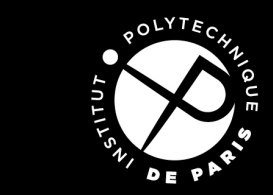
Mixtures of Gaussian Graphical Models with Constraints

Thomas Lartigue



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX034

Thèse de doctorat

Mixtures of Gaussian Graphical Models with constraints

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques Appliquées

Thèse présentée et soutenue à Biarritz (visioconférence), le 22 septembre 2020, par

THOMAS LARTIGUE

Composition du Jury :

Erwan Le Pennec Professeur, École polytechnique	Président
Sach Mukherjee Directeur de Recherche, DZNE	Rapporteur
Sophie Achard Directeur de Recherche, Laboratoire Jean Kuntzmann	Rapporteur
Etienne Birmelé Professeur, Université Paris-Descartes	Examinateur
Francis Bach Directeur de recherche, Inria	Examinateur
Stéphanie Allassonnière Professeur, Université Paris-Descartes	Directeur de thèse
Stanley Durrleman Directeur de Recherche, Inria	Co-directeur de thèse

À Pascal, Marine et Néis

Acknowledgements

I would first like to thank my thesis jury for participating in my defence. Their insightful questions, as well as their honest and helpful feedback, elevated the quality of this thesis. My special thoughts go to Sophie Achard and Sach Mukherjee, who dedicated much of their time to review the full manuscript. I also wanted to thank Sach in particular for inviting me to continue my work within his research group.

Thank you to my supervisors, Stéphanie Allassonnière and Stanley Durrleman, who gave me the chance to spend three amazing years doing exiting research. Their advice and their guidance allowed me to grow in more ways than one. I had a fun and unforgettable time.

Thank you to the laboratories that welcomed me: Aramis-lab at the Paris Brain Institute and CMAP at École polytechnique. Thank you in particular to the members of office 20.03 in CMAP for making my time there so much more enjoyable.

Finally, thank you to my family, who has always supported me unconditionally in all my undertakings.

Résumé en français

La description des co-variations entre plusieurs variables aléatoires observées est un problème délicat. Les réseaux de dépendance sont des outils populaires qui décrivent les relations entre les variables par la présence ou l'absence d'arêtes entre les nœuds d'un graphe. Bien entendu, il ne s'agit là que d'une vue simplifiée de la dynamique multidimensionnelle. Aucun graphe ne peut espérer décrire pleinement toutes les subtilités d'une distribution aléatoire multivariée. Il existe donc de nombreux types de réseaux différents qui codent chacun un type différent d'informations sur les variables. La structure de corrélation d'un vecteur aléatoire est une bonne candidate pour être représentée sous forme de graphe. Toutefois, dans les applications réelles, la matrice de corrélation du vecteur aléatoire a toutes les chances d'être entièrement connectée. En effet, deux événements ou mesures séparés sont susceptibles d'être liés l'un à l'autre, au moins de manière très diffuse, par une chaîne de plusieurs variables corrélées intermédiaires, dans une sorte d'"effet domino".
Dans un tel scénario, on peut préférer une description plus épurée de la dynamique, où seuls les "dominos" consécutifs sont marqués comme étant liés. Représenter uniquement les corrélations "directes" ou "explicites" entre les variables est précisément l'ambition des réseaux de corrélation conditionnelle. Dans ces réseaux, il existe un écart entre deux composantes Y_i et Y_j du vecteur aléatoire $Y \in \mathbb{R}^p$ si et seulement si $corr(Y_i, Y_j|(Y_k)_{k\neq i,j}) \neq 0$. L'idée est que s'il reste une corrélation entre deux caractéristiques après conditionnement par toutes les autres, alors les deux partagent un lien intime qui ne peut pas s'exprimer comme un simple transfert d'information de l'une à l'autre par des variables intermédiaires. Ces réseaux sont souvent étudiés sous l'hypothèse gaussienne et sont donc appelés "modèles graphiques gaussiens" (GGM).
Un seul réseau peut être utilisé pour représenter les tendances globales identifiées au sein d'un échantillon de données. Cependant, lorsque les données observées sont échantillonnées à partir d'une population hétérogène, il existe alors différentes sous-populations qui doivent toutes être décrites par leurs propres graphes. De plus, si les étiquettes des sous-populations (ou "classes") ne sont pas disponibles, des approches non supervisées doivent être mises en œuvre afin d'identifier correctement les classes et de décrire chacune d'entre elles avec son propre graphe.
Dans ce travail, nous abordons le problème relativement nouveau de l'estimation hiérarchique des GGM pour des populations hétérogènes non labellisées. Nous explorons plusieurs axes clés pour améliorer l'estimation des paramètres du modèle ainsi que l'identification non supervisée des sous-populations. Notre objectif est de s'assurer que les graphes de corrélations conditionnelles inférés sont aussi pertinents et interprétables que possible.

Premièrement - dans le cas d'une population simple et homogène - nous étudions les deux principaux paradigmes de l'état de l'art. Le premier de ces estimateurs parcimonieux est le maximum de vraisemblance pénalisé "graphical lasso" (GLASSO) [13,179]. Le second est l'agrégation de multiples régressions linéaires par nœud appelé "GGMselect" [56]. Dans ce chapitre, nous soutenons que le GLASSO appartient à une famille de "méthodes globales" qui considèrent le graphe entier en une seule fois, alors que GGMselect appartient aux "méthodes locales", qui construisent des graphes bord à bord autour de chaque nœud. Nous démontrons que les deux approches ont des défauts dans les paramètres de taille d'échantillon de haute dimension et de faible taille et proposons un algorithme composite pour canaliser leurs forces respectives dans une seule méthode, afin d'en corriger les faiblesses. Dans un second temps, nous appliquons les concepts de sélection de modèle développés pour notre algorithme composite afin de mettre au point une nouvelle méthode avec un critère de sélection explicite dans le cas particulier des graphes cordaux.

Pour les populations hétérogènes non labellisées, nous proposons d'estimer un mélange de GGM avec un algorithme espérance-maximisation (EM). Cependant, l'algorithme EM optimise une fonction non-convexe. Qui plus est, les réseaux de corrélations conditionnelles sont généralement étudiés pour décrire un assez grand nombre de variables. Par conséquent, plus la dimension est élevée, plus il est facile de tomber au cours de l'optimisation de la fonction objectif dans des extrema locaux sous-optimaux. L'initialisation devient alors extrêmement influente. Ainsi, afin d'améliorer les solutions de cet algorithme EM, et d'éviter de tomber dans ces extrema locaux, nous introduisons une version tempérée de cet algorithme EM, dont nous étudions aussi bien les garanties théoriques de convergence que le comportement et les performances empiriques.
Enfin, nous améliorons l'identification des variables cachées par l'algorithme EM en tenant compte des potentiels effets de co-facteurs externes sur les variables mesurées. En effet, une hypothèse cruciale derrière tout algorithme EM est que, dans l'espace des variables observées, les données sont organisées géométriquement en amas qui correspondent à des sous-populations intéressantes ou à des variables cachées. Cependant, dans certains cas, les amas de données sont plus corrélées avec des co-facteurs externes qui sont triviaux ou facilement observables, mais qui ont un impact fort sur les valeurs des variables observées. Cette situation est rendue encore plus complexe lorsque les effets des co-facteurs sur les variables sont hétérogènes au sein de la population. Pour résoudre cela, nous développons un mélange de modèles graphiques gaussiens conditionnels (CGGM). Les CGGM, introduits par [173] et [171], prennent en compte les effets des co-facteurs sur les variables et les suppriment afin de réajuster la position de chaque point de données. Cette correction regroupe les points de données appartenant à chaque sous-population et permet de rendre plus pertinente l'estimation ultérieure des paramètres.

Contents

Chapter 1

Introduction

Describing the co-variations between several observed random variables is a delicate problem. Dependency networks are popular tools that depict the relations between variables through the presence or absence of edges between the nodes of a graph. Of course, this is only a simplified view of the multidimensional dynamic. No single graph can hope to describe fully all the intricacies of a multivariate random distribution. Hence, many different kinds of networks exist that each encode a different type of information about the variables. The correlation structure of a random vector is a good candidate to be represented as a graph. However, in real applications, the correlation matrix of the random vector has every chance to be fully connected. Indeed, any two separate events or measures are likely to be linked to one another, at least in a very diffuse way, through a chain of several intermediary correlated variables, in some sort of "domino effect".
In such a scenario, one may prefer a sparser description of the dynamic, where only the consecutive "dominoes" are marked as connected. Representing only the "direct" or "explicit" correlations between the variables is precisely the ambition of the conditional correlation networks. In such networks, there is an edge between two components Y_i and Y_j of the random vector $Y \in \mathbb{R}^p$ if and only if $corr(Y_i, Y_j|(Y_k)_{k\neq i,j}) \neq 0$. The idea is that if there remains a correlation between two features after conditioning by all the others, then the two share an intimate connection that cannot be expressed as a simple transfer of information from one to the other through intermediary variables. A single network can be used to represent the overall tendencies identified within a data sample.
However, when the observed data is sampled from a heterogeneous population, then there exist different sub-populations that all need to be described through their own graphs. What is more, if the sub-population (or "class") labels are not available, unsupervised approaches must be implemented in order to correctly identify the classes and describe each of them with its own graph.
To estimate conditional correlation graphs from observed data, [35] introduce the "Covariance Selection" procedure. They propose to model the variables as a Gaussian vector $Y \sim \mathcal{N}(\mu, \Sigma)$ and make a sparse estimation of the inverse-covariance (or "precision") matrix $\Lambda := \Sigma^{-1}$. Indeed, within a Gaussian model, we have $corr(Y_i, Y_j|(Y_k)_{k\neq i,j}) = -\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}}$. Hence the sparsity of the precision matrix is the same as the sparsity of the conditional correlation network. Gaussian modelling has other advantages. First, for a general multivariate distribution, $corr(Y_i, Y_j|(Y_k)_{k\neq i,j})$ is a function of $(Y_k)_{k\neq i,j}$, not a constant. As a result, each edge weight of the conditional correlation graph is actually a function of the values in the vector. Having such a complex and unstable description would be a significant hindrance. Within the Gaussian model however, this function is a constant, as a consequence, conditional correlation graphs are only ever considered under the Gaussian assumption. A another advantage of Gaussian modelling is that un-correlation and independence are equivalent $corr(Y_i, Y_j|(Y_k)_{k\neq i,j}) = 0 \iff Y_i \perp\!\!\!\perp Y_j|(Y_k)_{k\neq i,j}$, which facilitate interpretation. This modelling of a random vector as following a multivariate normal distribution with sparse inverse-covariance, hence sparse conditional correlation graph, has been called "Gaussian Graphical Modelling".
Gaussian Graphical Models (GGM) have generated an extensive interest following [35]. For homogeneous populations, several different sparse estimation of the conditional correlation graph have been proposed. Some well known approaches include statistical tests [41], node-wise linear regres-

sions [115], penalised Maximum Likelihood Estimation (MLE) [13, 179], penalised reconstruction problem [18], and Bayesian techniques [94, 162].
Several authors considered heterogeneous labelled population, and designed methods to estimate jointly all the different sub-population networks, with elements of common structure. Many of them adapt the penalised MLE approach to the hierarchical case, as for instance [119] [30] and [172]. They each propose different regularisation that enforce different forms of common structure between the graphs.
Unlabelled heterogeneous populations have been less explored. Recent works have adapted the penalised MLE approach to the unsupervised case. They propose to estimate Mixtures of GGM with regularised Expectation-Maximisation (EM) algorithms. Recent examples that make use of penalties that encourage common structure include [52] and [61].

In this work, we tackle the fairly new problem of Hierarchical GGM estimation for unlabelled heterogeneous populations. We explore several key axes to improve the estimation of the model parameters as well as the unsupervised identification of the sub-populations. Our goal is to ensure that the inferred conditional correlation graphs are as relevant and interpretable as possible.
In chapter 2, we provide an extensive literature review that delves into the history of the many GGM techniques developed to describe homogeneous, heterogeneous and unlabelled populations. This notably includes the literature surrounding alternative models such as the Matrix Normal Graphical Models (MNGM), Conditional Gaussian Graphical Models (CGGM) and Dynamic Gaussian Graphical Models (DGGM). Moreover, we discuss key elements of the EM literature, since this optimisation method plays such an important role in the unsupervised approach to GGM.
In chapter 3, we explore the problem of model selection and validation criteria, particularly for the high dimension low sample size setting. In the simple case with homogeneous population, the most popular sparse GGM estimator is the penalised MLE of [179] and [13], which is colloquially referred to as the "Graphical Lasso" (GLASSO) estimator in reference to the work of [50]. The GLASSO estimator is not unique as it depends on an hyper-parameter: the penalty intensity. The necessity of choosing a value for this hyper-parameter naturally raises the question of model selection. Other works such as [56] have addressed with their GGMselect algorithm the problem of model selection in the case of the node-wise estimator of [115]. In this chapter, we argue that the GLASSO belongs to a family of "global methods" that considers the whole graph all at once, whereas GGMselect belongs to the "local methods", which build graphs edge by edge around each node. We demonstrate that the two approaches have flaws in the high dimension low sample size settings and propose a composite algorithm to channel their respective strengths into a single method. In particular, a key element of our method is the model selection procedure through the Out of Sample Cross Entropy or Kullback–Leibler (KL) divergence of the estimated distribution with regards to the empirical observed one. In addition to providing theoretical guarantees about our method, we show through many experiments that it outperforms GLASSO and GGMselect in the high dimension low sample size setting. In particular, we demonstrate that our selection criterion is more farsighted than the local proposed one in [56]. Even though this is done in the simple case for homogeneous populations, we acquire through this study valuable knowledge that we apply to the hierarchical case. In particular, the results of this chapter encourage us to confidently use the Out of Sample (OoS) KL divergence as a selection criteria or success metrics even when working with Hierarchical Models.
In chapter 4, we delve deeper into model selection within a more specific family of graphs: the chordal graphs. Chordal graphs, also called "decomposable graphs", are graphs where there is no cycle of more than three edges. They enjoy a certain popularity in graph theory since for any chordal graph, there exists a maximal prime sub-graphs decomposition into the set of maximal cliques $\mathcal{C}$ and the set of their intersections, the separator cliques $\mathcal{P}$. This is further developed in [91]. Interestingly, some authors, such as [57] and [14], have developed in interest for the estimation of conditional correlation graphs that remain within the chordal family. In this chapter, we apply the conclusions of chapter 3 and propose a model selection procedure that follows the KL divergence. Thanks to the properties of the chordal graphs, we are able to find, with an explicit formula, an unbiased estimator of the KL divergence of the proposed distribution with the real divergence. This exempts us from having to split the data into a training set and a validation set in order to compute the OoS KL divergence.

We prove a theoretical result on the selection with our new selection criterion and demonstrate with experiments it performs even better than the OoS KL divergence.
In chapter 5, we turn our gaze towards the EM algorithm, a crucial component of any unsupervised MLE approach to Mixtures of GGM. Networks are generally used to describe a somewhat large number of variables, and GGM are no exception. However, with mixtures, the EM algorithm optimises a non convex function. Hence, the higher the dimension, the easier it is to fall for sub-optimal local maxima and the more important the initialisation becomes. Escaping the initialisation in the context of non-convex optimisation is a very well known problem. The simulated annealing of [82] and later the parallel tempering (annealing MCMC) of [53, 149] were developed to address this issue. Adding a temperature that weakens at first the attractive power of the potential wells and overall makes the likelihood profile less radical allows the optimisation procedure to explore more before settling for the best local maximum encountered. Such a procedure can be used to improve the performances of the EM algorithm. Some works have introduced tempered EMs, such as [156] with their *deterministic annealing EM* and [120] who propose an alternative temperature profile for this algorithm. Both these works are very empirical and propose different temperature profile on the basis of their experimental successes. In this chapter, we prove a convergence theorem that provides guarantees for the tempered-EM (tmp-EM) under certain conditions on the temperature profile. We see that these conditions are very mild, which justifies the use of a much wider category of temperature profile than the previously proposed ones. Including for instance non-monotonous or oscillating profiles. Additionally, we provide a very extensive experimental study of the tmp-EM algorithm, confronting it to many adversarial scenarios and following many metrics. All of these demonstrate the amazing efficiency of the tempering when it comes to escaping the initialisation. In addition to all that, we introduce a new, more general framework of deterministic approximated EM algorithms that all benefit from the same convergence guarantees. The tmp-EM is but one of the methods that belong to this framework. In particular, we propose a new *Riemann approximation EM*, a deterministic alternative to the Monte Carlo EM [166] to deal with intractable E steps. Of course, the tmp-EM and the Riemann approximation EM can be combined into one method that fulfils both their objective and still benefits from the convergence guarantees.
In chapter 6, we examine how the cluster identification of the EM for Mixtures of GGM can be greatly improved with proper consideration of the potential effects of co-factors. Indeed, a crucial assumption behind any EM algorithm is that, in the feature space, the data is geometrically organised in clusters that correspond to interesting sub-populations or hidden variables. However, we argue that, with real data, the geometrical clusters have every chance of being correlated with external co-factors that are trivial or easily observable, but have a potent impact on the features' values. For instance, certain medical or biological measures are likely to be more correlated with gender or age group than with diagnosis or disease type. This situation is rendered more complex when the co-factors effects on the feature are heterogeneous over the population. As a consequence, we make use of the Conditional Gaussian Graphical Models (CGGM) introduced in [173] and [171], and notably used by [69] in the supervised hierarchical case. For an unlabelled population, we propose the Mixture of CCGM. This model takes into account the potentially heterogeneous effect of co-factors and remove them in order to readjust the position of each data point. This correction groups together the data points belonging to each sub-population and allows the subsequent parameter estimation to be more meaningful. We develop a regularised EM algorithm to compute a penalised MLE from this model. We show that this EM can be used with any of the state of the art structure-inducing penalties developed for the supervised case. We demonstrate with experiments that our EM with mixture of CGGM performs much better than the previously introduced EM with mixtures of GGM of [52] and [61]. Additionally, we prove a theorem that provides conditions on the regularisation in order for our EM to benefit from convergence guarantees. Moreover, we propose a tempered version of our EM and proves that it benefits from the guarantees of chapter 5. Additional experiments demonstrate how the tempering can improve even further the cluster recovery of the Mixture of CGGM.
Finally, in chapter 7, we synthetize our contributions. We widen the discussion by showcasing more grounded, pragmatic, applications of Hierarchical GGM and Mixtures of GGM to real problematics. We present in particular some of our recent clinical collaborations for ciliopathy and Cushing's

syndrome patients. In the end, we lay the groundwork for future works.

Chapter 2

Literature Review

In this chapter, we provide an extensive review of the literature surrounding the Gaussian Graphical Models (GGM) and conditional correlation networks. From the simple GGM that describe homogeneous populations to the unsupervised Mixtures of GGM for unlabelled heterogeneous populations. Given its omnipresence in the field of unsupervised Mixture of Gaussians inference, we also provide a rapid overview of the work done on the Expectation-Maximisation (EM) algorithm. Moreover, we discuss several alternative models such as the Matrix Normal Gaussian Graphical Models (MNGM) or Conditional Gaussian Graphical Models (CGGM) that each adapt the GGM concept for their own specific situation.

2.1 Simple models

The genesis of GGM can be found in [35]. In this work, Dempster seeks to describe a vector of random variables with their conditional correlation network. For that purpose, he proposes to model the variables as a Gaussian vector, Indeed, within the Gaussian model, both the sparsity and edge weights of the conditional correlation graph can be directly recovered from the inverse-covariance (precision) matrix. Moreover, as previously discussed, the Gaussian model makes the graph a constant instead of a function of the variables' values, a property so crucial that it makes Gaussian modelling inevitable for conditional correlation network analysis. To infer a conditional graph from data samples, Dempster proposes to make a sparse approximation of the variable's precision matrix. Indeed, un-constrained empirical estimators are rarely sparse, and fully connected graphs not very interesting. He calls "Covariance Selection" this procedure. Although the term "Gaussian Graphical Models" (GGM), popularised by later works, is the one that over the years came to be the default designation of this field of the study. These GGM have found application in many areas, such as biology and medicine, in particular genetic, see [167] for an early, very applied, example.
As an aside, note that although the GGM approach naturally links inverse-covariance matrix and conditional correlation graph, the transition from one to the other is not symmetrical. If the matrix is known, getting the graph is trivial, since they have the exact same sparsity. And even if one wishes to recover a weighted graph, the edge weights are related to the matrix coefficient with an explicit formula. However, estimating a sparse precision matrix from an unweighted conditional correlation graph is not so easy. When the edge weight information is not available, the sparse matrix has to be estimated from data. Works with a formal approach to graph theory such as [146] and [91] have introduced technical algorithms to compute, from an independent identically distributed (iid) data sample, the graph-constrained Maximum Likelihood Estimator of the precision matrix. Works such as [158] and then [157] provide extensive analyses of the properties of this constrained MLE.
The bulk of the literature however, focuses on designing methods to find the conditional correlation graph from data. Which they do either by working on the graph directly or by first estimating a sparse approximation of the inverse-covariance matrix. We present here some of the most notable, most influential works in that regard. The authors [41] use statistical tests to construct a graph edge by edge. With a penalised likelihood approach, the authors of [70] induce sparsity in

the Cholesky decomposition coefficient. Around the same time, [115] propose to solve sparse linear regression problems in parallel in which the value on each node is predicted by its neighbours. Shortly after, a very influential technique is introduced: the authors of [179] and [13] propose a convex, l_1-penalised maximum likelihood problem to estimate a sparse inverse-covariance matrix. The relaxed convex problem scales very well in high dimension. This technique is mostly known as the "Graphical LASSO" (GLASSO), however this name technically describes only one of the many numerical schemes later proposed to solve this l_1-penalised maximum likelihood problem, namely the method of [50]. Alternative algorithms that all solve this optimisation problem include the work of [29], [136], and [137], the Nesterov smooth gradient descent of [31] and [110], the Alternating Direction Method of Multipliers (ADMM) of [180] and [139], the Interior Point Method (IPM) of [99], the Sparse INverse COvariance selection algorithm (SINCO) of [140], the Newton method of [161] and [125], the Projected Subgradient Methods (PSM) of [43], the QUadratic approximation of Inverse Covariance matrices (QUIC) of [66], the Dual Primal Graphical LASSO (DP-GLASSO) of [114], the High-dimensional Undirected Graph Estimation package (HUGE) of [186], the Proximal Gradient descent of [40] and the Reproducing Kernel Hilbert Space (RKHS) method of [98]. Works such as [89] and [133] study the theoretical properties of the solution of the convex optimisation problem.
However, there are many other methods that estimate sparse graphs/matrix and that are not equivalent to the the convex l_1-penalised maximum likelihood problem. Some slight variants are still regularised MLE, for instance [45], who proposes an adaptative LASSO and a Smoothly Clipped Absolute Deviation (SCAD) penalty. Another example is the nonparanormal model of [105] and [104] that relaxes the Gaussian assumption. The *pathway Graphical Lasso* [58], that takes into account prior knowledge about sets connected components (pathways), is also a MLE. Other methods differ more radically from the regularised MLE, such as the Dantzig-type estimator of [177], the Constrained l_1-minimization for Inverse Matrix Estimation (CLIME) algorithm of [18] which solves a sparse matrix reconstruction problem and makes no Gaussian assumption. Other more recent original GGM estimators include the Sparse Column Inverse Operator (SICO) method [108], the Tuning-Insensitive Graph Estimation and Regression (TIGER) [106], the GGMselect [56], the scaled LASSO of [148] and the False discovery rate control based technique of [107]. When there is a latent structure within the graph, the authors of [7] developed the iterative SIMoNe (Statistical Inference for Modular Networks) algorithm.
There is also an abundant literature on Bayesian techniques for sparse inverse-covariance matrix estimation. These methods estimate the posterior distribution of the precision matrix, each with different, sparse priors. Popular priors used in the GGM context include the G-Wishart prior [94], the graphical LASSO prior [131,162], the continuous spike and slab prior [102,163] and the graphical horseshoe prior [100]. Recent works such as [51], [101] and [39] propose deterministic algorithms to explore efficiently the different possible posterior distributions.

These works, despite their methodological divergences, all address the same, plain, sparse Gaussian estimation problem. Some authors have deviated significantly more from this classical situation, and addressed altered problems.
The authors of [21] have introduced the problem of GGM estimation with latent variables. That is to say a subset of the variables in the graph are actually hidden. A framework that was later explored in more details by [112], [109] and [116]. In a similar fashion, the authors of [147] studied the case of GGM with missing values, while the very recent [10] addresses the problem of GGM censored values (i.e. where the too low and/or high values are capped at certain levels).
In order to describe random distributions with heavier tails, some works consider data following *Student's t-distribution* instead of the normal distribution. In particular, Finegold and Drton [47] adapt the Glasso of [50] into the *tlasso* and Ashurbekova et al [8] adapt the CLIME of [18] into the *tCLIME*.
Some authors alter the Gaussian model. The Matrix (or Matrix-variate) Normal distribution, already present in the work of [32] and more formally introduced in [60], describes a Gaussian vector that can naturally be reshaped into a matrix (usually with 2 entries, but any number is possible). This is a specific case of the Gaussian distribution with less degrees of freedom, since only the covari-

ance between the rows and the columns of the newly formed random matrix are free. The covariance matrix of the whole vector is the Kronecker product between the covariance matrices of the rows and of the columns. Although introduced in a more general context, this specialised Gaussian model adapted for the GGM case in [6] and [184]. These Matrix Normal Graphical Models (MNGM) were re-introduced by [93] who also provide a theoretical analysis. The authors of [174] make the argument that the MNGM are essential to study genomics data. They also provide a model selection procedure with theoretical bounds. The authors of [155] provide, among other things, a very similar bound on model selection with MNGM, which is supposedly better. Later works propose different MNGM estimators, the most notable examples are the bigraphical lasso [80], the Gemini [188] and the FDR-based method of [23]. Works such as [122] have proposed semi-parametric extensions of the MNGM.

Some authors have considered the case of GGM in the presence of additional external co-factors X. In the spirit of the Conditional Random Fields [88], they add these variables to the conditioning that defines each edge of the conditional correlation network. To estimate this newly defined graph with a Gaussian model, the authors of [173] introduce the Conditional Gaussian Graphical Model (CGGM) where the distribution of the features Y conditionally to the co-features X takes the form $\mathcal{N}(\beta X, \Sigma)$. They make no assumption on the distribution of X. They define a MLE estimator of the CGGM that is the solution of a bi-convex regularised optimisation problem. Several later works all consider the same estimator as [173], these include [19], [175] and [22]. A different work, [98], proposes a two stage estimator of the CGGM with RKHS. The authors of [164] estimate the model parameters with a series of penalised conditional regressions. Another brand of CGGM has been developed in parallel to the model of [173]. Introduced independently by [143], [171] and [181], these CGGM assume that the joint density of (Y, X) is not only known but also Gaussian. A more recent work by [26] estimates such a CGGM in the case where prior information on the structure is available. The authors of [11] adapt to the CGGM the problem with censored data that they introduced for GGM in [10]. They propose different methods for both brands of CGGM.

Some authors tackle the issue of time varying (or dynamic) networks. There are many possibilities when modelling the passing of time and the evolution of parameters, hence the work presented there is less unified. One of the first notable example is [145] that estimates time varying Bayesian networks. Then, many consecutive works consider a smoothly varying time, see [24, 83, 132, 144, 189]. Whereas other authors such as [84] study discrete time jumps. A recent work proposed with the *group-fused graphical lasso* to estimate discrete time varying networks [54]. There exist works such as [165] that consider varying graph edge weight, but fixed sparsity structure.

Consequent efforts have been made to develop, tune and adapt GGM techniques in order to describe homogeneous population under many different circumstances. Likewise, much effort has be made to transfer the know-how, the models and the algorithms from the homogeneous case to the case of a heterogeneous population with several, already identified and labelled, sub-populations to describe.

2.2 Hierarchical models

All the methods mentioned so far estimate only a single conditional correlation graph that is meant to correspond to the whole studied population. In the context of a heterogeneous population, authors have proposed Hierarchical Gaussian Graphical Models in order to allow each sub-population to be described with its own inverse-covariance matrix. In the supervised framework, where the population labels are known, this hierarchical problem is separable into as many simple problems as there are sub-populations. However, the different sub-populations may share core common elements which would be better identified if the dataset remained whole. Moreover, some sub-populations may contain too few data points to run a model successfully. These could benefit from the additional information contained within the data points belonging to other sub-populations, which are supposedly not completely alien with regards to each other. For these reasons, most authors have elected to estimate jointly all the parameters of their Hierarchical models and enforce and certain notion of common structure between them.

Early works such as [65, 160] estimate different matrices for each of the sub-population, but they all share the same sparsity structure. Most of the later works allow, and encourage, the different estimated matrices to have different sparsity structure, which in turn allows the different populations to have distinct conditional correlation networks. Some of the most influential works in this domain include [59], one of the pioneer of the joint Hierarchical GGM, who proposes a non-convex problem in which the common structure takes a multiplicative form. Later, lesser known work include [25] who imposed sign coherence between the different structures, [182] who used a fused nodewise lasso in a scenario with only 2 classes. A notable paper from Hara and Washio [62] treats the general case with K classes by considering an additive common structure and defining an estimator that is solution of a convex loss. A trend that many methods follow afterwards. The following later approaches became much more well known and influential. Zhu et al [192] design the *Structural Pursuit algorithm* with a concave l_1-type loss and prior on the groups that should have similar coefficients. The authors of [119], with their *node joint graphical lassos*, adapt the penalised maximum likelihood approach of [179] to the hierarchical case with a node based penalty. This is the more realised version of a prototype that they had proposed two years prior in [118]. Works such as [30] and [172] take a similar approach, with different penalties that enforce different types of structures, in their *joint Graphical LASSO* and *Fused Multiple Graphical Lasso* respectively. Departing from the regularised MLE, two later papers, [92] and [20], both generalise the *CLIME algorithm* [18] to the hierarchical case. In the same vein, [63] design a method with the Gaussian assumption relaxed (nonparanormal model) inspired by [20]. Actually, there are several works that take successful methods from the homogeneous case and adapt them to the hierarchical case. For instance, [134] uses statistical tests like [41]. Likewise, [111] proposes a neighbourhood selection approach with nodewise linear regressions similar to those of Meinshausen and Bühlmann [115].
Among the very recent works, many tackle tweaked problems, with additional objectives or constraints than simply estimating Hierarchical GGM. For instance in [183] there are both different populations and different data types. Which results in two "directions" of different covariance matrices. For the situation where there is prior knowledge about the interactions between some components, Wu et al [170] propose the *Weighted Fused Pathway Graphical Lasso* which generalises the pathway approach of [58] to the heterogeneous case.
Bayesian methods have also been implemented in the Hierarchical scenario. One of the first joint Bayesian estimator [130] use the G-Wishart prior. Afterwards [103] designed a joint Bayesian equivalent of the nodewise approach of Meinshausen and Bühlmann [115]. Then, [102] introduced the *Bayesian Joint spike-and-slab Graphical LASSO*. Other Bayesian works such as [151] consider the brand of models that express the common structure in a more complex, multiplicative, way. A recent work [142] takes a Bayesian approach to the problem of [183] where there are both different populations and different data types.
We discussed the methods that all estimate Hierarchical GGMs, albeit with some occasional additional constraints. Just like in the Homogeneous case, there are many opportunities to consider other models to describe a Heterogeneous population.

We find that, in the Heterogeneous case, the literature has yet to catch up when it comes the alternative models such as the MNGM and CGGM, and fewer works can be found.
The most notable Hierarchical MNGM example is Huang et al [67], who adapted some of the structure inducing penalties of [30] to the MNGM. A later work [191] also considers data following the Hierarchical MNGM model, but focuses on the estimation of the column-wise covariance matrix of the model only.
Chun et al [27] introduce a pseudo-Hierarchical CGGM model where the features Y can come from different "sources" (sub-populations) but not the co-features X, who are shared by all sources. They proposed a regularised MLE approach. Huan et al [69] later introduced a fully Hierarchical CGGM, where the pairs (Y, X) each belong to one of K sub-populations. Once again, they adapted some of the well known joint GGM penalties to the CGGM parameters. In a parallel work, they even proposed an extension to dynamic Hierarchical CGGM [68]. A very recent work [46] considers a model similar to Hierarchical CGGM, but only with a univariate co-factor.
A new method emerged to describe different sub-populations: the differential networks estimation,

where the inferred parameter is the difference between two precision matrices [185]. This method does not recover the common edges and focuses on describing the difference between groups. A later study introduces a differential estimator with lasso penalized D-trace loss [176]. A very recent work proposes a shrinkage-thresholing algorithm [152]. On a different note, the authors of [79] estimate sparse differential networks with a l_1 penalty in the context of Quadratic Discriminant Analysis. Some authors have considered variants of the joint differential networks analysis. In particular, we note applications to the Hierarchical MNGM [78], to the case with latent variables [124], and to the Hierarchial CGGM [123].

We see that numerous efforts have been made to describe Heterogeneous populations with Hierarchical GGM. All this work was done for labelled data, when the sub-population affiliations are known. For unsupervised heterogeneous populations, the literature is much less abundant.

2.3 Mixture models

Sometimes, the population is heterogeneous, and when the data arrives, the class labels are unknown. This is relatively uncharted territories with relation to the previously mentioned supervised hierarchical models.
For unlabelled populations, authors such as [187] and [86] have considered Mixtures of GGM. Since the MLE of a Mixture of Gaussian is a non convex problem, they propose Expectation-Maximisation (EM) algorithms [36] to find local likelihood maxima. The regularised MLE problems they consider include penalties that encourage the recovery of sparse precision matrices. However, they do not include any penalisation that would encourage the presence of common structure between the estimated matrices. Recent works like [52] and [61] correct this by using joint GGM penalties such as the Fused and Group Graphical LASSO penalties. In a similar fashion, the authors of [150] estimate a Bayesian Mixture of GGM. Other works such as [138] also propose a joint GGM estimation but adopt a completely different approach from the EM, and use instead a graph of proximity between the sub-populations.
We also mention some recent works that do not technically fit in the GGM context but face similar issues when inferring graphs for unlabelled populations. For instance, in the clustering problem of [48], the sparsity is imposed on the covariance and not on the precision matrix. They infer correlation graphs and not conditional correlation graphs. Maretic and Frossard [113] also focus on different matrices: the laplacian matrices, but they use the same kind of EM for Mixtures of sparse Gaussians. More recently, Ni et al [121] have worked with reciprocal graphs, which are a bit different since their are both directed and undirected. They develop several estimation techniques that do not make use if the EM algorithm.

To the best of our knowledge, there is no mention yet of unsupervised version of the alternative models such as the MNGM or the CGGM. This is one notable motivation behind our introduction of the Mixture of CGGM in Chapter 6. There is however another, very different, domain of research that makes use of models that exhibit some formal similarities with the Mixture of CGGM: the "Finite Mixture Regression models" (FMR). The FMR, see [37] or [81] for early examples of unpenalised FMR and penalised FMR respectively, consist of K parallel linear regressions of the form $Y^{(i)} = \beta_k X^{(i)}$, with unlabelled data $(Y^{(i)}, X^{(i)})$. The class label of each pair $(Y^{(i)}, X^{(i)})$ and the parameters β_k are estimated with an EM algorithm. Unlike our approach in Chapter 6, in FMR the predicted feature Y is usually one-dimensional, and the clustering is focused on identifying different linear models $X \mapsto Y$. There are some rare examples of FMR that consider a multivariate feature vector Y, such as [77]. Nevertheless, they consider no GGM-type penalty on the inverse-covariance of Y. This is very estranged from the GGM approach, that focuses on recovering the conditional covariance structure between the Y, and within which the inclusion of the effect of the co-feature X is mainly a tool to improve the clustering of the Y. The very recent FMR work of [129] portraits a one-dimensional feature Y and multidimensional co-features X with in-homogeneous generative models. Underlining the idea that the focus of this domain of research is on mixed linear regressions and not on graphical models for the features.

Many of these unsupervised methods make use of the EM algorithm to estimate a Mixture of GGM. In the following, we give a very brief overview of the main contributions in the EM literature, in particular those related to the convergence guarantees of the EM algorithm. The Expectation Maximisation algorithm was introduced by Dempster et al [36] to maximise non convex likelihood functions defined from inherent hidden variables. The algorithm is made of the iteration of an Expectation (E) step and a Maximisation (M) step. In addition to presenting the method, Dempster et al [36] provides convergence guarantees on the sequence of estimated parameters, namely that it converges towards a critical point of the likelihood function. Although their result was correct, their proof contained a mistake which was later corrected in [169]. The convergence guarantees of the algorithm were studied by Boyles [16]. On paper the EM algorithm can be applied with any likelihood function. In practice some likelihood functions can have problematic E step and/or M step. For thorny M steps, Wu [169] and Lange [90] proposed inexact optimisations with a coordinate descent and a gradient descent respectively. They both provide theoretical analysis of the convergence of their methods. Several later works tackle the case likelihood with intractable E steps. They replace the intractable E step by an approximation, usually relying Monte Carlo (MC) methods and Stochastic Approximations (SA). Notable examples include Delyon, Lavielle and Moulines [33] with the SAEM, Wei and Tanner [166] for the MC-EM, Fort and Moulines [49], the MCMC-EM, Kuhn and Lavielle [87] and Allassonnière, Kuhn and Trouvé [4], the MCMC-SAEM, and Chevalier and Allassonnière [2] for the Approximate SAEM. We highlight these contributions since they all come with their own theoretical convergence guarantees.

We brought to light that the literature on unsupervised GGM for Unlabelled Heterogeneous population is still new and growing. In this PhD thesis, we contribute to the development of this field by studying different key problematics. In particular, we delve into the EM algorithm for the Mixtures of GGM.

Chapter 3

Gaussian Graphical Model exploration and selection in high dimension low sample size setting

This Chapter has been published in IEEE-TPAMI. Ref: 10.1109/TPAMI.2020.2980542

Gaussian Graphical Models (GGM) are often used to describe the conditional correlations between the components of a random vector. In this Chapter, we compare two families of GGM inference methods: the nodewise approach of [115] and [56] and the penalised likelihood maximisation of [179] and [13]. We demonstrate on synthetic data that, when the sample size is small, the two methods produce graphs with either too few or too many edges when compared to the real one. As a result, we propose a composite procedure that explores a family of graphs with a nodewise numerical scheme and selects a candidate among them with an overall likelihood criterion. We demonstrate that, when the number of observations is small, this selection method yields graphs closer to the truth and corresponding to distributions with better KL divergence with regards to the real distribution than the other two. Finally, we show the interest of our algorithm on two concrete cases: first on brain imaging data, then on biological nephrology data. In both cases our results are more in line with current knowledge in each field.

3.1 Introduction

Dependency networks are a prominent tool for the representation and interpretation of many data types as, for example, gene co-expression [56], interactions between different regions of the cortex [17] or population dynamics. In those examples, the number of observations n is often small when compared to the number of vertices p in the network.
Conditional correlation networks are graphs where there exists an edge between two vertices if and only if the random variables on these nodes are correlated conditionally to all others. This structure can be more interesting than a regular correlation graph. Indeed, in real life, two phenomena, like the atrophy in two separate areas of the brain or two locations of bird migration, are very likely to be correlated. There almost always exists a "chain" of correlated events that "link", ever so slightly, any two occurrences. As a result, regular correlation networks tend to be fully connected and mostly uninformative. On the other hand, when intermediary variables explain the totality of the co-variations of two vertices, then these two are conditionally uncorrelated, removing their edge from the conditional correlation graph. The conditional correlation structure captures only the direct, explicit interactions between vertices. In our analyses, these interactions are the ones of most

interest.

