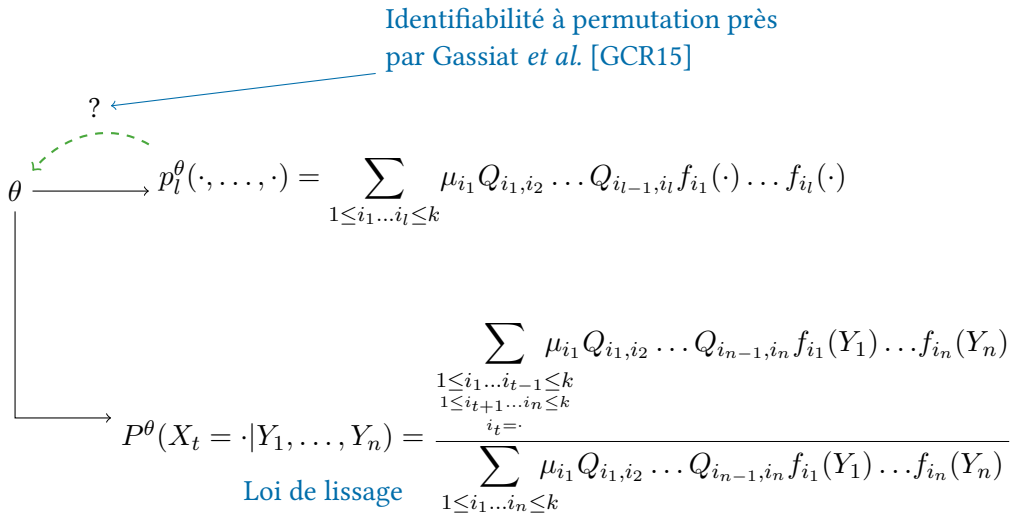


Figure A.3 – Obtenir de l'information sur θ ou les lois de lissage à partir de p_l^θ

expliciter un ensemble de paramètres pour lequel la divergence de Kullback associée au vrai paramètre était petite. Ceci est fait dans le Lemme 2.2.

Ainsi, j'ai obtenu que si Π_Q met du poids dans tous les voisinages de Q^* (voir l'hypothèse (A1.1a)) et que Π_f met du poids dans certains voisinages de f^* (voir les hypothèses (A1.1b), (A1.1c) et (A1.1d) pour avoir la description exacte de ces voisinages) alors la loi a posteriori est consistante en $\theta^* = (Q^*, f^*)$ par rapport à $\mathcal{T}_w$, voir le Théorème 2.1. J'ai aussi obtenu que si de plus Π_f ne met pas trop de poids sur des espaces trop gros (voir l'hypothèse (A1.2)) alors la loi a posteriori est consistante en $\theta^* = (Q^*, f^*)$ par rapport à $\mathcal{T}_l$, voir le Théorème 2.1.

Les deux topologies précédentes $\mathcal{T}_w$ et $\mathcal{T}_l$ concernaient la loi jointe marginale de l observations. Or on pourrait être intéressée par d'autres quantités comme le paramètre θ en lui-même ou les lois de lissage (i.e. la loi d'un état caché sachant les observations) par exemple. Ainsi j'ai cherché à comprendre ce que ça signifie sur Q et f ou sur $P^{(Q,f)}(X_t = \cdot|Y_1,\dots,Y_n)$ lorsque $p_l^{(Q,f)}$ est proche de $p_l^{(Q^*,f^*)}$ en norme L^1, voir la Figure A.3 pour une illustration. Ce problème n'est a priori pas facile car il est lié au problème d'identifiabilité des HMMs (i.e. de l'injectivité de $\theta \mapsto p_l^\theta$) qui est loin d'être un problème facile dans les HMMs. On parle de la résolution de ce problème dans les deux parties suivantes.

Consistance de l'a posteriori pour l'estimation de Q et f

Dans cette partie, on s'intéresse au problème de l'estimation du paramètre θ en lui-même, c'est-à-dire de la matrice de transition Q et des densités d'émission f_j, $j \le k$. Ainsi, on cherche à savoir si la loi a posteriori concentre sa masse autour des paramètres (Q, f) tels que Q est proche de Q^* et f est proche de f^*.

Obtenir ce type de consistance à partir de la consistance sur la loi des observations est intimement

lié à l'identifiabilité du modèle puisqu'on cherche à comprendre l'inverse de

$$\theta \in \Theta \mapsto p_l^\theta \tag{A.1}$$

si elle existe. L'existence de l'inverse est assurée par l'identifiabilité. Or Gassiat *et al.* [GCR15] ont montré qu'en supposant que k est connu, que Q est une matrice de rang plein et que les lois d'émission $f_1\lambda, \dots, f_k\lambda$ sont linéairement indépendantes ; si la loi de 3 observations $p_3^\theta \lambda^{\otimes 3}$ est égale à $p_3^{\tilde{\theta}} \lambda^{\otimes 3}$ alors il existe une permutation des états $\sigma \in \mathcal{S}_k$ telle que $Q_{i,j} = \tilde{Q}_{\sigma(i),\sigma(j)}$ et $f_i d\lambda = f_{\sigma(i)} d\lambda$ pour tout $1 \leq i, j \leq k$. On peut donc retrouver les paramètres à permutation des états près à partir de la loi jointe de 3 observations successives. Ainsi le modèle est identifiable à permutation des états près. Et on ne peut pas espérer retrouver exactement (Q, f) en toute généralité mais seulement à permutation des états cachés près, puisque les lois jointes de l observations, associées à des paramètres provenant d'une permutation des états cachés, sont les mêmes. Il en est de même pour la consistance. En effet, supposons que la loi a priori soit compatible avec la permutation des états cachés, c'est-à-dire

$$\Pi(U) = \Pi(\sigma U), \quad \forall U \subset \Theta, \quad \forall \sigma \in \mathcal{S}_k,$$
$$\sigma U = \{((Q_{\sigma(i),\sigma(j)})_{i,j}, (f_{\sigma(1)}, \dots, f_{\sigma(k)})) \in \Theta : \ (Q, f) \in U\},$$

où $\mathcal{S}_k$ est l'ensemble des permutations sur $\{1, \dots, k\}$. Alors la masse a posteriori d'un ensemble U de paramètres est aussi égale à la masse a posteriori des paramètres dans U pour lesquels les états cachés ont subi une permutation σ. Formellement,

$$\Pi(U|Y_1, \dots, Y_n) = \frac{\int_U p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)}{\int_\Theta p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)} = \frac{\int_U p_n^{\sigma\theta}(Y_1, \dots Y_n)\Pi(d\sigma\theta)}{\int_\Theta p_n^\theta(Y_1, \dots Y_n)\Pi(d\theta)} = \Pi(\sigma U|Y_1, \dots, Y_n),$$

pour toute permutation $\sigma \in \mathcal{S}_k$. Ainsi, le meilleur comportement de la loi a posteriori concernant la consistance serait que la loi a posteriori concentre sa masse en $\{\theta^*\}_{\mathcal{S}_k} = \cup_{\sigma \in \mathcal{S}_k} \sigma\{\theta^*\}$. Si la loi a priori est plus générale, lorsque le nombre d'observations augmente, la loi a priori devrait être "oubliée" et on devrait demander la concentration de la loi a posteriori autour du même ensemble $\{\theta^*\}_{\mathcal{S}_k}$.