A Gaussian Graphical Model (GGM) is a network whose values on the p vertices follow a Centred Multivariate Normal distribution in $\mathbb{R}^p$: $X \sim \mathcal{N}(0_p, \Sigma)$. This assumption is almost systematic when studying conditional correlation networks for three main reasons. First, it ensures that each conditional correlation $corr(X_i, X_j|(X_k)_{k\neq i,j})$ is a constant and not a function of the $p-2$ dimensional variable $(X_k)_{k\neq i,j}$; a crucial property allowing us to talk about a single graph and not a function graph. Second, it equates the notions of independence and un-correlation, in particular: $corr(X_i, X_j|(X_k)_{k\neq i,j}) = 0 \iff X_i \perp\!\!\!\perp X_j|(X_k)_{k\neq i,j}$. This makes interpretation much clearer. Finally, under the GGM assumption, we have the explicit formula: $corr(X_i, X_j|(X_k)_{k\neq i,j}) = -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$, where $K := \Sigma^{-1}$ is the inverse of the unknown covariance matrix. This means that the conditional correlations graph between the components of X is entirely described by a single matrix parameter, K. Moreover the graph and K have the exact same sparsity structure. With this property in mind, the author of [35] introduced the idea of Covariance Selection which consists of inferring - under a Gaussian assumption - a sparse estimation $\widehat{K}$ of K and interpreting its sparsity structure as a conditional dependency network.
Subsequently, many authors have proposed their own estimators $\widehat{K}$. In [115], a local edge selection approach that solves a LASSO problem on each node is introduced. It was noticeably followed by [55, 56], who developed the GGMselect algorithm, a practical implementation of this approach coupled with a model selection procedure. We call these methods "local", since they focus on solving problems independently at each node, and evaluating performances with an aggregation of nodewise metrics. Other works within the local paradigm have proposed Dantzing selectors [177], constrained l_1 minimisation [18], scaled LASSO [148], or merging all linear regression into a single problem [136]. On a different note, the authors of [179] and [13] considered a more global paradigm where the estimator is solution of a single l_1-penalised log-likelihood optimisation problem, that has the form of Eq. (3.1).

$$\widehat{K} := \underset{\widetilde{K}\succ 0}{\operatorname{argmax}}\, \mathcal{L}\left(\widetilde{K}\right) - \rho \sum_{i<j} \left|\widetilde{K}_{ij}\right| . \tag{3.1}$$

We call this point of view "global" since the likelihood estimates at once the goodness of fit of the whole proposed matrix. The introduction of problem (3.1) generated tremendous interest in the GGM community, and in its wake, many authors developed their own numerical methods to compute its solution efficiently. A few notable examples are block coordinate descent for the Graphical Lasso algorithm (GLASSO) of [50], Nesterov's Smooth gradient methods [31], Interior Point Methods (IPM) [99], Alternating Direction Methods of Multipliers (ADMM) [139, 180], Newton-CG primal proximal point [161], Newton's method with sparse approximation [66], Projected Subgradient Methods (PSM) [43], and multiple QP problems for the DP-GLASSO algorithm of [114]. The theoretical properties of the solutions to Eq. (3.1) are studied in [137], [89] and in [133]. Other methods within the global paradigm include [45], with penalties other than l_1 in (3.1), and [98], with a RKHS estimator.
More recent works have proposed more involved estimators, defined as modifications of already existing solutions and possessing improved statistical properties, such as asymptotic normality or better element-wise convergence. The authors of [135] and [74] adapted solutions of local regression problems including [115], whereas [76] modified the solutions of (3.1). In [75], the two approaches are unified with a de-biasing method applied to both local and global estimators.

In our applications - where the number of observations n is a fixed small number, usually smaller than the number of vertices p - we did not find satisfaction with the state of the art methods from either the local or the global approach. On one hand, GGMselect yields surprisingly too sparse graph, missing many of the important already known edges. On the other hand, the only solutions from the penalised likelihood problem (1) that are a decent fit for real distribution have so many edges that the information is hidden. To interpret a graph, one would prefer an intermediary number of edges. Additionally, the low sample size setting requires a method with non-asymptotic theoretical properties.

In this paper, we design a composite method, combining the respective strengths of the local and global approaches, with the aim of recovering graphs with a more reasonable amount of edges, that also achieves a better quantitative fit with the data. We also prove non-asymptotic oracle bounds in expectation and probability on the solution.

To measure the goodness of fit, many applications are interested in recovering the true graph structure and focus on the "sparsistency". In our case, the presence or absence of an edge is not sufficient information. The correlation amplitude is of equal interest. Additionally, we need the resulting structure to make sense as a whole, that is to say: describe a co-variation dynamic as close as possible to the real one despite being a sparse approximation. This means that edgewise coefficient recovery - as assessed by the l_2 error $\left\|K-\widehat{K}\right\|_F^2 = \sum_{i,j}(K_{i,j}-\widehat{K}_{i,j})^2$ for instance - which does not take into account the geometric structure of the graph as a whole is not satisfactory either. We want the distribution function described by the proposed matrix to be similar to the original distribution. The natural metric to describe proximity between distribution functions is Cross Entropy (CE) or, equivalently, the Kullback-Leibler divergence (KL). In the end, the CE between the original distribution and the proposed one - $\mathcal{N}\left(0,\widehat{K}^{-1}\right)$ - is our metric of choice. Other works, such as [95] and [190], have focused on the KL in the context of GGM as well.

In the following, we quantify the shortcomings of the literature's local and global methods when the data is not abundant. The GGMselect graphs are very sparse, but consistently and substantially outperform the solutions of Eq. (3.1) in terms of KL, regardless of the penalisation intensity ρ. In the KL/sparsity space, the solutions of GGMselect occupy a spot of high performing, very sparse solutions that the problem (3.1) simply does not reach. Additionally, the better performing solutions of (3.1) are so dense that they are excessively difficult to read. Subsequently, we demonstrate that despite its apparent success, the GGMselect algorithm is held back by its model selection criterion which is far too conservative and interrupts the graph exploration process too early. This results in graphs that are not only difficult to interpret but also perform sub-optimally in terms of KL.
With those observations in mind, we design a simple nodewise exploration numerical scheme which, when initialised at the GGMselect solution, is able to extract a family of larger, better performing graphs. We couple this exploration process with a KL-based model selection criterion to identify the best candidates among this family. This algorithm is composite insofar as it combines a careful local graph construction process with a perceptive global evaluation of the encountered graphs.
We prove non-asymptotic guarantees on the solution of the model selection procedure. We demonstrate with experiments on synthetic data that this selection procedure satisfies our stated goals. Indeed, the selected graphs are both substantially better in terms of distribution reconstruction (KL divergence), and much closer to the original graph than any other we obtain with the state of the art methods. Then, we put our method to the test with two experiments on real medical data. First on a neurological dataset with multiple modalities of brain imaging data, where $n < p$. Then on biological measures taken from healthy nephrology test subjects, with $p < n$. In both cases, the results of our method correspond more to the common understanding of the phenomena in their respective fields.

3.2 Covariance Selection within GGM

3.2.1 Introduction to Gaussian Graphical Models

Let S_p^+ and S_p^{++} be respectively the spaces of positive semi-definite and positive definite matrices in $\mathbb{R}^{p\times p}$. We model a phenomenon as a centred multivariate normal distribution in $\mathbb{R}^p$: $X \sim \mathcal{N}(0_p, \Sigma)$. To estimate the unknown covariance matrix $\Sigma \in S_p^{++}$, we have at our disposal an iid sample $\left(X^{(1)},...,X^{(n)}\right)$ assumed to be drawn from this distribution. We want our estimation to bring interpretation on the conditional correlations network between the components of X. No real network is truly sparse, yet it is natural to propose a sparse approximation. Indeed, this means recovering in

priority the strongest direct connections and privileging a simpler explanation of the phenomenon, one we can hope to infer even with a small amount of data. Sparsity in the conditional correlations structure is equivalent to sparsity in the inverse covariance matrix $K := \Sigma^{-1}$. Namely $K_{ij} = 0 \iff \text{Corr}\left(X_i, X_j | (X_k)_{k \neq i,j}\right) = 0$. As a consequence, our goal is to estimate from the dataset a covariance matrix $\widehat{\Sigma} \in S_p^{++}$ with both a good fit and a sparse inverse $\widehat{K}$. We say that $\widehat{\Sigma} := \widehat{K}^{-1}$ is "inverse-sparse".

In the following, we use the Cross Entropy to quantify the performances of a proposed matrix $\widehat{K}$. The CE, $H(p,q) = -\mathbb{E}_p\left[\log q(X)\right] = \int_x -p(x) ln(q(x)) \mu(\mathrm{d}x)$, is an asymmetric measure of the deviation of distribution q with regards to distribution p. The CE differs from the KL-divergence only by the term $H\left(p,p\right)$, which is constant when the reference distribution p is fixed. In GGM, the score $H(f_\Sigma, f_{\widehat{\Sigma}})$ represents how well the normal distribution with our proposed covariance $\widehat{\Sigma}$ is able to reproduce the true distribution $\mathcal{N}\left(0, \Sigma\right)$. We call this score the True CE of $\widehat{\Sigma}$. This metric represents a global paradigm where we explicitly care about the behaviour of the matrix as a whole. This is in contrast to a coefficient-wise recovery, for instance, which is a summation of local, nodewise, metrics. After removal of the additive constants, we get the simple formula (3.2) for the CE between two centred multivariate normal distributions $\mathcal{N}\left(0, \Sigma_1\right)$ and $\mathcal{N}\left(0, \Sigma_2\right)$.

$$H\left(\Sigma_1, \Sigma_2\right) := H\left(f_{\Sigma_1}, f_{\Sigma_2}\right) \equiv \frac{1}{2}\big(tr\left(\Sigma_1 K_2\right) - ln(|K_2|)\big)\,. \tag{3.2}$$

In the general case, the CE between a proposed distribution f_θ and an empirical distribution $\hat{f}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x=X^{(i)}}$ defined from data is the opposite of the log-likelihood:

$$H(\hat{f}_n, f_\theta) = -\frac{1}{n}\log p_\theta(X^{(1)}, ..., X^{(n)})\,.$$

In the GGM case, we denote the observed data $\underline{X} := \left(X^{(1)}, ..., X^{(n)}\right)^T \in \mathbb{R}^{n \times p}$, and set $S := \frac{1}{n}\underline{X}^T\underline{X} \in S_p^+$, the empirical covariance matrix. The opposite log-likelihood of any centred Gaussian $\mathcal{N}\left(0, \Sigma_2\right)$ satisfies:

$$H\left(S, \Sigma_2\right) := H\left(\hat{f}_n, f_{\Sigma_2}\right) \equiv \frac{1}{2}\big(tr\left(SK_2\right) - ln(|K_2|)\big)\,, \tag{3.3}$$

similar to Eq. (3.2). As a result, we adopt an unified notation. Details on calculations to obtain these formulas can be found in Section 3.10.1.

We use the following notations for matrix algebra, let A be a square real matrix, then: $|A|$ denotes the determinant, $\|A\|_* := tr\left(\left(A^TA\right)^{\frac{1}{2}}\right)$ the nuclear norm, $\|A\|_F := tr\left(\left(A^TA\right)\right)^{\frac{1}{2}} = \left(\sum_{i,j} A_{ij}^2\right)^{\frac{1}{2}}$ the Frobenius norm and $\|A\|_2 := \underset{x}{sup}\frac{\|Ax\|_2}{\|x\|_2} = \lambda_{max}(A)$ the spectral norm (operator norm 2) which is also the highest eigenvalue. We recall that when A is symmetrical positive, then $\|A\|_* = tr(A)$ and $\|A\|_F = tr(A^2)^{\frac{1}{2}}$. We also consider the scalar product $\langle A, B\rangle := tr\left(B^TA\right)$ on $\mathbb{R}^{p\times p}$.

3.2.2 Description of the state of the art

After its introduction, problem (3.1) became the most popular method to infer graphs from data with a GGM assumption. Reducing the whole inference process to a single loss optimisation is convenient. What is more, the optimised loss is a penalised version of the likelihood - which is an estimator of the True CE - hence the method explicitly takes into account the global performances of the solution. However, even though the l_1 penalty mechanically induces sparsity in the solution, it does not necessarily recover the edges that best reproduce the original distribution, especially when the data is limited. Indeed, the known "sparsitency" dynamics of the solutions of (3.1), see [89], always involve a large number of observations tending towards infinity. We demonstrate in this paper that, when the sample size is small, other methods recover consequently more efficient sparse structures, inaccessible to the l_1 penalised problem (3.1).

On the other hand, the local approach of [115] carefully assesses each new edge, focusing on making the most efficient choice at each step. We confirm that the latter approach yields better performance by comparing the solutions of problem (3.1) and GGMselect [56] on both synthetic and real data (Sections 3.4 and 3.5). However, the loss optimised in GGMselect, $Crit(\mathcal{G})$, see (3.4), is an amalgam of local nodewise regression score, with no explicit regard for the overall behaviour of the matrix:

$$Crit(\mathcal{G}) := \sum_{a=1}^{p} \left[\left\| X_a - \underline{X} \left[\widehat{\theta}_{\mathcal{G}} \right]_a \right\|_2^2 \left(1 + \frac{pen(d_a(\mathcal{G}))}{n - d_a(\mathcal{G})} \right) \right] , \tag{3.4}$$

where pen is a specific penalty function, $d_a(\mathcal{G})$ is the degree of the node a in the graph $\mathcal{G}$, X_a are all the observed values at node a, such that $\underline{X} = (X_1, ..., X_p) \in \mathbb{R}^{n \times p}$ is the full data, and:

$$\begin{aligned} \hat{\theta}_{\mathcal{G}} &:= \underset{\theta \in \Lambda_{\mathcal{G}}}{\operatorname{argmin}} \, \|\underline{X}(I_p - \theta)\|_F^2 \\ &= \underset{\theta \in \Lambda_{\mathcal{G}}}{\operatorname{argmin}} \sum_{a=1}^{p} \|X_a - \underline{X} \, [\theta]_a\|_2^2 \\ &= \left\{ \underset{\theta_a \in \Lambda_{\mathcal{G}}^a}{\operatorname{argmin}} \, \|X_a - \underline{X}\theta_a\|_2^2 \right\}_{a=1}^{p} , \end{aligned} \tag{3.5}$$

where $\Lambda_{\mathcal{G}}$ is the set of $p \times p$ matrices θ such that $\theta_{i,j}$ is non zero if and only if the edge (i, j) is in $\mathcal{G}$, and $\Lambda_{\mathcal{G}}^a$ is the set of vectors $\theta_a \in \mathbb{R}^p$ such that $(\theta_a)_i$ is non zero if and only if the edge (i, a) is in $\mathcal{G}$. Note that by convention, auto-edges (i, i) are never in the graph $\mathcal{G}$, and, in our work, $\mathcal{G}$ is always undirected. The full expression of pen can be found in Eq. 3 of [56]. It depends on a dimensionless hyper-parameter called K which the authors recommend to set equal to 2.5. We first tried other values without observing significant change, and decided to use the recommended value in every later experiment.

The expression (3.5) illustrates that each nodewise coefficients $\left[\hat{\theta}_{\mathcal{G}} \right]_a$ in the GGMselect loss are obtained from independent optimisation problems which each involve only the local sparsity of the graph in the vicinity of the node a, as seen in the definition of $\Lambda_{\mathcal{G}}^a$. In each parallel optimisation problem $\underset{\theta_a \in \Lambda_{\mathcal{G}}^a}{\operatorname{argmin}} \, \|X_a - \underline{X}\theta_a\|_2^2$, the rest of the graph is not constrained, hence is implicitly fully connected. In particular, the solutions of such problems involve an estimation of the covariance matrix between the rest of the vertices that is not inverse-sparse. This can bias the procedure towards the sparser graphs since it actually implicitly measures the performances of more connected graphs. Finally, the GGMselect model selection criterion (GGMSC) explicitly penalises the degree of each node in the graph making it so that string-like structures are preferred over hubs. Empirically, we observe that with low amounts of data, graphs with hubs are consistently dismissed by the GGMSC. Overall, we expect the selected solutions to be excessively sparse, which experiments on both synthetic and real data in Sections 3.4 and 3.5 confirm.

3.2.3 Graph constrained MLE

Even though a covariance matrix Σ uniquely defines a graph with its inverse K, the reciprocal is not true. To a given graph $\mathcal{G} := (V, E)$, with vertex set V and edge set E, corresponds a whole subset $\Theta_{\mathcal{G}}$ of S_p^{++}:

$$\Theta_{\mathcal{G}} := \left\{ \widetilde{\Sigma} \in S_p^{++} \middle| \forall i \neq j, \ (i, j) \notin E \Rightarrow \left(\widetilde{\Sigma}^{-1} \right)_{ij} = 0 \right\} .$$

When data is available, the natural matrix representing $\mathcal{G}$ is the constrained MLE:

$$\widehat{\Sigma}_{\mathcal{G}} := \underset{\widetilde{\Sigma} \in \Theta_{\mathcal{G}}}{\operatorname{argmax}} \, p_{\widetilde{\Sigma}}(X^{(1)}, ..., X^{(n)}) = \underset{\widetilde{\Sigma} \in \Theta_{\mathcal{G}}}{\operatorname{argmin}} \, H\left(S, \widetilde{\Sigma} \right) . \tag{3.6}$$

The existence of the MLE is not always guaranteed (see [35, 158]). When $n < p$, no MLE exists for the more connected graphs. However, in this paper, we design a procedure that can propose a MLE

for any n and any graph without computation errors. To tackle the issue of existence, we add a very small regularisation term to the empirical covariance matrix S. This leads to solving:

$$\widehat{\Sigma}_{\mathcal{G},\lambda} := \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S + \lambda I_p, \widetilde{\Sigma}\right). \tag{3.7}$$

λ is not a true hyper-parameter of the model. Its value is set once and for all, and as small as possible as long as the machine still recognises $S + \lambda I_p$ as invertible. Typical values range between 10^{-7} and 10^{-4}. This trick changes little for the already existing solutions. Indeed, if $\widehat{\Sigma}_{\mathcal{G}}$ solution of Eq. (3.6) exists, we observe empirically that for small values of λ: $\widehat{\Sigma}_{\mathcal{G}} \simeq \widehat{\Sigma}_{\mathcal{G},\lambda}$. On the other hand, if no solution $\widehat{\Sigma}_{\mathcal{G}}$ to Eq. (3.6) exists, then we now are able to propose a penalised MLE $\widehat{\Sigma}_{\mathcal{G},\lambda}$, thus avoiding degenerated computations. From now on, the MLE we use are always solutions of (3.7). We will omit the index λ and keep the notation $\widehat{\Sigma}_{\mathcal{G}}$ for the sake of simplicity.

3.2.4 Our composite algorithm

The exploration steps of our method are a variation of the local paradigm of [115]. First, we use the GGMselect solution as initialisation. Then we add edges one by one: at each step, for each vertex independently, we run a sparse linear regression using as predictors the vertices that are not among its neighbours yet, and as target the residual of the linear regression between the value on the vertex and its neighbours. With these regressions, each vertex proposes to add to the current graph an edge between them and their new best predictor. Here however, we deviate from the local paradigm by using a global criterion - the out of sample likelihood of the whole resulting new matrix - to evaluate each proposition and select one edge among these candidates. We end this exploration procedure after a fixed number of steps, the result is a family of gradually more connected graphs. The final selection step is done with a global metric: we pick, among the so constructed family, the graph minimising the Cross Validated (with fresh data) Cross Entropy. See Figure 3.1 for the details.
In the spirit of [74, 75, 76, 135], this method is designed to complete an already existing efficient, but sparse, solution. As a result, it is sensitive to the initial graph.

3.3 Oracle bounds on the model selection procedure

In this Section, we give non-asymptotic guarantees on the model selection step of our algorithm. We prove these results in Section 3.10. Using the statistical properties of our model selection criterion, in particular the absence of bias and convergence towards the oracle criterion, we describe the difference between the performance of the selected model and the oracle best performance ("regret"). This regret is dependent on the convergence of a Wishart random variable towards its expectation. As a result, we are able to prove non-asymptotic upper bounds in expectation and probability for the regret.

3.3.1 Framework

In this Section we define or recall the relevant concepts and notations. Let $\widetilde{\Sigma} \in \Theta_{\mathcal{G}}$ and $\widetilde{K} := \widetilde{\Sigma}^{-1}$. We recall and rephrase the definition, given in Eq. (3.7), of the constrained Maximum Likelihood Estimator we build from a given graph $\mathcal{G}$:

$$\begin{aligned}\widehat{\Sigma}_{\mathcal{G}}(S) &= \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S + \lambda I_p, \widetilde{\Sigma}\right) \\ &= \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S, \widetilde{\Sigma}\right) + \frac{\lambda}{2}\left\|\widetilde{K}\right\|_*.\end{aligned}$$

We use the Cross Validated Cross Entropy (CVCE) $H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ as a criterion to pick a graph $\widehat{\mathcal{G}}_{CV}$ among the ones encountered. This Cross Validated criterion uses the partition of the

Inputs: The *train* set are all the observations available for graph inference, Nb of steps T fixed in advance.
Start:
Run GGMselect on the *train* set to get the initial graph $\mathcal{G}_0 = (V, E_0)$;
Partition the *train* set into a *validation* set and *exploration* set;
for $t = 1, .., T$ **do**
  Partition randomly the *exploration* set into a *learning* set and an *evaluation* set;
  Compute the empirical covariance S^t_{eval} from the *evaluation* set;
  # We then "ask" each node for its desired next neighbour:
  for $a \in V$ vertex of $\mathcal{G}_{t-1}$ **do**
    Let $N_{t-1}(a)$ be the set of neighbours of a in $\mathcal{G}_{t-1}$ and $F_{t-1}(a) := V \setminus \{N_{t-1}(a) \cup \{a\}\}$ the remaining vertices;
    Run on the *learning* set the linear regression with the vector X_a of the values on a as the target, and the vectors $\{X_s | s \in N_{t-1}(a)\}$ on the neighbour nodes as predictors. Let $\tilde{X}_a$ be the residual of this regression;
    Run on the *learning* set one step of the LARS algorithm of [44], with $\tilde{X}_a$ as the target, and the remaining $\{X_s | s \in F_{t-1}(a)\}$ as predictors. Call $c_t(a) \in F_{t-1}(a)$ the index of the feature chosen by LARS;
  end for
  # We now have p potential new edges $\{(a, c_t(a))\}_{a \in V}$ *some of which can be identical*
  # We give priority to mutual selections: when $c_t(c_t(a)) = a$
  if $\{(a, c_t(a))\}_{c_t(c_t(a))=a} \neq \emptyset$ **then**
    Let $\mathcal{C} = \{(a, c_t(a))\}_{c_t(c_t(a))=a}$ be our set of candidate edges;
    # We keep only the mutual selections
  else
    Let $\mathcal{C} = \{(a, c_t(a))\}_{a \in V}$;
    # No mutual selection ⇒ keep the whole set
  end if
  for $c \in \mathcal{C}$ **do**
    Compute, with the *learning* set, the MLE $\widehat{\Sigma}^c_t$ from each new potential graph $\mathcal{G}^c_t := \mathcal{G}_{t-1} \cup c$;
  end for
  $c^* := \underset{c \in \mathcal{C}}{\operatorname{argmin}}\, H\left(S^t_{eval}, \widehat{\Sigma}^c_t\right)$;
  $\mathcal{G}_t := \mathcal{G}^{c^*}_t$;
  Compute, with the *exploration* set, the MLE $\widehat{\Sigma}_t$ from $\mathcal{G}_t$;
end for
Compute, with the *exploration* set, the MLE $\widehat{\Sigma}_0$ from $\mathcal{G}_0$;
Compute the empirical covariance S_{val} from the *validation* set;
$t^* := \underset{t=0,\dots,T}{\operatorname{argmin}}\, H\left(S_{val}, \widehat{\Sigma}_t\right)$;
$\widehat{\mathcal{G}} := \mathcal{G}_{t^*}$;
Return: Inferred graph $\widehat{\mathcal{G}}$.

Figure 3.1: Composite GGM estimation. We respectively identify with green or orange text the steps adhering to a local or global paradigm. Comments are in *blue*.

training set into a *validation* set - used to build the estimation S_{val} of the true matrix Σ - and an *exploration* set - used for the graph exploration process and to build the constrained MLE $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$ for each encountered graph $\mathcal{G}$. We compare the graph $\widehat{\mathcal{G}}_{CV}$ selected with CVCE with $\widehat{\mathcal{G}}^*$ selected with the True Cross Entropy $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ of the matrix $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$. We define formally those graphs: in Eq. (3.8) and (3.9):

$$\widehat{\mathcal{G}}^* \in \underset{\mathcal{G}\in\mathcal{M}}{\operatorname{argmin}} \left[H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)\right], \tag{3.8}$$

$$\widehat{\mathcal{G}}_{CV} \in \underset{\mathcal{G}\in\mathcal{M}}{\operatorname{argmin}} \left[H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)\right], \tag{3.9}$$

where we call $\mathcal{M}$ the family of graphs uncovered by the Composite algorithm.

Remark. With the data available, the ideal model selection would be made with True Cross Entropy $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right)$ of the matrix $\widehat{\Sigma}_{\mathcal{G}}(S_{train})$ built from the whole *train* set. Comparing ourselves to this criterion would allow to quantify the importance of having a balanced split between *validation* and *exploration* set. This is outside the scope of this Section. We just compare our $H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ to $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$. In this case, the convergence of S_{val} towards Σ is the only dynamic that matters.

3.3.2 Basic control

In this Section, we show a general upper bound on the regret, using only the properties of the model selection criterion, and not yet the properties of the estimators. From this point on, we generally do not highlight the dependency of $\widehat{\Sigma}_{\mathcal{G}}$ in S_{expl} to simplify notation. First of all, note that by definition we always have the lower bound on the difference of CE:

$$0 \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right).$$

The rest of the guarantees focus on the upper bounds for this difference.
From the observation that $H\left(\Sigma, \widehat{\Sigma}\right) = H\left(S, \widehat{\Sigma}\right) + \frac{1}{2}\left\langle \Sigma - S, \widehat{K}\right\rangle$, we get the control (3.10) on the regret $H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)$:

$$H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) \leq \frac{1}{2}\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}} - \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle, \tag{3.10}$$

where all the MLE $\widehat{\Sigma}_{\mathcal{G}}$ depend only on $\mathcal{G}$ and S_{expl}. The random variable $\widehat{\mathcal{G}}^*$ is a function of S_{expl} only, whereas $\widehat{\mathcal{G}}_{CV}$ depends on both S_{val} and S_{expl}. Since S_{val} and S_{expl} are independent, then:

$$\mathbb{E}\left[\left\langle S_{val}, \widehat{K}_{\widehat{\mathcal{G}}^*}(S_{expl})\right\rangle \middle| S_{expl}\right] = \left\langle \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}(S_{expl})\right\rangle.$$

In the end, with $e := \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right]$ the expected regret, we have:

$$0 \leq e \leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right]. \tag{3.11}$$

3.3.3 Control in expectation

In this Section, we use the sparsity properties of the estimator $\widehat{K}_{\widehat{\mathcal{G}}_{CV}}$ as well as the statistical properties of $\Sigma - S_{val}$ to obtain a more explicit control on the expected regret. In addition, we use a known concentration result to obtain an alternative control in expectation. The result (3.11) is completely agnostic of the way the matrices $\widehat{K}_{\mathcal{G}} \in S_p^{++}$ are defined as long as they depend on S_{expl} only. To get an order of this control, however, we use the assumption that $\widehat{\Sigma}_{\mathcal{G}}$ is the graph

constrained MLE defined in (3.7). Let us first notice that we can ensure $\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}$ thanks to our penalised definition of (3.7). Let $\Sigma_\infty := \max_{i,j} |\Sigma_{ij}|$. We call $E_{\max}$ the union of the maximal edge sets in $\mathcal{M}$, and $d_{\max} = |E_{max}| \leq \frac{p(p-1)}{2}$ its cardinal. We underline here that, by convention, conditional correlation graphs do not contain self loops, hence the edge sets E never include any of the pairs $\{(i,i)\}_{i=1,\dots,p}$. We then get the control (3.12) by using Cauchy-Schwartz's inequality in (3.11).

Proposition 1. *With the previously introduced notations, if the set E_{max} is independent of the exploration empirical matrix S_{expl}, we have:*

$$0 \leq e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{(p + 2d_{max})^{\frac{1}{2}} p}{\sqrt{n_{val}}} . \tag{3.12}$$

In the case of our Composite procedure, by construction E_{max} is a random variable depending on the exploration set. However (3.12) still holds by replacing d_{max} with $\mathbb{E}[d_{max}]$:

$$0 \leq e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{(p + 2\mathbb{E}[d_{max}])^{\frac{1}{2}} p}{\sqrt{n_{val}}} . \tag{3.13}$$

We can get an alternative order of the control by using known concentrations inequalities.

Proposition 2. *By using the Theorem 4 of [85], we get:*

$$0 \leq e \leq c \frac{\lambda_{max}(\Sigma)}{\lambda} p \left(\sqrt{\frac{p}{n_{val}}} \vee \frac{p}{n_{val}} \right) . \tag{3.14}$$

Where c is a constant independent of the problem.

In the end, with (3.13) and (3.14), we have two different upper bounds on e and can use the minimum one depending on the situation.

3.3.4 Control in probability

In this Section, we use the sparsity properties of the estimator $\widehat{K}_{\widehat{\mathcal{G}}_{CV}}$ as well as the concentration properties of $\Sigma - S_{val}$ around 0 to obtain a control in probability (concentration inequality) on the regret. In addition to the controls in expectation we got in (3.11) and (3.12), there is in the CVCE a concentration dynamic based on the convergence rate of a Wishart random matrix towards its average. We call $\Pi_{\max}$ the orthogonal projection on the set of edges $E_{\max} \cup \{(i,i)\}_{i=1}^p$. That is to say, for any matrix $M \in \mathbb{R}^{p\times p}$, $\Pi_{\max}(M)_{i,j} = M_{i,j}\mathbb{1}_{(i,j)\in E_{\max}\cup\{(i,i)\}_{i=1}^p}$. Let $W := K^{\frac{1}{2}} S_{val} K^{\frac{1}{2}}$. Then $n_{val} W \sim \mathcal{W}_p(n_{val}, I_p)$ is a standard Wishart random variable depending only on the *validation* data, hence independent of every matrix $\widehat{K}_{\mathcal{G}}$. Let $P := \mathbb{P}\left(\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_*}\right)\right| \leq \delta\right)$ be the probability that the regret is small. We get two different lower bounds (3.15) and (3.16) on P.

Proposition 3. *With the previously introduced notations, the two following inequalities hold:*

$$P \geq \mathbb{P}\left(\|W - I_p\|_F \leq \frac{\delta}{\max_{\mathcal{G}} \left\|\Sigma^{\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{\frac{1}{2}}\right\|_F} \right) , \tag{3.15}$$

$$P \geq \mathbb{P}\left(\|\Pi_{max}(S_{val} - \Sigma)\|_F \leq \frac{\delta}{\max_{\mathcal{G}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F} \right) . \tag{3.16}$$

Moreover, the results (3.15) and (3.16) hold when every probability is taken conditionally to the exploration data or, equivalently here, conditionally to S_{expl}.

If we work conditionally to the *exploration* data, then $\max_{\mathcal{G}} \left\| \Sigma^{\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{\frac{1}{2}} \right\|_F$, $\max_{\mathcal{G}} \left\| \widehat{K}_{\mathcal{G}} \right\|_F$ and E_{max} are constants of the problem. In that case, the lower bound in (3.15) only depends on the dynamic of a standard Wishart $\mathcal{W}_p(n_{val}, I_p)$. Similarly, the lower bound in (3.16) only depends on the convergence dynamic of some coefficients of S_{val} towards the corresponding ones in Σ.
The bound in (3.16) has a less general formulation than (3.15), since the $S_{val} \mapsto \Sigma$ is a more specific dynamic than $W \mapsto I_p$. On the other hand, only the diagonal coefficients and those in $E_{\max}$ need to be close, which can make a huge difference if p is very large and $\mathcal{M}$ contains only sparse graphs and make the bound (3.16) tighter.

3.4 Experiments on synthetic data

We show in this Section the shortcomings of the global problem (3.1) of [179] and [13] and of the local approach of [115] and [56] on synthetic data. We demonstrate that - when the data is not abundant - the solutions of GGMselect consistently reproduce the true distribution much better than any solution of the global problem (3.1). In addition to being outperformed in KL divergence, the best solutions of (3.1) are also very connected, consequently more than the real graph. However, we also illustrate that the solutions of GGMselect are always very sparse, regardless of the real graph. In the end, we demonstrate that our selection criterion improves both the distribution reproduction and the graph recovery of the previous two methods.

3.4.1 The solutions missed by the global paradigm: a comparison of GLASSO and GGMselect

We start by comparing the two state of the art global and local paradigms, and show that the global paradigm misses crucial solutions when the number of observations is small. We use the scikit learn, see [127], implementation of the GLASSO of [50] to solve problem (3.1) for any penalisation level ρ and the R implementation of GGMselect, see [56], to represent the [115] approach.
We use an inverse-sparse covariance matrix Σ fixed once and for all to generate a matrix of observations $\underline{X}$. The same observations are provided to the two methods. On Figure 3.2, we compare the True CE $H(\Sigma, \widehat{\Sigma})$ of each estimated matrix as a function of the number of non-zero, off diagonal coefficients in their inverse $\widehat{K}$ (complexity of the model). The green dot is the MLE - computed as in (3.7) - under the constraints of the GGMselect graph. In the case of GLASSO, different solutions are obtained by changing the level of penalisation ρ in Eq. (3.1). We call those solutions $\widehat{\Sigma}_\rho$, indexed by their penalisation intensity ρ. They are represented by the blue curve on Figure 3.2. All of them are inverse-sparse and define a graph we call $\mathcal{G}(\rho)$. The orange curve is the path of the MLEs $\widehat{\Sigma}_{\mathcal{G}(\rho)}$ - computed as in (3.7) - refitted from those same graphs without the l_1 penalty of problem (3.1). They have the same inverse-sparsity as their raw solution counterparts, but do not have the extra-penalisation on the non-zero coefficients that every LASSO solution bears.
The three columns correspond to graphs with different connectivity - illustrated by a random example on top of each column - and the two rows have different graph sizes, $p = 30$ and $p = 50$ respectively. For each simulation, the two methods were given the same $n = 30$ observations to work with, and each figure represents the average and standard deviation of 100 simulations.

We notice that the GGMselect solution is always very sparse. When the true graph is sparse, GGMselect outperforms the penalised likelihood problem (3.1) regardless of the penalty intensity. For large connected graphs, the most connected solutions of (3.1) can perform better than the GGMselect solution. However GGMselect is consistently better than the equally sparse problem (3.1) solution. The failure of GLASSO to reach the spot of GGMselect in the performances/complexity with any penalisation intensity - even when the MLE is refitted from the GLASSO graph without penalty - indicates that when n is small, the l_1 penalised likelihood problem (3.1) has difficulties selecting the most efficient edges. Additionally, the better performing solutions of GLASSO have many edges - usually much more than the real graph - which draws the focus away from the relevant ones and makes it difficult to get a qualitative reading of the graph.

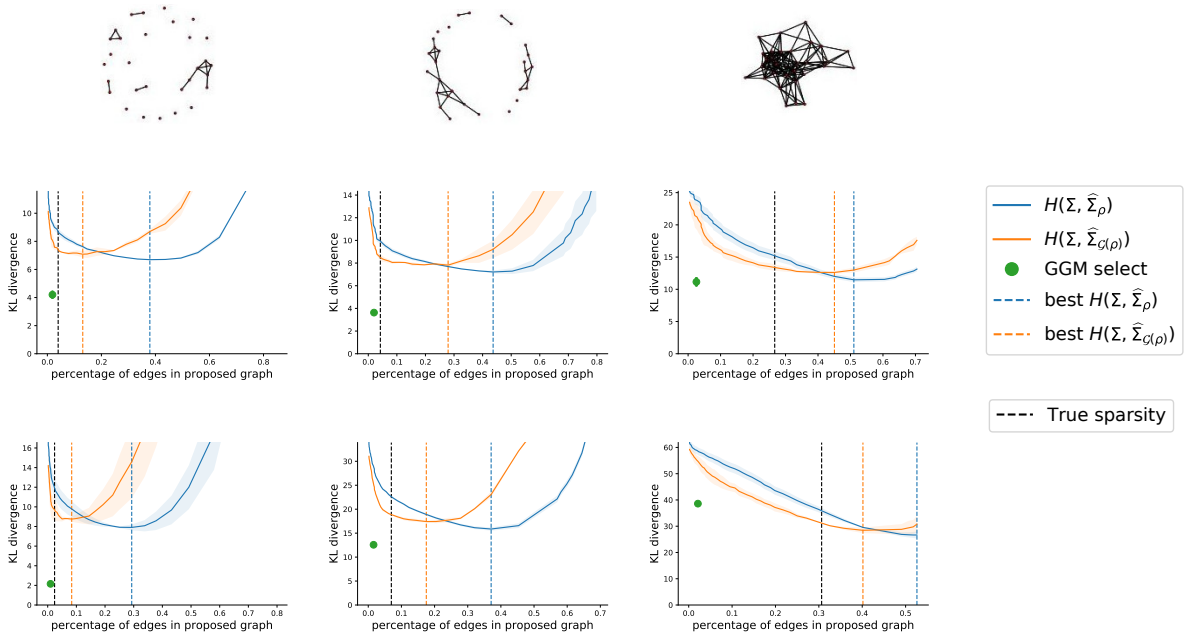


Figure 3.2: Average performances as a function of the complexity for: the MLE from the GGMselect graph (green), the GLASSO solutions (blue) and the MLEs from the GLASSO graphs (orange). The average is taken over 100 simulations. In each simulation, $n = 30$ data points are simulated from a given true graph, different for each subfigure. The two rows of subfigures correspond to two different graph sizes, $p = 30$ and $p = 50$ vertices respectively. The three columns correspond to true graphs with different connectivity. At the top of each column, a graph illustrates the typical connectivity of the true graphs in said column.

When the number of observations is small, it seems that GGMselect's numerical scheme allows it to find high performing sparse graphs that problem (3.1) never can. This is the type of solution we want, and the main reason why we choose to initialise our composite method from this point.

3.4.2 Conservativeness of the GGMselect criterion: an example with a hub

We identified that GGMselect produced high quality, very sparse solutions. We argue here that they might be too sparse for their own good.
As discussed in Section 3.2.2, the numerical scheme of the GGMselect algorithm is based on a nodewise approach, and so is its model selection criterion. It penalises independently the degree of every node in the proposed graph. This makes it very unlikely to select graphs with a hub, i.e. a central node connected to many others. However recovering hubs is very important in conditional correlation networks. Genetic regulation networks for instance often feature hubs. With synthetic data, $n = 30, p = 30$, we encounter a "soft cap" effect, where it becomes very hard for GGMselect to propose a graph including a node of degree higher than 3. The penalty for such a node being too large to be compensated by the improved goodness of fit. On the other hand, we see on Figure 3.3 that the Cross Validated Cross Entropy selects a graph which features the entire hub, and is in addition closer to the real graph regarding the remaining edges. Indeed, in the example of Figure 3.3, other edges than the ones forming the hub are also ignored by GGMselect. With such a behaviour of the model selection criterion when the number of observations n is small, the GGMselect graphs are hard to interpret, with many key connections potentially missing.
Such observations motivated us to replace the GGMselect criterion with the Cross Validated Cross Entropy for graph selection. The next subsection proposes a quantitative comparison of the graphs

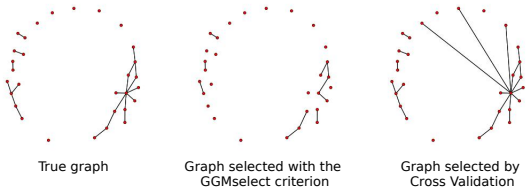


Figure 3.3: Graph selection in the presence of a hub. The first figure is the true graph. The second and third are the graphs respectively selected by the GGMSC and CVCE on the same fixed graph path going from the fully sparse to the fully connected, via the GGMselect graph and the true graph

selected by these two metrics.