On cherche alors à étudier la consistance par rapport à la topologie $\mathcal{T}_{Q,f}$ qui est le produit de la topologie associée à la norme sup sur les matrices de transition et la topologie associée à la convergence en loi (associée à une distance d_{weak}) sur les lois d'émission le tout à permutation des état cachés près. J'ai obtenu que la consistance de la loi a posteriori par rapport à D_l (avec $l \geq 3$) en θ^* plus les hypothèses d'identifiabilité en le paramètre θ^* implique que la loi a posteriori est consistante par rapport à $\mathcal{T}_{Q,f}$. Voir le Théorème 2.3. Le transfert de la consistance d'une topologie à une autre a été obtenue grâce à des arguments de continuité.

Consistance de la loi a posteriori pour les lois de lissage

Les modèles de Markov cachés à espace d'états finis sont souvent utilisés pour classer les observations suivant les états cachés qui leur correspondent. Dans ce but, on peut chercher à estimer les lois de lissage c'est-à-dire les lois d'un état caché sachant les observations

$$P^{\theta}(X_t = \cdot | Y_1, \dots, Y_n).$$

Mieux, on peut s'intéresser à la loi d'un ensemble fini d'états cachés consécutifs sachant les observations, c'est-à-dire la loi de lissage m-jointe :

$$P^{\theta}((X_1, \dots, X_m) = (\cdot, \dots, \cdot) | Y_1, \dots, Y_n), \quad m \in \mathbb{N} \text{ fixed }, n \geq m.$$

Dans le Chapitre 2, on étudie aussi la consistance a posteriori par rapport à la loi de lissage m-jointe, i.e., on veut savoir si la loi a posteriori se concentre autour des paramètres pour lesquels la loi de lissage m-jointe associée est proche de la vraie (à permutation près). Ce type de consistance mène à des lois a posteriori qui permettent de bien classer les observations suivant l'état caché correspondant.

J'ai montré que, dans le cas particulier où les observations sont discrètes, sous les hypothèses d'identifiabilité en θ^*, si la loi a posteriori est consistante en θ^* par rapport à D_l alors la loi a posteriori concentre sa masse en les paramètres θ pour lesquels la loi de lissage m-jointe associée est proche de la vraie à permutation près.

On notera que l'estimation des lois de lissage a été étudiée ultérieurement dans De Castro *et al.* [DGC15] d'un point de vue fréquentiste. Dans De Castro *et al.* [DGC15], la distance en variation totale entre deux lois de lissage associées à deux paramètres $\tilde{\theta}$ et θ est contrôlée par la norme de Frobenius $\|\tilde{Q} - Q\|_{FB}$ et la norme L^1 : $\|\tilde{f}_j - f_j\|_{L^1}, j \leq k$. Ceci permet de montrer qu'à partir d'un estimateur consistant de la matrice de transition et des estimateurs consistants, par rapport à la norme L^1 des lois densités d'émission, on peut construire un estimateur consistant des lois de lissage. Pour en déduire un résultat Bayésien, on aurait besoin d'un contrôle Bayésien de $\|\tilde{f}_j - f_j\|_{L^1}$. À ma connaissance, un tel contrôle n'existe que dans le cas d'observations discrètes étudié dans le Chapitre 2. En effet dans ce cas, la topologie associée à la convergence en loi est la même que celle associée à la norme L_1 et les Théorèmes 2.1 et 2.3 nous donnent un contrôle Bayésien de $\|\tilde{f}_j - f_j\|_{L^1}$. On peut alors en déduire un résultat de consistance sur les lois de lissage. On obtient alors un résultat dans le même cadre que le Théorème 2.8.

Application à différents modèles et lois a priori

Dans la Partie 2.3, je propose des modèles et lois a priori concrets pour lesquels la loi a posteriori associée est consistante par rapport aux différentes topologies décrites précédemment. Je considère :

- des observations continues, avec des lois d'émission i.i.d. selon un mélange de Gaussiennes sous la loi a priori, voir la Partie 2.3.1,
- des observations continues, avec des lois d'émission translatées $f_j = g(\cdot - m_j)$ et g est distribuée selon un mélange de Gaussiennes sous la loi a priori, voir la Partie 2.3.2,
- des observations discrètes, avec des lois d'émission i.i.d. selon un processus de Dirichlet, voir la Partie 2.3.3.

Limitation des résultats du Chapitre 2

• Une des hypothèses, portant sur le support de Π_Q, utilisée pour obtenir la consistance par rapport à D_l, nécessite de connaître une minoration des éléments de la vraie matrice de transition. Cette hypothèse permet de contrôler les propriétés de mélange des HMMs, pour s'assurer de l'existence de certains tests (ceux construits dans Gassiat and Rousseau [GR14]). Dans le Chapitre 3 (sur les vitesses de concentration), cette hypothèse n'est pas faite, mais des hypothèses plus fortes sur f^* et Π_f sont utilisées pour obtenir une vitesse.

Perspectives du Chapitre 2

• À ma connaissance, on ne connaît pas d'ensemble d'hypothèses qui implique la consistance de la loi a posteriori par rapport à la norme L_1 sur les lois d'émission. Cette perspective est intéressante car elle permettrait d'assurer un bon classement des observations en utilisant De Castro *et al.* [DGC15], en plus d'assurer plus finement une estimation consistance des lois d'émission.

• Un autre projet serait d'étudier la consistance de la loi a posteriori lorsque le nombre d'états k n'est pas connu ni borné et que les lois d'émission vivent dans un espace de dimension infinie. Ce cadre a été étudié par Gassiat and Rousseau [GR14] et van Havre *et al.* [HRWM16] lorsque les lois d'émission sont paramétrées de façon paramétrique. Mélanger les techniques de preuve de Gassiat and Rousseau [GR14], van Havre *et al.* [HRWM16] et le Chapitre 2 devrait mener à des résultats positifs.