3.4.3 The short-sightedness of the local model selection: a comparison of the GGMselect criterion and the CVCE

In this Section, we compare solely the model section metrics - and not the graph exploration schemes - on a fixed, shared, family of graphs. We demonstrate that our global approach to model selection yields graphs much closer to the original one and that reproduces the true distribution much better than the GGMselect criterion, which rejects the better, more connected graphs.
We compare the graphs selected by our Cross Validated CE (CVCE) and the GGMSC when shown the same family of candidate graphs. We consider a given true graph ($p = 30$). We compute once and for all one GGMselect solution with $n = 30$ observations drawn from this graph. With these key graphs in hand, we build manually (without the exploration scheme of Figure 3.1) a deterministic sequence of graphs. Starting from the Fully Sparse with no edges, we add one by one, and in an arbitrary order, the edges needed to reach the GGMselect graph. From there, in the same manner, we add the missing edges and remove the excess edges to reach the true graph. Finally, we add - still one by one, still in an arbitrary order - the remaining edges until the Fully Connected graph, with all possible edges. All the encountered graphs in this sequence constitute the fixed family of candidates to be assessed by the model selection criteria. For each simulation, we generate n observations and use them to compute the GGMSC and CVCE along the path. We make 1000 of those simulations. The GGMSC uses the full data freely, while the CVCE must split the n points into the *exploration* covariance S_{expl}, to compute the graph constrained MLE $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$, and a *validation* covariance S_{val} to evaluate them. This leads to different results depending on the split size. Let S_{train} be the empirical covariance matrix built with the full data. We assess the performances of each graph $\mathcal{G}$ with the True CE (TCE) of the MLE built from S_{train} under the constraints of $\mathcal{G}$: $H(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. Since there is a known true Σ we actually compute the True KL $KL(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. This metric differs from the TCE only by a constant, hence is equivalent when ranking methods, but offers a sense of scale since the proximity to 0 in KL is meaningful. Figure 3.4 illustrates the behaviour on one simulation. The most noticeable trend is that the GGMSC (in green) advocates a much earlier stop than the CVCE (in red), which stops almost on the same graph as the TCE (in blue). Additionally, on that run, the graph selected by the CVCE is actually the true graph (in grey). Figure 3.5 represents the results over all simulations. We compare the average and standard deviation of the performances (true KL, on the y axis) and complexity (number of edges, x axis) of the models selected by the CVCE with different *exploration/validation* splits (in shades of red), GGMSC (in green) and with the TCE (in blue). The three columns represent different number of available observations ($n = 25, 40, 100$) and the second row is a zoomed in view of the first. This quantitative analysis confirms that the GGMSC selects graphs that are way too sparse even when shown more complex graphs with better performances. With the performances measured in KL, relative improvement is meaningful, and we see the CVCE improving the GGMSC choice by a factor from 2 to 5, and being much closer to the oracle solution in terms of KL. Additionally, the graphs selected by CVCE are also much closer to the original one. This is especially true when a large fraction

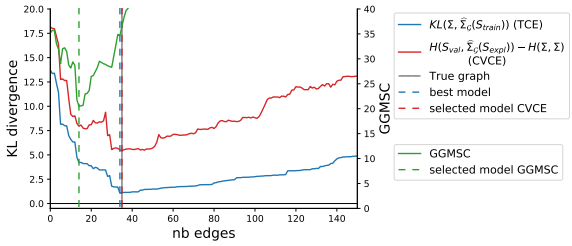


Figure 3.4: On a single simulation: evolution of and model selected by GGMSC (green), CVCE (red) and TCE (blue) along the fixed deterministic path. The true graph's position on that path is represented by a vertical grey line. GGMSC stops early whereas CVCE selects the true graph (the vertical grey line and the dashed red one are the same). Moreover, the CVCE graph is very close to the best graph in terms of True Cross Entropy.

of the data (35% or 40% of the *training* data) is kept in the *validation* set. The same results are observed with two other oracle metrics: the l_2 recovery of the True Σ, $\left\|\Sigma - \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right\|_F$, and the oracle nodewise regression l_2 recovery $\left\|\Sigma^{\frac{1}{2}}(I_p - \Theta_{\mathcal{G}}(\underline{X}_{train}))\right\|_F$ (the oracle metric of the GGMselect authors [56]). Those metrics also reveal that when the *validation* set is small (20%), the variance of the performances of CVCE increases and it can become less reliable depending on the metric. The Figures and details on these two metrics can be found in supplementary materials.
This experiment illustrated how the model selection criterion of GGMselect can actually be very conservative, and even though the numerical scheme of the method explores interesting graph families, the model selection criterion might dismiss the more complex, better performing ones on them. This leads us to believe we can make substantial improvements by using the CVCE on a path built using the GGMselect solution as initialisation.

3.4.4 Execution time comparison

In this Section, we compare the runtimes of GLASSO, GGMselect and the Composite method for several values of p. For each p, 20 simulation are made, with $n = p/2$ observations each. This number of observations is an arbitrary heuristic to have both $n < p$ and n increasing with p. Table 3.1 synthesises the results. The runtime and complexity of the Composite method depend linearly on the number of steps chosen by the user. As seen in Figure 3.1, this number of steps is the number of graphs that are constructed and evaluated. Ideally, this sequence of graphs should be just long enough to see the Oracle (or Out of Sample) performance improve as much as they can, and stop when they start deteriorating, when the point of overfitting is reached. In this experiment, the number of steps is chosen according to an heuristic depending on the number of edges in the initialisation graph with regards to p. The average number of steps over the simulations is also recorded in Table 3.1.
The Composite method and GGMselect both include a model selection step, however GLASSO just returns one solution of Eq. (3.1) for one given value of the penalty parameter ρ. As a result, all three methods are not strictly comparable. This was corrected in this experiment: for every simulation, the GLASSO is run on a grid of ρ with as many values as the number of estimated graphs by the Composite method. We call this the "grid GLASSO".

Table 3.1 shows that GGMselect is faster than the other two methods by 1 and 2 orders of magnitude in average. The Composite method is faster than the grid of GLASSOs when the dimension is small, but suffers when the dimension goes above $p = 100$. The Composite algorithm has indeed a high complexity in p, it runs $p \times n_{steps}$ ordinary linear regression with $p - 2$ features and computes then evaluates $(p + 1) \times n_{steps}$ graph constrained MLE of size $p \times p$ each.

The algorithmic of GGMselect and GLASSO were very well optimised by their respective authors.

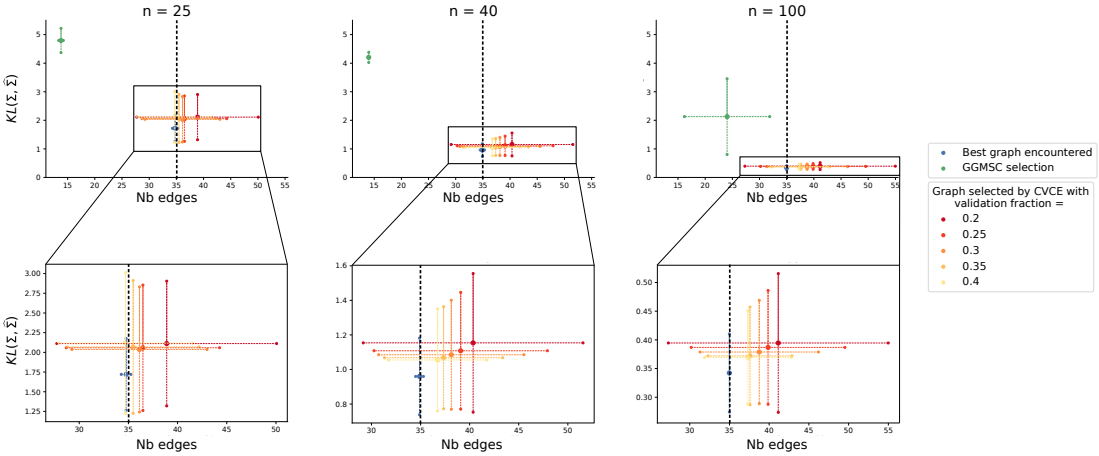


Figure 3.5: Average KL divergence (y axis) and complexity (x axis) of the models selected with GGMSC (green), CVCE (shades of red) and TCE (blue) on synthetic data. The sparsity level of the true graph is represented by a black dashed vertical line. The second row offers a zoomed in view of the boxed areas to focus on the CVCE and TCE models. The graphs selected by the CVCE are much closer to the best in True Cross Entropy in terms of performance and edge structure than the GGMSC one. Moreover, they are also very close to the true graph used in the simulation, even when the sample size is small.

Table 3.1: Average and (standard deviation) of the execution times of different GGM methods. The grid GLASSO compute solutions for as many values of the penalty parameter ρ as there are estimated graphs (steps) in the Composite method. The last column presents the average of this number of steps/number of estimated graphs. The number of observations is $n = p/2$.

p	GGMsel (fast)	grid GLASSO	Composite	nb steps
30	0.19 (0.07)	14.9 (8.60)	3.09 (1.80)	8.4
50	0.39 (0.03)	62.1 (32.9)	16.6 (8.20)	14.9
100	1.66 (0.66)	247 (135)	226 (138)	26.3
300	25.8 (1.04)	1470 (775)	6847 (1453)	40

This shows in the very fast GGMselect computations, making it a very efficient initialisation for our Composite method. However, the implementation of the Composite, see Figure 3.1, is naive and sequential. By running the linear regressions and LARS in parallel, and not re-calculating the MLE for the same graph several times, the performance would be greatly improved and closer to GLASSO.

3.5 Experiments on real data with the Composite GGM estimation algorithm

In this Section, we present two experiments with our composite method on real data. First, we demonstrate on brain imaging data from a cohort of Alzheimer's Disease patients that it recovers the known structures better than the classical local and global methods, while also having a better Out of Sample goodness of fit with the data. Then, we showcase how it is able to describe known dynamics between factors involved in Adrenal steroid synthesis on a database of Nephrology test subjects.

3.5.1 Experiment on Alzheimer's Disease patients

We first confirm our previous observations and demonstrate the performances of the complete numerical scheme of our composite procedure on real medical data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We have $p = 343$ features, $n = 92$ different patients. The first 240 features are measures of atrophy (MRI) and glucose consumption (PET) in the 120 areas of the cortex defined by the AAL2 map. The next 98 are two descriptors of the diffusion, fractional anisotropy and mean diffusivity, followed in the 49 regions of the JHU ICBM-DTI-81 white matter atlas. The rest of the features are basic descriptions of the patient.

Experiment

First we need a new evaluation metric. Indeed, with real data, we do not know the real covariance matrix. So we cannot anymore compute the True Cross Entropy to evaluate the inferred matrices. To replace the TCE, we keep $n = 18$ patients aside as a *test* set to define a test empirical covariance matrix S_{test}, whereas the $n = 74$ patients left constitute the *train* set, used to define S_{train}. We evaluate an inverse-sparse covariance matrix built from S_{train} with the negative Out of Sample Likelihood (OSL): $H(S_{test}, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. The OSL is less absolute than the True CE, but still quantifies with no bias the goodness of fit for real data. Additionally, we cannot use a KL divergence for scale reference anymore, see Section 3.10.1 for more details.
The experiment run on the ADNI database is very simple: we compute the GGMselect solution and build our Composite GGM estimation procedure from it. To be fair, we also evaluate every graph our procedure encounters with the GGMSC, giving GGMselect a chance to change its mind if one of the new graphs were to fit its criterion better. In addition, we used the GLASSO algorithm of [50] to get the solutions of (3.1) for different penalty intensity.

Comparison of GLASSO and GGMselect

We confirm the observations and conclusions of Section 3.4.1. Figure 3.6 shows that, even with varying penalty intensity, GLASSO does not encounter any solution with an OSL as good as GGMselect. This indicates that the optimisation problem (3.1) cannot find high-performing sparse graphs in this concrete setting either. The path of GLASSO is interrupted before its completion as we have computational error with the scikit learn package at low penalty levels. We encounter such errors eventually no matter how we regularise and precondition the empirical covariance S. This means we do not get to see the more connected solutions of the GLASSO. This is not a problem since we already go far enough in the GLASSO path to reach unacceptably complex graphs: 6% of the $\sim$ 59000 possible edges, i.e. 3500 edges for a graph with 343 nodes. By stopping early, we only consider the reasonable solutions of the GLASSO. In that case, GGMselect has a clear advantage, proposing a solution with a better Out of Sample fit with the data and only 281 edges.

Comparison of GGMselect and the Composite GGM estimation algorithm

We represent the selected graphs on left panel of Figure 3.7, with the same conventions as Figure 3.5. Once again the GGMSC (green) selects a sparse model, with 281 edges over the $\sim$ 60k possible. All the reasonable *validation* fractions (from 10% to 30%) of the CVCE (shades of red) select one out of two graphs, with both better OSL than the GGMSC one and closer to the OSL-optimum on the path (blue). Those two graphs have 589 or 813 edges respectively. This indicates that many conditional correlations were potentially missed by GGMSC, and that the CVCE graphs may propose a more complete interpretation.

For a full comparison of the thee methods, the right panel of Figure 3.7 is a zoomed out view that also includes the best model obtainable with problem (3.1) in terms of OSL (purple point). As we have seen, it is a very complex model with many edges. We visualise the successive improvements in Out of Sample Likelihood made first by GGMselect, with a sparser solution, then with our Composite GGM estimation procedure, with a more complete model. This experiment demonstrates the

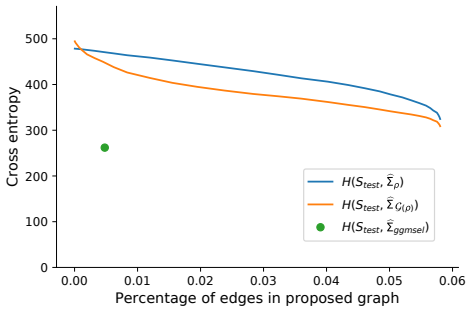


Figure 3.6: Out of sample performances as a function of the complexity for: the MLE from the GGMselect graph (green), the GLASSO solutions (blue) and the MLEs refitted from the GLASSO graphs (orange).

quantitative benefits of running the Composite algorithm in a High Dimension Low Sample Size setting.

In addition to those *quantitative* improvements, our method allows for a better *qualitative* interpretation of the disease. Figure 3.8 represents, using the Colin 27 brain image of [64] and the MRView software of [154], the graphs selected by GGMSC and CVCE (589 edges version), as well as the best GLASSO graph in OSL ($\sim$ 3500 edges). We recall that each of the methods estimates a large graph with $p = 343$ vertices, a mix of different modalities measured in different areas of the cortex. The full graph cannot be displayed on an image of the cortex. For the sake of clarity, we only represent sub-parts of this one graph. On Figure 3.8, only edges in-between the 120 MRI measures are represented. Additional views of the cortex can be found in supplementary materials. The GGMselect network is mostly composed of inter-hemispheric connections between symmetrical areas (hidden by the perspective in Figure 3.8, see the supplementary materials for different views). These mainly reflect the symmetry of the atrophy pattern and are less informative for understanding disease process. The intra-hemispheric connections have a better interpretation potential to explain the pathology. Our algorithm reveals many more of these correlations - for instance in parietal areas, which are thought to be key hubs in the disease process - promising a more interesting description of the pathology. The GLASSO solution on the other hand, proposes many edges, making even this simple sub-graph unreadable. Similar observations can be made for connections in-between PET measures (see supplementary materials).
Additionally, Figure 3.9 shows that the GGMselect graph features absolutely no edge between MRI and PET measures, effectively proposing a model in which there is no correlation whatsoever between anatomical and functional variables, a very unlikely and unsatisfactory description. Our method on the contrary recovers a reasonable amount of edges between those two modalities. GLASSO recovers a similar number of edges in this sub-part of the graph. However, Figure 3.8 shows that it does so while having an extremely large number of edges in other regions of the graphs. Sparser GLASSO solution on the other hand, behave similarly to GGMselect and recover no edge linking MRI and PET measures, see supplementary materials. Of all these solutions, the Composite method proposes the most balanced.
These results suggest that our approach could be an interesting tool to study inter-regional and inter-modality dependencies in Alzheimer's Disease. This would need to be confirmed with larger populations of patients and more extensive experiments, which is out of the scope of the present paper and is left for future work.

3.5.2 Experiments on neprhology patients

In this Section, we compare qualitatively the methods in an environment with $p < n$. Although the Composite procedure was developed specifically for the case $n < p$, we demonstrate here that it still

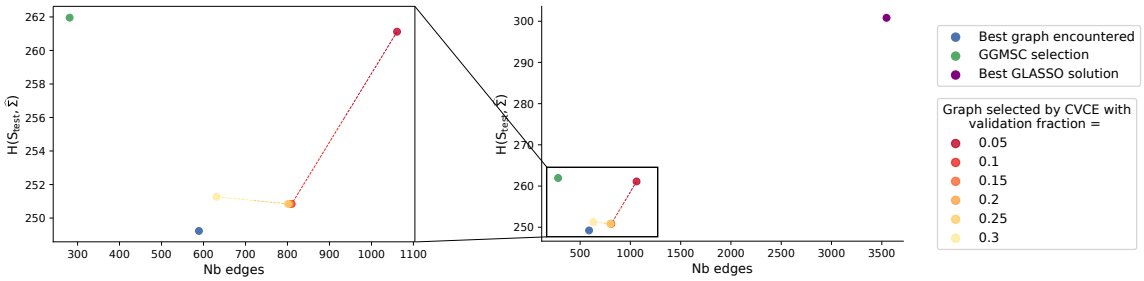


Figure 3.7: Out of Sample Likelihood (y axis) and complexity (x axis) of models selected by GGMSC (green), CVCE (shades of red) and OSL (blue) on real data. The right picture offers a zoomed out view to include the model selected by OSL on the GLASSO path (purple). The left figure corresponds to the boxed area of the right figure.

holds up to the state of the art outside of its intended application framework. We use a dataset of variables relevant to the adrenal steroidogenesis on a cohort of healthy test subjects.
Adrenal steroid synthesis in childhood is a complex process involving an enzymatic cascade that transforms cholesterol into mineralocorticoids, glucocorticoids or androgens, depending on the enzymatic equipment of each zona of the adrenal gland. Even though most important ways of adrenal steroidogenesis are known, we now assess new related metabolite that may ask new questions regarding adrenal steroidogenesis. Thus, we analysed a pediatric cohort of $n = 172$ healthy volunteers aged from 3 months to 16 years old with blood count and LC-MS/MS adrenal steroid profile analysis ($p = 35$).

Figure 3.10 represents the matrices of pairwise conditional correlations corresponding to the GGMselect solution (left), the Composite solution (middle) and a sparse GLASSO solution (right). The rest of the path of GLASSO solution can be found in the supplementary materials. The other solutions contain many more edges than any of the three matrices here.
The models proposed by the three matrices have been compared to literature data for hematological parameters and steroidogenesis analysis. Regarding hematological analysis, both the Composite and GGMselect models confirm well known relations such as strong direct positive links between hemoglobin concentration (Hb) and red cells count (RBC); between hemoglobin concentration and mean corpuscular volume (WCV); between white cells (WBC) and platelet counts (PC); and a strong negative link between red cells count and mean corpuscular volume; between white cells count and age. The GLASSO solution did not show any of them.
Regarding steroid metabolism, 11-β1 hydroxylase (11 Ohase B1) and 21 hydroxylase (21 Ohase) activities, the Composite method and GGMselect reach the same conclusion: there is a strong positive direct link between enzymatic activities and the concentration of their corresponding alternate product. This is in accordance with common description of adrenal steroidogenesis process: decreased activity leads to an accumulation product of the alternative pathway. The GLASSO solution failed to show these relations. In the same way, GGMselect and the Composite method exhibit a negative link between the lack of 11-β HSD type 2 (11b HSD2) activity (that catabolizes cortisol into cortisone) and the concentration of its product, cortisone (e). The sparse GLASSO fails to underline this link. All these data tend to show a better interpretation of steroids profile with the GGMselect and Composite solutions. Interestingly, these models also underline a new link: a strong positive link between 18-hydroxycorticosterone (18ohb) and 18-hydroxycortisol (18ohf) concentrations, two steroids that are supposed to be independently produced in two different zonas of the adrenal gland. This result could imply an alternative pathway in adrenal steroidogenesis that needs to be explored.
The GGMselect and Composite graphs are mostly identical, although some of the conditional correlations are weaker in the Composite matrix. Among the subtle differences, two edges that are coherent with the state of the art, and are present in the GGMselect graph, were alleviated in

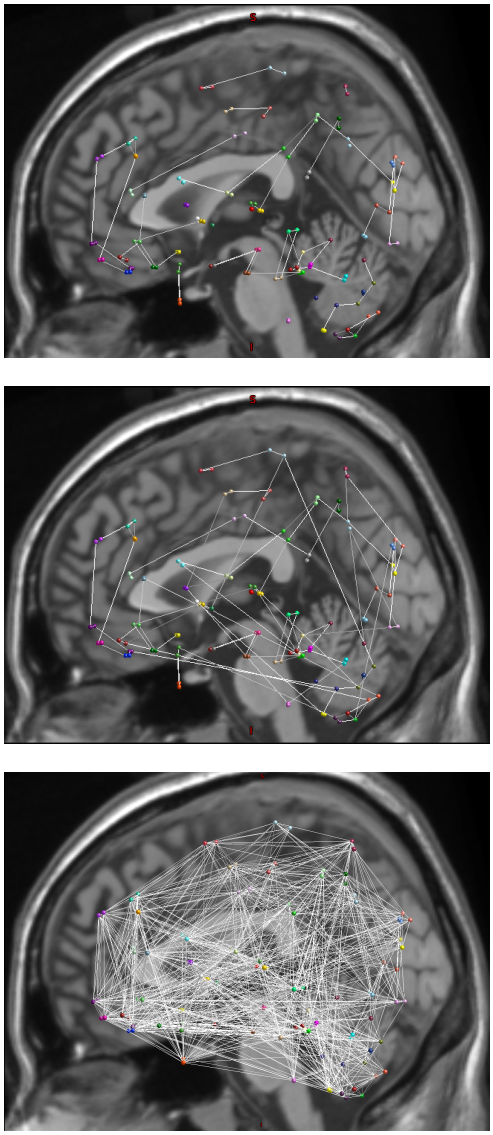


Figure 3.8: Selected edges by GGMselect (up), our Composite method (mid) and the best Out of Sample GLASSO (down) in-between MRI measures. The perspective of the sagittal view hides the many edges between symmetrical regions. GLASSO proposes too many to allow for interpretation.

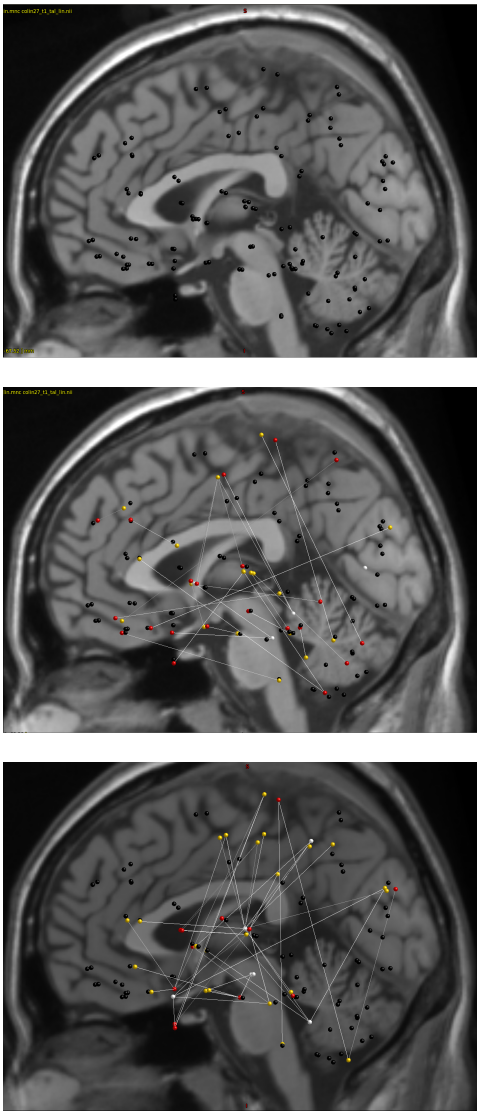


Figure 3.9: Selected edges by GGMselect (up), our Composite method (mid) and the best Out of Sample GLASSO (down) between PET (yellow) and MRI (red) measures. GGMselect finds no connection in this sub-part of the graph, although one may expect some.

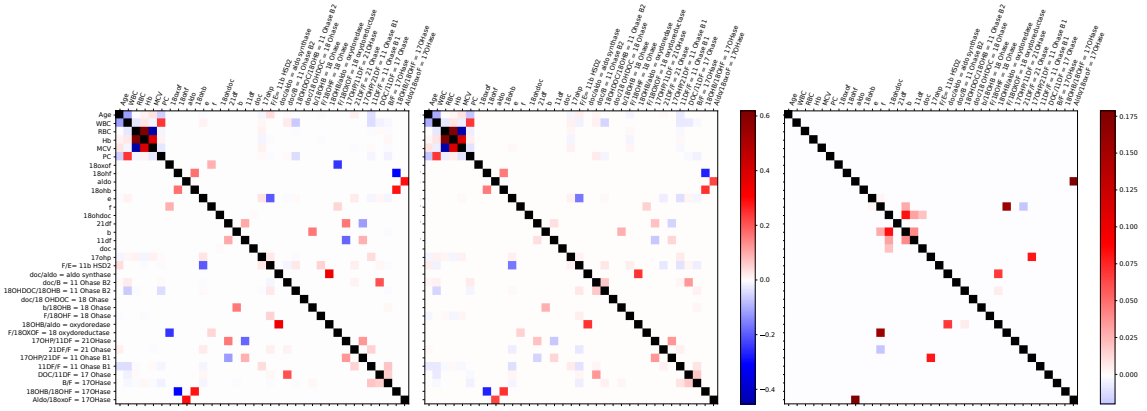


Figure 3.10: Conditional covariance matrix between the 35 variables measured on the cohort. The positive correlations are in red and the negative in blue. The diagonal coefficients are ignored in this study. GGMselect (left) and Composite (middle) share the same colour scale. The rightmost figure corresponds to one of the sparsest GLASSO solution.

the Composite matrix (resulting in invisible connections in Figure 3.10): the link between the 18-oxocortisol (18oxof) and cortisol (f) concentrations, and the very strong negative link between the ratio cortisol/18-oxocortisol (F/18oxof) and 18-oxocortisol. The other very few additions and removals in the Composite model are hard to validate or disprove with the current state of the art. From a medical analysis point of view, all these results are preliminary and will have to be confirmed by more in depth studies. From a purely machine learning point of view, this example illustrates that the Composite method behaves appropriately when $p < n$. In this example, the GGMselect solution seems already acceptable, and the Composite procedure does not deviate too much from it.

To summarise these experimental studies, Section 3.5.1 showed the quantitative and qualitative improvements made by the Composite method on real data, in the High Dimension Low Sample size setting ($n < p$) the method was designed for. In this Section, with enough data available ($p < n$), hence outside the intended area of application, the qualitative analysis suggests that, running the Composite procedure does not provide additional benefits, but does not cause any loss either.

3.6 Additional Oracle metrics

In this section, we present the results, evaluated with new oracle metrics, of the experiment on synthetic data of section 4.3 "Model selection on a deterministic path with the GGMselect criterion and with the Cross Validated Cross Entropy" of the main paper. The new metrics are the widely used l_2 reconstruction of the True Σ, $\left\|\Sigma - \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right\|_F$, and the oracle nodewise regression l_2 recovery $\left\|\Sigma^{\frac{1}{2}}(I_p - \Theta_{\mathcal{G}}(\underline{X}_{train}))\right\|_F$ (the oracle metric of the GGMselect authors [56]). Figure 3.11 shows the results in terms of performance and sparsity over 1000 simulations. In the top row, the performances are measured with the KL, in the middle row with the matrix l_2 and in the bottom row, with the local regression l_2. The observations wit the new metrics are the same as with the KL: the solution selected by the Out of Sample criterion of the Composite method (shades of red) is better in average (an in some case significantly better) than GGMselect (in green) according to the metric in question. Moreover, these solution are closer in terms of number of edges to the best (in blue) and real (with 35 edges) graphs. We observe that this particularly true when the fraction of data reserved for the *validation* set is large (35% or 40% of the *training* data, the lighter nuances of red in the figure, closest to yellow). Indeed, the graphs selected with a smaller *validation* set (20%) have a larger variance in terms of their size, and in terms of their performances when the metric is not the KL. They are also further away from the real graph in average and have the worst average

performances of the solutions selected by CVCE. They are still better in average than the solutions selected with the GGMSC though.

3.7 Cortex visualisation

We display different perspectives of the graphs proposed by GGMselect (281 edges) and the Composite method ($\sim$ 600 edges) as well as two GLASSO solutions on the Alzheimer's Disease data (ADNI) of Section 5.1 "Experiment on Alzheimer's Disease patients" of the main paper. The first GLASSO solution corresponds to a value of the penalty parameter ρ that gives it a number of edges similar to GGMselect (364 here). This allow to visually compare the sparse GLASSO graphs with GGMselect. The second GLASSO solution is the best encountered on the GLASSO path in terms of Out of Sample KL. It is much larger, with $\sim$ 3500 edges.
As explain in section 5.1 of the main paper, each graph is made of 343 nodes, representing both different areas of the brain and different measured modalities. To visually represent such a complex graph, we choose to display different subsets of its many edges. The following Figure 3.12 to 3.17 correspond to three of these subsets of edges.
The sub-graph containing only the inferred conditional correlations in-between MRI measures are represented on Figure 3.12 for the GGMselect and Composite solutions, and Figure 3.13 for the two GLASSO solutions. The best GLASSO has so many edges that the graph is hard to interpret. The other graphs possess many connections between symmetric areas of the cortex. With the Composite graph having comparatively more intra-hemispheric edges, and the sparse GLASSO comparatively less edges overall, on this part of the graph.
The sub-graph containing only the inferred conditional correlations in-between PET measures are represented on Figure 3.14 for the GGMselect and Composite solutions, and Figure 3.15 for the two GLASSO solutions. The observations are mostly the same, but the sparse GLASSO has many more edges than between the MRI measure. It has even more edges than GGMselect and Composite on this part of the graph.
Finally on Figure 3.16 and 3.17, we represent the inter-modality edges between MRI (red) and PET (yellow) nodes. We observe that neither GGMselect nor the sparse GLASSO put any edges in this part of the graph, both methods hence excluding inter-modality conditional correlation from the estimated model. On the other hand, the best GLASSO solution has as many edges as the Composite method on this sub-part of the graph, despite having a considerably larger amount of edges everywhere else.

3.8 GLASSO solutions on the nephrology patients

In this section, we display, see Figure 3.18, the path of GLASSO solutions applied to the neprhology patients of Section 5.2 "Experiments on neprhology patients" of the main paper. The bottom left matrix, $\rho = 0.8$, is the one compared to the GGMselect and Composite solution in the main paper. It was chosen as a representative because the other GLASSO solution have many more edges than GGMselect and Composite. This decision allowed us to compare the first edges selected by GLASSO on its path of solutions as ρ decreases with the edges highlighted by GGMselect. The medical analysis in the main paper concluded that the GGMselect edges were much more consistent with the domain knowledge. The other GLASSO solutions displayed here feature some of the important edges missed by the first, sparse, solution, but these edges appear later in the path, alongside many other a priory irrelevant edges.

3.9 Conclusion

When it came to inferring conditional covariance graphs from a small number of observations, we were dissatisfied with the state of the art GGM methods. In this paper, we quantified the shortcomings in terms of goodness of fit, distribution reconstruction and interpretability of the local approach of [115]

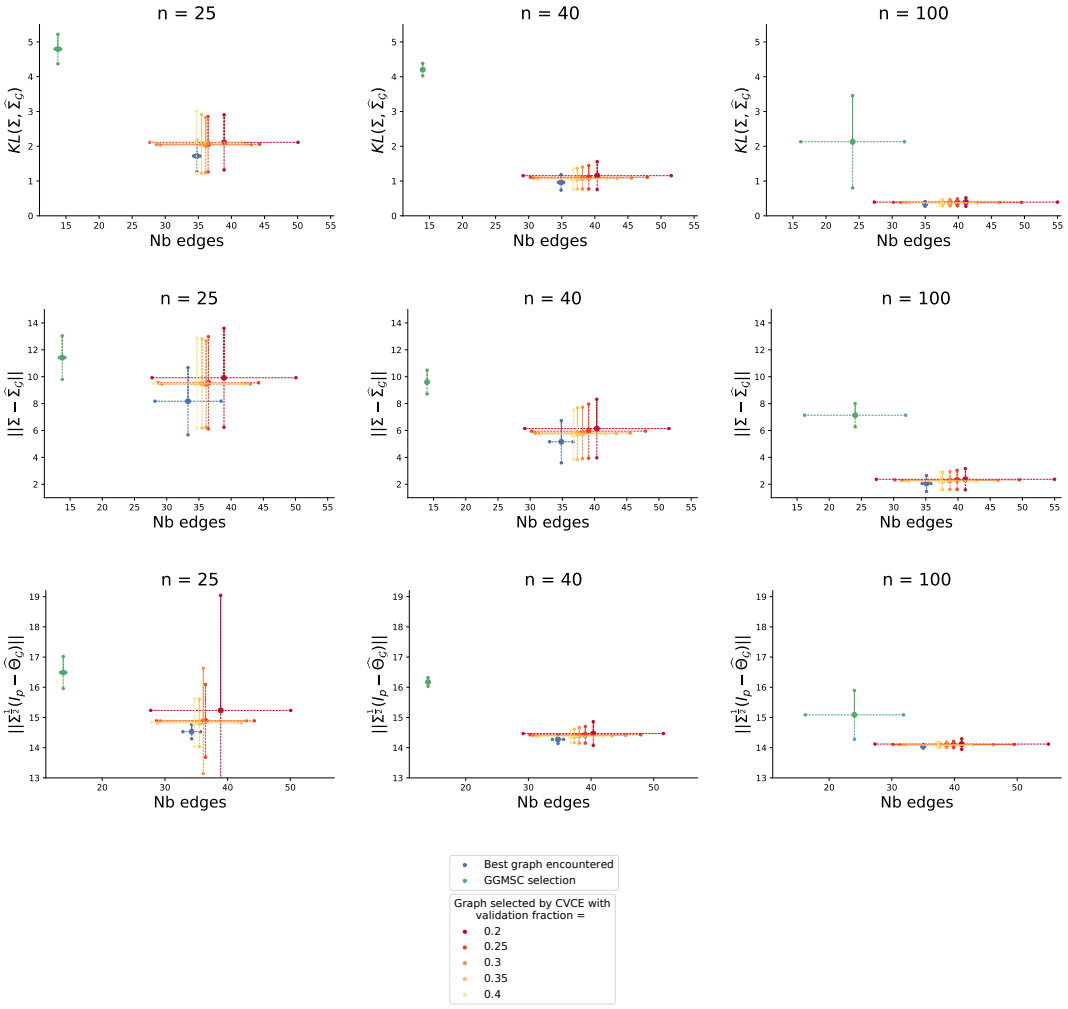


Figure 3.11: Results of the experiment of section 4.3 of the main paper evaluated with the KL (top), as well as two alternative metrics: the l_2 recovery of Σ (middle) and the Oracle metric of GGMselect, the nodewise l_2 recovery (bottom). The behaviour and conclusion are the same as with the KL. We also observe that when the validation set is too small (only 20% of the training set), there is a lot of variance on the selection by CVCE, and the performances suffer.

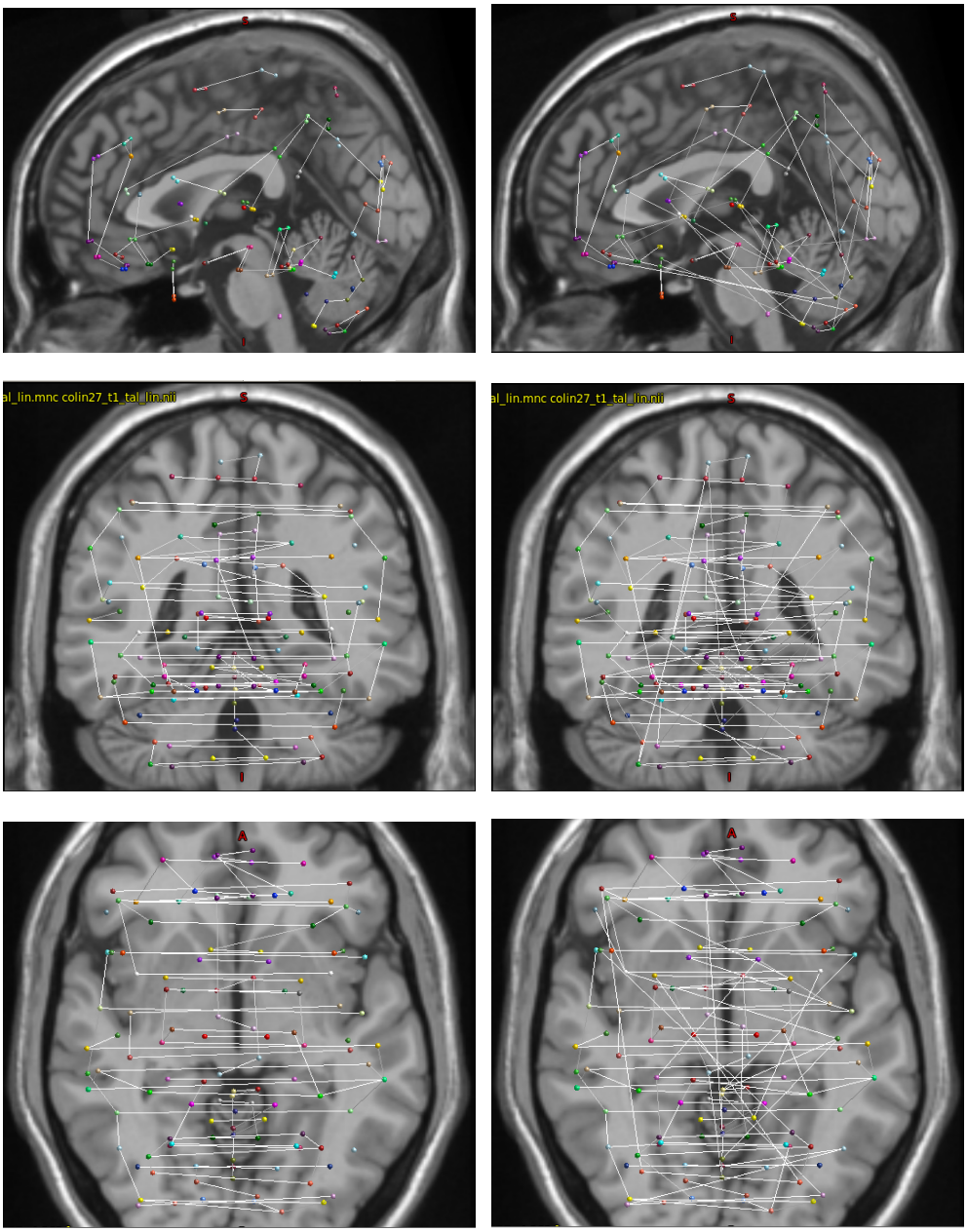


Figure 3.12: **I.1 Sub-graph in-between MRI measures from the GGMselect (left) and Composite (right) solutions.** The GGMselect full graph has 281 edges in total, and the Composite around 600. The sagittal, frontal and transverse views of the Cortex are displayed. With both methods, many of the connections are inter-hemispheric, between symmetrical areas. Although the Composite solution proposes a number of new, intra-hemispheric, edges.