1.2.2 Contribution 2 : Vitesse de concentration de la loi a posteriori dans les modèles de Markov cachés non paramétriques à espace d'état fini, Chapitre 3, Vernet [Ver15a]

Ce projet a été mené dans le même cadre que le projet précédent. On a voulu pousser l'étude précédente afin de comprendre à quelle vitesse la loi a posteriori se concentrait. Cette contribution est détaillée dans le Chapitre 3, elle est aussi disponible sur arXiv : Vernet [Ver15a].

Le projet, évoqué dans la Partie 1.2.1 précédente, concernait l'étude de la consistance de la loi a posteriori. On voulait déterminer des hypothèses sous lesquelles la loi a posteriori se concentrait

autour du vrai paramètre lorsque le nombre d'observations tendait vers l'infini. Dans ce projet, on s'intéresse à la vitesse à laquelle la loi a posteriori se concentre. Formellement on dit que la loi a posteriori se concentre avec une vitesse $\epsilon_n \to 0$ en θ^*, par rapport à une pseudo-métrique d sur Θ s'il existe une constante $M > 0$ telle que

$$\Pi\left(\{\theta : \ d(\theta, \theta^*) > M\epsilon_n\} \mid Y_1, \dots Y_n\right) \to 0, \qquad \text{in } P^{\theta^*}\text{-probability.}$$

Les résultats de vitesse permettent de comparer des lois a priori. C'est un critère d'optimalité. On dira que la loi a posteriori se concentre à une vitesse minimax lorsque la loi a posteriori se concentre avec la meilleure vitesse possible. L'étude de la vitesse de concentration permet aussi de mieux comprendre le rôle joué par la loi a priori.

Tout comme l'étude de la consistance, l'analyse de la vitesse de concentration nécessite de choisir une pseudo-métrique. Dans ce projet, j'ai utilisé $D_l(\theta, \tilde{\theta}) = \|p_l^\theta - p_l^{\tilde{\theta}}\|_{L_1}$. Je rappelle ici que la topologie induite par D_l est intéressante dans le but d'estimer la loi des observations et donc aussi dans un but de prédiction. C'est aussi une première étape pour obtenir une vitesse de concentration par rapport à une métrique sur les lois d'émission.

Vitesse de concentration par rapport à D_l

Mon but était d'obtenir des hypothèses explicites et réalisables sur Π_Q, Π_f, Q^* et f^* impliquant l'obtention de vitesse. Dans ce but, j'ai utilisé Ghosal and van der Vaart [GV07a], qui donne un théorème général permettant d'obtenir des vitesses de concentration (voir le Théorème 1.3) et j'ai explicité ses hypothèses dans le cas des HMMs. Les vitesses de concentration sont plus difficiles à obtenir que la consistance de l'a posteriori. En effet, l'obtention de vitesse demande un contrôle plus fin du voisinage de type Kullback autour du vrai paramètre. Cela demande donc une meilleure compréhension de la vraisemblance autour du vrai paramètre. J'ai construit de nouveaux contrôles des ces "voisinages" aidés par des résultats sur les HMMs paramétriques comme Douc and Matias [DM01] et Douc *et al.* [DMR04], voir les Lemmes 3.2 et 3.3. Obtenir des hypothèses satisfaites par des lois a priori usuelles m'a demandé beaucoup de travail.

Pour finir, j'ai obtenu un théorème général (Théorème 3.1) qui associe la vitesse de concentration par rapport à D_l à la loi a priori $(\Pi_Q, \Pi_f^{[k]})$ et au vrai paramètre (Q^*, f^*). La vitesse atteinte a la forme suivante $\epsilon_n / \underline{q}_n$ où ϵ_n dépend du côté "non paramétrique" du modèle, à savoir $\Pi_f^{(k)}$ et f^* quant au taux $\underline{q}_n$ il dépend de Π_Q. Ainsi le taux ϵ_n est détérioré par $\underline{q}_n$, c'est-à-dire par la liberté donnée à Π_Q en ce qui concerne les propriétés de mélange de la chaîne de Markov associée à Q.

Application à différents modèles et lois a priori

J'ai appliqué le théorème général dont je parle dans la partie précédente à différents cadres. Il aboutit à des vitesses minimax à une puissance de $\log(n)$ près, dans différents modèles et pour différentes lois a priori, voir la Partie 3.4.

En particulier, des vitesses minimax (à une puissance de $\log(n)$ près) sont obtenues dans le cas d'observations discrètes avec des lois d'émission (qui sont donc des lois de probabilité sur $\mathbb{N}$) i.i.d. selon un processus de Dirichlet sous la loi a priori. Plus précisément, une vitesse $1/\sqrt{n}$ à une puissance de $\log(n)$ près a été obtenue. Voir la Partie 3.4.1.

De plus des vitesses de concentration adaptatives (c'est-à-dire minimax pour différents sous ensembles de paramètres, la loi a posteriori s'adapte alors à la régularité des données) sont atteintes dans le cas d'observations continues et des densités d'émission i.i.d. selon un mélange de Gaussienne par Processus de Dirichlet sous la loi a priori. Ainsi une vitesse proportionnelle à $n^{-\beta/(2\beta+1)}$, à une puissance de $\log(n)$ près, est obtenue lorsque les densités d'émission appartiennent à une classe de fonctions de type β-Hölder dans la Partie 3.4.2.

Dans les deux cas précédents, on a obtenu des vitesses minimax (à une puissance de $\log(n)$ près) pour peu que Π_Q pénalise suffisamment (i.e. ne mette pas beaucoup de poids dans) le voisinage de la frontière de $\Delta_k^k := \{Q \in [0,1]^{k\times k} : \sum_{j=1}^k Q_{i,j} = 1, \forall 1 \leq i \leq k\}$, l'ensemble des matrices de transition. De manière générale, si $\Pi_f^{(k)} = \Pi_f^{\otimes k}$, avec Π_f qui induit une concentration minimax de la loi a posteriori par rapport à la norme L_1 sur les densités dans le cas de l'estimation de densité avec des observations i.i.d., on s'attend alors à ce que la la loi a posteriori se concentre à une vitesse minimax dans le cadre des HMMs pour peu que Π_Q pénalise suffisamment le voisinage de la frontière de Δ_k^k.

On peut remarquer que les vitesses minimax obtenues pour une classe de densités d'émission et $\Pi_f^{(k)} = \Pi_f^{\otimes k}$ sont les mêmes que dans le cadre de l'estimation de densité avec des observations i.i.d., par rapport à la norme L_1 et la même classe de densités. Ainsi dans nos exemples, la dépendance générée par les HMMs sur les observations ne détériore pas la vitesse minimax, comparé au cadre i.i.d.. La même remarque est faite dans De Castro *et al.* [DGLar] et Bonhomme *et al.* [BJR16a] où des vitesses d'estimateurs fréquentistes sont considérées.