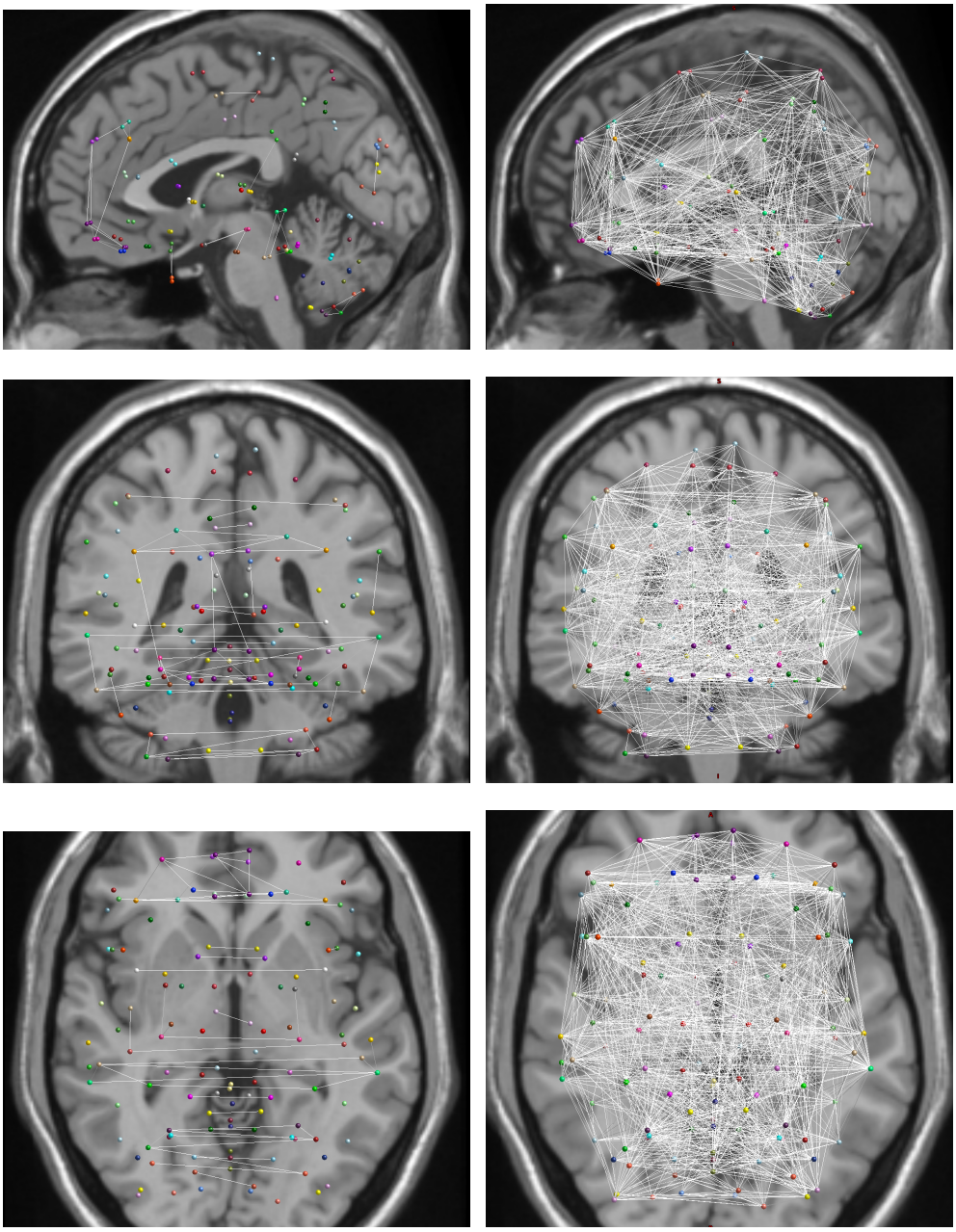


Figure 3.13: **I.2 Sub-graph in-between MRI measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** The full graph of the sparse GLASSO solution has 364 edges in total. A number chosen to be close to the GGMselect solution. The best OSL GLASSO solution features 3546 edges in total. The sparse solution features mostly inter-hemispheric connections between symmetrical areas. The larger solution, with better performances, is mostly unreadable.

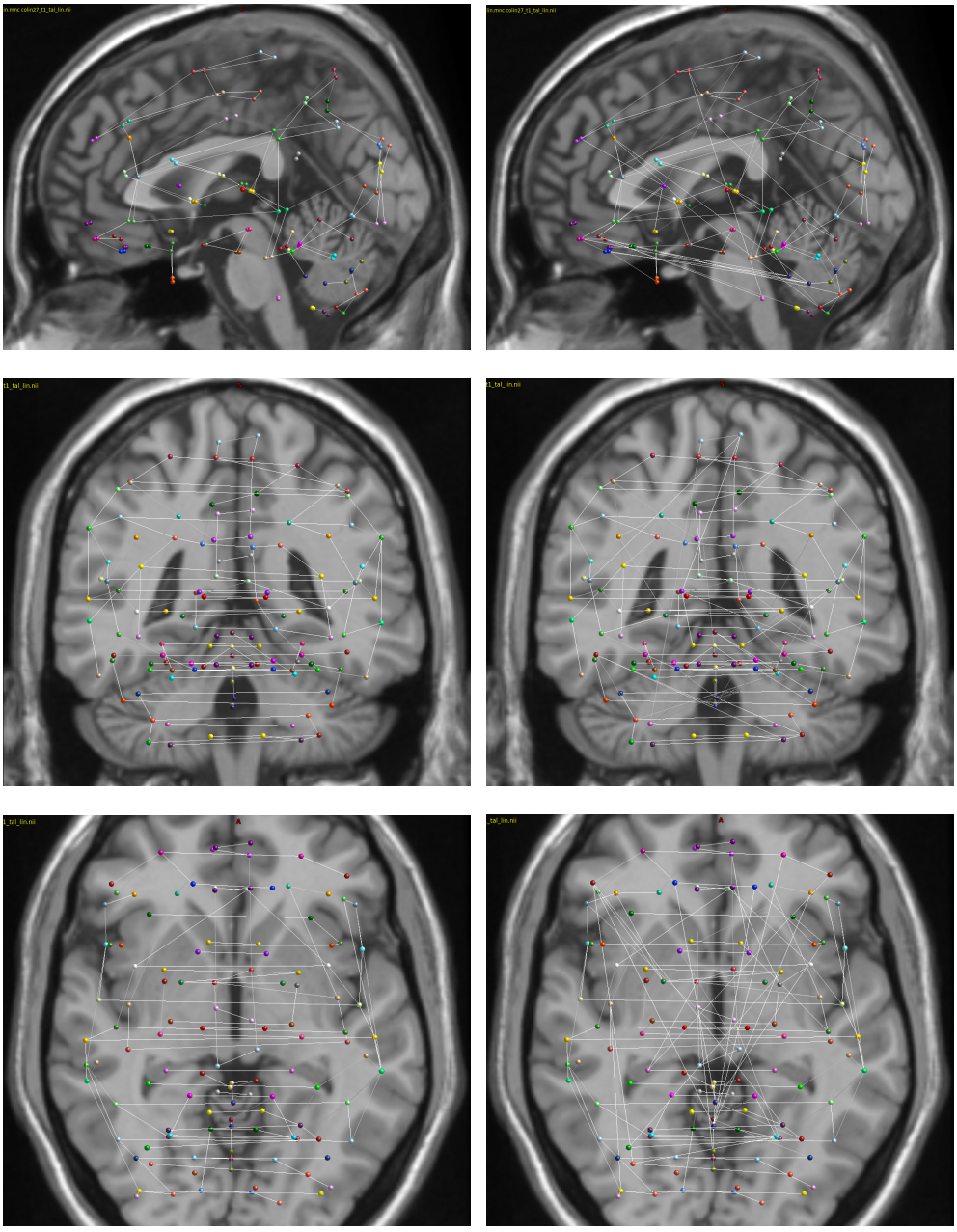


Figure 3.14: **II.1 Sub-graph in-between PET measures from the GGMselect (left) and Composite (right) solutions.** The observations are similar to the MRI sub-graph: many inter-hemispheric connections between symmetric regions, with new intra-hemispheric connections in the Composite solution.

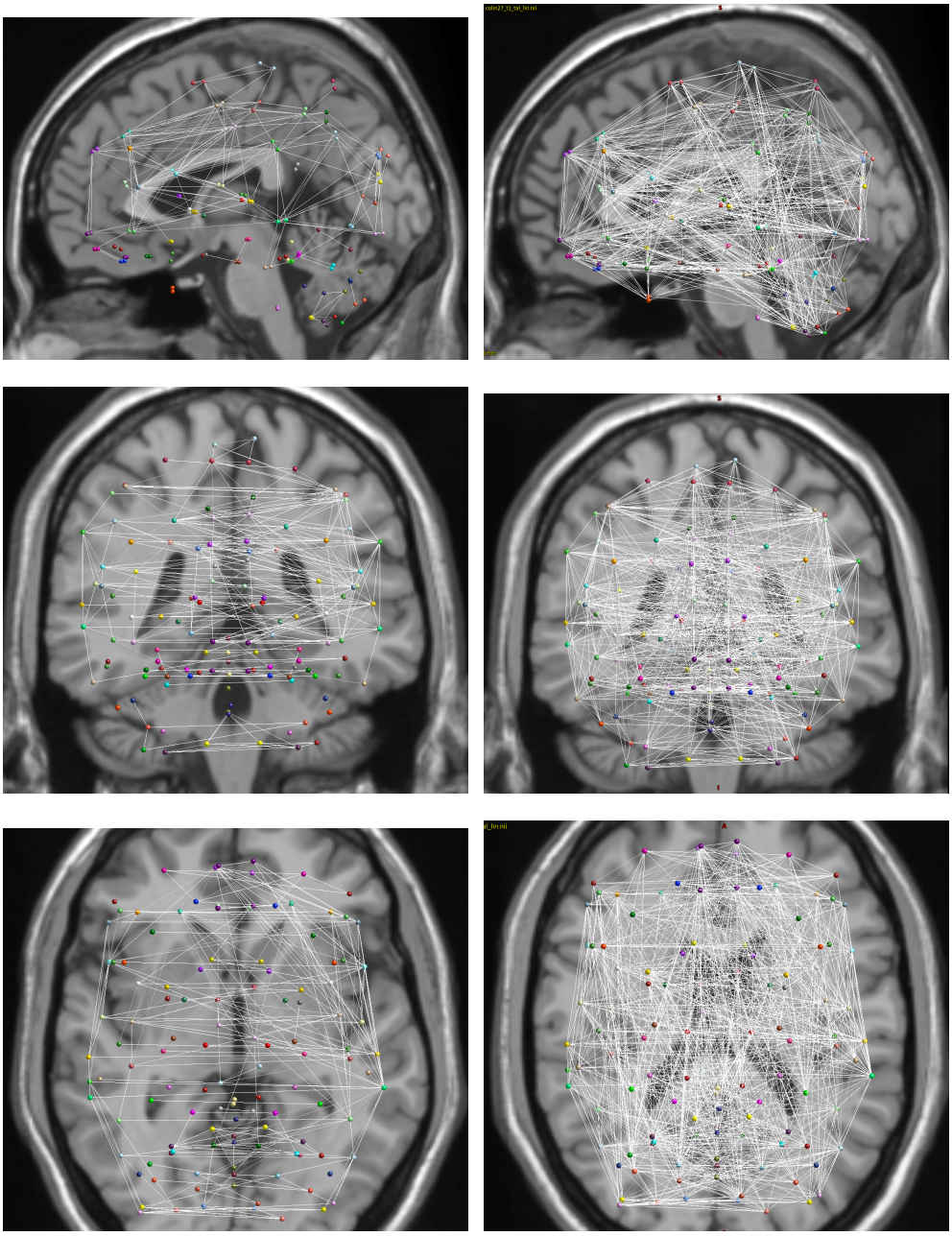


Figure 3.15: **II.2 Sub-graph in-between PET measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** This figure reveals that the sparse GLASSO solution possesses more edges in-between PET measures than in-between MRI measures. The larger GLASSO is still mostly unreadable.

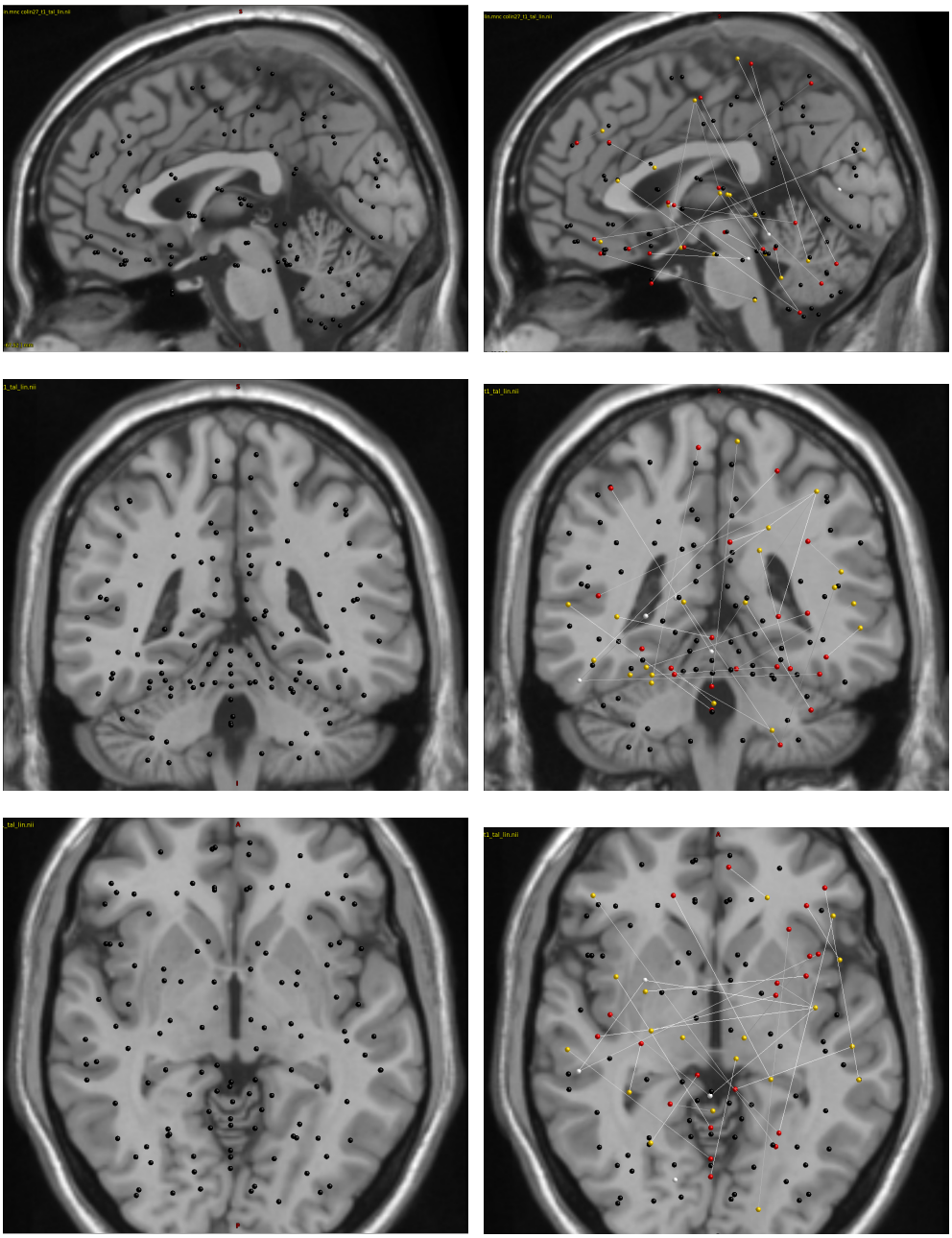


Figure 3.16: **III.1 Sub-graph of the edges between MRI (red) and PET (yellow) measures from the GGMselect (left) and Composite (right) solutions.** Unlike the Composite method, the GGMselect graph proposes no edge between any MRI and PET measures.

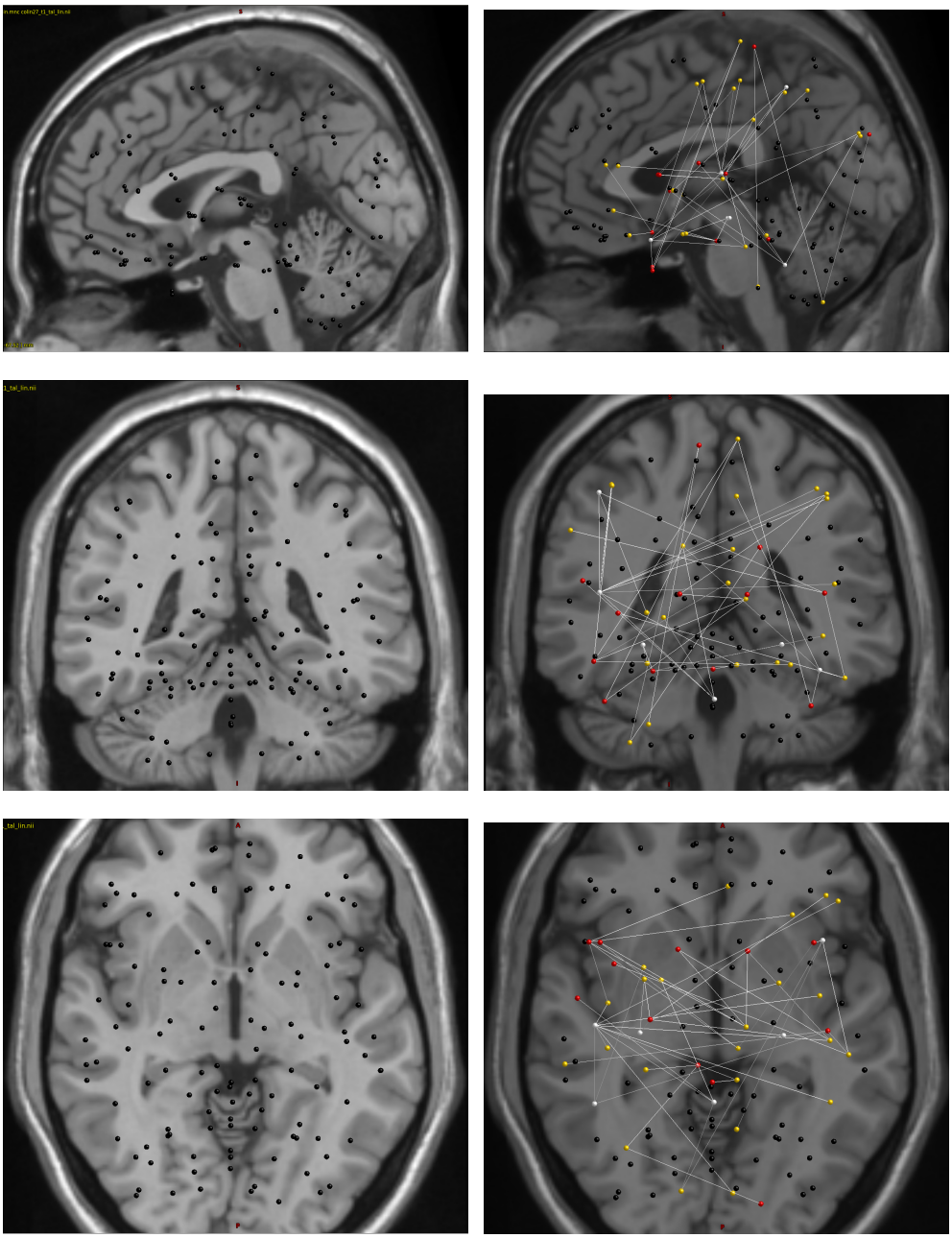


Figure 3.17: **III.2 Sub-graph of the edges between MRI (red) and PET (yellow) measures from a sparse GLASSO solution (left) and the best Out of Sample GLASSO solution in KL (right).** Like with GGMselect, there is no edge in this part of the sparse GLASSO graph. The larger GLASSO solution has edges in this sub-part of the graph. Unlike in the other regions however, the large GLASSO features a number of edges similar to the corresponding Composite method sub-graph.

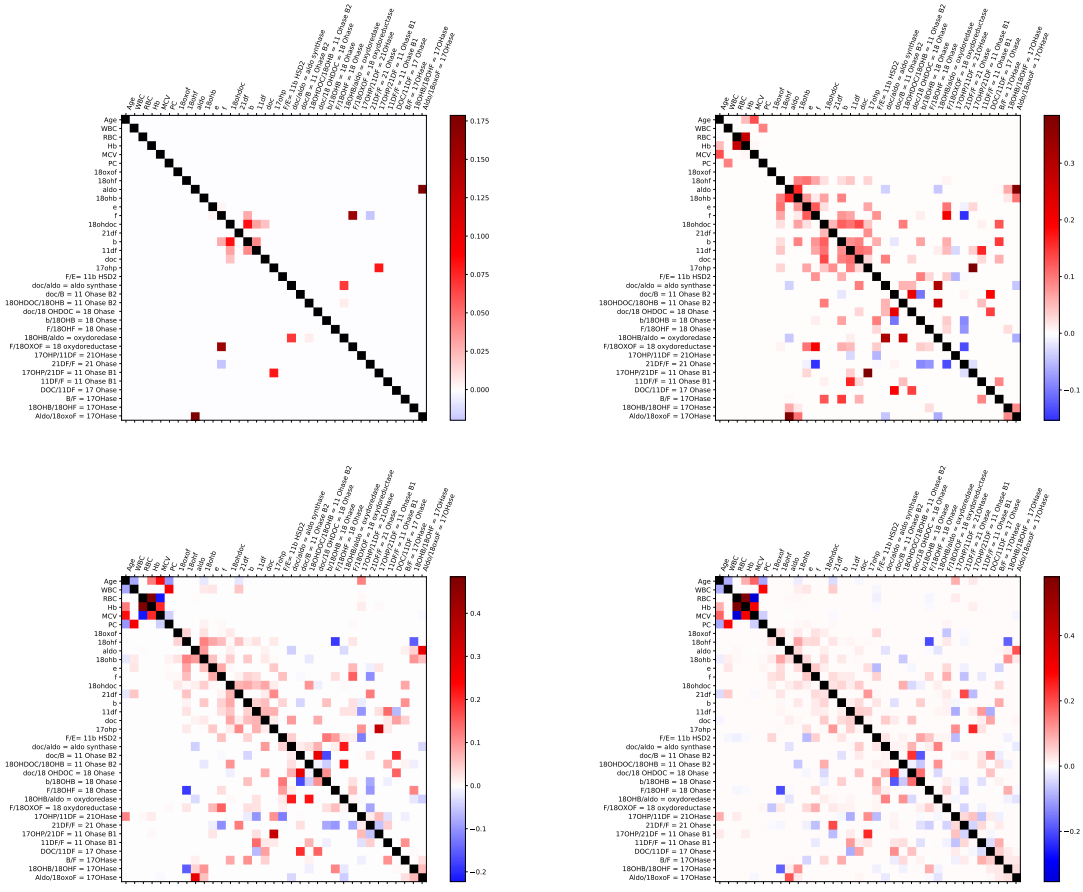


Figure 3.18: Matrices of the pairwise conditional correlations corresponding to the GLASSO solutions applied to the nephrolgy patients. The value considered for the parameter ρ are 0.8 (top left), 0.4 (top right), 0.2 (bottom left) and 0.1 (bottm right)

and the global optimisation problem of [13, 179]. *We proposed a method composed of a structure learning algorithm coupled with model selection criterion.* In the latter, the structure learning steps are a variation of the parallel nodewise linear regressions of [115] and the model selection steps guided by out of sample versions of the likelihood optimised in [179] and [13]. The validity of our method was demonstrated on synthetic and real data when $n < p$. Quantitatively, it consistently reached consequently lower KL divergences and better sparsistency than the aforementioned state of the art paradigms. A qualitative analysis on a neurological data set of real data, revealed that it better recovered the known dynamics of the field. An additional real data experiment, with $p < n$, suggested that the method did not cause any loss when used outside the intended scope of application. In the future, optimising the numerical scheme will allow us to make further quantitative improvements. Such as lower execution times and better performances with less reliance on the initialisation.

3.10 Proofs of the main results

3.10.1 Basic Cross Entropy calculus for Gaussian vectors

In this Section, we offer details and commentary on the Cross Entropy manipulation with normal distributions and prove (3.2) and (3.3).
The formula of the Cross Entropy $H(p,q)$ is given by:

$$H(p,q) := -\mathbb{E}_p\left[\log q(X)\right] = \int_x -p(x)ln(q(x))\mu(\mathrm{d}x)\,.$$

The likelihood p_θ of a parametric distribution f_θ with iid observations $(X^{(1)},...,X^{(n)})$ is given by:

$$p_\theta(X^{(1)},...,X^{(n)}) = \prod_{i=1}^{n} f_\theta(X^{(i)})\,.$$

Let $\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x=X^{(i)}}$ be the empirical distribution of the sample $(X^{(1)},...,X^{(n)})$, we see the connection between CE and likelihood:

$$H\left(\hat{f}_n, f_\theta\right) = -\frac{1}{n}\sum_{i=1}^{n}\log(f_\theta(X_i)) = -\frac{1}{n}\log p_\theta(x_1,...,x_n)\,.$$

Proof of (3.2) *and* (3.3) . In the case of Centered Multivariate Gaussians, let $H(\Sigma_1,\Sigma_2) := H(f_{\Sigma_1}, f_{\Sigma_2})$ and let us omit the constant $\frac{p}{2}ln(2\pi)$ from the calculations:

$$\begin{aligned}
H(\Sigma_1,\Sigma_2) &\equiv \int_X f_{\Sigma_1}(x)\left(-\frac{1}{2}ln(|K_2|) + \frac{1}{2}X^T K_2 X\right)\mathrm{d}X \\
&= -\frac{1}{2}ln(|K_2|) + \frac{1}{2}\int_X f_{\Sigma_1}(x)\left\langle XX^T, K_2\right\rangle \mathrm{d}X \\
&= -\frac{1}{2}ln(|K_2|) + \frac{1}{2}\left\langle \int_X f_{\Sigma_1}(x) XX^T \mathrm{d}X, K_2\right\rangle \\
&= -\frac{1}{2}ln(|K_2|) + \frac{1}{2}\left\langle \Sigma_1, K_2\right\rangle\,.
\end{aligned}$$

In the end, we get (3.2):

$$\boxed{H(\Sigma_1,\Sigma_2) \equiv \frac{1}{2}\left(\left\langle \Sigma_1, K_2\right\rangle - ln(|K_2|)\right)\,.}$$

With the observed data $\underline{X} := (X_1,...,X_n)^T \in \mathbb{R}^{n\times p}$, let $S := \frac{1}{n}\underline{X}\,\underline{X}^T \in S_p^+$, the empirical covariance

matrix. The log likelihood of any centred Gaussian distribution f_{Σ_2} is given by:

$$\begin{aligned} H\left(\hat{f}_n, f_{\Sigma_2}\right) &\equiv \frac{1}{2n}\sum_{i=1}^{n}\left(-ln(|K_2|) + X_i^T K_2 X_i\right) \\ &= -\frac{1}{2}ln(|K_2|) + \left\langle \sum_{i=1}^{n}\frac{X_i X_i^T}{2n}, K_2 \right\rangle \\ &= -\frac{1}{2}ln(|K_2|) + \frac{1}{2}\left\langle S, K_2\right\rangle \ , \end{aligned}$$

where, as in (3.2), we omit the constant term $\frac{p}{2}ln(2\pi)$ from the calculations. In the end, we get (3.3):

$$\boxed{H\left(\hat{f}_n, f_{\Sigma_2}\right) \equiv \frac{1}{2}\left(\left\langle S, K_2\right\rangle - ln(|K_2|)\right) \ .}$$

□

The likelihood $H\left(\hat{f}_n, f_{\Sigma_2}\right)$ follows a similar formula as the Cross Entropy between two normal distributions (3.2). When S defines a non degenerate normal distribution, what we actually have is $H\left(\hat{f}_n, f_{\Sigma_2}\right) = H\left(f_S, f_{\Sigma_2}\right)$. However, when $n < p$, S is singular and the density f_S is not defined. The formula (3.3) still holds though, and we write $H\left(S, \Sigma_2\right) := H\left(\hat{f}_n, f_{\Sigma_2}\right)$ since the formula is the same as (3.2) for $H\left(\Sigma_1, \Sigma_2\right)$.

Remark. When the density f_S does exist, we have equality in the CE $H\left(\hat{f}_n, f_{\Sigma_2}\right) = H\left(f_S, f_{\Sigma_2}\right)$, but not in the Entropies $H\left(\hat{f}_n, \hat{f}_n\right) \neq H\left(f_S, f_S\right)$, as a consequence the KL divergences are different as well: $KL\left(\hat{f}_n, f_{\Sigma_2}\right) \neq KL\left(f_S, f_{\Sigma_2}\right)$. In practice $KL\left(f_S, f_{\Sigma_2}\right) << KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ and $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ will never reach 0, since a normal distribution will tend to be closer to another normal distribution than to an empirical one, this is particularly true with n small and Σ_2 close to S. As a result, $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ offers a poor sense of scale, since the value 0 cannot be used as a reference. For this reason, when we represent $H\left(f_{S_{test}}, f_{\Sigma_2}\right)$ as we do in Figure 3.7, we do not use it under the form of a KL with 0 as its minimum for scale reference - as we do on synthetic data in Figure 3.5 - since the only KL we can compute is the mostly irrelevant $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$.

3.10.2 Preliminary results for the model selection guarantees

To prove the controls we stated in Sections 3.3.2, 3.3.3 and 3.3.4, we need the two following lemmas.

Lemma 3.10.1. *Let $S^{(\lambda)} := S + \lambda I_p$. With $\widehat{K}_{\mathcal{G}} := \widehat{\Sigma}_{\mathcal{G}}^{-1}$, where $\widehat{\Sigma}_{\mathcal{G}}$ is defined as in (3.7), we have:*

$$\forall \mathcal{G} \in \mathcal{M}, \quad \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}}\right\rangle = p\,. \tag{3.17}$$

Proof. Let $\Pi_{\mathcal{G}}$ be the orthogonal projection on the edge set $E_{\mathcal{G}} \cup \{(i,i)\}_{i=1}^{p}$. That is to say, for any matrix $M \in \mathbb{R}^{p\times p}$, $\Pi_{\mathcal{G}}(M)_{i,j} = M_{i,j}\mathbb{1}_{(i,j)\in E_{\mathcal{G}}\cup\{(i,i)\}_{i=1}^{p}}$. A property of the MLE is that $\Pi_{\mathcal{G}}(\widehat{\Sigma}_{\mathcal{G}}) = \Pi_{\mathcal{G}}(S^{(\lambda)})$, i.e. the matrices have the same values on the diagonal and the edge set, see [35]. Additionally, note that, because of the sparsity of $\widehat{K}_{\mathcal{G}}$, for any matrix M, we have $\left\langle M, \widehat{K}_{\mathcal{G}}\right\rangle = \left\langle \Pi_{\mathcal{G}}(M), \widehat{K}_{\mathcal{G}}\right\rangle$. Then:

$$\begin{aligned} \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}}\right\rangle &= \left\langle \Pi_{\mathcal{G}}(S^{(\lambda)}), \widehat{K}_{\mathcal{G}}\right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}}\right\rangle &= \left\langle \Pi_{\mathcal{G}}(\widehat{\Sigma}_{\mathcal{G}}), \widehat{K}_{\mathcal{G}}\right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}}\right\rangle &= \left\langle \widehat{\Sigma}_{\mathcal{G}}, \widehat{K}_{\mathcal{G}}\right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}}\right\rangle &= p\,. \end{aligned}$$

□

Lemma 3.10.2. *With $\widehat{K}_{\mathcal{G}} := \widehat{\Sigma}_{\mathcal{G}}^{-1}$, where $\widehat{\Sigma}_{\mathcal{G}}$ is defined as (3.7), we have:*

$$\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}\,.$$

Proof. We have:

$$\begin{aligned}\left\langle S + \lambda I_p, \widehat{K}_{\mathcal{G}} \right\rangle &= p \\ \left\langle S, \widehat{K}_{\mathcal{G}} \right\rangle + \lambda tr(\widehat{K}_{\mathcal{G}}) &= p \\ tr\left(\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}}\right) + \lambda tr(\widehat{K}_{\mathcal{G}}) &= p\,.\end{aligned}$$

Since $\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}} \in S_p^+$, we have $tr\left(\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}}\right) \geq 0$ and $\lambda tr(\widehat{K}_{\mathcal{G}}) \leq p$, i.e.

$$\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}\,.$$

□

3.10.3 Bounds in expectation for the CVCE solutions

We prove the results of Sections 3.3.2 and 3.3.3.

Proof of (3.11), (3.12), (3.13), *and* (3.14) . We want to control the expected regret

$$e := \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right]\,.$$

First, note that by definition of $\widehat{\mathcal{G}}^*$, we have

$$0 \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\,.$$

So the lower bound:

$$0 \leq e\,,$$

is guaranteed.
From the definition of $\widehat{\mathcal{G}}_{CV}$ (3.9), we get:

$$H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) \leq H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\,.$$

We have for any $\widetilde{\Sigma} \in S_p^{++}$, with $\widetilde{K} := \widetilde{\Sigma}^{-1}$:

$$H\left(S_{val}, \widetilde{\Sigma}\right) = H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{2}\left\langle S_{val} - \Sigma, \widetilde{K}\right\rangle\,.$$

Hence:

$$\begin{aligned}H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) &+ \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle \\ &- \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\,.\end{aligned} \tag{3.18}$$

Since $K_{\widehat{\mathcal{G}}^*}$ is defined from S_{expl} uniquely, and independently of S_{val}, we get

$$\begin{aligned}\mathbb{E}\left[\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle \middle| S_{expl}\right] &= \left\langle \mathbb{E}\left[S_{val} - \Sigma | S_{expl}\right], \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle \\ &= 0\,.\end{aligned} \tag{3.19}$$

From (3.18) and (3.19) we get:

$$\mathbb{E}\left[H\left(\Sigma,\widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right)\right] \leq \mathbb{E}\left[H\left(\Sigma,\widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right] + \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right].$$

Which is exactly the result of Eq. (3.11):

$$\boxed{e \leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right].}$$

As we discussed in Section 3.3.3, to obtain Eq. (3.11), we only used the definitions of $\widehat{\mathcal{G}}_{CV}$ for the upper bound and $\widehat{\mathcal{G}}^*$ for the lower bound. Since we assume nothing on the model family $\mathcal{M}$, those bounds are somewhat optimal in terms of the available information. Additionally, (3.11) is actually independent of how the symmmetric positive matrices $\{\widehat{\Sigma}_{\mathcal{G}}\}_{\mathcal{G}\in\mathcal{M}}$ are defined as long as they are function only of S_{expl}. They do not need to be associated with a different graph each, or with any graph for that matter. They do not need to be solutions of the MLE problem (3.7) and could be for example all the solutions on the path of solution of the l_1-penalised likelihood optimisation problem (3.1).
To get a more explicit control on the CVCE however, we need the assumption that $\widehat{\Sigma}_{\mathcal{G}}$ is the constrained MLE defined in (3.7).
Let $\Sigma_\infty := \max_{i,j}|\Sigma_{ij}|$. We call $E_{\max}$ the union of the maximal edge sets in $\mathcal{M}$, $d_{\max} = |E_{max}| \leq \frac{p(p-1)}{2}$ its cardinal and Π_{max} the orthogonal projection on $E_{\max}\cup\{(i,i)\}_{i=1}^p$. We have:

$$\begin{aligned}
e &\leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\langle \Pi_{\widehat{\mathcal{G}}_{CV}}\left(\Sigma - S_{val}\right), \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\left\|\Pi_{\widehat{\mathcal{G}}_{CV}}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left\|\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\|_F^2\right]^{\frac{1}{2}} \\
&\leq \frac{1}{2}\mathbb{E}\left[\left\|\Pi_{max}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left\|\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\|_*^2\right]^{\frac{1}{2}} \\
&\leq \frac{1}{2}\left(\sum_{i=1}^p \mathbb{E}\left[\left(\Sigma^{ii} - S_{val}^{ii}\right)^2\right]\right. \\
&\qquad \left. + \sum_{(i,j)\in E_{max}} \mathbb{E}\left[\left(\Sigma^{ij} - S_{val}^{ij}\right)^2\right]\right)^{\frac{1}{2}}\frac{p}{\lambda} \\
&\leq \frac{1}{2}\left(\frac{2\Sigma_\infty^2}{n_{val}}\left(p + 2d_{max}\right)\right)^{\frac{1}{2}}\frac{p}{\lambda}.
\end{aligned}$$

From which we finally get the result of (3.12):

$$\boxed{e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}}\frac{\left(p+2d_{max}\right)^{\frac{1}{2}}p}{\sqrt{n_{val}}}.}$$

If $E_{\max}$ is dependent on the *exploration* data - because the graph family $\mathcal{M}$ was built from S_{expl} for instance - we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\Pi_{max}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}} \\
&\quad = \left(\sum_{i=1}^{p}\mathbb{E}\left[\left(\Sigma^{ii} - S_{val}^{ii}\right)^2\right]\right. \\
&\qquad \left. + \mathbb{E}\left[\sum_{i,j\in E_{max}}\mathbb{E}\left[\left(\Sigma^{ij} - S_{val}^{ij}\right)^2 \middle| S_{expl}\right]\right]\right)^{\frac{1}{2}} \\
&\quad \leq \left(\frac{2\Sigma_\infty^2}{n_{val}}\left(p + 2\mathbb{E}\left[d_{max}\right]\right)\right)^{\frac{1}{2}}.
\end{aligned}
$$

We get the control (3.13), the same as (3.12) but with an additional expectation term:

$$
\boxed{e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}}\frac{\left(p + 2\mathbb{E}\left[d_{max}\right]\right)^{\frac{1}{2}}p}{\sqrt{n_{val}}}.}
$$

In order to prove (3.14), we start by showing how the regret is bounded by operator norm $\|\Sigma - S_{val}\|_2$. By tracial matrix Holder inequality:

$$
\begin{aligned}
\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle &\leq \|\Sigma - S_{val}\|_2 \left\|\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\|_* \\
&= \|\Sigma - S_{val}\|_2 \, tr\left(\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right) \\
&\leq \frac{\|\Sigma - S_{val}\|_2}{\lambda}p\,.
\end{aligned}
$$

Then, using (3.11), we get:

$$
e \leq \mathbb{E}\left[\|\Sigma - S_{val}\|_2\right]\frac{p}{2\lambda}\,. \tag{3.20}
$$

To prove (3.14), we first recall Theorem 4 of [85]:

Theorem 4 of [85]. *Let $X_1, X_2, ..., X_n$ be i.i.d. weakly square integrable centered random vectors in a separable Banach space with norm $\|.\|$ and Σ be their covariance operator. If X is Gaussian, then there exist an absolute constant c, independent of the problem, such that:*

$$
\mathbb{E}\left[\left\|\widehat{\Sigma} - \Sigma\right\|\right] \leq c\,\|\Sigma\|\,max\left(\sqrt{\frac{\mathbb{E}\left[\|X\|\right]^2}{n\,\|\Sigma\|}}, \frac{\mathbb{E}\left[\|X\|\right]^2}{n\,\|\Sigma\|}\right), \tag{3.21}
$$

where $\|.\|$ for operators denotes the operator norm associated with the vector norm $\|.\|$, that is to say:

$$
\|\Sigma\| = \sup_{\|u\|=1}\|\Sigma u\|\,.
$$

In our case, $X \sim \mathcal{N}\left(0_p, \Sigma\right)$ is a Gaussian vector in the Banach space $\mathbb{R}^p$, with the euclidean norm $\|X\|_2$, that verifies the integrability properties of the Theorem and whose covariance operator is the covariance matrix Σ. Hence the theorem can be applied. The operator norm for a symmetric positive matrix Σ associated with the euclidean norm is also called the spectral norm, since it corresponds to the highest eigenvalue: $\|\Sigma\|_2 = \lambda_{max}(\Sigma)$.
For a Gaussian vector: $Z \sim \mathcal{N}\left(0_p, I_p\right)$, we have:

$$
\mathbb{E}\left[\|Z\|_2\right] \leq \sqrt{p}.
$$

Since $K^{\frac{1}{2}}X \sim \mathcal{N}\left(0_p, I_p\right)$, and

$$
\begin{aligned}
\|X\|_2 &= \left\|\Sigma^{\frac{1}{2}}K^{\frac{1}{2}}X\right\|_2 \\
&\leq \left\|\Sigma^{\frac{1}{2}}\right\|_2\left\|K^{\frac{1}{2}}X\right\|_2,
\end{aligned}
$$

we have:

$$\mathbb{E}\left[\|X\|_2\right] \leq \left\|\Sigma^{\frac{1}{2}}\right\|_2 \sqrt{p}\,.$$

Since $\|\Sigma\|_2 = \lambda_{max}(\Sigma)$, we have by definition, $\left\|\Sigma^{\frac{1}{2}}\right\|_2 = \|\Sigma\|_2^{\frac{1}{2}}$. In the end, when we apply (3.21) to our case, we get:

$$\mathbb{E}\left[\|S_{val} - \Sigma\|_2\right] \leq c\lambda_{max}(\Sigma) max\left(\sqrt{\frac{p}{n_{val}}}, \frac{p}{n_{val}}\right). \tag{3.22}$$

We apply this concentration result on (3.20) to obtain (3.14):

$$\boxed{e \leq c\frac{\lambda_{max}(\Sigma)}{\lambda} p\left(\sqrt{\frac{p}{n_{val}}} \vee \frac{p}{n_{val}}\right).}$$

□

3.10.4 Bounds in probability for the CVCE solutions

We prove the results of Section 3.3.4.