Cette contribution concerne les vitesses de concentration. Or, si la loi a posteriori se concentre à une certaine vitesse tendant vers 0 par rapport à D_l alors la loi a posteriori est aussi consistante. On peut alors utiliser les résultats du Chapitre 2 et montrer que la loi a posteriori est alors aussi consistante pour la topologie $\mathcal{T}_{Q,f}$ (utile dans le cadre de l'estimation de θ) sous condition d'identifiabilité du vrai paramètre.

Perspectives au Chapitre 3

• L'hypothèse faite sur Π_Q, concernant la pénalisation des matrices de transition qui sont trop proches de la frontière de Δ_k^k, est plus faible que celle supposée pour obtenir la consistance de la loi a posteriori. Malgré tout, cette hypothèse est encore forte et n'est pas vérifiée par les lois utilisées en pratique. Il serait intéressant de savoir à quel point cette hypothèse est nécessaire.

• Une perspective à ce travail est l'obtention d'une vitesse de concentration par rapport à la norme L_1 sur les lois d'émission à partir de la vitesse par rapport à D_l. Ce transfert est plus

difficile dans le cas de la vitesse de concentration que dans le cas de la consistance. En effet il demande une compréhension plus fine (que la continuité) de l'inverse (à permutation près) de $\theta \mapsto p_l^\theta$. Un transfert similaire a été fait dans De Castro *et al.* [DGLar] en considérant la norme L_2 et non L_1. Cet article semble être une bonne base de travail pour ce problème.

1.2.3 Contribution 3: Estimation semi-paramétrique efficace et sélection de modèle pour les modèles de mélange multidimensionnels (travail en collaboration avec E. Gassiat et J. Rousseau), Chapitre 4, Gassiat *et al.* [GRV16]

Ma dernière contribution concerne un problème semi-paramétrique. C'est un travail en collaboration avec mes deux directrices de thèse Elisabeth Gassiat (Paris-Sud University) et Judith Rousseau (CEREMADE). Les détails de ce projet sont rédigés dans le Chapitre 4 et sont aussi disponibles sur arXiv : Gassiat *et al.* [GRV16].

On a cherché à étudier l'efficacité asymptotique pour une composante du paramètre, à savoir le paramètre de mélange dans les méthodes de mélange.

On remarquera que ce paramètre remplace la matrice de transition dans le cadre des HMMs.Cette étude est donc aussi une première étape pour comprendre l'estimation semi-paramétrique des matrices de transition dans les HMMs.

On précise ici le cadre des résultats obtenus dans le Chapitre 4. Comme précédemment les états cachés X_t vivent dans un espace d'état fini $\{1, \dots, k\}$ où k est connu. Mais ici ces états sont i.i.d. selon une loi $\sum_{i=1}^k \mu_i \delta_i$. De plus, les observations Y_t, $t \in \mathbb{N}$ vivent dans $[0,1]^3$. Sachant les états cachés $(X_t)_{t\in\mathbb{N}}$, les observations sont toujours indépendantes avec Y_t qui ne dépend que de X_t. Mais de plus sachant X_t, les trois composantes $Y_{t,1}$, $Y_{t,2}$ et $Y_{t,3}$ de l'observation Y_t sont indépendantes avec pour lois respectives $f_{X_t,1}d\lambda$, $f_{X_t,2}d\lambda$ et $f_{X_t,3}d\lambda$. Ce modèle peut être visualisé Figure A.4. On remarquera que ce modèle est identifiable à permutation des états cachés près sous des hypothèses naturelles (voir la Partie 1.3.1 pour plus de précisions).

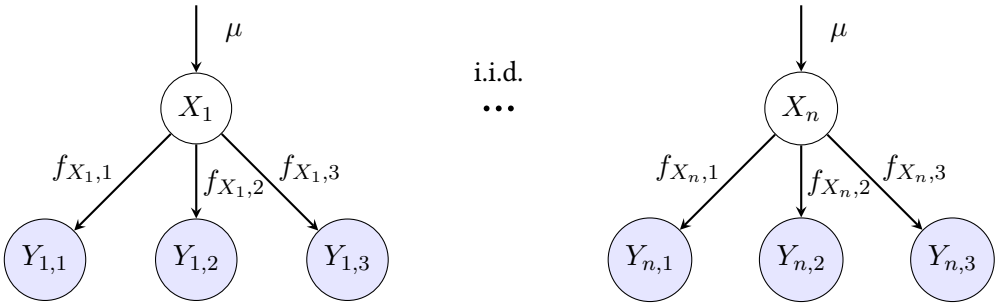


Figure A.4 – Visualisation du modèle de mélange multidimensionnel.

Efficacité asymptotique

Pour obtenir des estimateurs réguliers efficaces, on utilise des modèles d'approximation. À savoir, on projette orthogonalement dans L^2 les densités d'émission sur l'ensemble des histogrammes associés à une partition quelconque $\mathcal{I}_M = \{I_1, \dots, I_M\}$ de $[0,1]$. On obtient alors un modèle paramétrique et les paramètres de ce modèle sont le paramètre $\mu \in \Delta_k$, qui détermine la loi des états cachés, et $\omega_M \in (\Delta_M)^{3k}$ qui paramètre les lois d'émission. La loi d'une observation est alors

$$g_{\mu,\omega;M}(y)\lambda(dy) = \sum_{j=1}^{k} \mu_j \prod_{c=1}^{3} f_{j,c;M}(y_c)\lambda(dy_c),$$

où $\mathbf{f}_M = (f_{j,c;M})_{j\leq k, 1\leq c\leq 3}$, $f_{j,c;M} = \sum_{m=1}^{M} (\omega_{j,c,m;M}/|I_m|)\mathbb{1}_{I_m}$, $j \leq k, 1 \leq c \leq 3$. Enfin, on considère une famille de partitions $\mathcal{I}_M$, $M \in \mathbb{N}$, indexée par le nombre d'éléments dans la partition, associée à une famille de modèles d'approximation.

Dans ces modèles d'approximation, on peut déterminer un maximum de vraisemblance $(\widehat{\theta}_M, \widehat{\omega}_M)$ qui, à permutation près, est asymptotiquement normal (pour le modèle d'approximation) en (θ^*, ω^*), où $\omega^*_{i,j,m} = \int_{I_m} f^*_{j,c} d\lambda$. Ainsi $\widehat{\theta}_M$ est régulier et est asymptotiquement Gaussien autour de θ^*, mais a pour variance asymptotique l'inverse de l'information de Fisher associée au modèle d'approximation, qui peut être différent de l'information de Fisher efficace pour le modèle semi-paramétrique complet. Or, en raffinant la partition suffisamment doucement lorsque le nombre d'observations augmente, on obtient un estimateur $\widehat{\theta}_{M_n}$ régulier efficace (pour le modèle complet) de θ^*, voir le Théorème 4.5.

Plus précisément, on obtient tout d'abord que lorsque la partition est raffinée, l'information de Fisher, associé au modèle d'approximation, augmente ; voir la Proposition 4.2. De plus, lorsque la partition est raffinée telle que le sup de la taille des ensembles des partitions tend vers zéro, alors l'information de Fisher associée aux modèles d'approximation tend vers l'information de Fisher efficace associée au modèle semi-paramétrique complet ; voir le Lemme 4.3. Pour finir, on prouve l'existence d'un raffinement M_n de l'ensemble des partitions tel que la suite associée de maximum de vraisemblance θ_{M_n} est régulière efficace dans le modèle semi-paramétrique complet; voir le Théorème 4.5.

On applique la même méthode dans le cadre Bayésien. C'est-à-dire, si on a une famille de lois a priori $(\Pi_M)_M$, une loi pour chaque modèle associé à une partition $\mathcal{I}_M$, qui sont absolument continues par rapport à la mesure de Lebesgue et qui ont une densité continue et positive sur leur ensemble de définition ; alors en raffinant la partition suffisamment lentement, on peut obtenir un théorème de type Bernstein von Mises. Plus formellement, il existe un raffinement L_n de l'ensemble des partitions tel que la suite de lois a posteriori $\Pi_{L_n}(|Y_1, \dots, Y_n)$ vérifie un théorème de Bernstein von Mises ; voir le Théorème 4.5.

Sélection de modèle

Les deux résultats précédents sont des résultats d'existence mais ne sont pas constructifs. C'est-à-dire qu'ils ne donnent pas d'indice sur à quel point le raffinement M_n doit se faire lentement. Dans la Partie 4.3.1, on obtient que si le raffinement M_n est fait trop vite dans le cas de l'estimateur du maximum de vraisemblance, alors la suite d'estimateurs du maximum de vraisemblance $\widehat{\theta}_{M_n}$ tend presque sûrement vers le poids uniforme et n'est donc même pas consistante.

On propose une procédure pour sélectionner le raffinement d'une famille de partitions s'appuyant sur la validation croisée. Dans le Théorème 4.8, on expose une inégalité oracle pour le risque de l'estimateur sélectionné, associé à a_n ($<< n$) observations alors que l'on utilise n observations pour sélectionner le modèle. Ce choix de raffinement pourrait mener à une sélection trop conservatrice. Nous pensons que ce conservatisme ne devrait pas modifier les qualités asymptotiques de l'estimateur sélectionné.

Enfin, on applique notre critère de sélection à des simulations. Nous avons été surprises par le fait que notre procédure "conservatrice" avait de bonnes propriétés même à horizon fini (i.e. lorsque n est fixé). Voir la Partie 4.4.

Perspectives du Chapitre 4

• Nous aimerions obtenir une vitesse de raffinement explicite (sur M_n) qui assure une efficacité asymptotique.

• Il serait aussi intéressant de généraliser les résultats de cette dernière contribution au cas des HMMs. La généralisation des résultats du Chapitre 4 au cas des HMMs semble loin d'être évidente. En effet, les propriétés du maximum de vraisemblance et la détermination de l'information de Fisher dans les HMMs paramétriques ne sont pas immédiates, voir Douc and Matias [DM01], Douc *et al.* [DMR04], Douc *et al.* [DMOH11] et les références dans Cappé *et al.* [CMR05] par exemple.

• Enfin, dans le cadre Bayésien, on peut se demander s'il existe une loi a priori qui mène à une loi a posteriori ayant des propriétés optimales de concentration à la fois sur le paramètre décrivant le modèle latent (avec un résultat de type BvM) et sur les lois d'émission (en vitesse).

1.3 Résumé de mes contributions

Les résultats que j'ai obtenus durant ma thèse sur les modèles de Markov cachés non paramétriques et les modèles semi-paramétriques de mélange multidimensionnels à espace d'états fini sont résumés dans le Tableau 1.3.

estimation	Consistance de la loi a posteriori, Chapitre 2	vitesse de concentration, Chapitre 3	efficacité asymptotique, Chapitre 4
de la densité p_l^θ	✓	✓	
du paramètre décrivant la loi du modèle latent μ ou Q	✓		✓
des densités d'émission $f_1, \dots, f_k$	✓ (en topologie faible)		
des lois de lissage $P(X_t = \cdot \vert Y_1, \dots, Y_n)$	✓ (quand les observations sont discrètes)		
	dans les HMMs non paramétriques		dans les modèles de mélange semi-paramétriques

Les cases avec le signe ✓ correspondent à des problèmes que j'ai étudiés et pour lesquels j'ai obtenu des résultats positifs. Les parenthèses permettent de préciser une restriction aux résultats. À ma connaissance, les résultats correspondant aux deuxième et troisième colonnes sont les premiers résultats obtenus sur le comportement asymptotique de la loi a posteriori dans le cadre des modèles de Markov caché à espace d'état fini. De même, les résultats correspondant à la quatrième colonne, sont à ma connaissance les premiers résultats sur l'efficacité asymptotique obtenus dans le cadre des modèles de mélange semi-paramétriques multidimensionnels.

BIBLIOGRAPHY

[AHL16] G. Alexandrovich, H. Holzmann, and A. Leister, "Nonparametric identification and maximum likelihood estimation for hidden Markov models", *Biometrika*, vol. 103, no. 2, pp. 423–434, 2016.

[AMR09] E. S. Allman, C. Matias, and J. Rhodes, "Identifiability of parameters in latent structure models with many observed variables", *The Annals of Statistics*, vol. 37, pp. 3099–3132, 2009.

[AGH+14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models", *Journal of Machine Learning Research (JMLR)*, vol. 15, pp. 2773–2832, 2014.

[And10] T. Ando, *Bayesian model selection and statistical modeling*, ser. Statistics: Textbooks and Monographs. CRC Press, Boca Raton, FL, 2010, pp. xiv+286.

[AGR13] J. Arbel, G. Gayraud, and J. Rousseau, "Bayesian optimal adaptive estimation using a sieve prior", *Scandinavian Journal of Statistics*, vol. 40, no. 3, pp. 549–570, 2013.

[Arl14] S. Arlot, "Contributions to statistical learning theory: estimator selection and change-point detection", Habilitation à diriger des recherches, Habilitation à diriger des recherches, University Paris Diderot, Dec. 2014.