Proof of (3.15) *and* (3.16). We want to lower bound the probability that the regret is small: $P := \mathbb{P}\left(\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\right)$. The concentration dynamic driving the results comes from the convergence of random Wishart matrix S_{val} towards its average Σ, which is made stronger by the number of observations n_{val} in the *validation* set. Since:

$$\begin{aligned}\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq & \\ & \left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right)\right| \\ & + \left|H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right|,\end{aligned}$$

then

$$\begin{aligned}\forall \mathcal{G} \in \mathcal{M},\ \left|H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}\right)\right| \leq \frac{\delta}{2} \\ \implies \left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\,.\end{aligned}$$

Since:

$$H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}\right) = \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle,$$

then

$$\begin{aligned}\forall \mathcal{G} \in \mathcal{M},\ \left|\left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle\right| \leq \delta \\ \implies \left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\,.\end{aligned} \tag{3.23}$$

From the logical implication (3.23), we can take two path to derive two different bounds: one with a more general expression, and a more precise one taking into consideration the sparsity of the models. For the first one, note that $S_{val} = \Sigma^{\frac{1}{2}} W \Sigma^{\frac{1}{2}}$ where $n_{val} W \sim \mathcal{W}_p\left(I_p, n_{val}\right) p$ is a standard Wishart matrix. Then we have:

$$\begin{aligned}\forall \mathcal{G},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle &= \left\langle W - I_p, \Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\rangle \\ &\leq \|W - I_p\|_F \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F \\ &\leq \|W - I_p\|_F \max_{\mathcal{G} \in \mathcal{M}} \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F.\end{aligned}$$

We plug this result into (3.23) to obtain:

$$\begin{aligned}
&\|W - I_p\|_F \max_{\mathcal{G}\in\mathcal{M}} \left\|\Sigma^{-\frac{1}{2}}\widehat{K}_{\mathcal{G}}\Sigma^{-\frac{1}{2}}\right\|_F \leq \delta \\
\Longrightarrow\ &\forall \mathcal{G}\in\mathcal{M},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle \leq \delta \\
\Longrightarrow\ &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\,.
\end{aligned}$$

We end up with the control (3.15) by taking the probability in the previous expression:

$$\boxed{P \geq \mathbb{P}\left(\|W - I_p\|_F \leq \frac{\delta}{\max\limits_{\mathcal{G}\in\mathcal{M}} \left\|\Sigma^{-\frac{1}{2}}\widehat{K}_{\mathcal{G}}\Sigma^{-\frac{1}{2}}\right\|_F}\right)\,.}$$

For the second result, let $\Pi_{\mathcal{G}}$ and Π_{max} be the orthogonal projections on the edge sets $E_{\mathcal{G}}\cup\{(i,i)\}_{i=1}^p$ and $E_{max}\cup\{(i,i)\}_{i=1}^p$ respectively. We have:

$$\begin{aligned}
\forall \mathcal{G},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle &= \left\langle \Pi_{\mathcal{G}}(S_{val} - \Sigma), \widehat{K}_{\mathcal{G}}\right\rangle \\
&\leq \|\Pi_{\mathcal{G}}(S_{val} - \Sigma)\|_F \left\|\widehat{K}_{\mathcal{G}}\right\|_F \\
&\leq \|\Pi_{max}(S_{val} - \Sigma)\|_F \max_{\mathcal{G}\in\mathcal{M}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F\,.
\end{aligned}$$

Hence we get, from (3.23), the logical implication:

$$\begin{aligned}
&\|\Pi_{max}(S_{val} - \Sigma)\|_F \max_{\mathcal{G}\in\mathcal{M}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F \leq \delta \\
\Longrightarrow\ &\forall \mathcal{G},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}}\right\rangle \leq \delta \\
\Longrightarrow\ &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\,.
\end{aligned}$$

From which we get the control (3.16) by taking the probability of the events:

$$\boxed{P \geq \mathbb{P}\left(\|\Pi_{max}(S_{val} - \Sigma)\|_F \leq \frac{\delta}{\max\limits_{\mathcal{G}\in\mathcal{M}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F}\right)\,.}$$

We underline that we obtain the two controls (3.15) and (3.16) directly from logical implications. Hence, they remain true when every probability is taken conditionally to any random variable, for instance the *exploration* data set, or the sufficient statistic built from it: S_{expl}. □

Remark. Since $\forall \mathcal{G}\in\mathcal{M},\ \left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}$, both $\max\limits_{\mathcal{G}\in\mathcal{M}} \left\|\Sigma^{-\frac{1}{2}}\widehat{K}_{\mathcal{G}}\Sigma^{-\frac{1}{2}}\right\|_F$ and $\max\limits_{\mathcal{G}\in\mathcal{M}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F$ are bounded random variables. They depend only on the *exploration* empirical covariance S_{expl} and can be seen as constants of the problem if working conditionally to the *exploration* set. Likewise, Π_{max} is a deterministic function conditionally to S_{expl}.

Chapter 4

Model selection without Cross Validation for chordal graphs

4.1 Introduction

Dependency networks are a prominent tool for the representation and interpretation of many data types as, for example, gene regulation, [56], interactions between different regions of the cortex, [17] or social interaction in a large population. In those examples, the number of observations n is often small when compared to the number of vertices p in the network. A common and interesting dynamic to represent as a network is the conditional correlation structure. In a conditional correlation network, two vertices are connected by an edge if and only if the associated random variables are correlated conditionally to all others. This notion of correlation can be refereed to as "direct" or "explicit" correlations, they are usually more insightful than regular correlations. Indeed, any two real life phenomena, like the atrophy in two separate areas of the brain or two locations of bird migration, are very likely to be correlated. The reason being that these kind of phenomena are smooth, hence there often exists a path of highly correlated neighbouring vertices linking any two points of the graph. As a consequence the regular correlation network is most of the time fully connected and uninteresting. On the other hand, there is no conditional correlation between two vertices if their co-variations are entirely explained by intermediary variables. Only the direct, explicit interactions remain as edges in a conditional correlation network, which leads to a richer and less trivial analysis.

A Gaussian Graphical Model (GGM) is a network whose values on the p vertices follow a Centred Multivariate Normal distribution in $\mathbb{R}^p$: $X \sim \mathcal{N}(0_p, \Sigma)$. Under this assumption, the conditional correlations between the components of X have the sparsity of the inverse of the unknown covariance matrix $K := \Sigma^{-1}$. Dempster [35] introduced the famous Covariance Selection procedure, which infers from data a conditional correlation network through a sparse approximation of the precision matrix K. Many subsequent works developed methods to estimate a sparse precision matrix [13, 179], or even directly a sparse graph [41, 115] within the Gaussian model. For methods that return an unweighted conditional correlation graph, the corresponding precision matrix can be estimated through a graph constrained Maximum Likelihood Estimation (MLE) problem, see [91].

Graphs with no cycle of length longer than three are called "chordal", or "decomposable". Some authors have proposed GGM inference methods where the estimated graph is constrained to be chordal, see [38, 57, 72], or [14]. With such graphs, there is an explicit formula for the constrained MLE, allowing for better theoretical results and more efficient computations.

In a model selection context, one is presented with a family of candidate graphs among which a choice must be made. Our goal is to design model selection criterion, backed by theoretical guarantees, to make the best possible choice among any proposed chordal graphs.

In our paradigm, the oracle best graph is the one whose associated precision matrix $\widehat{K}$ by constrained MLE defines the candidate distribution $\mathcal{N}\left(0_p, \widehat{K}^{-1}\right)$ that best fits the true distribution $\mathcal{N}(0_p, \Sigma)$. As argued in Chapter 3, local metrics, like coefficient-wise recovery, do not explicitly assess this behaviour as they consider each node independently to estimate the whole network. On the other hand, the Cross Entropy (CE) - or equivalently the Kullback-Leibler divergence (KL) - is a global metric designed to quantify the proximity between distributions. This is once again the metric we adopt to evaluate the adequacy between any covariance matrix and the real one, the metric to minimise to find the best approximation among the propositions. In the end, our objective in terms of model selection is to find the graph that achieves the lowest KL with regards to the real distribution.

In the following, we find an explicit formula for an unbiased estimator of the CE with regards to the true distribution. With it, we define the Unbiased Chordal Explicit Estimator (UCEE), a new selection criterion for chordal graphs. Unlike the CVCE of Chapter 3, this criterion does not need to split the data set into a training and a validation set. In addition to using the data to its fullest potential, this also alleviates the need to tune the hyper-parameter that is the size of the train/test split. We prove non asymptotic bounds on the performances of the model selected by the UCEE. We compare both criteria empirically and demonstrate that the UCEE selects better models than CVCE on synthetic and real data. On the hippocampus data we use, we are able to recover graphs that take into consideration the long distance conditional correlations between the deformations.

4.2 Presentation of the problem, the tools and the approach

In this section we introduce step by step the stakes as well as the core notations of this Chapter, as well as many properties and remarks about the Cross Entropy that were not included in Chapter 3.

4.2.1 Model selection with Cross Entropy in Gaussian Graphical Models

Let $p \in \mathbb{N}^*$, $[\![1, p]\!]$ is the set of integers between 1 and p inclusively. We call S_p^+ and S_p^{++} the sets of positive semidefinite and definite matrices in $\mathbb{R}^{p\times p}$ respectively. We refer to matrices in S_p^{++} as "covariance matrices". We assume we have a covariance matrix $\Sigma \in S_p^{++}$ and study X, the centred Gaussian random variable in $\mathbb{R}^p$ with covariance Σ. We note: $X \sim \mathcal{N}(0_p, \Sigma)$, where 0_p is the vector of zeros of size p. The Gaussian vector X can be written component-wise as:

$$X = (X^{(1)}, ..., X^{(p)})^T \in \mathbb{R}^p \,.$$

We note $X^{-(ij)} = \left\{X^{(k)}\right\}_{k\in[\![1,p]\!], k\neq i,j}$ the vectors of all the components of X excluding $X^{(i)}$ and $X^{(j)}$. We call "conditional correlation" between $X^{(i)}$ and $X^{(j)}$ their correlation conditionally to all other components: $corr(X^{(i)}, X^{(j)}|X^{-(ij)})$. The conditional correlations between the component of a centred Gaussian vector $X \sim \mathcal{N}(0_p, \Sigma)$ are directly connected to the matrix $K := \Sigma^{-1}$. We call K the "inverse covariance matrix", it is also sometimes referred to as the precision matrix or the concentration matrix. We have the explicit relation:

$$corr(X^{(i)}, X^{(j)}|X^{-(ij)}) = \frac{-K_{ij}}{K_{ii}K_{jj}} \,.$$

This means in particular that K and the conditional correlation structure of X have the same sparsity:

$$corr(X^{(i)}, X^{(j)}|X^{-(ij)}) = 0 \iff K_{ij} = 0 \,.$$

Hence, if one takes interest in the conditional correlation network between the components of a random vector X, modelling this X as multivariate Gaussian $\mathcal{N}(0_p, \Sigma)$ allows to read the pursued structure directly from $K = \Sigma^{-1}$. This is why the Gaussian assumption is very adapted and very popular in that context. When K is sparse, we say that Σ is "inverse sparse".

When many of the conditional correlations in a random vector X are null, in other words when the conditional correlation structure is sparse, it can be interesting to represent this structure as a graph $m = (V, E_m)$, with vertex set V and edge set E_m. Traditionally, V is the set of components of X, and there is an edge if E_m if and only if the two corresponding components are conditionally correlated. By convention E_m contains no self loops and is not oriented. If X is Gaussian vector with inverse covariance matrix K, we have:

$$\forall (i,j) \in [\![1,p]\!]^2 \,,\ i \neq j, \quad (i,j) \notin E_m \iff K_{ij} = 0 \,.$$

As a consequence, to any graph m naturally corresponds the set Θ_m of the covariance matrices whose inverse verify the sparsity of the edge set E_m:

$$\Theta_m := \left\{ \widetilde{\Sigma} \in S_p^{++} \middle| \forall i \neq j,\ (i,j) \notin E_m \implies \left(\widetilde{\Sigma}^{-1}\right)_{ij} = 0 \right\} . \tag{4.1}$$

If $\Sigma \in \Theta_m$ and $X \sim \mathcal{N}(0_p, \Sigma)$, then the graph m describes the conditional correlation structure of X. However, two phenomena are rarely truly conditionally uncorrelated. Likewise, no real life covariance matrix Σ is truly inverse sparse. To describe the conditional correlation structure of a random vector X in an insightful and understandable way, we have to propose a good inverse sparse approximation of Σ, where the weakest, more superfluous, conditional correlations are removed and only the most potent ones remain.

To define a best approximation Σ_m of the real Σ whose inverse $K_m := \Sigma_m^{-1}$ corresponds to a given sparse conditional correlation graph m, we need to choose a notion of distance or deviation d and minimise it over the set of candidate matrices Θ_m:

$$\Sigma_m \in \underset{\widetilde{\Sigma} \in \Theta_m}{\operatorname{argmin}}\, d(\Sigma, \widetilde{\Sigma}) \,. \tag{4.2}$$

We want the solution Σ_m of the optimisation problem (4.2) to be the matrix of Θ_m that best reproduces the behaviour of the real Σ. That is to say that we want the distribution $\mathcal{N}(0_p, \Sigma_m)$ to be as close to $\mathcal{N}(0_p, \Sigma)$ as possible. A natural metric to express proximity between distributions is the Cross Entropy (CE) - or equivalently the Kullback-Leibler (KL) divergence. The CE $H(p, q)$ measures how well distribution q reproduces p. As a consequence, the CE is the adequacy metric we want to optimise. Since Σ is invertible, a result from [35] shows that when Σ_m is defined as follows, it always exists and is unique:

$$\Sigma_m := \underset{\widetilde{\Sigma} \in \Theta_m}{\operatorname{argmin}}\, H\left(f_\Sigma, f_{\widetilde{\Sigma}}\right) \,, \tag{4.3}$$

where f_Σ is the probability density function (pdf) of $\mathcal{N}(0_p, \Sigma)$.

Now that we can define the best representative Σ_m for any graph m, the question of which graph is the best arises. By definition of the problem, the fully connected graph, whose representative is Σ, is the best, reproducing perfectly the behaviour of Σ. We always have $H(\Sigma, \Sigma) \leq H(\Sigma, \Sigma_m)$. More generally, for any two graphs m_1 and m_2 with a relation of inclusion: $E_{m_1} \subset E_{m_2}$, we have $\Theta_{m_1} \subset \Theta_{m_2}$. This means that the problem (4.3) is less constrained for m_2 than for m_1, as a consequence: $H(\Sigma, \Sigma_{m_2}) \leq H(\Sigma, \Sigma_{m_1})$. In other words, more connected graphs are always better since they correspond to less constrained optimisation problems.
As a consequence, defining the best graph m as the one producing the best matrix Σ_m in terms of CE is quite meaningless. This all comes from the fact that the matrices Σ_m are directly computed from Σ. With the full knowledge of Σ, better connected graphs always correspond to matrices resembling more closely Σ and there is no advantage to being sparse. In a practical scenario however, Σ is unknown and must be inferred from observations. In that case sparse estimates can perform better than more connected ones, especially when the number of observations is low. A balance needs to be found between having enough edges and having too many, and the best edges must be kept in priority. This is the framework we want to work with and the problem we want to solve.

We assume that we observe an independent identically distributed (iid) sample of size n drawn from the distribution $\mathcal{N}(0_p, \Sigma)$: $\underline{X} = (X_1, ..., X_n)^T \in \mathbb{R}^{n\times p}$. We call $S := \frac{1}{n}\sum_{k=1}^n X_k X_k^T = \frac{1}{n}\underline{X}^T\underline{X} \in S_p^+$ the empirical covariance matrix. If $n < p$, then, with probability 1, the rank of S is n, which implies that $S \notin S_p^{++}$. Since $S_{ij} = \frac{1}{n}\sum_{k=1}^n X_k^{(i)} X_k^{(j)}$, then, $\mathbb{E}[S_{ij}] = Cov(X^{(i)}X^{(j)}) = \Sigma_{ij}$ and $S_{ij} \underset{n\to\infty}{\to} \Sigma_{ij}$. As a result, S is an unbiased, convergent estimator of Σ. By definition, the matrix nS follows a centred Wishart distribution with n degrees of freedom and scale matrix Σ. We note $nS \sim \mathcal{W}_p(\Sigma, n)\, p$. In the scenario where empirical data is all there is at our disposal, then we have no knowledge of the real matrix Σ or of the pdf f_Σ. Hence Σ_m defined in Eq. (4.3) are inaccessible. To a graph m, we can now only associate the matrix $\widehat{\Sigma}_m$ that best fits the empirical distribution $\hat{f}_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x=X_i}$:

$$\widehat{\Sigma}_m := \underset{\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right). \tag{4.4}$$

We remark that $H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right) = -\frac{1}{n}\sum_{i=1}^n log\left(f_{\widetilde{\Sigma}}(X_i)\right)$ is the negative log-likelihood of the observations $(X_1, ..., X_n)$ under the Gaussian model of parameter matrix $\widetilde{\Sigma}$. In other words, the problem of minimising the CE with relation to an empirical distribution is equivalent to the problem of maximising the likelihood of the corresponding observations.

In order to simplify the notations, we first notice that, for any two covariance matrices Σ_1 and Σ_2, the CE between the pdf of $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ can be expressed as:

$$H(f_{\Sigma_1}, f_{\Sigma_2}) = \frac{1}{2}\big(\langle\Sigma_1, K_2\rangle - log(|K_2|) + p\, log(2\pi)\big),$$

where $\langle\Sigma_1, \Sigma_2\rangle := tr(\Sigma_1\Sigma_2)$ is the Frobenius scalar product on the matrix space. Similarly, the negative log-likelihood, which is the cross entropy between a normal distribution and the empirical distribution, can be written:

$$H\left(\hat{f}_n, f_{\Sigma_2}\right) = \frac{1}{2}\big(\langle S, K_2\rangle - log(|K_2|) + p\, log(2\pi)\big).$$

Since the formulas are the same and only dependent in the covariance matrices, we adopt a unified notation in Eq. (4.5). We also removed the additive constant.

$$\begin{aligned} H(\Sigma_1, \Sigma_2) &:= \frac{1}{2}\big(\langle\Sigma_1, K_2\rangle - log(|K_2|)\big), \\ H(S, \Sigma_2) &:= \frac{1}{2}\big(\langle S, K_2\rangle - log(|K_2|)\big). \end{aligned} \tag{4.5}$$

As a consequence the definition (4.3) of Σ_m can simply be rewritten as:

$$\Sigma_m := \underset{\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(\Sigma, \widetilde{\Sigma}\right), \tag{4.6}$$

and the definition (4.4) of $\widehat{\Sigma}_m$ as:

$$\widehat{\Sigma}_m := \underset{\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(S, \widetilde{\Sigma}\right). \tag{4.7}$$

We must address a technical issue here. Unlike Σ_m, the existence of $\widehat{\Sigma}_m$ in Eq. (4.7) is not guaranteed for any graph when n is too small, see [35] and [158]. To propose a matrix for any graph and any number of observation n, we define a penalised MLE that we will use instead:

$$\widehat{\Sigma}_{m,\lambda} := \underset{\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(S + \lambda I_p, \widetilde{\Sigma}\right) = \underset{\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(S, \widetilde{\Sigma}\right) + \frac{\lambda}{2}\left\|\widetilde{K}\right\|_*. \tag{4.8}$$

As long as $\lambda > 0$, $S^{(\lambda)} := S + \lambda I_p \in S_p^{++}$ and $\widehat{\Sigma}_{m,\lambda}$ exists. For the sake of simplicity, we keep noting $\widehat{\Sigma}_m$ without the λ index throughout the whole paper. The only exception is Section 4.2.2, where we

first show results for the unpenalised $\widehat{\Sigma}_m$ before generalising them to $\widehat{\Sigma}_{m,\lambda}$, hence the need for the distinction.

With this new estimable representative $\widehat{\Sigma}_m$ of any graph m, we are now fully equipped to look for the best graph. The CE $H\left(\Sigma, \widehat{\Sigma}_m\right)$ still represents how well the distribution $\mathcal{N}\left(0_p, \widehat{\Sigma}_m\right)$ reproduces the behaviour of $\mathcal{N}(0_p, \Sigma)$. However this time the CE is not mechanically lower for the most connected graphs, since each $\widehat{\Sigma}_m$ is defined by an optimisation problem on $S + \lambda I_p$ (Eq. (4.8)) and not on Σ directly, unlike the matrices Σ_m (Eq. (4.6)).
A finite data set only offers a partial vision of the phenomenon it is generated from, as a result there is only a limited quantity of information we can extract from a finite number of observations. With that in mind, solving (4.8) with a too connected graph m requires to infer too many coefficients - one per edge - with relation to this available quantity of information and will result in a poor or even pathological matrix $\widehat{\Sigma}_m$ with a high divergence $H\left(\Sigma, \widehat{\Sigma}_m\right)$. On the opposite side of the spectrum, inference with too small graphs produces weak MLEs that do not fully exploit the amount of data available and will also have a high divergence with the real Σ. In the middle ground, there are graphs with just enough edges to extract all the information present in the observations. Those graphs produce the least pathological, most relevant MLEs $\widehat{\Sigma}_m$, with the lower divergence $H\left(\Sigma, \widehat{\Sigma}_m\right)$. This is why we set the CE $H\left(\Sigma, \widehat{\Sigma}_m\right)$ between the MLE and the true matrix as our metric to assess the quality of a graph: this deviation quantifies how complete and relevant of an interpretation the graph can offer **with the fixed, finite data set at hand**. And this is the dynamic we want to capture: minimising the deviation $H(\Sigma, \Sigma_m)$ would yield the graphs that perform the best given all the necessary information, and those are systematically the more connected ones. On the other hand, minimising the deviation $H\left(\Sigma, \widehat{\Sigma}_m\right)$ will allow us to identify the graphs that can handle a finite data set of observations and still produce good matrices.

The model selection framework we consider is the following: let $\mathcal{M}$ be a set of graphs, which we call a family of models, we want to select among them the one realising the best performances given the data. That is to say the graph $\hat{m}^*$ whose associated MLE has the lowest deviation from the real covariance matrix:

$$\hat{m}^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(\Sigma, \widehat{\Sigma}_m\right) \right] . \tag{4.9}$$

Even though all of the $\widehat{\Sigma}_m$ are calculable from the data, the CE with Σ - which we call the "True CE" (TCE) - is not, because Σ is still unknown. We can minimise instead the In Sample Negative Log-likelihood (ISNL), $H\left(S^{(\lambda)}, \widehat{\Sigma}_m\right)$, and get model $\hat{m}_0$, defined by:

$$\hat{m}_0 \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(S^{(\lambda)}, \widehat{\Sigma}_m\right) \right] . \tag{4.10}$$

However, the ISNL is overly optimistic and mechanically favours the most connected graphs, in the same way that the CE $H(\Sigma, \Sigma_m)$ of the matrices Σ_m built from Σ and not from data is always improved for larger graphs. We call m^* the graph obtained from optimising the latter criterion:

$$m^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H(\Sigma, \Sigma_m) \right] . \tag{4.11}$$

Both m^* and $\hat{m}_0$ are mechanically among the graphs in $\mathcal{M}$ with maximal edge sets. Moreover if there is in $\mathcal{M}$ a maximum graph containing all the other, which is a common occurrence, then automatically $m^* = \hat{m}_0$. To avoid systematically selecting very connected graphs, we defined an estimable model $\hat{m}_{\text{pen}}$ which optimises a penalised ISNL:

$$\hat{m}_{\text{pen}} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(S^{(\lambda)}, \widehat{\Sigma}_m\right) + \text{pen}(m) \right] . \tag{4.12}$$

The design of such penalties $\text{pen}(m)$ with the goal of making the realised TCE $H\left(\Sigma, \widehat{\Sigma}_{\hat{m}_{\text{pen}}}\right)$ as close to the optimal TCE $H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$ as possible is the stake of this paper. Our approach will consist in using $\text{pen}(m)$ to correct the bias between the ideal target function to optimise - the True CE $H\left(\Sigma, \widehat{\Sigma}_m\right)$ - and the observed ISNL $H\left(S^{(\lambda)}, \widehat{\Sigma}_m\right)$. That is to say, we want: $\mathbb{E}\left[H\left(S^{(\lambda)}, \widehat{\Sigma}_m\right) + \text{pen}(m) - H\left(\Sigma, \widehat{\Sigma}_m\right)\right] = 0$.

We consider the case when the family of graphs $\mathcal{M}$ is constituted only of chordal, also known a "decomposable", graphs. A very powerful result gives us an explicit formula for $\widehat{\Sigma}_m$, which we use in Section 4.3 to design an unbiased criterion.

Remark. The parallels between covariance selection and variable selection. The covariance selection we undertake here, under the Gaussian Graphical Model $X \sim \mathcal{N}(0, \Sigma)$, shares similarities with the variable selection under the linear model $y = \beta^T x + \epsilon \sim \mathcal{N}\left(\beta^T x, \sigma^2\right)$. The following points explain in more details the parallels that are drawn in table 4.2.1:

- We call f_Σ the pdf of X and $f_\beta^{(x,y)} = f^{(x)} f_\beta^{(y|x)}$ the joint pdf of (x, y), which we cut in two to simplify since we assume that x does not depend on β, hence only the conditional pdf $f_\beta^{(y|x)}$ carries the dependency.
- In both cases the true value of the parameter can be expressed from the moments of the random variables: $K = \Sigma^{-1} = \mathbb{E}\left[XX^T\right]^{-1}$ and $\beta = \mathbb{E}\left[xx^T\right]^{-1}\mathbb{E}\left[xy\right]$.
- With no knowledge of those moments but with observations, one can search for the MLEs instead: $\widehat{K} := \underset{\widetilde{K} \in S_p^{++}}{\text{argmin}}\, H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right)$ and $\widehat{\beta} := \underset{\widetilde{\beta} \in \mathbb{R}^p}{\text{argmin}}\, H\left(\hat{f}_n^{(x,y)}, f_{\widetilde{\beta}}^{(x,y)}\right) = \underset{\widetilde{\beta} \in \mathbb{R}^p}{\text{argmin}}\, H\left(\hat{f}_n^{(x,y)}, f_{\widetilde{\beta}}^{(y|x)}\right)$
- For those two MLEs we have explicit formulas: $\widehat{K} = \left(\frac{1}{n}\sum_{k=1}^n X_k X_k^T\right)^{-1}$ and $\widehat{\beta} = \left(\frac{1}{n}\sum_{k=1}^n x_k x_k^T\right)^{-1}\left(\frac{1}{n}\sum_{k=1}^n x_k y_k^T\right)$
- When $n < p$ the MLEs are not defined anymore: there are too many degrees of freedom and too few information to infer a parameter of that size without constraints.
- Inferring less coefficients in the parameter by searching for it inside a subset with fixed sparsity can make the problem well defined again. Indeed, even when $n < p$, the MLE $\widehat{K}_m = \underset{\widetilde{K}, \widetilde{\Sigma} \in \Theta_m}{\text{argmin}}\, H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right)$ can exist if the edge set E_m is small enough. In a similar fashion, with I_m a subset of $[\![1, p]\!]$, we can define the sparse parameter set $\Lambda_m := \{\beta \in \mathbb{R}^p | i \notin I_m \implies \beta_i = 0\}$, and $\widehat{\beta}_m := \underset{\widetilde{\beta} \in \Lambda_m}{\text{argmin}}\, H\left(\hat{f}_n^{(x,y)}, f_\beta^{(y|x)}\right)$. For $\widehat{\beta}_m$ there even an explicit existence condition: $|I_m| \leq n$ and an explicit formula $\widehat{\beta}_m = \left(\frac{1}{n}\sum_{k=1}^n x_k^{(m)} x_k^{(m)^T}\right)^{-1}\left(\frac{1}{n}\sum_{k=1}^n x_k^{(m)} y_k^T\right)$, where $x_k^{(m)} \in \mathbb{R}^{|I_m|}$ is the restriction of x_k to its components in I_m.
- Even when $\widehat{K}_m$ and $\widehat{\beta}_m$ exist, they can still be close to degenerate if they contain too many coefficients with regards to the data. Searching for a balanced feature or edge set is the common stake of variable and covariance selection.
- Hence, in variable selection, one is looking for the feature set $I_{\hat{m}^*}$ that will produce the best MLE. In the case that we use the CE with the real pdf $f_\beta^{(x,y)}$ to define this notion of "best", we have: $\hat{m}^* := \underset{m}{\text{argmin}}\, H\left(f_\beta^{(x,y)}, f_{\hat{\beta}_m}^{(x,y)}\right) = \underset{m}{\text{argmin}}\, \mathbb{E}\left[\left(x^T(\beta - \hat{\beta}_m)\right)^2\right]$. If we apply a similar reasoning to the covariance selection problem, and look in that case for the edge set E_m producing the best MLE in terms of CE f_Σ, we get Eq. (4.9), defining the ideal model.

	Covariance selection	Feature selection
Model	$X \sim \mathcal{N}(0_p, \Sigma)$	$x \in \mathbb{R}^p,\ y \sim \mathcal{N}\left(x^T\beta, \sigma^2\right)$
pdf	$f_\Sigma(X) = \frac{1}{\sqrt{2\pi\lvert\Sigma\rvert}} e^{-\frac{1}{2}X^T\Sigma^{-1}X}$	$f_\beta^{(x,y)}(x,y) = f^{(x)}(x)\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(y-x^T\beta\right)^2}$
Link Parameter/Moment	$K = \Sigma^{-1} = \mathbb{E}\left[XX^T\right]^{-1}$	$\beta = \mathbb{E}\left[xx^T\right]^{-1}\mathbb{E}\left[xy\right]$
Negative Log-likelihood	$H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right) = H\left(S, \widetilde{\Sigma}\right)$	$H\left(\hat{f}_n^{(x,y)}, f_{\widetilde{\beta}}^{(x,y)}\right) \equiv \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\widetilde{\beta})^2$
Unconstrained MLE	$\widehat{K} = \underset{\widetilde{K}\in S_p^{++}}{\operatorname{argmin}} H\left(\hat{f}_n, f_\Sigma\right)$ $\widehat{K} = \left(\frac{1}{n}\sum_{k=1}^n X_k X_k^T\right)^{-1} = S^{-1}$	$\widehat{\beta} = \underset{\widetilde{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} H\left(\hat{f}_n^{(x,y)}, f_{\widetilde{\beta}}^{(x,y)}\right)$ $\widehat{\beta} = \left(\frac{1}{n}\sum_{k=1}^n x_k x_k^T\right)^{-1}\left(\frac{1}{n}\sum_{k=1}^n x_k y_k^T\right)$
Constrained MLE	$\widehat{K}_m = \underset{\widetilde{K},\widetilde{\Sigma}\in\Theta_m}{\operatorname{argmin}} H\left(\hat{f}_n, f_{\widetilde{\Sigma}}\right)$	$\widehat{\beta}_m = \underset{\widetilde{\beta}\in\Lambda_m}{\operatorname{argmin}} H\left(\hat{f}_n^{(x,y)}, f_{\widetilde{\beta}}^{(y\lvert x)}\right)$
Best estimable model	$\hat{m}^* \in \underset{m\in\mathcal{M}}{\operatorname{argmin}}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right]$	$\hat{m}^* = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\,\mathbb{E}\left[\left(x^T(\beta - \hat{\beta}_m)\right)^2\right]$

Table 4.1: Summary of the parallels between covariance and feature selection. The left column recaps the formalism we introduced to tackle the problem of covariance selection. The right columns represent their equivalents in the classical feature selection with a linear model.

4.2.2 Some observations and results on the Cross Entropy

Notations. For any two elements $\Sigma_1, \Sigma_2 \in S_p^+$, the Frobenius scalar product is:

$$\langle \Sigma_1, \Sigma_2 \rangle = tr\left(\Sigma_1\Sigma_2\right) = \sum_{i,j=1}^p \Sigma_1^{ij}\Sigma_2^{ij}\,.$$

We also recall that, when $\Sigma_2 \in S_p^{++}$, and $K_2 := \Sigma_2^{-1}$, the formula for the Cross Entropy without additive constant is:

$$H\left(\Sigma_1, \Sigma_2\right) = \frac{1}{2}\big(\,\langle \Sigma_1, K_2\rangle - log\left(\lvert K_2\rvert\right)\big)\,.$$

We call Π_m the orthogonal projection on $\left\{M \in \mathbb{R}^{p\times p}\middle|\forall i \neq j,\ (i,j) \notin E_m \implies M_{ij} = 0\right\}$, the matrix space with the sparsity of the edge set E_m. In other words:

$$\Pi_m(M)_{ij} = \begin{cases} M_{ij} & if\ i = j\ or\ (i,j) \in E_m \\ 0 & \text{otherwise} \end{cases}$$

General results. Let Σ, Σ_1 and $\Sigma_2 \in S_p^{++}$ be three covariance matrices with $K_1 := \Sigma_1^{-1}$ and $K_2 := \Sigma_2^{-1}$, and $S \in S_p^+$ a semi definite positive matrix not necessarily invertible. Then:

$$H\left(\Sigma, \Sigma_1\right) = H\left(\Sigma, \Sigma_2\right) + \frac{1}{2}\left(\langle \Sigma, K_1 - K_2\rangle - log\left(\frac{\lvert K_1\rvert}{\lvert K_2\rvert}\right)\right)\,, \tag{4.13}$$

$$H\left(\Sigma, \Sigma_1\right) = H\left(S, \Sigma_1\right) + \frac{1}{2}\langle \Sigma - S, K_1\rangle\,. \tag{4.14}$$

The following lemma illustrates a very simple property of the Cross Entropy when the second argument is an inverse sparse matrix.