[AC10] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection", *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[BJRP13] F. Balabdaoui, H. Jankowski, K. Rufibach, and M. Pavlides, "Asymptotics of the discrete log-concave maximum likelihood estimator and related applications", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 4, pp. 769–790, 2013.

[BB95] P. Barbe and P. Bertail, *The Weighted Bootstrap*, ser. Lecture Notes in Statistics. Springer, 1995, vol. 98.

[BSW99] A. Barron, M. J. Schervish, and L. Wasserman, "The consistency of posterior distributions in nonparametric problems", *The Annals of Statistics*, vol. 27, no. 2, pp. 536–561, 1999.

[Bar88] A. Barron, "The exponential convergence of posterior probabilities with implications for bayes estimators of density functions", Technical report, Apr. 1988.

[BMM12] J.-P. Baudry, C. Maugis, and B. Michel, "Slope heuristics: overview and implementation", *Statistics and Computing*, no. 22, pp. 455–470, 2012.

[BP66] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains", *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[BK12a] M. Beal and P. Krishnamurthy, "Gene expression time course clustering with countably infinite hidden Markov models", *ArXiv preprint arXiv:1206.6824*, 2012.

[BK12b] P. J. Bickel and B. J. K. Kleijn, "The semiparametric Bernstein-von Mises theorem", *The Annals of Statistics*, vol. 40, no. 1, pp. 206–237, 2012.

[BKRW05] P. J. Bickel, C. Klaassen, Y. Ritov, and J. A. Wellner, "Semiparametric inference and models", Mimeo, University of California, Berkeley, Tech. Rep., 2005.

[BKRW98] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York, 1998, pp. xxii+560, Reprint of the 1993 original.

[BJR16a] S. Bonhomme, K. Jochmans, and J.-M. Robin, "Estimating multivariate latent-structure models", *The Annals of Statistics*, vol. 44, no. 2, pp. 540–563, 2016.

[BJR16b] ——, "Non-parametric estimation of finite mixtures from repeated measurements", *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 78, no. 1, pp. 211–229, 2016.

[Bon11] D. Bontemps, "Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors", *The Annals of Statistics*, vol. 39, no. 5, pp. 2557–2584, 2011.

[BZH+13] D. L. Borchers, W. Zucchini, M. P. Heide-Jørgensen, A. Cañadas, and R. Langrock, "Using hidden Markov models to deal with availability bias on line transect surveys", *Biometrics*, vol. 69, no. 3, pp. 703–713, 2013.

[BV10] L. Bordes and P. Vandekerkhove, "Semiparametric two-component mixture model with a known component: an asymptotically normal estimator", *Mathematical Methods of Statistics*, vol. 19, no. 1, pp. 22–41, 2010.

[BDV06] L. Bordes, C. Delmas, and P. Vandekerkhove, "Semiparametric estimation of a two-component mixture model where one component is known", *Scandinavian Journal of Statistics. Theory and Applications*, vol. 33, no. 4, pp. 733–752, 2006.

[BMV06] L. Bordes, S. Mottelet, and P. Vandekerkhove, "Semiparametric estimation of a two-component mixture model", *The Annals of Statistics*, vol. 34, no. 3, pp. 1204–1232, Jun. 2006.

[BG09] S. Boucheron and E. Gassiat, "A Bernstein-von Mises theorem for discrete probability distributions", *Electronic Journal of Statistics*, vol. 3, pp. 114–148, 2009.

[BL06] M. A. Brookhart and M. J. van der Laan, "A semiparametric model selection criterion with applications to the marginal structural model", *Computational Statistics & Data Analysis*, vol. 50, no. 2, pp. 475–498, 2006.

[CD11] A. Canale and D. B. Dunson, "Bayesian kernel mixtures for counts", *Journal of the American Statistical Association*, vol. 106, no. 496, 2011.

[CMR05] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. Springer, 2005.

[Cas12a] I. Castillo, "A semiparametric Bernstein–von Mises theorem for Gaussian process priors", *Probability Theory and Related Fields*, vol. 152, no. 1-2, pp. 53–99, 2012.

[Cas12b] ——, "Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors", *Sankhya A*, vol. 74, no. 2, pp. 194–221, 2012.

[CR15] I. Castillo and J. Rousseau, "A Bernstein–von Mises theorem for smooth functionals in semiparametric models", *The Annals of Statistics*, vol. 43, no. 6, pp. 2353–2383, 2015.

[CR08] T. Choi and R. V. Ramamoorthi, "Remarks on consistency of posterior distributions", in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, ser. Inst. Math. Stat. Collect. Vol. 3, Inst. Math. Statist., Beachwood, OH, 2008, pp. 170–186.

[CH08] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2008, vol. 27, pp. xviii+312.

[CC00] L. Couvreur and C. Couvreur, "Wavelet-based non-parametric hmm's: theory and applications", in *ICASSP'00*, vol. 1, 2000, pp. 604–607.

[DH09] P. De Blasi and N. L. Hjort, "The Bernstein-von Mises theorem in semiparametric competing risks models", *Journal of Statistical Planning and Inference*, vol. 139, no. 7, pp. 2316–2328, 2009.

[DGLar] Y. De Castro, E. Gassiat, and C. Lacour, "Minimax adaptive estimation of nonparametric hidden Markov models", *Journal of Machine Learning Research (JMLR)*, To appear.

[DGC15] Y. De Castro, E. Gassiat, and S. L. Corff, "Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models", *ArXiv preprint arXiv:1507.06510*, 2015.

[DF86] P. Diaconis and D. Freedman, "On the consistency of Bayes estimates", *The Annals of Statistics*, pp. 1–26, 1986.

[DM01] R. Douc and C. Matias, "Asymptotics of the maximum likelihood estimator for general hidden Markov models", *Bernoulli*, vol. 7, pp. 381–420, 2001.

[DMR04] R. Douc, E. Moulines, and T. Rydén, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime", *The Annals of Statistics*, vol. 32, no. 5, pp. 2254–2304, 2004.

[DMOH11] R. Douc, E. Moulines, J. Olsson, and R. van Handel, "Consistency of the maximum likelihood estimator for general hidden Markov models", *The Annals of Statistics*, vol. 39, no. 1, pp. 474–513, 2011.

[Dud02] R. M. Dudley, *Real analysis and probability*. Cambridge University Press, 2002, vol. 74.

[DL14] T. Dumont and S. Le Corff, "Nonparametric regression on hidden phi-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure", *Arxiv preprint arXiv:1209.0633*, 2014.

[DHKR13] C. Durot, S. Huet, F. Koladjo, and S. Robin, "Least-squares estimation of a convex discrete distribution", *Computational Statistics & Data Analysis*, vol. 67, pp. 282–298, 2013.