Lemma 4.2.1. *With defintion (4.1) for Θ_m, we have: $\forall M \in S_p^{++}$, $\forall A_m \in \Theta_m$:*

$$H\left(M, A_m\right) = H\left(\Pi_m(M), A_m\right) . \tag{4.15}$$

In other words:

$$\forall M \in S_p^{++}, \forall A_m \in \Theta_m, \quad \cdot \mapsto H\left(\cdot, A_m\right) \text{ is constant on } \Pi_m^{-1}(M) . \tag{4.16}$$

Proof.

$$\begin{aligned}
H\left(M, A_m\right) &= \frac{1}{2}\Big(\left\langle M, A_m^{-1} \right\rangle - log\left(\left|A_m^{-1}\right|\right) \Big) \\
&= \frac{1}{2}\left(\sum_{i,j \in [\![1,p]\!]^2} M^{ij} {A_m^{-1}}^{ij} - log\left(\left|A_m^{-1}\right|\right) \right) \\
&= \frac{1}{2}\left(\sum_{i,j \in E_m} M^{ij} {A_m^{-1}}^{ij} - log\left(\left|A_m^{-1}\right|\right) \right) \\
&= \frac{1}{2}\Big(\left\langle \Pi_m(M), A_m^{-1} \right\rangle - log\left(\left|A_m^{-1}\right|\right) \Big) \\
&= H\left(\Pi_m(M), A_m\right) .
\end{aligned}$$

The key element of the proof being that $\left\langle M, A_m^{-1} \right\rangle = \left\langle \Pi_m(M), A_m^{-1} \right\rangle$. □

Results on the MLE. In this section, we prove very general results when $\widehat{\Sigma}_m$ is the regular, unpenalised MLE defined in (4.7).
For any graph m, [35] shows that the unpenalised MLE (4.7) exists when $\exists \widetilde{S} \in S_p^{++}$, $\widetilde{S}_{|E_m} = S_{|E_m}$. In that case, we additionally have:

$$\begin{aligned}
\Pi_m(\widehat{\Sigma}_m) &= \Pi_m(S) , \\
\Pi_m(\widehat{K}_m) &= \widehat{K}_m .
\end{aligned} \tag{4.17}$$

Since $\Sigma \in S_p^{++}$ already, with the same arguments we have that the best sparse approximation of Σ defined in (4.6) always exists for any graph m and that:

$$\begin{aligned}
\Pi_m(\Sigma_m) &= \Pi_m(\Sigma) , \\
\Pi_m(K_m) &= K_m .
\end{aligned} \tag{4.18}$$

As a consequence of lemma 4.2.1, under the assumption that $\widehat{K}_m$ exists - i.e. that $\exists \widetilde{S} \in \Pi_m^{-1}(S) \cap S_p^{++}$ - we get from Eq. (4.17) and (4.18) that:

$$\begin{aligned}
H\left(\Sigma, \widehat{\Sigma}_m\right) &= H\left(\Sigma_m, \widehat{\Sigma}_m\right) , \\
H\left(\Sigma, \Sigma_m\right) &= H\left(\Sigma_m, \Sigma_m\right) , \\
H\left(S, \widehat{\Sigma}_m\right) &= H\left(\widehat{\Sigma}_m, \widehat{\Sigma}_m\right) , \\
H\left(S, \Sigma_m\right) &= H\left(\widehat{\Sigma}_m, \Sigma_m\right) .
\end{aligned} \tag{4.19}$$

Some of these divergences have very simple formulas:

$$\begin{aligned}
H\left(S, \widehat{\Sigma}_m\right) &= H\left(\widehat{\Sigma}_m, \widehat{\Sigma}_m\right) \\
&= \frac{1}{2}\left(tr\left(\widehat{\Sigma}_m \widehat{K}_m\right) - log\left(\left|\widehat{K}_m\right|\right)\right) \\
&= \frac{1}{2}\left(tr\left(I_p\right) - log\left(\left|\widehat{K}_m\right|\right)\right) \\
&= \frac{1}{2}\left(p - log\left(\left|\widehat{K}_m\right|\right)\right) .
\end{aligned} \tag{4.20}$$

$$
\begin{aligned}
H\left(\Sigma, \Sigma_m\right) &= H\left(\Sigma_m, \Sigma_m\right) \\
&= \frac{1}{2}\left(tr\left(\Sigma_m K_m\right) - log\left(|K_m|\right)\right) \\
&= \frac{1}{2}\left(tr\left(I_p\right) - log\left(|K_m|\right)\right) \\
&= \frac{1}{2}\left(p - log\left(|K_m|\right)\right) .
\end{aligned}
\tag{4.21}
$$

We also have a formula to link the divergences between $\widehat{\Sigma}_m$ and the two reference matrices S and Σ:

$$
\begin{aligned}
H\left(\Sigma, \widehat{\Sigma}_m\right) &= H\left(\Sigma_m, \widehat{\Sigma}_m\right) \\
&= \frac{1}{2}\left(tr\left(\Sigma \widehat{K}_m\right) - log\left(\left|\widehat{K}_m\right|\right)\right) \\
&= \frac{1}{2}\left(tr\left(\Sigma \widehat{K}_m\right) - p\right) + \frac{1}{2}\left(p - log\left(\left|\widehat{K}_m\right|\right)\right) \\
&= \frac{1}{2}\left(tr\left(\Sigma \widehat{K}_m\right) - p\right) + H\left(\widehat{\Sigma}_m, \widehat{\Sigma}_m\right) \\
&= \frac{1}{2}\left(tr\left(\Sigma \widehat{K}_m\right) - p\right) + H\left(S, \widehat{\Sigma}_m\right) ,
\end{aligned}
\tag{4.22}
$$

where we used Eq. (4.20) to identify the formula of $H\left(S, \widehat{\Sigma}_m\right)$. This result can also be obtained as an application of (4.14) with the lemma 4.8.1 presented in appendix.

All of the results presented in this section - from (4.17) to (4.22) - come simply from the definition of Σ_m (4.6) as an approximation of Σ on one hand, and of $\widehat{\Sigma}_m$ (4.7) as an approximation of S on the other. The penalised $\widehat{\Sigma}_{m,\lambda}$ are defined from $S^{(\lambda)} = S + \lambda I_p$ in a similar fashion. As a consequence, EQ. (4.17) to (4.22) hold if we replace $\widehat{\Sigma}_m$ by $\widehat{\Sigma}_{m,\lambda}$ and S by $S^{(\lambda)}$ in all the formulas. In that case, the existence of $\widehat{\Sigma}_{m,\lambda}$ is not in question, since $S^{(\lambda)} \in S_p^{++}$.

The last connection we can describe is between $H\left(S, \Sigma_m\right)$ and $H\left(\Sigma, \Sigma_m\right)$. We prove that there actually is a very general and insightful dynamic between $H\left(S, \widetilde{\Sigma}\right)$ and $H\left(\Sigma, \widetilde{\Sigma}\right)$ for any positive definite matrix $\widetilde{\Sigma}$ independent of S. We give an explicit formula in proposition 4, whose proof is in appendix.

Proposition 4. *Let $nS = \sum_{i=1}^{n} X_i X_i^T \sim \mathcal{W}_p\left(\Sigma, n\right) p$, then for any matrix $\widetilde{\Sigma} \in S_p^{++}$ independent of S, with $\widetilde{K} := \widetilde{\Sigma}^{-1}$, we have:*

$$
H\left(S, \widetilde{\Sigma}\right) = H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{2}\left(\sum_{i=1}^{p} \lambda_i(\widetilde{K}^{\frac{1}{2}} \Sigma \widetilde{K}^{\frac{1}{2}}) \, \frac{\chi_n^{2\,(i)} - n}{n}\right),
\tag{4.23}
$$

where, conditionally to $\widetilde{\Sigma}$, $\left(\chi_n^{2\,(i)}\right)_{i \in [\![1,p]\!]}$ are iid χ^2 random variables with n degrees of freedom, and $\lambda_i(\widetilde{K}^{\frac{1}{2}} \Sigma \widetilde{K}^{\frac{1}{2}})$ are the eigenvalues of $\widetilde{K}^{\frac{1}{2}} \Sigma \widetilde{K}^{\frac{1}{2}}$.
In particular we have:

$$
\mathbb{E}\left[H\left(S, \widetilde{\Sigma}\right)\right] = \mathbb{E}\left[H\left(\Sigma, \widetilde{\Sigma}\right)\right] .
\tag{4.24}
$$

This proposition can be applied in particular to $\widetilde{\Sigma} = \Sigma_m$ since $\Sigma_m \in S_p^{++}$ is a constant matrix,

hence independent of S, to get:

$$\begin{aligned} H\left(\widehat{\Sigma}_m, \Sigma_m\right) &= H\left(S, \Sigma_m\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{p}\lambda_i \frac{\chi_n^{2\,(i)} - n}{n}\right) + H\left(\Sigma, \Sigma_m\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{p}\lambda_i \frac{\chi_n^{2\,(i)} - n}{n}\right) + H\left(\Sigma_m, \Sigma_m\right), \end{aligned} \tag{4.25}$$

where $\lambda_i := \lambda_i(K_m^{\frac{1}{2}}\Sigma K_m^{\frac{1}{2}})$ are the eigenvalues of $K_m^{\frac{1}{2}}\Sigma K_m^{\frac{1}{2}}$.

In the proposition 4, we used that nS is a Wishart matrix, as a consequence we cannot directly apply the results to $S^{(\lambda)}$. In (4.26), we do the additional work necessary to get the equivalent of the relation (4.25) for $\widehat{\Sigma}_{m,\lambda}$ and $S^{(\lambda)}$

$$\begin{aligned} H\left(\widehat{\Sigma}_{m,\lambda}, \Sigma_m\right) &= H\left(S^{(\lambda)}, \Sigma_m\right) \\ &= \frac{1}{2}\left(tr\left(S^{(\lambda)}K_m\right) - log\left(|K_m|\right)\right) \\ &= \frac{1}{2}\left(tr\left(SK_m\right) - log\left(|K_m|\right)\right) + \frac{\lambda}{2}tr\left(K_m\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{p}\lambda_i \frac{\chi_n^{2\,(i)} - n}{n}\right) + H\left(\Sigma, \Sigma_m\right) + \frac{\lambda}{2}tr\left(K_m\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{p}\lambda_i \frac{\chi_n^{2\,(i)} - n}{n}\right) + H\left(\Sigma_m, \Sigma_m\right) + \frac{\lambda}{2}tr\left(K_m\right). \end{aligned} \tag{4.26}$$

4.3 Model selection criterion with explicit formula for chordal graphs

In this section we propose an estimator of the CE when the graphs we are looking for are chordal. With those graphs, the existence of a closed form expression for the MLE, allows us to find an explicit formula to compute and correct the bias of the In Sample Negative Log-likelihood. We prove guarantees on the performances of the models selected by the new criterion.

Properties of chordal graphs. Chordal graphs are graphs where there is no cycle of more than three edges. They are also called decomposable, since there exists a maximal prime sub-graphs decomposition into the set of maximal cliques $\mathcal{C}$ and the set of their intersections, the separator cliques $\mathcal{P}$, see [91] for more details. When m is chordal we have an explicit formula (4.27) for the inverse of the MLE defined in (4.7) from its decomposition $(\mathcal{C}_m, \mathcal{P}_m)$, see [91].

$$\widehat{K}_m := \widehat{\Sigma}_m^{-1} = \sum_{c\in\mathcal{C}_m}\left[\left(S_{cc}\right)^{-1}\right]^0 - \sum_{p\in\mathcal{P}_m}\left[\left(S_{pp}\right)^{-1}\right]^0. \tag{4.27}$$

The notation $\left[\left(S_{cc}\right)^{-1}\right]^0$ indicates a matrix in $\mathbb{R}^{p\times p}$ where the coefficients of the square sub-matrix defined by clique c are those of S_{cc}^{-1} and the rest are 0. We get as well a necessary and sufficient condition for the existence of $\widehat{\Sigma}_m$ (see [158]): $n > |c|_{max}^{(m)}$, where $|c|_{max}^{(m)}$ is the size of the largest clique in m.

Thanks to the chordal hypothesis, we know whether $\widehat{\Sigma}_m$ defined in Eq. (4.7) exists, and we do not need to penalise the MLE as Eq. (4.8) to be safe. Hence in this whole section, we set $\lambda = 0$ and use the unpenalised $\widehat{\Sigma}_m$ which we compute directly with the formula (4.27).

Presentation. In this section we propose a criterion to optimise for model selection and offer guarantees on its choices. The ideal, but inaccessible quantity to minimise is the true CE, $H\left(\Sigma, \widehat{\Sigma}_m\right)$. From the observation: $H\left(\Sigma, \widehat{\Sigma}_m\right) = H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2}\left\langle \Sigma - S, \widehat{K}_m \right\rangle$, we see that $\left\langle \Sigma, \widehat{K}_m \right\rangle$ is the only un-estimable term in this quantity. With (4.27), we can express this scalar product, and get, under the assumption that $n > |c|^{(m)}_{max} + 1$, an explicit formula for its expectation:

$$\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_m \right\rangle\right] = \sum_{c \in \mathcal{C}_m} \frac{n\,|c|}{n - |c| - 1} - \sum_{p \in \mathcal{P}_m} \frac{n\,|p|}{n - |p| - 1} =: f(m)\,. \tag{4.28}$$

The term $f(m)$ that emerges does not depend on Σ anymore, it is only a function of the graph m. With $f(m)$, we correct the bias of the In Sample Negative Log-likelihood $H\left(S, \widehat{\Sigma}_m\right)$ and define our estimator H_m:

$$H_m := H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2}(f(m) - p) \equiv H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2} f(m) \quad \textbf{(UCEE)}\,. \tag{4.29}$$

This estimator is unbiased by construction:

$$\mathbb{E}\left[H_m - H\left(\Sigma, \widehat{\Sigma}_m\right)\right] = 0\,. \tag{4.30}$$

We call H_m the Unbiased Chordal Explicit Estimator (UCEE). In Eq. (4.29), the bias correction term $f(m)$ acts as a penalty that balances the optimistic In Sample Negative Log-likelihood $H\left(S, \widehat{\Sigma}_m\right)$. Its formula (4.28) shows that it does so by taking into account the very structure of the graph, and not simply its size. Additionally, we stress out the fact that the UCEE is completely non-parametric. The estimator can be used directly without the need to optimise any hyperparameter to get the best performances out of it.

Guarantees. From definitions (4.12) and (4.29), we get a natural control on the performance of UCEE's choice with the best expected score for any given model:

$$\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right]\,. \tag{4.31}$$

To remove the dependency in $\hat{m}$ in Eq (4.31), we upper bound even further the inequality and get the control (4.32) that has a more general expression. We have a similar bound, see Eq (4.33), when working the best model on the data, $\hat{m}^*$, as a reference.

$$\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right] \right\rangle\right|\right]\,, \tag{4.32}$$

$$\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)\right] + \mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right] \right\rangle\right|\right]\,. \tag{4.33}$$

Let $\mathcal{C}_{max}$ and $\mathcal{P}_{max}$ be deterministic cliques sets such that $\forall m \in \mathcal{M}, \mathcal{C}_m \in \mathcal{C}_{max}$ and $\mathcal{P}_m \in \mathcal{P}_{max}$. Let $|c|_{max} := \max_{m \in \mathcal{M}} |c|^{(m)}_{max}$. If $n > |c|_{max} + 3$, then Eq (4.34) and Eq (4.35) are upper bounds with explicit orders of the controls (4.32) and (4.33) respectively.

Proposition 5. *With the previously introduced notations, if $n > |c|_{max} + 3$, then the following inequalities hold:*

$$\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] - \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] \leq \sum_{c \in \mathcal{C}_{max}} \sqrt{\frac{n^3}{2}} \frac{\sqrt{|c|}}{(n - |c| - 3)^2} + \sum_{p \in \mathcal{P}_{max}} \sqrt{\frac{n^3}{2}} \frac{\sqrt{|p|}}{(n - |p| - 3)^2}\,, \tag{4.34}$$

$$\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)\right] \leq \sum_{c \in \mathcal{C}_{max}} \sqrt{2n^3} \frac{\sqrt{|c|}}{(n - |c| - 3)^2} + \sum_{p \in \mathcal{P}_{max}} \sqrt{2n^3} \frac{\sqrt{|p|}}{(n - |p| - 3)^2}\,. \tag{4.35}$$

Sketch of proof. We expose here the main steps necessary to prove equations (4.34) and (4.35). The complete proof can be found in appendix. We already have (4.32) and (4.33), hence we just need to prove that $\mathbb{E}\left[\max_{m\in\mathcal{M}}\left|\left\langle\Sigma,\widehat{K}_m-\mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right|\right]\leq\sum_{c\in\mathcal{C}_{max}}\sqrt{2n^3}\frac{\sqrt{|c|}}{(n-|c|-3)^2}+\sum_{p\in\mathcal{P}_{max}}\sqrt{2n^3}\frac{\sqrt{|p|}}{(n-|p|-3)^2}$. To that end:

- We start by showing in Lemma 4.8.3 in appendix which provides the formula for the variance of the trace of an inverse Wishart
- We then show that $\max_{m\in\mathcal{M}}\left|\left\langle\Sigma,\widehat{K}_m-\mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right|\leq\sum_{c\in\mathcal{C}_{max}}\left|\left\langle\Sigma_{cc},(S_{cc})^{-1}-\mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right|+\sum_{p\in\mathcal{P}_{max}}\left|\left\langle\Sigma_{pp},(S_{pp})^{-1}-\mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right|$
- Then that for any clique we find a Wishart W such that $\mathbb{E}\left[\left|\left\langle\Sigma_{cc},(S_{cc})^{-1}-\mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right|\right]\leq Var\left[tr\left(W^{-1}\right)\right]^{\frac{1}{2}}$
- The formula of Lemma 4.8.3 allows us to conclude

4.4 Implementation principles

For the sake of illustrating the performances of our model selection criteria, we must test them on real graph families. The algorithmic of our method is very simple: for every graph m in a family $\mathcal{M}$ of undirected chordal networks, we compute the MLE $\widehat{\Sigma}_m$ as defined in (4.8), then the UCEE (4.29). We finally pick the model $\hat{m}$ in $\mathcal{M}$ with the smallest UCEE. We can then compare this choice $\hat{m}$ to models proposed by other model selection procedures among $\mathcal{M}$.
In this Chapter, we heavily focus on the model selection aspect of the problem not on the graph exploration numerical scheme that builds the family $\mathcal{M}$. We adopt the same procedures as Chapter 3. With synthetic data, where everything is known, we manually make deterministic graph families. With real data, we simply reuse the graph exploration procedure of the Composite algorithm 3.1 of Chapter 3.

4.4.1 A preliminary question: how to efficiently evaluate the many members of a graph family

To compute the UCEE for every model $m\in\mathcal{M}$, it is necessary to compute every MLE $\widehat{K}_m$. This process can be sped up by taking into consideration the structure of $\mathcal{M}$. Iterative Proportional Scaling (IPS) [34, 91] is a classical algorithm that starts from any matrix in Θ_m and converges towards $\widehat{K}_m$ for a given empirical covariance S_1 and the graph m. Initialising the algorithm with matrix close to $\widehat{K}_m$ can significantly speed up the process. If $\mathcal{M}=\{m_1,...,m_N\}$ is an increasing sequence of graphs: $E_{m_1}\subset...\subset E_{m_N}$, then a very natural way to speed up the estimation of the family $\{\widehat{K}_m\}_{m\in\mathcal{M}}$ is to start with $\widehat{K}_{m_1}\in\Theta_{m_1}\subset\Theta_{m_2}$ and use it as initialisation for the computation of $\widehat{K}_{m_2}$. Then, iteratively use $\widehat{K}_{m_i}$ as initialisation for $\widehat{K}_{m_{i+1}}$ until every MLE is computed. If there is only a few edges of difference between the two consecutive models, not only do $\widehat{K}_{m_i}$ and $\widehat{K}_{m_{i+1}}$ have a similar number of non-zero coefficients, but those coefficients should be close. As a consequence, the number of iteration to get $\widehat{K}_{m_{i+1}}$ can be greatly reduced.
This can be done even if $\mathcal{M}$ is not a single increasing sequence of graphs: in the general case, $\mathcal{M}$ is composed of several increasing graph paths each leading to one of the maximal graphs in $\mathcal{M}$, the ensemble constituting an inclusion tree. Starting from the root of the tree, the MLEs just have to be propagated along the different branches until the leaves. In our study, the model families $\mathcal{M}$ will by design mostly be simple increasing paths of graphs.

We illustrate the time gain with synthetic data. We take an increasing family of 300 models with one edge of difference between every two consecutive graphs. We compute each MLE with either

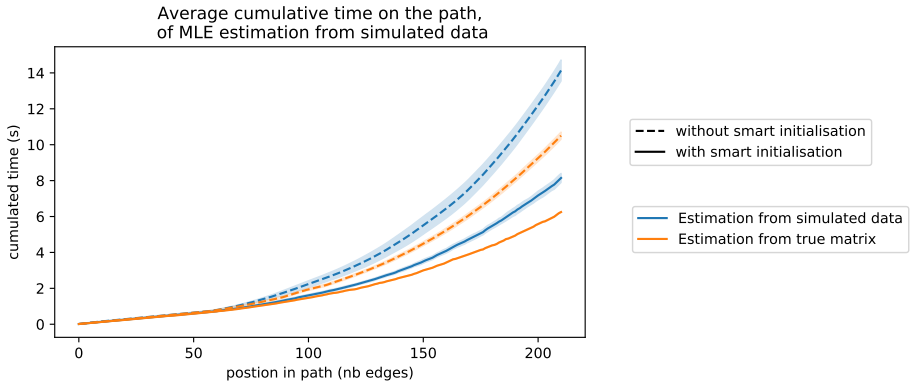


Figure 4.1: Cumulative time to compute the MLEs. In blue, the MLE are computed from empirical data, in orange from the real matrix directly. The dotted lines represent IPS initialised with the sparser previously computed MLE (smart initialisation), whereas the full lines represent IPS initialised with a diagonal matrix (neutral initialisation).

the previous one (smart initialisation) or a diagonal matrix (neutral initialisation). Many simulations are made and aggregated. On Figure 4.1 we represent, with standard deviation, the average cumulative estimation time of $\{\widehat{K}_m\}_m$ as a function of the number of models m estimated. The estimation can be done from simulated data (blue) or directly from the true matrix (orange). The smart initialisation significantly reduces the computation times.

4.4.2 Algorithms

In this section we detail the algorithms we use to construct model families and showcase the performances of our model selection criterion. After tackling the questions of the previous section, we see the advantage of building families as paths of graphs, with gradually increasing edge sets. As we described, we can take advantage of this inclusion structure to reduce the MLE computation time. As mentioned, we work like in Chapter 3. With synthetic data, in a controlled environment, we use our knowledge of the real graph to build deterministic paths. With real data, we use the exploration scheme of the Composite algorithm 3.1.

Numerical scheme in a simulated environment

In a simulated environment, we give ourselves a true covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ whose inverse K has the sparsity pattern of a fixed graph m_0. We can simulate without limitation from $\mathcal{N}(0_p, \Sigma)$.

Family building. We use the knowledge we have in this synthetic environment to manually craft an interesting family $\mathcal{M}$. We construct, edge by edge, a path of chordal graphs from the Fully Sparse (FS), with no edge, to the Fully Connected (FC), with all possible edges, via m_0, the true graph. Having m_0 as an imposed intermediary point forces $\mathcal{M}$ to contain good graphs.
The resulting family $\mathcal{M}$ is a path increasing for the inclusion: the ideal case for efficient computations of the MLEs.

Simulations and family evaluation. For each simulation, we generate n observations. Let S be the empirical covariance matrix built with all of them. For every graph m of the family $\mathcal{M}$, we compute the MLE $\widehat{\Sigma}_m(S)$, then the associated value of the UCEE. The UCEE is an estimation of the True Cross Entropy (TCE) $H(\Sigma, \widehat{\Sigma}_m(S))$ of this MLE. This Cross Entropy represents the performances we are concretely able to achieve with that data and that graph m.

To assess the perceptiveness of the UCEE as a model selection criterion, we compare it to the (1-fold) Cross Validated Cross Entropy (CVCE):

$$H\left(S_2, \widehat{\Sigma}_m(S_1)\right) \textbf{ (Cross Validated Cross Entropy)} \tag{4.36}$$

To calculate the CVCE, we split the n available points into an *inference* set and a *validation* set. The n_1 points of the *inference* set are used to build an *inference* covariance S_1, which is then used to compute the MLE $\widehat{\Sigma}_m(S_1)$ for each graph $m \in \mathcal{M}$. On the other hand, the n_2 points of the *validation* set define a *validation* covariance S_2, which serves as a replacement for Σ in the Cross Entropy formula. Indeed, different split sizes define different versions of the CVCE and lead to different results. We use Cross Validation as a benchmark for comparison since it is a very general, tried and tested, method for model selection.
We compare on each simulation the performances in terms of True Cross Entropy of the models selected by the different versions of CVCE, by our UCEE and by the TCE. We make N of those simulations.

Our numerical scheme in an uncontrolled environment

When the phenomenon is observed and not simulated, we do not know the real graph anymore. In that case we must work from the data to construct $\mathcal{M}$.

The major changes. With real data, we do not know the real covariance matrix or its graph, which is probably not truly sparse. So we cannot anymore compute the Oracle Criterion or make a deterministic path through the real graph. To replace the TCE, we keep n_{test} data points aside as a test set to define a test empirical covariance matrix S_{test}. Meanwhile, the n_{train} points left are used to make a train empirical covariance matrix S_{train}. With those two, we compute the Out of Sample Negative Log-likelihood (OSNL) $H(S_{test}, \widehat{\Sigma}_m(S_{train}))$. The OSNL is less absolute than the TCE, but at least we still have a metric. Not knowing the true graph is more troublesome: we need to explore an interesting graph family to showcase the performances of the UCEE model selection criterion. Hence we design our own numerical scheme. We take a naive nodewise approach similar to [115]. Starting from the fully sparse graph, we add edges one by one. At each step, to decide which will be the new edge, we solve p linear prediction problem in parallel - one per vertex - with the Least Angle Regression (LARS) [44] algorithm. For implementation details, see the composite algorithm 3.1 of Chapter 3.

4.5 Results

With the algorithmic of Section 4.4, we construct families of chordal graph on both synthetic data and real hippocampus data. We illustrate the quality of the models selected by UCEE (4.29) against those selected by the Cross Validated Cross Entropy (CVCE) on those families. On one hand, the Cross Validated Cross Entropy (CVCE), a general criterion that makes no use of the formula for the MLE. On the other hand, our UCEE takes into account the additional information about the graph structure.

4.5.1 Synthetic Data

Setup. We generate a covariance matrix $\Sigma \in S_p^{++}$ of size $p = 30$ and whose inverse is sparse with a chordal pattern. We build a family of chordal graphs with the procedure of Section 4.4.2. For each simulation we generate n observations from $\mathcal{N}(0_p, \Sigma)$ and use them to compute the UCEE and CVCE on for every graph in $\mathcal{M}$. While the UCEE uses straightforwardly all the observations in its formula, we recall that CVCE splits n into n_1, for train, and n_2, for validation. Once again we try different values of n and $\alpha = \frac{n_2}{n}$, the fraction of observations put in the validation set of the CVCE. For each scenario, we make $N = 500$ simulations.

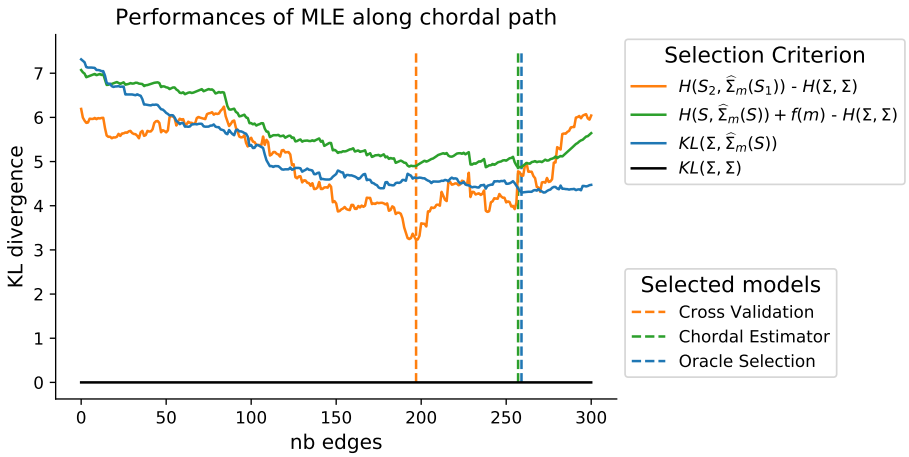


Figure 4.2: On a single simulation: evolution of and model selected by CVCE (orange), UCEE (green) and TCE (blue) along the fixed deterministic path of chordal graphs. The models selected by UCEE and TCE are very similar.

Illustration of the experiment on a single simulation. On Figure 4.2, we follow each criterion along the graph path. We have $n = 60$ total observations, 20% of which went in the validation set for CVCE. We display the KL divergence (CE minus the constant entropy $H(\Sigma, \Sigma)$) to give a sense of scale: a KL of 0 means a perfect reconstruction of the distribution, hence the relative differences in KL are meaningful. On this simulation, we see UCEE (green) and TCE (blue) reaching their minima on very similar graphs, whereas the CVCE (orange) selects a less connected graph. It seems UCEE is a better estimator of TCE and selects better graphs than the Cross Validation.

Quantitative analysis : aggregated metrics over multiple trials. To confirm this intuition, we aggregate the results of several simulations. We take a grid of different total data set sizes ($n =$ 40, 60, 80, 100) and fractions of data used in the validation set of CVCE (0.05, 0.1, 0.15, 0.2, 0.3, 0.4). We run 500 simulations for each combination.
We describe the models selected by a certain criterion over all simulations by their performances in KL divergence ($KL(\Sigma, \widehat{\Sigma}_m)$) and their complexities (number of edges). We represent on Figure 4.3 the averages and standard deviations of those quantities, with the KL on the y axis and the complexity on the x axis.
Each sub-figure corresponds to a different data set size. On each one, we can observe the path of the CVCE solutions with different validation set sizes (shades of red). This allows us to identify, in function of the total number of observations, the ideal balance between train and validation size to get the best performances out of the CVCE. However, we see that for each n, the solution from the UCEE (green) consistently beats CVCE regardless of the train/validation ratio. It is generally closer to the optimal graph on the path (blue) both in terms of size and performances. This confirms the observed trend on a single simulation.
As expected, the UCEE, with its explicit formula, makes good use of the very specific structure of the problem and displays more perceptiveness in its selection. However, despite being agnostic to the properties of the graphs in $\mathcal{M}$, the CVCE is not far behind, the relative difference in KL between the selected models of the two methods being is quite small.

4.5.2 Hippocampus Data

The UCEE showed promising behaviour on synthetic data. We put it to the test with real neurological data.

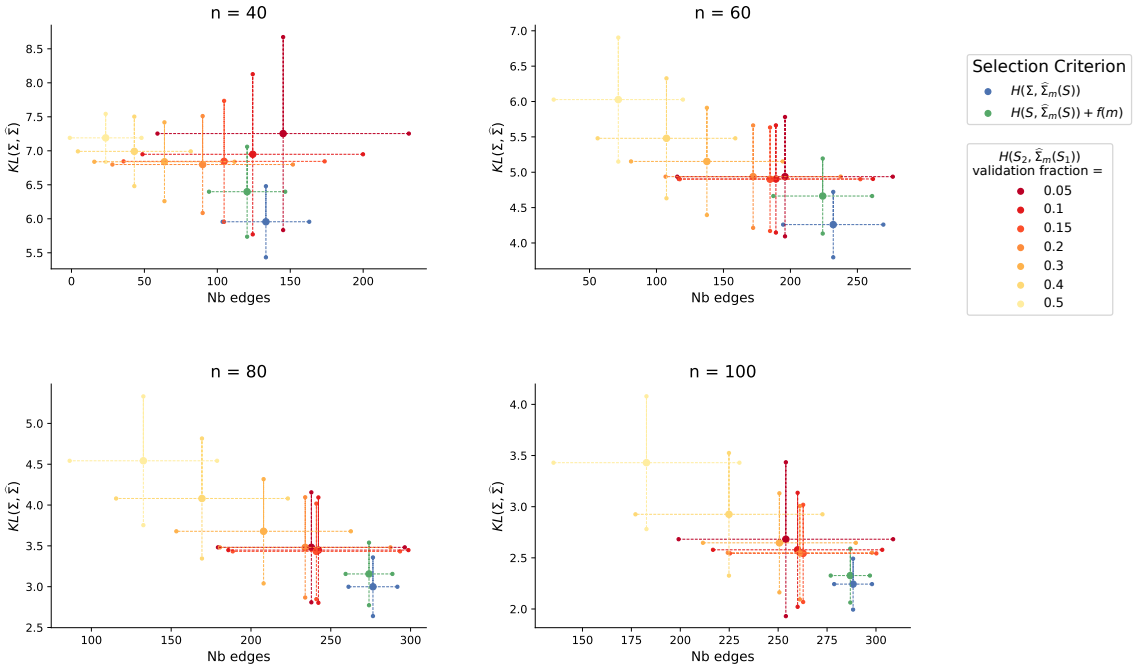


Figure 4.3: Average KL divergence (y axis) and complexity (x axis) of the models selected with CVCE (shades of red), UCEE (green) and TCE (blue).

Presentation of the data. We use a dataset containing measurements of the alteration of hippocampi from 101 patients where about a half are MCI or AD patients. These alterations are estimated by quantifying the deformation of a template hippocampus to each subject one. These deformations are performed by a diffeomorphic registration technique from [117]. Inferring a conditional correlation network between the areas of the hippocampus subject to deformation will help us better understand their spatial dependencies and the pattern of degeneration.
Looking for a chordal conditional correlation network between those measures is natural. There is an innate spatial organisation of the nodes, with good reasons for many of the geographical neighbours to be conditionally correlated, since the deformation is a continuous phenomenon, impacting in a similar fashion the nodes in proximity to one another. A chordal graph having no long chord-less cycle imposes a similar notion of coherence, where neighbours are connected in chunks. Two distant areas will have a tendency to be connected by thick strings of interconnected neighbours instead of thin strings of consecutive, semi-isolated nodes.

We have $p = 290$ measured areas and 101 patients. A previous work, [3], already studied the problem of inferring such networks on this dataset, but without specifically looking for chordal networks. In particular, they propose to use a variation of GGMselect from [56] where a prior graph that accounted for local correlations is used to exhibit long distance conditional correlations which did not appear otherwise. The idea was to project the data on the orthogonal of their prior and estimate the remaining edges.

Setup. To compute the OSNL $H(S_{test}, \widehat{\Sigma}_m(S_{full\ train}))$, we set aside once and for all $n_{test} = 20$ patients to build the test covariance matrix S_{test}, and use the remaining $n_{full\ train} = 81$ patients to make the train covariance matrix $S_{full\ train}$. Those $n_{full\ train}$ are the ones we let the various methods use to infer graphs and for model selection.
We use the chordal variant of the procedure described in Section 4.4.2 to build a chordal family $\mathcal{M}$. After excluding the patients used to build the graph path, 45 patients remain in the full train to be

used by the model selection methods. We then compute for all the graphs in this family the UCEE, using all those 45 points, and the CVCE, splitting those points into pure train and validation. As usual, we try out different validation fraction α for the CVCE.

Selected Models. We represent the selected graphs by each criterion on Figure 4.4 with the same conventions as the Figure 4.3. Here again, the UCEE model (green) outperforms all the CVCE models (shade of red) but is on par with the best ones (validation fractions of 0.05 and 0.1). This time the UCEE solution is significantly more connected than the best one (blue).
The strength of UCEE in this context is the absence of hyper-parameter. With Cross Validation we do not know in advance the best train/validation split. Identifying it with only the full training data would require an additional study, with an additional decomposition of the data into test and training sets. This can be especially troublesome if n is small and offers no guarantee of actually finding the right fraction in the end. The UCEE on the other hand can be used directly with its parameter-less formula and reaches the same performances levels as the best splits in CVCE.

Comparison with previous works on long distance connections. On the left panel of Figure 4.5, we visualise the conditional correlation network of deformations in the hippocampus selected by UCEE. On the right figure, we display only the long distance connections from this network, pruning our estimate by removing the edges of close points (with respect to the Euclidean distance). A connection is considered to be long distance when the distance between two nodes is among the 25% longest in the hippocampus. This shows us how the UCEE was able to select a graph featuring not only the obvious co-occurrences of deformations between neighbours, but also some more subtle long distance conditional correlations. The authors of [3] were also able to infer long distance conditional correlations in the hippocampus with this dataset. However they require to have a prior graph which, as mentioned by the authors, may be of different forms, leading potentially to very different results. On the other hand, we were able, by exploring a family of chordal graphs, to find and select graphs with both short distance and long distance conditional correlations without the need of such a process.

Splitting the patients dataset according to the diagnosis. 57 patients of this database are "control" and have not contracted the disease, the remaining 45 are either MCI or AD ans show signs of the disease. In order to understand the effect and signs of the disease on the hippocampus, we split the data in two and make one model for each population. The left graph of Figure 4.6, is the inferred conditional network for the control patients, whereas the right graph is the inferred network for the MCI and AD patients. To see more clearly, we represent on figure 4.7 the long distance correlations only. Once again, this means we only display an edge of the graph if the distance between the two corresponding nodes is among the 25% largest in the hippocampus.
It appears on Figure 4.6 that the estimated graph for control patients has many more edges than the MCI+AD patient one. This shows a first difference: the variability in terms of shape of the hippocampus in control population is very structured and smooth in particular in the head part of the anatomical structure where the network is highly dense.
When removing the local conditional correlations which although interesting are less informative, the pruned networks show again different patterns for the two sub-populations. Indeed, the density pattern of the edges looks uniform for the MCI-AD population whereas the control group shows many more conditional correlations of the deformation in the top part of the shape. When pruning the graph, this difference remains.
To compare with the previous study made in [3], the results are coherent. Their algorithm was based on GGMselect leading to very sparse selected graph. The introduction of prior graph enabled to make the long distance conditional correlation appear. Thank to our criterion, we can take into account both the local smoothness of the deformation we are analysing imposing a chordal structure and the long distance conditional correlations that appear since our criterion selects denser graphs. However, the conclusions are quite similar on the whole population where one can see the important coupling of the deformation of the top head and external side of the tail of the shape. Concerning the sub-population analysis, the results were quite poor for the previous work whereas we are able

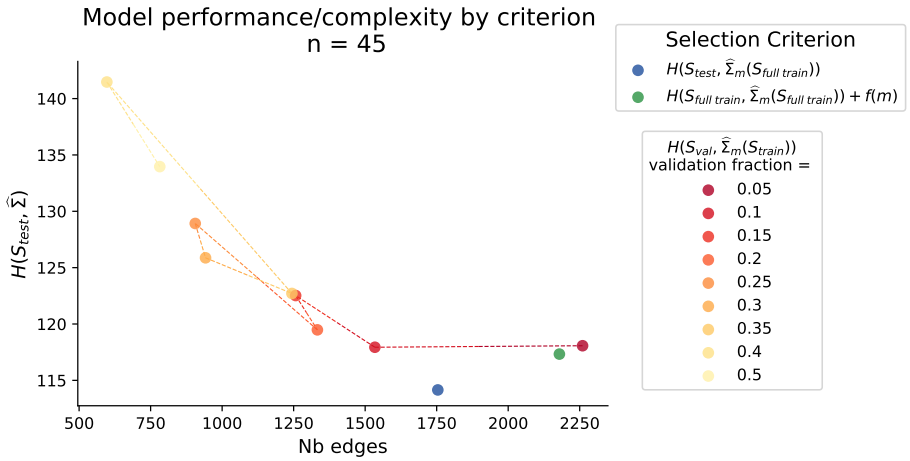


Figure 4.4: KL divergence (y axis) and complexity (x axis) of the models selected with CVCE (shades of red), UCEE (green) and TCE (blue).

to highlight very different graphs with many different edges in these two groups.

We are aware that the comparison of these graph shows very preliminary results that will need to be confirmed by a larger analysis on an extended database.

4.6 Discussion: the difficulty of comparing oneself to Σ_{m^*}

We recall the respective definitions of the models m^* and $\hat{m}^*$:

$$m^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(\Sigma, \Sigma_m\right) \right] ,$$
$$\hat{m}^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(\Sigma, \widehat{\Sigma}_m\right) \right] .$$

From the family of models $\mathcal{M}$, we can define $\Theta_{\mathcal{M}} := \{\Theta_m\}_{m \in \mathcal{M}}$, the matrix space that any proposed covariance matrix has to belong to. Throughout this paper, we used $\widehat{\Sigma}_{\hat{m}^*}$, which by definition of $\hat{m}^*$ is the best MLE in $\Theta_{\mathcal{M}}$, as the reference to get all our controls. From the definition of m^*, we see that Σ_{m^*} is actually the closest matrix to Σ of all the elements of $\Theta_{\mathcal{M}}$. We already discussed the idea that, since Σ_{m^*} is constructed from Σ, which is equivalent to having at our disposal an infinite amount of observation, setting Σ_{m^*} as reference can be somewhat unfair and misguided. In this section, we illustrated how, without any additional hypothesis on $\mathcal{M}$, we cannot theoretically discriminate good and bad matrices by comparing them to Σ_{m^*}.
To make this argument, we only use the definition of each matrix Σ_{m^*}, $\widehat{\Sigma}_{\hat{m}^*}$, $\widehat{\Sigma}_{\hat{m}_0}$ and their associated models m^*, $\hat{m}^*$, $\hat{m}_0$. Since we make no assumption on the models that are in $\mathcal{M}$, we have no way to compare m^* and $\hat{m}^*$ other than with their respective definitions. The only things we can tell of those models without any further hypothesis are:

- $\forall m \in \mathcal{M}, \quad H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \Sigma_m\right)$,
- $\forall m \in \mathcal{M}, \quad H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_m\right)$,
- m^* is one of the maximal graphs of $\mathcal{M}$.

Since we do not know what other models are in $\mathcal{M}$, we loose no information in our task to compare m^* and $\hat{m}^*$ by just evaluating the two previous equations in $\hat{m}^*$ and m^* respectively.

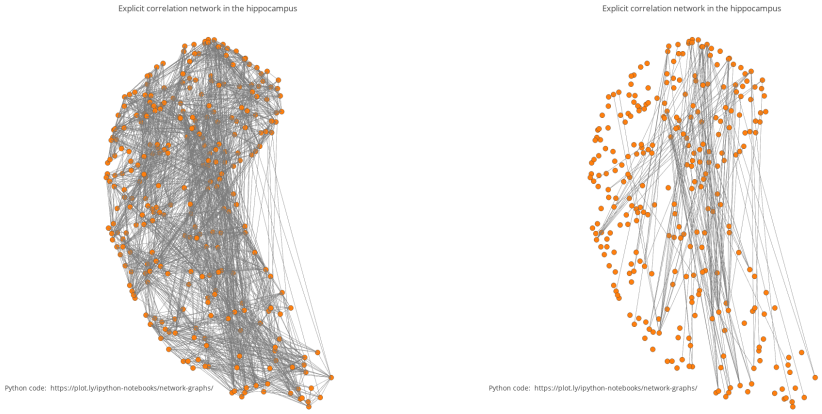


Figure 4.5: **Left**: Spatial representation of the edges and nodes of the conditional correlation network selected by the UCEE. **Right**: Spatial representation of the long distance connections only, in the conditional correlation network selected by the UCEE.

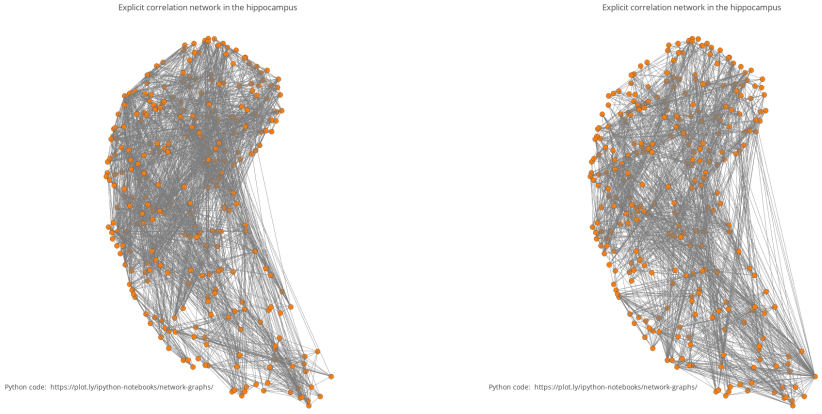


Figure 4.6: Spatial representation of the edges and nodes of the conditional correlation networks selected by the UCEE on control patients **(left)** and on MCI and AD patients **(right)**.

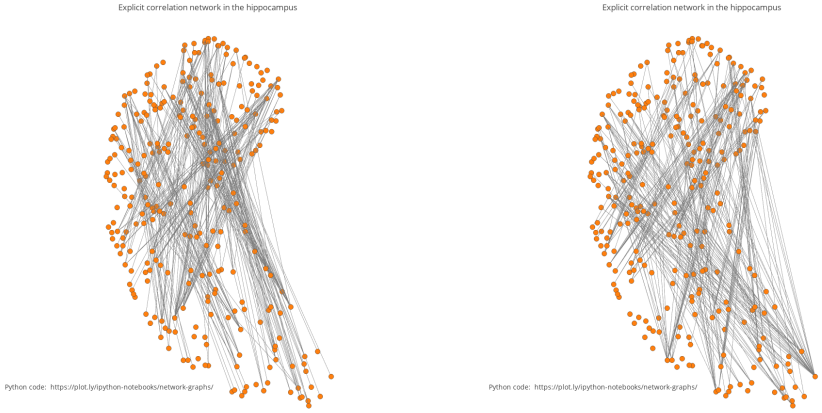


Figure 4.7: Spatial representation of the long distance connections only, in the conditional correlation networks selected by the UCEE on control patients **(left)** and on MCI and AD patients **(right)**.

In the end, the two following inequalities can be considered optimal, in the sense that we cannot find any tighter bound without additional assumptions:

$$H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{m^*}\right), \tag{4.37}$$

$$H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \Sigma_{\hat{m}^*}\right). \tag{4.38}$$

We introduce a quantity defined as:

$$A_m := H\left(\Sigma, \widehat{\Sigma}_m\right) - H\left(\Sigma, \Sigma_m\right). \tag{4.39}$$

By definition of Σ_m and since $\widehat{\Sigma}_m \in \Theta_m$, we have:

$$\forall m \in \mathcal{M}, \quad A_m \geq 0.$$

We use (4.37) and (4.38) to get an upper and a lower control of $H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$ by the optimal $H\left(\Sigma, \Sigma_{m^*}\right)$. We start from (4.38):

$$\begin{aligned}
& H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \Sigma_{\hat{m}^*}\right) \\
\iff \quad & H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) + H\left(\Sigma, \Sigma_{\hat{m}^*}\right) - H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \\
\iff \quad & H\left(\Sigma, \Sigma_{m^*}\right) + H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) - H\left(\Sigma, \Sigma_{\hat{m}^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)
\end{aligned}$$

We can then work with both the lower and upper bound at the same time:

$$\begin{aligned}
H\left(\Sigma, \Sigma_{m^*}\right) + H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) - H\left(\Sigma, \Sigma_{\hat{m}^*}\right) &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{m^*}\right) \\
H\left(\Sigma, \Sigma_{m^*}\right) + H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) - H\left(\Sigma, \Sigma_{\hat{m}^*}\right) &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \leq H\left(\Sigma, \Sigma_{m^*}\right) + H\left(\Sigma, \widehat{\Sigma}_{m^*}\right) - H\left(\Sigma, \Sigma_{m^*}\right) \\
H\left(\Sigma, \Sigma_{m^*}\right) + A_{\hat{m}^*} &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) \leq H\left(\Sigma, \Sigma_{m^*}\right) + A_{m^*}
\end{aligned}$$

We get in the end the control:

$$H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) - H\left(\Sigma, \Sigma_{m^*}\right) \in \left[A_{\hat{m}^*}, A_{m^*}\right] \tag{4.40}$$

We worked with equivalences only from the inequalities (4.37) and (4.38). Hence, the control (4.40) is also optimal in terms of the assumptions made.

Remark.

$$\begin{aligned} & A_{\hat{m}^*} \leq A_{m^*} \\ \iff & H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) - H\left(\Sigma, \Sigma_{\hat{m}^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{m^*}\right) - H\left(\Sigma, \Sigma_{m^*}\right) \\ \iff & H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) + H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{m^*}\right) + H\left(\Sigma, \Sigma_{\hat{m}^*}\right) \end{aligned}$$

Which is true by definition of m^* and $\hat{m}^*$. This only works with those two specific models.

Remark. $A_{\hat{m}^*} \geq 0$ means that $H\left(\Sigma, \Sigma_{m^*}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$ always, i.e. non matrix buildable from S or $S^{(\lambda)}$ can ever beat the best matrix build from Σ. This is obvious when you write that $\forall m \in \mathcal{M},\ H\left(\Sigma, \widehat{\Sigma}_m\right) \geq H\left(\Sigma, \Sigma_m\right) \geq H\left(\Sigma, \Sigma_{m^*}\right)$. Alternatively, remember that Σ_{m^*} is the best matrix in $\Theta_{\mathcal{M}}$ and that $\widehat{\Sigma}_{\hat{m}^*} \in \Theta_{\mathcal{M}}$.

Remark. Since the lower bound of the interval (4.40) solely comes from the definition of m^*, we actually have:

$$\forall m \in \mathcal{M}, \quad 0 \leq A_m \leq H\left(\Sigma, \widehat{\Sigma}_m\right) - H\left(\Sigma, \Sigma_{m^*}\right)$$

And the particularity of $\hat{m}^*$ is that it is the only model to get A_{m^*} as a guarantied upper bound.

We now put in perspective the upper bound we found. We know that m^* is a maximal graph: not the kind we want to infer. Indeed, because of the limited data, we need a balance in the number of selected edges. Since n is small and $\mathcal{M}$ potentially contains large graphs, $\widehat{\Sigma}_{m^*}$ is a priori a very poor matrix with a large Cross Entropy $H\left(\Sigma, \widehat{\Sigma}_{m^*}\right)$. Actually in most cases, maximal graphs like m^* have no unpenalised MLE and we are able to define $\widehat{\Sigma}_{m^*}$ only because of the penalisation in λ. In that case $H\left(\Sigma, \widehat{\Sigma}_{m^*}\right)$ will diverge when $\lambda \mapsto 0$. Hence when λ is very small, which it always is by design, $H\left(\Sigma, \widehat{\Sigma}_{m^*}\right)$ is often pathologically large. In a word, just as $H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$ represents the best performances reachable with a MLE, $H\left(\Sigma, \widehat{\Sigma}_{m^*}\right)$ is a reference for the worst performances a MLE can achieve within $\mathcal{M}$.

From Eq. (4.40), we have that the optimal upper control we can find on the best reachable performances $H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$ with $H\left(\Sigma, \Sigma_{m^*}\right)$ is A_{m^*}. However, we have by definition of A_m:

$$H\left(\Sigma, \widehat{\Sigma}_{m^*}\right) - H\left(\Sigma, \Sigma_{m^*}\right) = A_{m^*}\,.$$

That is to say that A_{m^*} is also the exact order of the control of the "worst" performances $H\left(\Sigma, \widehat{\Sigma}_{m^*}\right)$ with $H\left(\Sigma, \Sigma_{m^*}\right)$. In other words the controls we found for the best and the "worst" model are the exact same. Moreover, we showed that we cannot find a tighter control on the good model $\hat{m}^*$ without additional assumptions. As a consequence, it is pointless to set the true lowest possible CE on $\mathcal{M}$ $H\left(\Sigma, \Sigma_{m^*}\right)$ as a reference to assess the performance of any model m since even the best model $\hat{m}^*$ cannot be differentiated from m^*, one of the worst models.

The intuition behind this discussion is that the difference between the optimal CE $H\left(\Sigma, \Sigma_{m^*}\right)$ and the performances $H\left(\Sigma, \widehat{\Sigma}_m\right)$ of any model m is so large that it is approximately the same for any model m making it difficult to discriminate between models. On the other hand, $H\left(\Sigma, \widehat{\Sigma}_m\right)$ is a reachable CE for the MLE defined from models in $\mathcal{M}$, much closer to their own performances. Hence, setting it as a reference allows us to better identify different order of control for different models. Which is why this is the CE we use in our guarantees throughout the paper.

4.7 Conclusion

We tackled the problem of model selection with Gaussian Graphical Models. To assess the relevance of any new graph, we proposed to use the Cross Entropy of the Constrained Maximum Likelihood Estimator (MLE) with relation to the real covariance matrix. This metric quantifies how well the proposed matrix reproduces the law of the real once, we chose it because it explicitly describes the global quality of said proposition, unlike local metrics such as coefficients recovery. To provide a numerically stable method under all circumstances, we adopted a penalised definition of the MLE. Since the real matrix is by nature unknown, the Cross Entropy had to be estimated before being used for model selection. We proposed a new unbiased estimator of this deviation.
In the case of chordal graphs, we were able to capitalise on the existence of a closed from expression for the MLE. We proposed an unbiased estimator of the Cross Entropy with an explicit formula and no need for additional data: the Unbiased Chordal Explicit Estimator (UCEE). We proved theoretical bounds on the performances of the selected models by the criterion.

We compared empirically the UCEE to the Cross Validation Cross Entropy (CVCE) of Chapter 3. The CVCE is completely general estimator, agnostic to the graph properties or the formula, but has proven its worth on general graphs. On synthetic data, we demonstrated how the UCEE consistently selects better performing models that are closer to the optimum. The CVCE, however, is not far behind, with some of its train/validation splits reaching similar performances. On real data, UCEE and the de facto optimal train/validation split for the CVCE achieved equivalent Out of Sample performances.
Overhaul, if we consider CVCE with its best split size, both methods performed similarly. However, this optimal split of Cross Validation is an unknown hyper-parameter that remains to be found by fine tuning, whereas UCEE, in addition to performing well, has a non-parametric formula, saving the trouble of running an additional study. This makes, in our opinion, UCEE the most relevant criterion to use with chordal graphs.
The UCEE was tested on a database of diffeomorphic deformations of the hippocampus anatomical shape. The results show that the deformation is not simple and that the atrophy is not random but has a particular structural pattern. In addition, we were able to recover previous results based on GGMselect where a prior informative graph was added to the model. For population comparison, we were able to highlight differences in the deformation pattern. These will have to be further investigated in order to exhibit potentially new pathological effect of the disease.

4.8 Proofs

In this section we prove the different results presented in the paper.

4.8.1 Lemmas

We start with two lemmas needed for the subsequent proofs.

Lemma 4.8.1. *For any $\lambda \geq 0$, let $S_1^{(\lambda)} = S_1 + \lambda I_p$. When $\widehat{K}_m$ as defined in (4.8) exists, we have:*

$$\forall m \in \mathcal{M}, \quad \left\langle S_1^{(\lambda)}, \widehat{K}_m \right\rangle = p. \tag{4.41}$$

Proof. Let Π_m be the orthogonal projection on the edge set E_m. A property of the MLE is that $\Pi_m(\widehat{\Sigma}_m) = \Pi_m(S_1^{(\lambda)})$, i.e. the matrices have the same values on the edge set [35]. Additionally, because of the sparsity of $\widehat{K}_m$, we have from lemma 4.2.1 that for any matrix M, $\left\langle M, \widehat{K}_m \right\rangle =$

$\left\langle \Pi_m(M), \widehat{K}_m \right\rangle$. Then:

$$\begin{aligned}
\left\langle S_1^{(\lambda)}, \widehat{K}_m \right\rangle &= \left\langle \Pi_m(S_1^{(\lambda)}), \widehat{K}_m \right\rangle \\
\left\langle S_1^{(\lambda)}, \widehat{K}_m \right\rangle &= \left\langle \Pi_m(\widehat{\Sigma}_m), \widehat{K}_m \right\rangle \\
\left\langle S_1^{(\lambda)}, \widehat{K}_m \right\rangle &= \left\langle \widehat{\Sigma}_m, \widehat{K}_m \right\rangle \\
\left\langle S_1^{(\lambda)}, \widehat{K}_m \right\rangle &= p\,.
\end{aligned}$$

□

Lemma 4.8.2. *For any $\lambda > 0$, with $\widehat{K}_m$ as defined in (4.8), we have:*

$$\left\| \widehat{K}_m \right\|_* \leq \frac{p}{\lambda}\,.$$

Proof. We have:

$$\begin{aligned}
\left\langle S_1 + \lambda I_p, \widehat{K}_m \right\rangle &= p \\
\left\langle S_1, \widehat{K}_m \right\rangle + \lambda tr(\widehat{K}_m) &= p \\
tr\left(\widehat{K}_m^{\frac{1}{2}} S_1 \widehat{K}_m^{\frac{1}{2}} \right) + \lambda tr(\widehat{K}_m) &= p\,.
\end{aligned}$$

Since $\widehat{K}_m^{\frac{1}{2}} S_1 \widehat{K}_m^{\frac{1}{2}} \in S_p^+$, we have $tr\left(\widehat{K}_m^{\frac{1}{2}} S_1 \widehat{K}_m^{\frac{1}{2}} \right) \geq 0$ and $\lambda tr(\widehat{K}_m) \leq p$, i.e.

$$\left\| \widehat{K}_m \right\|_* \leq \frac{p}{\lambda}\,.$$

□

4.8.2 Proof of proposition 4 (Section 4.2.2)

Let $\lambda_i := \lambda_i(\widetilde{K}^{\frac{1}{2}} \Sigma \widetilde{K}^{\frac{1}{2}}) = \lambda_i(\Sigma^{\frac{1}{2}} \widetilde{K} \Sigma^{\frac{1}{2}})$.

$$H\left(S, \widetilde{\Sigma}\right) = H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{2}\left\langle S - \Sigma, \widetilde{K} \right\rangle\,.$$

First: $\left\langle \Sigma, \widetilde{K} \right\rangle = tr\left(\Sigma \widetilde{K}\right) = tr\left(\Sigma^{\frac{1}{2}} \widetilde{K} \Sigma^{\frac{1}{2}}\right) = \sum_{i=1}^p \lambda_i$.

We have $S = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, with $\forall i, X_i \sim \mathcal{N}(0_p, \Sigma)$ iid. We can write $X_i = \Sigma^{\frac{1}{2}} N_i$ with $\forall i, N_i \sim \mathcal{N}(0_p, I_p)$ iid. Let $N := (N_1, ..., N_n)^T \in \mathbb{R}^{n \times p}$, then $S = \frac{1}{n}\Sigma^{\frac{1}{2}} N^T N \Sigma^{\frac{1}{2}}$. Additionaly, we define the decomposition in orthonormal basis: $\Sigma^{\frac{1}{2}} \widetilde{K} \Sigma^{\frac{1}{2}} = P^T D P$. With $P = (P_1, ..., P_p)^T \in \mathbb{R}^{p \times p}$ the orthonormal transfer matrix, and $D = diag(\lambda_1, ..., \lambda_p)$. We have:

$$
\begin{aligned}
\left\langle S, \widetilde{K} \right\rangle &= \frac{1}{n} tr\left(\Sigma^{\frac{1}{2}} N^T N \Sigma^{\frac{1}{2}} \widetilde{K}\right) \\
&= \frac{1}{n} tr\left(N^T N \, \Sigma^{\frac{1}{2}} \widetilde{K} \Sigma^{\frac{1}{2}}\right) \\
&= \frac{1}{n} tr\left(N^T N \, P^T D P\right) \\
&= \frac{1}{n} tr\left(P N^T N P^T \, D\right) \\
&= \frac{1}{n} \sum_{i=1}^{p} \left(P N^T N P^T\right)_{ii} \, \lambda_i \\
&= \frac{1}{n} \sum_{i=1}^{p} P_i^T N^T N P_i \, \lambda_i \\
&= \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n} (P_i^T N_j)^2 \, \lambda_i \, .
\end{aligned}
$$

We now work conditionally to $\widetilde{\Sigma}$, that is to say conditionally to P. Since P is orthonormal, we have that each component $P_i^T N_j$ follows a standard normal distribution $\mathcal{N}(0,1)$. Additionally, $\left\{P_i^T N_j\right\}_{(i,j)\in[\![1,p]\!]\times[\![1,n]\!]} \in \mathbb{R}^{np}$ is a Gaussian vector $\sim \mathcal{N}(0_{np}, I_{np})$ (see Eq. (4.42)). This implies that all the Gaussian normal variables $P_i^T N_j$ are iid. As a consequence, $\forall i \in [\![1,p]\!]$, $\chi_n^{2\,(i)} := \sum_{j=1}^{n} (P_i^T N_j)^2$ follows a chi square distribution with n degrees of freedom and all the $\chi_n^{2\,(i)}$ are independent.

$$
\left\{P_i^T N_j\right\}_{(i,j)\in[\![1,p]\!]\times[\![1,n]\!]} = \begin{bmatrix}
P_1 & 0 & \cdots & 0 \\
0 & P_1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & P_1 \\
P_2 & 0 & \cdots & 0 \\
0 & P_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & P_2 \\
\vdots & \vdots & \vdots & \vdots \\
P_p & 0 & \cdots & 0 \\
0 & P_p & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & P_p
\end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{bmatrix} \sim \mathcal{N}(0_{np}, I_{np}) \, . \tag{4.42}
$$

In the end, we have:

$$
\left\langle S, \widetilde{K} \right\rangle = \frac{1}{n} \sum_{i=1}^{p} \lambda_i \, \chi_n^{2\,(i)} \, . \tag{4.43}
$$

And

$$
\begin{aligned}
H\left(S, \widetilde{\Sigma}\right) &= H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{2}\left\langle S - \Sigma, \widetilde{K} \right\rangle \\
&= H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{n} \sum_{i=1}^{p} \lambda_i \, \chi_n^{2\,(i)} \, - \sum_{i=1}^{p} \lambda_i \\
&= H\left(\Sigma, \widetilde{\Sigma}\right) + \sum_{i=1}^{p} \lambda_i \, \frac{\chi_n^{2\,(i)} - n}{n} \, .
\end{aligned}
$$

This concludes the proof.

4.8.3 Chordal graph selection with the UCEE (Section 4.3)

Formula and absence of bias of the UCEE (Eq. (4.28) and Eq. (4.30))

From the chordal formula of the MLE (4.27), we get:

$$\left\langle \Sigma, \widehat{K}_m \right\rangle = \sum_{c \in \mathcal{C}_m} \left\langle \Sigma_{cc}, (S_{cc})^{-1} \right\rangle - \sum_{p \in \mathcal{P}_m} \left\langle \Sigma_{pp}, (S_{pp})^{-1} \right\rangle . \tag{4.44}$$

Under the assumption that $n > |c|_{max} + 1$, we can take the expectation of the inverse Wishart matrices in that relation and prove Eq. (4.28):

$$\begin{aligned}
\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_m \right\rangle\right] &= \sum_{c \in \mathcal{C}_m} \left\langle \Sigma_{cc}, \frac{n}{n - |c| - 1} (\Sigma_{cc})^{-1} \right\rangle - \sum_{p \in \mathcal{P}_m} \left\langle \Sigma_{pp}, \frac{n}{n - |p| - 1} (\Sigma_{pp})^{-1} \right\rangle \\
&= \sum_{c \in \mathcal{C}_m} \frac{n\,|c|}{n - |c| - 1} - \sum_{p \in \mathcal{P}_m} \frac{n\,|p|}{n - |p| - 1} =: f(m) .
\end{aligned}$$

We use this $f(m)$ as our bias correction term to define the UCEE:

$$H_m = H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2}(f(m) - p) \equiv H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2} f(m) .$$

Using the definition of $f(m)$ and the fact that $H\left(\Sigma, \widehat{\Sigma}_m\right) = H\left(S, \widehat{\Sigma}_m\right) + \frac{1}{2}(\left\langle \Sigma, \widehat{K}_m \right\rangle - p)$, we prove Eq. (4.30), the absence of bias property of the UCEE:

$$\mathbb{E}\left[H_m - H\left(\Sigma, \widehat{\Sigma}_m\right)\right] = \frac{1}{2}\left(f(m) - \mathbb{E}\left[\left\langle \Sigma, \widehat{K}_m \right\rangle\right]\right) = 0 .$$

Controls on the UCEE solutions (Eq. (4.31), (4.32) and (4.33))

By definition (4.12) of $\hat{m}$ we get a control (4.45). Note that Since $\hat{m}$ is a random variable $\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle\right] \neq f(\hat{m})$, but instead: $f(\hat{m}) = \left\langle \Sigma, \mathbb{E}\left[\widehat{K}_m\right]|_{m=\hat{m}} \right\rangle$.

$$\forall m \in \mathcal{M},\ H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right) \leq H\left(\Sigma, \widehat{\Sigma}_m\right) + \frac{1}{2}\left(\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right) - \frac{1}{2}\left(\left\langle \Sigma, \widehat{K}_m \right\rangle - f(m)\right) . \tag{4.45}$$

From (4.45) we can get two controls: one with the best expected score of any fixed model: $\min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right]$, an another with the best score on the data $\min_{m \in \mathcal{M}} H\left(\Sigma, \widehat{\Sigma}_m\right) = H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)$. For the first result, recall that $\forall m \in \mathcal{M},\ \mathbb{E}\left[\left\langle \Sigma, \widehat{K}_m \right\rangle\right] = f(m)$, then by taking the expectation in (4.45), we get both Eq. (4.31) and Eq. (4.32):

$$\begin{aligned}
&\forall m \in \mathcal{M},\ \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right] \\
&\quad\Longrightarrow\ \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right] \\
&\qquad\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\max_{m \in \mathcal{M}} \left(\left\langle \Sigma, \widehat{K}_m \right\rangle - f(m)\right)\right] \\
&\qquad\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_m \right\rangle - f(m)\right|\right] \\
&\qquad\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] \leq \min_{m \in \mathcal{M}} \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_m\right)\right] + \frac{1}{2}\mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right] \right\rangle\right|\right] .
\end{aligned}$$

To prove the control with the best model, Eq. (4.33), with fixed data $\hat{m}^*$, we apply (4.45) to $m = \hat{m}^*$:

$$
\begin{aligned}
H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right) &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) + \frac{1}{2}\left(\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right) - \left(\frac{1}{2}\left\langle \Sigma, \widehat{K}_{\hat{m}^*} \right\rangle - f(\hat{m}^*)\right) \\
H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right) &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) + \frac{1}{2}\left|\left\langle \Sigma, \widehat{K}_{\hat{m}} \right\rangle - f(\hat{m})\right| + \frac{1}{2}\left|\left\langle \Sigma, \widehat{K}_{\hat{m}^*} \right\rangle - f(\hat{m}^*)\right| \\
H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right) &\leq H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right) + \max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_{m} \right\rangle - f(m)\right| \\
\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] &\leq \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)\right] + \mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_{m} \right\rangle - f(m)\right|\right] \\
\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}}\right)\right] &\leq \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\hat{m}^*}\right)\right] + \mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_{m} - \mathbb{E}\left[\widehat{K}_m\right] \right\rangle\right|\right] .
\end{aligned}
$$

Order of the controls (Eq. (4.34) and (4.35))

We start by showing the following lemma for a standard inverse Wishart distribution:

Lemma 4.8.3. *Let $n\,W \sim \mathcal{W}_p\left(I_p, n\right)$ p follow a centred Wishart distribution. Then:*

$$
Var\left[tr\left(W^{-1}\right)\right] = \frac{2n^2(n-1)p}{(n-p)(n-p-1)^2(n-p-3)} . \tag{4.46}
$$

Proof. Let $D = \mathrm{diag}(W^{-1}) \in \mathbb{R}^p$ be the vector of diagonal coefficients of W^{-1}. Then $tr\left(W^{-1}\right) = \mathbb{1}^T.D$, and:

$$
\begin{aligned}
Var\left[tr\left(W^{-1}\right)\right] &= Var\left[\mathbb{1}^T.D\right] \\
&= \mathbb{1}^T.\mathrm{Cov}(D).\mathbb{1} \\
&= \sum_{i,j \in [\![1,p]\!]^2} \mathrm{Cov}(D_i, D_j) .
\end{aligned}
$$

Since $\frac{W^{-1}}{n} \sim \mathcal{W}_p^{-1}\left(I_p, n\right)$ is an inverse Wishart, we have the formula:

$$
\mathrm{Cov}(D_i, D_j) = n^2 \frac{2 + 2(n-p-1)\mathbb{1}_{i=j}}{(n-p)(n-p-1)^2(n-p-3)} .
$$

Then:

$$
\begin{aligned}
Var\left[tr\left(W^{-1}\right)\right] &= \sum_{i,j \in [\![1,p]\!]^2} n^2 \frac{2 + 2(n-p-1)\mathbb{1}_{i=j}}{(n-p)(n-p-1)^2(n-p-3)} \\
&= 2n^2 p \frac{p + (n-p-1)}{(n-p)(n-p-1)^2(n-p-3)} \\
&= \frac{2n^2(n-1)p}{(n-p)(n-p-1)^2(n-p-3)} .
\end{aligned}
$$

□

Theorem 4.8.4. *If $n > |c|_{max} + 3$, then:*

$$
\mathbb{E}\left[\max_{m \in \mathcal{M}} \left|\left\langle \Sigma, \widehat{K}_{m} - \mathbb{E}\left[\widehat{K}_m\right] \right\rangle\right|\right] \leq \sum_{c \in \mathcal{C}_{max}} \sqrt{2n^3} \frac{\sqrt{|c|}}{(n - |c| - 3)^2} + \sum_{p \in \mathcal{P}_{max}} \sqrt{2n^3} \frac{\sqrt{|p|}}{(n - |p| - 3)^2} . \tag{4.47}
$$

Proof. For any $m \in \mathcal{M}$, we have, from (4.44):

$$\begin{aligned}\left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right| &= \left|\sum_{c\in\mathcal{C}_m}\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle - \sum_{p\in\mathcal{P}_m}\left\langle \Sigma_{pp}, (S_{pp})^{-1} - \mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right| \\ &\leq \sum_{c\in\mathcal{C}_m}\left|\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right| + \sum_{p\in\mathcal{P}_m}\left|\left\langle \Sigma_{pp}, (S_{pp})^{-1} - \mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right| \\ &\leq \sum_{c\in\mathcal{C}_{max}}\left|\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right| + \sum_{p\in\mathcal{P}_{max}}\left|\left\langle \Sigma_{pp}, (S_{pp})^{-1} - \mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right| .\end{aligned}$$

Since this is true for any model m, we have:

$$\max_{m\in\mathcal{M}}\left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right| \leq \sum_{c\in\mathcal{C}_{max}}\left|\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right| + \sum_{p\in\mathcal{P}_{max}}\left|\left\langle \Sigma_{pp}, (S_{pp})^{-1} - \mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right| .$$

The sets $\mathcal{C}_{max}$ and $\mathcal{P}_{max}$ are not random variables, meaning that the expectation symbol can enter the sums:

$$\begin{aligned}\mathbb{E}\left[\max_{m\in\mathcal{M}}\left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right|\right] \leq &\sum_{c\in\mathcal{C}_{max}}\mathbb{E}\left[\left|\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right|\right] \\ &+ \sum_{p\in\mathcal{P}_{max}}\mathbb{E}\left[\left|\left\langle \Sigma_{pp}, (S_{pp})^{-1} - \mathbb{E}\left[(S_{pp})^{-1}\right]\right\rangle\right|\right] .\end{aligned} \tag{4.48}$$

We control each term of the sums separately with Jensen's Inequality:

$$\mathbb{E}\left[\left|\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle\right|\right] \leq \mathbb{E}\left[\left\langle \Sigma_{cc}, (S_{cc})^{-1} - \mathbb{E}\left[(S_{cc})^{-1}\right]\right\rangle^2\right]^{\frac{1}{2}} .$$

Which can be rewritten with the trace instead of the scalar product:

$$\mathbb{E}\left[\left|tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right) - \mathbb{E}\left[tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right)\right]\right|\right] \leq \mathbb{E}\left[\left(tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right) - \mathbb{E}\left[tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right)\right]\right)^2\right]^{\frac{1}{2}} . \tag{4.49}$$

We have:

$$tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right) = tr\left(\left(\Sigma_{cc}^{-\frac{1}{2}} S_{cc} \Sigma_{cc}^{-\frac{1}{2}}\right)^{-1}\right) .$$

Let $W := \Sigma_{cc}^{-\frac{1}{2}} S_{cc} \Sigma_{cc}^{-\frac{1}{2}}$. Since $nS_{cc} \sim \mathcal{W}_p\left(\Sigma_{cc}, n\right)|c|$, then $nW \sim \mathcal{W}_p\left(I_{|c|}, n\right)|c|$. Then (4.49) becomes:

$$\begin{aligned}\mathbb{E}\left[\left|tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right) - \mathbb{E}\left[tr\left(\Sigma_{cc}\left(S_{cc}\right)^{-1}\right)\right]\right|\right] &\leq \mathbb{E}\left[\left(tr\left(W^{-1}\right) - \mathbb{E}\left[tr\left(W^{-1}\right)\right]\right)^2\right]^{\frac{1}{2}} \\ &= Var\left[tr\left(W^{-1}\right)\right]^{\frac{1}{2}} \\ &= \left(\frac{2n^2(n-1)|c|}{(n-|c|)(n-|c|-1)^2(n-|c|-3)}\right)^{\frac{1}{2}} ,\end{aligned}$$

where we applied the previous lemma. We can plug this result in (4.48) to get:

$$\begin{aligned}\mathbb{E}\left[\max_{m\in\mathcal{M}}\left|\left\langle \Sigma, \widehat{K}_m - \mathbb{E}\left[\widehat{K}_m\right]\right\rangle\right|\right] &\leq \sum_{c\in\mathcal{C}_{max}}\left(\frac{2n^2(n-1)|c|}{(n-|c|)(n-|c|-1)^2(n-|c|-3)}\right)^{\frac{1}{2}} \\ &\quad + \sum_{p\in\mathcal{P}_{max}}\left(\frac{2n^2(n-1)|p|}{(n-|p|)(n-|p|-1)^2(n-|p|-3)}\right)^{\frac{1}{2}} \\ &\leq \sum_{c\in\mathcal{C}_{max}}\sqrt{2n^3}\frac{\sqrt{|c|}}{(n-|c|-3)^2} + \sum_{p\in\mathcal{P}_{max}}\sqrt{2n^3}\frac{\sqrt{|p|}}{(n-|p|-3)^2} .\end{aligned} \tag{4.50}$$

□

In the end, we directly apply this theorem to the controls (4.32) and (4.33) to get the upper bounds (4.34) and (4.35) of their orders:

$$\mathbb{E}\left[H\left(\Sigma,\widehat{\Sigma}_{\hat{m}}\right)\right]-\min_{m\in\mathcal{M}}\mathbb{E}\left[H\left(\Sigma,\widehat{\Sigma}_{m}\right)\right]\leq\sum_{c\in\mathcal{C}_{max}}\sqrt{\frac{n^3}{2}}\frac{\sqrt{|c|}}{(n-|c|-3)^2}+\sum_{p\in\mathcal{P}_{max}}\sqrt{\frac{n^3}{2}}\frac{\sqrt{|p|}}{(n-|p|-3)^2}\,,$$

$$\mathbb{E}\left[H\left(\Sigma,\widehat{\Sigma}_{\hat{m}}\right)-H\left(\Sigma,\widehat{\Sigma}_{\hat{m}^*}\right)\right]\leq\sum_{c\in\mathcal{C}_{max}}\sqrt{2n^3}\frac{\sqrt{|c|}}{(n-|c|-3)^2}+\sum_{p\in\mathcal{P}_{max}}\sqrt{2n^3}\frac{\sqrt{|p|}}{(n-|p|-3)^2}\,.$$

Chapter 5

Deterministic Approximate EM algorithm; Application to the Riemann approximation EM and the tempered EM

This Chapter has been submitted for review.

The Expectation Maximisation (EM) algorithm is widely used to optimise non-convex likelihood functions with hidden variables. Many authors modified its simple design to fit more specific situations. For instance the Expectation (E) step has been replaced by Monte Carlo (MC) approximations, Markov Chain Monte Carlo approximations, tempered approximations... Most of the well studied approximations belong to the stochastic class. By comparison, the literature is lacking when it comes to deterministic approximations. In this paper, we introduce a theoretical framework, with state of the art convergence guarantees, for any deterministic approximation of the E step. We analyse theoretically and empirically several approximations that fit into this framework. First, for cases with intractable E steps, we introduce a deterministic alternative to the MC-EM, using Riemann sums. This method is easy to implement and does not require the tuning of hyper-parameters. Then, we consider the tempered approximation, borrowed from the Simulated Annealing optimisation technique and meant to improve the EM solution. We prove that the the tempered EM verifies the convergence guarantees for a wide range of temperature profiles. We showcase empirically how it is able to escape adversarial initialisations. Finally, we combine the Riemann and tempered approximations to accomplish both their purposes, and prove that the resulting algorithm still benefits from the convergence guarantees.

5.1 Introduction

The Expectation Maximisation (EM) algorithm was introduced by Dempster et al [36] to maximise likelihood functions $g(\theta)$ defined from inherent hidden variables z that were non-convex and had intricate gradients and Hessians. In addition to presenting the method, Dempster et al [36] provides convergence guarantees on the sequence of estimated parameters $\{\theta_n\}_n$, namely that it converges towards a critical point of the likelihood function. More convergence guarantees were studied by Boyles [16]. Some likelihood functions are too complex to apply Dempster's raw version of the EM. As a consequence, authors of later works have proposed alternative versions, usually with new con-

vergence guarantees. On the one hand, when the maximisation step (M step) is problematic, other optimisation methods such as coordinate descent [169] or gradient descent [90] have been proposed. On the other hand, several works introduce new versions of the algorithm where the expectation step (E step), which can also be intractable, is approximated. Most of them rely on Monte Carlo (MC) methods and stochastic approximations to estimate this expectation. Notable examples include Delyon, Lavielle and Moulines [33] with the SAEM, Wei and Tanner [166] for the MC-EM, Fort and Moulines [49], the MCMC-EM, Khun and Lavielle [87], the MCMC-SAEM, and Chevalier and Allassonnière [2] for the Approximate SAEM. All these variants come with their own theoretical convergence guarantees for the models of the exponential family. These stochastic approximations constitute an extensive catalogue of methods. Indeed, there are many possible variants of MCMC samplers that can be considered, as well as the additional parameters, such as the "burn-in" period length and the gain sequence decrease, that have to be set. All these choices have an impact on the convergence of the EM and making the appropriate one for each problem can be overwhelming, see [15, 96, 97], among others, for discussions on tuning the MC-EM alone. On several cases, one might desire to dispose of a simpler method, possibly non-stochastic, and non-parametric to run an "EM-like" algorithm for models with no closed forms. However the literature is lacking in that regards. The Quasi-Monte Carlo EM, introduced by Pan and Thompson [126], is a deterministic version of Monte Carlo EM, however theoretical guarantees are not provided. In that vein, Jank [73] introduces the randomised Quasi-Monte Carlo EM, which is not deterministic, and does not have theoretical guarantees either.

Other types of deterministic approximations of the E step have been proposed with the aim to improve the solutions of the algorithm. One notable example is the tempering (or "annealing") of the conditional probability function. Instead of making the problem tractable, the tempering approximation is used to find better local maxima of the likelihood profile during the optimisation process, in the spirit of the simulated annealing [82] and parallel tempering (annealing MCMC) [53, 149]. The deterministic annealing EM was introduced by Ueda and Nakano [156] with a decreasing temperature profile; another temperature profile was proposed in [120]. Contrary to most of the studies on stochastic approximations, these two works do not provide theoretical convergence guarantees for the proposed tempered methods. Which, as a consequence, does not provide insight on the choice of the temperature scheme. Moreover, the tempered methods do not allow the use of the EM in case of an intractable E step. In their tempered SAEM algorithm, Chevallier and Allassonnière [2] combine the stochastic and tempering approximations, which allows the EM to run, even with an intractable E step, while benefiting from the improved optimisation properties brought by the tempering. In addition, theoretical convergence guarantees are provided. However, this method is once again stochastic and parametric.

Overall, most of the literature on approximated E steps focuses on stochastic approximations that estimate intractable conditional probability functions. The few purely deterministic approximations proposed, such as the tempered/annealed EM, are used for other purposes, improving the optimisation procedure, and lack convergence guarantees.

In this paper, we propose a new, unified class of EM with deterministic approximations of the E step. We prove that members of this class benefit from the state of the art theoretical convergence guarantees of [33, 90, 169], under mild regularity conditions on the approximation. Then, we provide examples of approximations that fall under this framework and have practical applications. First, for E steps without closed form, we propose to use Riemann sums to estimate the intractable normalising factor. This "Riemann approximation EM" is a deterministic, less parametric, alternative to the MC-EM and its variants. Second, we prove that the deterministic annealed EM (or "tempered EM") of [156] is a member of our general deterministic class as well. We prove that the convergence guarantees are achieved with almost no condition of the temperature scheme, justifying the use of a wider range of temperature profile than those proposed in [156] and [120]. Finally, since the Riemann and tempered approximations are two separate methods that fulfil very different practical purposes, we also propose to associate the two approximations in the "tempered Riemann approximation EM" when both their benefits are desired.