[Fer74] T. S. Ferguson, "Prior distributions on spaces of probability measures", *The Annals of Statistics*, vol. 2, pp. 615–629, 1974.

[FJSW09] E. B. Fox, M. I. Jordan, E. B. Sudderth, and A. S. Willsky, "Sharing features among dynamical systems with beta processes", in *Advances in Neural Information Processing Systems*, 2009, pp. 549–557.

[FSJW11] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization", *The Annals of Applied Statistics*, pp. 1020–1056, 2011.

[Fre65] D. A. Freedman, "On the asymptotic behavior of Bayes estimates in the discrete case. II", *Annals of Mathematical Statistics*, vol. 36, pp. 454–456, 1965.

[GCR15] E. Gassiat, A. Cleynen, and S. Robin, "Inference in finite state space non parametric hidden Markov models and applications", English, *Statistics and Computing*, pp. 1–11, 2015.

[GPS13] E. Gassiat, D. Pollard, and G. Stoltz, "Revisiting the van Trees inequality in the spirit of Hajek and Le Cam", *Unpublished manuscript*, 2013.

[GRV16] E. Gassiat, J. Rousseau, and E. Vernet, "Efficient semiparametric estimation and model selection for multidimensional mixtures", *ArXiv preprint arXiv:1607.05430*, 2016.

[GR14] E. Gassiat and J. Rousseau, "About the posterior distribution in hidden Markov models with unknown number of states", *Bernoulli*, vol. 20, no. 4, pp. 2039–2075, 2014.

[GR16] ——, "Nonparametric finite translation hidden Markov models and extensions", *Bernoulli*, vol. 22, no. 1, pp. 193–212, 2016.

[GCSR14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 2014, vol. 2.

[GGV00] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart, "Convergence rates of posterior distributions", *The Annals of Statistics*, vol. 28, no. 2, pp. 500–531, 2000.

[GV07a] S. Ghosal and A. W. van der Vaart, "Convergence rates of posterior distributions for non-i.i.d. observations", *The Annals of Statistics*, vol. 35, no. 1, pp. 192–223, 2007.

[GV07b] ——, "Posterior convergence rates of Dirichlet mixtures at smooth densities", *The Annals of Statistics*, vol. 35, no. 2, pp. 697–723, 2007.

[GR03] J. Ghosh and R.V. Ramamoorthi, *Bayesian Nonparametrics*. Springer, 2003.

[GS08] M. C. de Gunst and O. Shcherbakova, "Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels", *Mathematical Methods of Statistics*, vol. 17, no. 4, pp. 342–356, 2008.

[HJW14] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss", *ArXiv preprint arXiv:1411.1467*, 2014.

[HY01] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length", *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[HRWM16] Z. van Havre, J. Rousseau, N. White, and K. Mengersen, "Overfitting hidden markov models with an unknown number of states", *ArXiv preprint arXiv:1602.02466*, 2016.

[HRS+15] M. Hoffmann, J. Rousseau, J. Schmidt-Hieber, *et al.*, "On adaptive posterior concentration rates", *The Annals of Statistics*, vol. 43, no. 5, pp. 2259–2295, 2015.

[HH13] D. Hohmann and H. Holzmann, "Semiparametric location mixtures with distinct components", *Statistics*, vol. 47, no. 2, pp. 348–362, 2013.

[HWY16] H. Hu, Y. Wu, and W. Yao, "Maximum likelihood estimation of the mixture of log-concave densities", *Computational Statistics & Data Analysis*, vol. 101, pp. 137–147, 2016.

[HWH07] D. R. Hunter, S. Wang, and T. P. Hettmansperger, "Inference for mixtures of symmetric distributions", *The Annals of Statistics*, vol. 35, no. 1, pp. 224–251, Feb. 2007.

[JW09] H. K. Jankowski and J. A. Wellner, "Estimation of a discrete monotone distribution", *Electronic Journal of Statistics*, vol. 3, p. 1567, 2009.

[Joc15] M. Jochmann, "Modeling US inflation dynamics: a Bayesian nonparametric approach", *Econometric Reviews*, vol. 34, no. 5, pp. 537–558, 2015.

[Kim06] Y. Kim, "The Bernstein-von Mises theorem for the proportional hazard model", *The Annals of Statistics*, vol. 34, no. 4, pp. 1678–1700, 2006.

[KL04] Y. Kim and J. Lee, "A Bernstein-von Mises theorem in the nonparametric right-censoring model", *The Annals of Statistics*, vol. 32, no. 4, pp. 1492–1512, 2004.

[KRV10] W. Kruijer, J. Rousseau, and A. W. van der Vaart, "Adaptive Bayesian density estimation with location-scale mixtures", *Electronic Journal of Statistics*, vol. 4, pp. 1225–1257, 2010.

[Kru77] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics", *Linear Algebra and Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[LKSD15] R. Langrock, T. Kneib, A. Sohn, and S. L. DeRuiter, "Nonparametric inference in hidden Markov models using p-splines", *Biometrics*, 2015.

[LY00] L. Le Cam and G. Yang, *Asymptotics in Statistics. Some Basic Concepts, Second Edition*. Springer-Verlag, New-York, 2000.

[LMMR09] K. Lee, J.-M. Marin, K. Mengersen, and C. Robert, "Bayesian inference on finite mixtures of distributions", in *Perspectives in mathematical sciences. I*, ser. Stat. Sci. Interdiscip. Res. Vol. 7, World Sci. Publ., Hackensack, NJ, 2009, pp. 165–202.

[Lef03] F. Lefèvre, "Non-parametric probability estimation for HMM-based automatic speech recognition", *Computer Speech & Language*, vol. 17, no. 2, pp. 113–136, 2003.

[LJWC14] S. W. Linderman, M. J. Johnson, M. A. Wilson, and Z. Chen, "A nonparametric Bayesian approach to uncovering rat hippocampal population codes during spatial navigation", *ArXiv preprint arXiv:1411.7706*, 2014.

[MY15] Y. Ma and W. Yao, "Flexible estimation of a semiparametric two-component mixture model with one parametric component", *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 444–474, 2015.

[MZ97] I. L. MacDonald and W. Zucchini, *Hidden Markov and other models for discrete-valued time series*. London, UK: Chapman and Hall/CRC, 1997.

[MZ09] ——, *Hidden Markov models for time series: an introduction using R*. London, UK: Chapman and Hall/CRC, 2009.

[Mas07] P. Massart, *Concentration inequalities and model selection*, ser. Lecture Notes in Mathematics. Berlin: Springer, 2007, vol. 1896, pp. xiv+337, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[MW00] B. McNeney and J. A. Wellner, "Application of convolution theorems in semiparametric models with non-i.i.d. data", *Journal of Statistical Planning and Inference*, vol. 91, no. 2, pp. 441–480, 2000, Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

[Rho10] J. A. Rhodes, "A concise proof of kruskal's theorem on tensor decomposition", *Linear Algebra and Appl.*, vol. 432, no. 7, pp. 1818–1824, 2010.

[Rio00] E. Rio, "Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes", *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, vol. 330, no. 10, pp. 905–908, 2000.

[RB90] Y. Ritov and P. J. Bickel, "Achieving information bounds in non and semiparametric models", *The Annals of Statistics*, vol. 18, no. 2, pp. 925–938, 1990.

[RR12a] V. Rivoirard and J. Rousseau, "Bernstein-von Mises theorem for linear functionals of the density", *The Annals of Statistics*, vol. 40, no. 3, pp. 1489–1523, 2012.

[RR12b] ——, "Posterior concentration rates for infinite dimensional exponential families", *Bayesian Analysis*, vol. 7, no. 2, pp. 311–333, 2012.

[Rob01] C. Robert, *The Bayesian Choice*, second. New York: Springer-Verlag, 2001.

[Rou15] J. Rousseau, "On the frequentist properties of Bayesian nonparametric methods", *Annual statistical reviews*, 2015, to appear.

[RM11] J. Rousseau and K. Mengersen, "Asymptotic behaviour of the posterior distribution in overfitted mixture models", *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 73, no. 5, pp. 689–710, 2011.

[Sch65] L. Schwartz, "On Bayes procedures", *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 4, pp. 10–26, 1965.

[Scr14] C. Scricciolo, "Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures", *Bayesian Analysis*, vol. 9, no. 2, pp. 475–520, 2014.

[STG13] W. Shen, S. T. Tokdar, and S. Ghosal, "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures", *Biometrika*, vol. 100, no. 3, pp. 623–640, 2013.

[She02] X. Shen, "Asymptotic normality of semiparametric and nonparametric posterior distributions", *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 222–235, 2002.

[SW01] X. Shen and L. Wasserman, "Rates of convergence of posterior distributions", *The Annals of Statistics*, vol. 29, no. 3, pp. 687–714, 2001.

[SS05] E. M. Stein and R. Shakarchi, *Real analysis*, ser. Princeton Lectures in Analysis, III. Princeton University Press, Princeton, NJ, 2005, pp. xx+402, Measure theory, integration, and Hilbert spaces.

[Tok06] S. T. Tokdar, "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression", *Sankhya. The Indian Journal of Statistics*, vol. 68, no. 1, pp. 90–110, 2006.

[Vaa98] A. W. van der Vaart, *Asymptotic statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998, vol. 3, pp. xvi+443.

[VZ08] A. W. van der Vaart and J. H. van Zanten, "Rates of contraction of posterior distributions based on gaussian process priors", *The Annals of Statistics*, pp. 1435–1463, 2008.

[VZ09] A. W. van der Vaart and J. H. van Zanten, "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth", *The Annals of Statistics*, vol. 37, no. 5B, pp. 2655–2675, 2009.

[Vaa02] A. W. van der Vaart, "Semiparametric statistics", in *Lectures on probability theory and statistics (Saint-Flour, 1999)*, ser. Lecture Notes in Math. Vol. 1781, Springer, Berlin, 2002, pp. 331–457.

[Ver15a] E. Vernet, "Non parametric hidden markov models with finite state space: posterior concentration rates", *ArXiv preprint arXiv:1511.08624*, 2015.

[Ver15b] ——, "Posterior consistency for nonparametric hidden Markov models with finite state space", *Electronic Journal of Statistics*, vol. 9, pp. 717–752, 2015.

[WLM03] J. P. Whiting, M. F. Lambert, and A. V. Metcalfe, "Modelling persistence in annual australian point rainfall", *Hydrology and Earth System Sciences*, vol. 7, no. 2, pp. 197–211, 2003.

[XYW14] S. Xiang, W. Yao, and J. Wu, "Minimum profile Hellinger distance estimation for a semiparametric mixture model", *Canadian Journal of Statistics*, vol. 42, no. 2, pp. 246–267, 2014.

[YPRH11] C. Yau, O. Papaspiliopoulos, G. Roberts, and C. Holmes, "Bayesian non-parametric hidden Markov models with applications in genomics", *Journal of the Royal Statistical Society*, vol. 73, pp. 37–57, 2011.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modèles de mélange et de Markov caché non paramétriques : propriétés asymptotiques de la loi a posteriori et efficacité

Mots-clés : Statistiques non et semi-paramétriques, statistiques Bayésiennes, statistiques asymptotiques, modèle de Markov caché, modèle de mélange

Résumé : Les modèles latents sont très utilisés en pratique, comme en génomique, économétrie, reconnaissance de parole, étude de population... Comme la modélisation paramétrique des lois d'émission, c'est-à-dire les lois d'une observation sachant l'état latent, peut conduire à de mauvais résultats en pratique, un récent intérêt pour les modèles latents non paramétriques est apparu dans les applications. Or ces modèles ont peu été étudiés en théorie. Dans cette thèse je me suis intéressée aux propriétés asymptotiques des estimateurs (dans le cas fréquentiste) et de la loi a posteriori (dans le cadre Bayésien) dans deux modèles latents particuliers : les modèles de Markov cachés et les modèles de mélange. J'ai tout d'abord étudié la concentration de la loi a posteriori dans les modèles non paramétriques de Markov cachés. Plus précisément, j'ai étudié la consistance puis la vitesse de concentration de la loi a posteriori. Enfin je me suis intéressée à l'estimation efficace du paramètre de mélange dans les modèles semi-paramétriques de mélange.

Title : Nonparametric Mixture Models and Hidden Markov Models: Asymptotic Behaviour of the Posterior Distribution and Efficiency

Keywords : Non and semiparametric statistics, Bayesian statistics, asymptotic statistics, hidden Markov model, mixture model

Abstract : Latent models have been widely used in diverse fields such as speech recognition, genomics, econometrics. Because parametric modeling of emission distributions, that is the distributions of an observation given the latent state, may lead to poor results in practice, in particular for clustering purposes, recent interest in using nonparametric latent models appeared in applications. Yet little thoughts have been given to theory in this framework. During my PhD I have been interested in the asymptotic behaviour of estimators (in the frequentist case) and the posterior distribution (in the Bayesian case) in two particuliar nonparametric latent models: hidden Markov models and mixture models. I have first studied the concentration of the posterior distribution in nonparametric hidden Markov models. More precisely, I have considered posterior consistency and posterior concentration rates. Finally, I have been interested in efficient estimation of the mixture parameter in semiparametric mixture models.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France