In section 5.2, we introduce our general class of deterministic approximated versions of the EM algorithm and prove their convergence guarantees, for models of the exponential family. We discuss

the "Riemann approximation EM" in section 5.3, the "tempered EM" in section 5.4, and their association, "tempered Riemann approximation EM", in section 5.5.
We demonstrate empirically that the Riemann EM converges properly on a model with and an intractable E step, and that adding the tempering to the Riemann approximation allows in addition to get away from the initialisation and recover the true parameters. On a tractable Gaussian Mixture Model, we compare the behaviours and performances of the tempered EM and the regular EM. In particular, we illustrate that the tempered EM is able to escape adversarial initialisations, and consistently reaches better values of the likelihood than the unmodified EM, in addition to better estimating the model parameters.

5.2 Deterministic Approximate EM algorithm and its convergence for the curved exponential family

5.2.1 Context and motivation

In this section, we propose a new class of deterministic EM algorithms with approximated E step. This class of algorithms is general and includes both methods that estimate intractable E steps as well as methods that strive to improve the algorithm's solution. We prove that members of this class benefit from the same convergence guarantees that can be found in the state of the art references [33, 90, 169] for the classical EM algorithm, and under similar model assumptions. The only condition on the approximated distribution being that it converges towards the real conditional probability distribution with a certain l_2 regularity. Like the authors of [2, 33, 49], we work with probability density functions belonging to the curved exponential family. The specific properties of which are given in the hypothesis M1 of theorem 5.2.1.

The general framework of the EM is the following: a random variable x has a probability density function with natural parameter $\theta \in \Theta \subset \mathbb{R}^l$. We observe independent and identically distributed (iid) realisations of the distribution: $(x_1, ..., x_n)$ and wish to maximise with respect to θ the resulting likelihood, which is noted $g(\theta)$. In the notations and the discourse, we mostly ignore x as a variable since the observations $(x_1, ..., x_n)$ are supposed fixed throughout the reasoning. We assume there exists a hidden variable z informing the behaviour of the observed variable x such that $g(\theta)$ is the integral of the complete likelihood $h(z;\theta)$: $g(\theta) = \int_z h(z;\theta)\mu(dz)$, with μ the reference measure. The conditional density function of z is then $p_\theta(z) := h(z;\theta)/g(\theta)$.
The foundation of the EM algorithm is that while $ln\, g(\theta)$ is hard to maximise in θ, the functions $\theta \mapsto ln\, h(z;\theta)$ and even $\theta \mapsto \mathbb{E}_z\left[ln\, h(z;\theta)\right]$ are easier to work with because of the information added by the hidden variable z (or its distribution). In practice however, the actual value of z is unknown and its distribution $p_\theta(z)$ dependent on θ. Hence, the EM was introduced in [36] as the two-stages procedure starting from an initial point θ_0 and iterated over the number of steps n:

(E) With the current parameter θ_n, calculate the conditional probability $p_{\theta_n}(z)$;

(M) To get θ_{n+1}, maximise in $\theta \in \Theta$ the function $\theta \mapsto \mathbb{E}_{z\sim p_{\theta_n}(z)}\left[ln\, h(z;\theta)\right]$;

Which can be summarised as:

$$\theta_{n+1} := T(\theta_n) := \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{z\sim p_{\theta_n}(z)}\left[ln\, h(z;\theta)\right] . \tag{5.1}$$

Where we call T the point to point map in Θ corresponding to one EM step. We will not redo the basic theory of the exact EM here, but this procedure noticeably increase $g(\theta_n)$ at each new step n. However, in some cases, one may prefer or have to use an approximation of $p_{\theta_n}(z)$ instead of the exact analytical value. The authors of [2, 33, 49, 87] for instance cannot compute this probability in closed form and resort to stochastic approximation instead. The authors of [120, 156] use a deterministic tempered approximation to reach better critical points. Finally the authors of [2] combine the two approaches, with a stochastic tempered approximation.
In the following, we consider a deterministic approximation of $p_\theta(z)$ noted $\tilde{p}_{\theta,n}(z)$ which depends

on the current step n and on which we make no assumption at the moment. The resulting steps, defining the "Approximate EM", can be written under the same form as eq. (5.1):

$$\theta_{n+1} := F_n(\theta_n) := \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{z\sim\tilde{p}_{\theta_n,n}(z)}\left[ln\, h(z;\theta)\right] . \tag{5.2}$$

Where $\{F_n\}_{n\in\mathbb{N}}$ is the sequence of point to point maps in Θ associated with the sequence of approximations $\{\tilde{p}_{\theta,n}(z)\}_{n\in\mathbb{N}}$. As done in [49] with their stochastic approximation, we add a slight modification in order to ensure the desired convergence guarantees: truncation with increasing compact sets. Assume that you dispose of an increasing sequence of compacts $\{K_n\}_{n\in\mathbb{N}}$ such that $\cup_{n\in\mathbb{N}}K_n = \Theta$ and $\theta_0 \in K_0$. Define $j_0 := 0$. Then, the transition $\theta_{n+1} = F_n(\theta_n)$ is accepted only if $F_n(\theta_n)$ belongs to the current compact K_{j_n}, otherwise the sequence is reinitialised at θ_0. The steps of this algorithm, called "Stable Approximate EM", can be written as:

$$\begin{cases} if F_n(\theta_n) \in K_{j_n}, \text{ then} & \theta_{n+1} = F_n(\theta_n), \text{ and } j_{n+1} := j_n \\ if F_n(\theta_n) \notin K_{j_n}, \text{ then} & \theta_{n+1} = \theta_0, \text{ and } j_{n+1} := j_n + 1 \end{cases} \tag{5.3}$$

This re-initialisation of the EM sequence may seem like a hurdle, however, this truncation is mostly a theoretical requirement. In practice, the first compact K_0 is taken so large that it covers the most probable areas of Θ and the algorithms eq. (5.2) and eq. (5.3) are identical as long as the sequence $\{\theta_n\}_n$ does not diverge towards the border of Θ.

5.2.2 Theorem

In the following, we will state the convergence theorem of Equation (5.3) and provide a brief description of the main steps of the proof.

Theorem 5.2.1 (Convergence of the Stable Approximate EM)**.** *Let $\{\theta_n\}_{n\in\mathbb{N}}$ be a sequence of the Stable Approximate EM defined in Equation (5.3). Let us assume two sets of hypotheses:*

- ***The** $M1-3$ **conditions of [49].***

  M1. *$\Theta \subseteq \mathbb{R}^l$, $\mathcal{X} \subseteq \mathbb{R}^d$ and μ is a σ-finite positive Borel measure on $\mathcal{X}$. Let $\psi : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$ and $S : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}^q$. Define $L : \mathcal{S} \times \Theta \to \mathbb{R}$ and $h : \mathcal{X} \times \Theta \to \mathbb{R}_+ \setminus \{0\}$:*

  $$L(s;\theta) := \psi(\theta) + \langle s, \phi(\theta)\rangle\,, \quad h(z;\theta) := exp(L(S(z);\theta))\,.$$

  M2. *Assume that*

  (a) ψ and ϕ are continuous on Θ;*

  (b) for all $\theta \in \Theta$, $\bar{S}(\theta) := \int_z S(z)p_\theta(z)\mu(dz)$ is finite and continuous on Θ;

  (c) there exists a continuous function $\hat{\theta} : \mathcal{S} \to \Theta$ such that for all $s \in \mathcal{S}$, $L(s;\hat{\theta}(s)) = \sup_{\theta\in\Theta} L(s;\theta)$;

  (d) g is positive, finite and continuous on Θ and, for any $M > 0$, the level set $\{\theta \in \Theta, g(\theta) \geq M\}$ is compact.

  M3. *Assume either that:*

  (a) The set $g(\mathcal{L})$ is compact or

  (a′) for all compact sets $K \subseteq \Theta$, $g\left(K \cap \mathcal{L}\right)$ is finite.

- ***The conditions on the approximation.*** *Assume that $\tilde{p}_{\theta,n}(z)$ is deterministic. Let $S(z) = \{S_i(z)\}_{i=1,\dots,q}$. For all indices i, for any compact set $K \subseteq \Theta$, one of the two following configurations holds:*

  $$\int_z S_i^2(z)dz < \infty \text{ and } \sup_{\theta\in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\,\infty}{\longrightarrow} 0\,. \tag{5.4}$$

  Or

  $$\sup_{\theta\in K} \int_z S_i^2(z)p_\theta(z)dz < \infty \text{ and } \sup_{\theta\in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0\,. \tag{5.5}$$

Then,

(i) *(a) With probability 1,* $\underset{n\infty}{lim}\, j_n < \infty$ *and* $\underset{n\in\mathbb{N}}{sup} \|\theta_n\| < \infty$*;*

(b) $g(\theta_n)$ *converges towards a connected component of* $g(\mathcal{L})$*.*

(ii) If, additionally, $g\left(\mathcal{L} \cap Cl\left(\{\theta_n\}_{n\in\mathbb{N}}\right)\right)$ *has an empty interior, then:*

$$g(\theta_n) \underset{n\infty}{\longrightarrow} g^*\,,$$
$$d(\theta_n, \mathcal{L}_{g^*}) \underset{n\infty}{\longrightarrow} 0\,.$$

Where $\mathcal{L} := \{\theta \in \Theta | \nabla g(\theta) = 0\}$ and $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L} | g(\theta) = g^*\}$.

Remark.

- M2(a) is modified with regards to [49], we remove the hypothesis that S has to be a continuous function of z that is not needed when the approximation is not stochastic. We call M2 (a*) this new sub-hypothesis.
- The condition $\int_z S_i^2(z)dz < \infty$ of the condition eq. (5.4) can seem hard to verify since S is not integrated against a probability function. However, when z is a finite variable, as is the case for finite mixtures, this integral becomes a finite sum.
- The two sufficient conditions eq. (5.4) and eq. (5.5) involve a certain form of integral l_2 convergence of $\tilde{p}_{\theta,n}$ towards p_θ. If the hidden variable z is continuous, this excludes countable (and finite) approximations such as sums of Dirac functions, since they have a measure of zero. In particular, this excludes Quasi-Monte Carlo approximations. However, one look at the proof of the theorem (in section 5.6.1) or at the following sketch of proof reveals that having for any compact set K, $\underset{\theta\in K}{sup} \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\| \underset{n\infty}{\longrightarrow} 0$ is actually a sufficient condition to benefit from the results of theorem 5.2.1. This condition can be verified by finite approximations.

5.2.3 Sketch of proof

The detailed proof of this results can be found in section 5.6.1, we propose here a abbreviated version where we highlight the key steps.
The proof of theorem 5.2.1 follows the same steps as the proof of theorem 3 in [49]. theorem 5.2.1 is the direct consequence of the application of two intermediary propositions introduced and proven in [49]. They are called Propositions 9 and 11 by the authors, and are stated as follows:

Proposition 6 ("Proposition 9")**.** *Let* $\Theta \subseteq \mathbb{R}^l, K$ *compact* $\subset \Theta, \mathcal{L} \subseteq \Theta$ *such that* $\mathcal{L} \cap K$ *compact. Let us assume*

- WC^0 *Lyapunov function with regards to* $(T, \mathcal{L})$*.*
- $\exists u_n \in K^{\mathbb{N}}$ *such that* $|W(u_{n+1}) - W \circ T(u_n)| \underset{n\infty}{\longrightarrow} 0$

Then

- $\{W(u_n)\}_{n\in\mathbb{N}}$ *converges towards a connected component of* $W(\mathcal{L} \cap K)$
- *If* $W(\mathcal{L}\cap K)$ *has an empty interior, then* $\{W(u_n)\}_n$ *converges towards* w^* *and* $\{u_n\}_n$ *converges towards the set* $\mathcal{L}_{w^*} \cap K$

$$\mathcal{L}_{w^*} = \{\theta \in \mathcal{L} | W(\theta) = w^*\}$$

Proposition 7 ("Proposition 11")**.** *Let* $\Theta \subseteq \mathbb{R}^l, T$ *and* $\{F_n\}_n$ *point to point maps on* Θ*. Let* $\{\theta_n\}_n$ *be the sequence defined by the Stable Approximate EM with likelihood* g *and approximate maps sequence* $\{F_n\}_n$*. Let* $\mathcal{L} \subset \Theta$*. We assume*

- *the $A1-2$ conditions of Proposition 10 of [49].*
  - ***(A1)*** *There exists W, a C^0 Lyapunov function with regards to $(T, \mathcal{L})$ such that $\forall M > 0$, $\{\theta \in \Theta, W(\theta) > M\}$ is compact, and:*
$$\Theta = \cup_{n\in\mathbb{N}} \left\{\theta \in \Theta | W(\theta) > n^{-1}\right\} .$$
  - ***(A2)*** *$W(\mathcal{L})$ is compact OR* ***(A2')*** *$W(\mathcal{L} \cap K)$ is finite for all compact $K \subseteq \Theta$.*
- $\forall u \in K_0, \quad \lim_{n\,\infty} |W \circ F_n - W \circ T|(u) = 0$
- $\forall$ *compact* $K \subseteq \Theta, \quad \lim_{n\,\infty} |W \circ F_n(u_n) - W \circ T(u_n)| \mathbb{1}_{u_n \in K} = 0$

Then
$$\textit{With probability 1, } \operatorname*{limsup}_{n\,\infty} j_n < \infty \textit{ and } \{u_n\}_n \textit{ compact sequence}$$

For the proofs of these two results, see [49]. The proof of theorem 5.2.1 is structured as follows: verifying the conditions of proposition 7, applying proposition 7, verifying the conditions of proposition 6 and finally applying proposition 6.

Verifying the conditions of proposition 7. We first make explicit which object of our model plays which part in the Proposition. Let g be the likelihood function of a model of the curved exponential family.

- The set of its critical points is called $\mathcal{L}$: $\mathcal{L} := \{\theta \in \Theta | \nabla g(\theta) = 0\}$.
- We call T the point to point map describing the transition between θ_n and θ_{n+1} in the exact EM algorithm, that is to say $T := \hat{\theta} \circ \bar{S}$.
- The general properties of the EM tell us that its stationary points are the critical points of g: $\mathcal{L} = \{\theta \in \Theta | T(\theta) = \theta\}$. Additionally, we have that g is a C^0 Lyapunov function associated to $(T, \mathcal{L})$, hence it is fit to play the part of W from proposition 7.
- Let $\{\theta_n\}_n$ be the sequence defined by the Stable Approximate EM, and $\{F_n\}_n$ the corresponding sequence of point to point maps.

With this setup, the assumptions M1-3 of theorem 5.2.1 directly imply that A1 and A2 or A2' are verified.

We need to prove that the last two conditions for proposition 7 are verified:
$$\forall \theta \in K_0, \quad \lim_{n\,\infty} |g \circ F_n - g \circ T|(\theta) = 0\,, \tag{5.6}$$
$$\forall \text{ compact } K \subseteq \Theta, \quad \lim_{n\,\infty} |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0. \tag{5.7}$$

We denote $\tilde{S}_n(\theta_n)$ the approximated E step in the Stable Approximate EM (so that $F_n = \hat{\theta} \circ \tilde{S}_n$). By using uniform continuity properties on compacts, we first obtain that
$$\forall \text{ compact } K, \ \sup_{\theta\in K} \left\| \tilde{S}_n(\theta) - \bar{S}(\theta) \right\| \underset{n\,\infty}{\longrightarrow} 0\,, \tag{5.8}$$

is a sufficient condition to obtain both eq. (5.6) and eq. (5.7), and conclude that we can apply proposition 7. Writing $\tilde{S}_n$ and $\bar{S}$ as integrals in z makes it clear that the two hypothesis eq. (5.4) and eq. (5.5) of theorem 5.2.1 are both sufficient to have eq. (5.8). Which concludes this section of the Proof.

Applying proposition 7. Since we verify all the condition of proposition 7, we can apply its conclusion:
$$\text{With probability 1, } \operatorname*{limsup}_{n\,\infty} j_n < \infty \text{ and } \{\theta_n\}_n \text{ compact sequence } ,$$

which is specifically the result $(i)(a)$ of theorem 5.2.1.

Verifying the conditions of proposition 6. With proposition 6, we prove the remaining points of theorem 5.2.1: $(i)(b)$ and (ii).
For the application of proposition 6:

- $Cl\left(\{\theta_n\}_n\right)$ plays the part of the compact K
- $\{\theta \in \Theta | \nabla g(\theta) = 0\} = \{\theta \in \Theta | T(\theta) = \theta\}$ plays the part of the set $\mathcal{L}$
- The likelihood g is the C^0 Lyapunov function with regards to $(T, \mathcal{L})$
- $\{\theta_n\}_n$ is the K valued sequence (since K is $Cl\left(\{\theta_n\}_n\right)$).

The last condition that remains to be shown to apply proposition 6 is that:

$$\lim_{n\,\infty} |g(\theta_{n+1}) - g \circ T(\theta_n)| = 0\,.$$

We have more or less already proven that, in the previous section, with $F_n(\theta_n)$ in place of θ_{n+1}. The only indices where $F_n(\theta_n) \neq \theta_{n+1}$ are when the value of the sequence j_n experiences an increment of 1.

$$|g(\theta_{n+1}) - g \circ T(\theta_n)| = |g(\theta_0) - g \circ T(\theta_n)| \mathbb{1}_{j_{n+1}=j_n+1} \; + \; |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{j_{n+1}=j_n}\,.$$

We have proven with proposition 7 that there is only a finite number of such increments and that $Cl(\{\theta_k\}_k)$ is a compact. Since θ_n is always in $Cl(\{\theta_k\}_k)$ by definition, we can apply to $K := Cl(\{\theta_k\}_k)$ the result:

$$\forall \text{ compact } K \subseteq \Theta, \quad \lim_{n\,\infty} |g \circ F_n(\theta_n) - g \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0\,,$$

that we proved in order to verify proposition 7, and get the needed condition:

$$\lim_{n\,\infty} |g(\theta_{n+1}) - g \circ T(\theta_n)| = 0\,.$$

Applying proposition 6 Since we verify all we need to apply the conclusions of proposition 6:

- $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a connected component of $g(\mathcal{L} \cap Cl(\{\theta_n\}_n)) \subset g(\mathcal{L})$.
- If $g(\mathcal{L} \cap Cl(\{\theta_n\}_n))$ has an empty interior, then $\{g(\theta_n)\}_{n\in\mathbb{N}}$ converges towards a $g^* \in \mathbb{R}$ and $\{\theta_n\}_n$ converges towards $\mathcal{L}_{g^*} \cap Cl(\{\theta_n\}_n)$. Where $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L} | g(\theta) = g^*\}$

Which are both respectively exactly $(i)(b)$ and (ii) of theorem 5.2.1 and concludes the proof of the theorem.

5.3 Riemann approximation EM

5.3.1 Context and motivation

In this section, we introduce one specific case of Approximate EM useful in practice: approximating the conditional probability density function $p_\theta(z)$ at the E step by a Riemann sum, in the scenario where the hidden variable z is continuous and bounded. We call this procedure the "Riemann approximation EM". After motivating this approach, we prove that it is an instance of the Approximate EM algorithm and verifies the hypotheses of theorem 5.2.1, therefore benefits from the convergence guarantees.
When the conditional probability $p_\theta(z)$ is a continuous function, and even if $h(z;\theta)$ can be computed point by point, a closed form may not exist for the re-normalisation term $g(\theta) = \int_z h(z;\theta)dz$. In that case, this integral is usually approximated stochastically with a Monte Carlo estimation, see for instance [2,33,49]. When the dimension is reasonably small, a deterministic approximation through

Riemann sums can also be performed. Unlike the stochastic methods, which often require to define and tune a Markov Chain, the Riemann approximation involves almost no parameter. The user only needs to choose the position of the Riemann intervals, a choice which is very guided by the well known theories of integration (Lagrange, Legendre...).
We introduce the Riemann approximation as a member of the Approximate EM class. Since z is supposed bounded in this section, without loss of generality, we will assume that z is a real variable and $z \in [0,1]$. We recall that $p_\theta(z) = h(z;\theta)/g(\theta) = h(z;\theta)/\int_z h(z;\theta)dz$. Instead of using the exact joint likelihood $h(z;\theta)$, we define a sequence of step functions $\left\{\tilde{h}_n\right\}_{n\in\mathbb{N}^*}$ as: $\tilde{h}_n(z;\theta) := h(\lfloor\varphi(n)z\rfloor/\varphi(n);\theta)$. Where φ is a strictly increasing function from $\mathbb{N}^* \to \mathbb{N}^*$, so that $\varphi(n) \underset{n\to\infty}{\longrightarrow} \infty$. For the sake of simplicity, we will take $\varphi = Id$, hence $\tilde{h}_n(z;\theta) = h(\lfloor nz\rfloor/n;\theta)$. The following results, however, can be applied to any strictly increasing function φ. With these steps functions, the re-normalising factor $\tilde{g}_n(\theta) := \int_z \tilde{h}_n(z;\theta)dz$ is now a finite sum: $\tilde{g}_n(\theta) = \frac{1}{n}\sum_{k=0}^{n-1} h(\lfloor kz\rfloor/n;\theta)$. The approximate conditional probability $\tilde{p}_n(\theta)$ is then naturally defined as: $\tilde{p}_n(\theta) := \tilde{h}_n(z;\theta)/\tilde{g}_n(\theta)$. Thanks to the replacement of the integral by the finite sum, this deterministic approximation is much easier to compute than the real conditional probability.

5.3.2 Theorem and proof

We state and prove the following theorem for the convergence of the EM with a Riemann approximation.

Theorem 5.3.1. *Under conditions $M1-3$ of theorem 5.2.1, and when z is bounded, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{h(\lfloor nz\rfloor/n;\theta)}{\int_{z'} h(\lfloor nz'\rfloor/n;\theta)dz'}$, which we call "Riemann approximation EM", verifies the remaining conditions of applicability of theorem 5.2.1 as long as $z \mapsto S(z)$ is continuous.*

Proof. This is the detailed proof of theorem 5.3.1.
The conditions $M1-3$ on the model are already assumed to be verified. In order to apply theorem 5.2.1, we need to verify either Equation (5.4) or eq. (5.5). Here, with $z \mapsto S(z)$ continuous, we prove Equation (5.4):

$$\int_z S_i^2(z)dz < \infty \text{ and } \forall \text{compact } K \subseteq \Theta, \quad \sup_{\theta\in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\,\infty}{\longrightarrow} 0\,.$$

Since z is bounded (and assumed to be in $[0,1]$ for simplicity) and S is continuous, the first part of the condition is easily verified: $\int_{z=0}^1 S_i^2(z)dz < \infty$. Only the second part remains to be proven.
First we note that $h(z;\theta) = exp(\psi(\theta) + \langle S(z), \phi(\theta)\rangle)$ is continuous in (z,θ), hence uniformly continuous on the compact set $[0,1]\times K$. Additionally, we have:

$$0 < m := \min_{(z,\theta)\in[0,1]\times K} h(z;\theta) \leq h(z;\theta) \leq \max_{(z,\theta)\in[0,1]\times K} h(z;\theta) =: M < \infty\,.$$

Where m and M are constants independent of z and θ. This also means that $m \leq g(\theta) = \int_{z=0}^1 h(z;\theta) \leq M$. Moreover, since $\tilde{h}_n(z;\theta) = h\left(\lfloor nz\rfloor/n;\theta\right)$, then we also have $\forall z \in [0,1], \theta \in K, n \in \mathbb{N}$, $m \leq \tilde{h}_n(z;\theta) \leq M$ and $m \leq \tilde{g}_n(\theta) = \int_{z=0}^1 \tilde{h}_n(z;\theta) \leq M$.

Since h is uniformly continuous, $\forall\epsilon > 0, \exists\delta > 0, \forall(z,z') \in [0,1]^2, (\theta,\theta') \in K^2$:

$$|z - z'| \leq \delta \text{ and } \|\theta - \theta'\| \leq \delta \implies |h(z;\theta) - h(z';\theta')| \leq \epsilon\,.$$

By definition, $\lfloor nz\rfloor/n - z \leq 1/n$. Hence $\exists N \in \mathbb{N}, \forall n \geq N, \lfloor nz\rfloor/n - z \leq \delta$. As a consequence:

$$\forall\epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall(z,\theta) \in [0,1]\times K, \quad \left|h(z;\theta) - \tilde{h}_n(z;\theta)\right| \leq \epsilon\,.$$

In other words, $\{\tilde{h}_n\}_n$ converges uniformly towards h. Let ϵ be given, we assume that $n \geq N$, then $\forall (z, \theta) \in [0,1] \times K$:

$$\begin{aligned}
\tilde{p}_{\theta,n}(z) - p_\theta(z) &= \frac{\tilde{h}_n(z;\theta)}{\int_z \tilde{h}_n(z;\theta)dz} - \frac{h(z;\theta)}{\int_z h(z;\theta)dz} \\
&= \frac{\tilde{h}_n(z;\theta) - h(z;\theta)}{\int_z \tilde{h}_n(z;\theta)dz} + h(z;\theta)\frac{\int_z \left(h(z;\theta) - \tilde{h}_n(z;\theta)\right)dz}{\int_z h(z;\theta)dz \int_z \tilde{h}_n(z;\theta)dz} \\
&\leq \frac{\epsilon}{m} + M\frac{\epsilon}{m^2} \\
&= \epsilon\frac{m+M}{m^2}\,.
\end{aligned}$$

Hence:

$$\forall n \geq N, \quad \sup_{\theta \in K} \int_{z=0}^{1} \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \leq \epsilon^2 \left(\frac{m+M}{m^2}\right)^2 ,$$

By definition, this means that $\sup_{\theta \in K} \int_{z=0}^{1} \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\,\infty}{\longrightarrow} 0$. The last hypothesis needed to apply theorem 5.2.1. Which concludes the proof. □

5.3.3 Application to a Gaussian model with the Beta prior

We demonstrate the interest of the method on an example with a continuous bounded random variable following a Beta distribution $z \sim Beta(\alpha, 1)$, and an observed random variable following $x \sim \mathcal{N}(\lambda z, \sigma^2)$. In other words, with $\epsilon \sim \mathcal{N}(0,1)$ independent of z:

$$x = \lambda z + \sigma\epsilon\,.$$

This results in a likelihood belonging to the exponential family:

$$h(z;\theta) = \frac{\alpha z^{\alpha-1}}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\lambda z)^2}{2\sigma^2}\right).$$

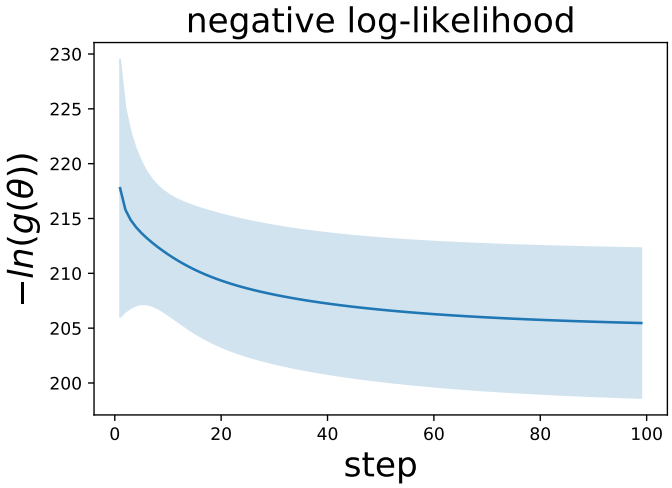
Since z is bounded, and everything is continuous in the parameter $(\alpha, \lambda, \sigma^2)$, this model easily verifies each of the conditions M1-3. The E step with this model involves the integral $\int_z z^\alpha exp\left(-\frac{(x-\lambda z)^2}{2\sigma^2}\right)dz$, a fractional moment of the Gaussian distribution. Theoretical formulas exists for these moments, see [168], however they involve Kummer's confluent hypergeometric functions, which are infinite series. Instead, we use the Riemann approximation to run the EM algorithm with this model: $\tilde{h}_n(z;\theta) := h(\lfloor\varphi(n)z\rfloor/\varphi(n);\theta)$. As done previously, we take, without loss of generality, $\varphi(n) := n$ for the sake of simplicity. The E step only involves the n different values taken by the step function probabilities $h(\lfloor nz\rfloor/n;\theta)$:

$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right) = \frac{h^{(i)}(\frac{k}{n};\theta)}{\frac{1}{n}\sum_{l=0}^{n-1} h^{(i)}(\frac{l}{n};\theta)}\,.$$

Where the exponent (i) indicates the index of the observation x_i. The M step is then written as:

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right)\int_{z=k/n}^{(k+1)/n} ln(z)dz\,, \\
\hat{\lambda} &= \frac{\sum_{i=1}^{N}\sum_{k=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right)\int_{z=k/n}^{(k+1)/n} x_i z dz}{\sum_{i=1}^{N}\sum_{k=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right)\int_{z=k/n}^{(k+1)/n} z^2 dz}\,, \\
\hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{n-1} \tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right)\hat{\lambda}^2\int_{z=k/n}^{(k+1)/n}\left(z - \frac{x_i}{\hat{\lambda}}\right)^2 dz\,.
\end{aligned} \tag{5.9}$$

Where N is the total number of observations: $x := (x_1, ..., x_N)$ iid. We test this algorithm on synthetic data. With real values $\alpha = 2, \lambda = 5, \sigma^2 = 1.5$, we generate a dataset with $N = 100$ observations and run the Riemann EM with random initialisation. This simulation is ran 2000 times. We observe that the Riemann EM is indeed able to increase the likelihood, despite the EM being originally intractable. On fig. 5.1, we display the average trajectory, with standard deviation, of the negative log-likelihood $-ln\left(g(\theta)\right)$ during the Riemann EM procedure. The profile is indeed decreasing. The standard deviation around the average value is fairly high, since each run involves a different dataset and a different random initialisation, hence different value of the likelihood, but the decreasing trend is the same for all of the runs. We also display the average relative square errors on the parameters at the end of the algorithm. They are all small, with reasonably small standard deviation, which indicates that the algorithm consistently recovers correctly the parameters.



(a)

Metric	Average (std)
$\frac{(\alpha-\hat{\alpha})^2}{\alpha^2}$	0.129 (0.078)
$\frac{(\lambda-\hat{\lambda})^2}{\lambda^2}$	0.048 (0.058)
$\frac{(\sigma-\hat{\sigma})^2}{\sigma^2}$	0.108 (0.119)

(b)

Figure 5.1: (Right). Average values, with standard deviation, over 2000 simulations of the negative log-likelihood along the steps of the Riemann EM. The Riemann EM increases the likelihood. (Left). Average and standard deviation of the relative parameter reconstruction errors at the end of the Riemann EM.

5.4 Tempered EM

5.4.1 Context and motivation

In this section, we consider another particular case of Deterministic Approximate EM: the Tempered EM (tmp-EM). We first motivate this algorithm. Then, we prove that under mild conditions, it verifies the hypothesis of theorem 5.2.1, hence has the state of the art EM convergence guarantees. In particular, we prove that the choice of the temperature profile is almost completely free.
When optimising a non-convex function, following the gradients leads to one of the local extrema closest to the initialisation. If the method was allowed to explore more the profile of the function to be optimised, it would encounter points with better values and areas with stronger gradients missed because of its early commitment to one of the nearest potential wells.
A very well known way to encourage such an exploratory behaviour is the tempering, also called annealing. In its simplest form, the function to optimised g is elevated to the power $g^{\frac{1}{T_n}}$, with T_n a temperature tending towards 1 as the number n of steps of the procedure increases. This manipulation equalises the value of the function in the different points of the space, renders the gradients less strong, and makes the potential wells less attractive the higher the temperature T_n is. As a result, the optimisation procedure is not incited to limit itself to its starting region. Additionally, the general shape of the function g, in particular the hierarchy of values, is still preserved, meaning

that the early course of the algorithm is still made on a function replicating the reality. As T_n gets closer to 1, the optimised function becomes identical to g and the potential wells become attractive again. By this point, the assumption is that the algorithm will be in a better place than it was at the initialisation.
These concepts are put in application in many state of the art procedures. The most iconic maybe being the Simulated Annealing, introduced and developed in [1,82,159], where in particular $T_n \longrightarrow 0$ instead of 1. It is one of the few optimisation technique proven to find global optimum of non-convex functions. The Parallel Tempering (or Annealing MCMC) developed in [53, 71, 149] also makes use of these ideas to improve the MCMC simulation of a target probability distribution. The idea of applying a tempering to a classical EM was introduced in the Deterministic Annealed EM of [156] with a specific decreasing temperature scheme. Another specific, non-monotonous, temperature scheme was later proposed by [120]. In both cases, theoretical convergence guarantees are lacking. In [2], tempering is applied to the SAEM, and convergence guarantees are provided with any temperature scheme for this algorithm.
Here, we introduce the tmp-EM as a specific case of the Approximate EM of section 5.2. We use the approximated distribution: $\tilde{p}_{n,\theta}(z) := p_\theta^{\frac{1}{T_n}}(z) / \int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz' = h(z;\theta)^{\frac{1}{T_n}} / \int_{z'} h(z';\theta)^{\frac{1}{T_n}} dz'$ (renormalised to sum to 1). Unlike [156] and [120], we do not specify any temperature scheme T_n, and prove in the following theorem 5.4.1 that, under very mild conditions on the model, any sequence $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$ guarantees the state of the art convergence.

Remark. Elevating $p_\theta(z)$ to the power $\frac{1}{T_n}$, as is done here and in [120, 156], is not equivalent to elevating to the power $\frac{1}{T_n}$ the objective function $g(\theta)$, which would be expected for a typical annealed or tempered optimisation procedure. It is not equivalent either to elevating to the power $\frac{1}{T_n}$ the intermediary function $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ that is optimised in the M step. Instead, the weights $p_{\theta_n}(z)$ (or equivalently, the terms $h(z;\theta_n)$) used in the calculation of $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ are the tempered terms. This still results in the desired behaviour and is only a more "structured" tempering. Indeed, with this tempering, it is the estimated distribution of the hidden variable z that are made less unequivocal, with weaker modes, at each step. This forces the procedure to spend more time considering different configurations for those variables. Which renders as a result the optimised function $\mathbb{E}_{z\sim p_{\theta_n}(z)}[h(z;\theta)]$ more ambiguous regarding which θ is the best, just as intended.

5.4.2 Theorem

We now give the convergence theorem for the Approximate EM with the tempering approximation. In particular, this result highlights that there are almost no constraints on the temperature profile to achieve convergence.

Theorem 5.4.1. *Under conditions $M1-3$ of theorem 5.2.1, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'}$, which we call "Tempered EM", verifies the remaining conditions of applicability of theorem 5.2.1 as long as $T_n \underset{n\infty}{\longrightarrow} 1$ and for any compact $K \in \Theta$, $\exists \epsilon \in]0,1[$, $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$:*

- $\underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z)dz < \infty$
- $\forall i \in [\![1,q]\!]$, $\underset{\theta\in K}{sup} \int_z S_i^2(z) p_\theta^\alpha(z)dz < \infty$

Where $\overline{\mathcal{B}}(1,\epsilon)$ is the closed ball centered in 1 and with radius ϵ in $\mathbb{R}$, and the index i of $S_i(z)$ indicates each of the real component of the $S(z) \in \mathcal{S} \subset \mathbb{R}^q$. The conditions on the integrability of $p_\theta^\alpha(z)$ and $S_i^2(z)p_\theta^\alpha(z)$ brought by the tempering are very mild. Indeed, in section 5.4.4, we will show classical examples that easily verify the much stronger conditions: for any compact $K \in \Theta, \forall \alpha \in \mathbb{R}_+^*$,

$$\underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z)dz < \infty \quad ,$$

$$\forall i \in [\![1,q]\!], \quad \underset{\theta\in K}{sup} \int_z S_i^2(z) p_\theta^\alpha(z)dz < \infty \quad .$$

5.4.3 Sketch of proof

The detailed proof of theorem 5.4.1 can be found in section 5.6.2, we propose here a abbreviated version.
In order to apply theorem 5.2.1, we need to verify five conditions. The three inevitable are M1, M2 and M3. The last two can either be that, $\forall$ compact $K \in \Theta$:

$$\int_z S_i^2(z)dz < \infty \text{ and } \sup_{\theta \in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\,\infty}{\longrightarrow} 0\,.$$

Or

$$\sup_{\theta \in K} \int_z S_i^2(z) p_\theta(z)dz < \infty \text{ and } \sup_{\theta \in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0\,.$$

The hypothesis of theorem 5.4.1 already include M1, M2, M3 and:

$$\forall \text{ compact } K \in \Theta,\ \forall i, \quad \sup_{\theta \in K} \int_z S_i^2(z) p_\theta(z)dz < \infty\,.$$

As a result, to apply theorem 5.2.1, it is sufficient to verify that, with the tempering approximation, we have:

$$\sup_{\theta \in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0\,.$$

The proof of theorem 5.4.1 revolves around proving this result.
With a Taylor development in $\left(\frac{1}{T_n} - 1\right)$, which converges toward 0 when $n \to \infty$, we control the difference $\left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2$:

$$\left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z)\right)^2 \leq 2\left(\frac{1}{T_n} - 1\right)^2 p_\theta(z)^2 \Bigg(\left(ln\, p_\theta(z)\, e^{a(z,\theta,T_n)}\right)^2 A(\theta,T_n) + B(\theta,T_n)\Bigg)\,.$$

The terms $A(\theta,T_n)$, $B(\theta,T_n)$ and $a(z,\theta,T_n)$ come from the Taylor development. With the previous inequality, we control the integral of interest:

$$\begin{aligned}\int_z \frac{\left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z)\right)^2}{p_\theta(z)} dz \leq & 2\left(\frac{1}{T_n} - 1\right)^2 A(\theta,T_n) \int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z)dz \\ & + 2\left(\frac{1}{T_n} - 1\right)^2 B(\theta,T_n)\,.\end{aligned} \tag{5.10}$$

$A(\theta,T_n)$ and $B(\theta,T_n)$ have upper bounds involving $\int_z p_\theta(z)^{\frac{1}{T_n}} ln\, p_\theta(z)$. Similarly, the term $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z)$ is bounded by terms involving $\int_z p_\theta(z)^{\frac{2}{T_n}-1} ln^2\, p_\theta(z)dz$.

Thanks to the hypothesis of the theorem, we prove that for any $\alpha \in \overline{\mathcal{B}}(1,\epsilon)$ and $\theta \in K$ the two terms, $\int_z p_\theta(z)^\alpha ln\, p_\theta(z)$ and $\int_z p_\theta(z)^\alpha ln^2\, p_\theta(z)$ are upper bounded by a constant C independent of θ and α.

Since $T_n \underset{n \mapsto \infty}{\longrightarrow} 1$, then when n is large enough, $\frac{1}{T_n} \in \overline{\mathcal{B}}(1,\epsilon)$ and $\frac{2}{T_n} - 1 \in \overline{\mathcal{B}}(1,\epsilon)$ meaning that the previous result applies to the three terms $A(\theta,T_n)$, $B(\theta,T_n)$ and $\int_z p_\theta(z) e^{2a(z,\theta,T_n)} ln^2\, p_\theta(z)dz$: they are upper bounded by constants C_1, C_2 and C_3 respectively, all independent of θ and T_n.

The inequality eq. (5.10) then becomes:

$$\int_z \frac{1}{p_\theta(z)} \left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z) \right)^2 dz \leq 2 \left(\frac{1}{T_n} - 1 \right)^2 C_1 C_2 + 2 \left(\frac{1}{T_n} - 1 \right)^2 C_3 \,.$$

By taking the supremum in $\theta \in K$ and the limit when $n \longrightarrow \infty$, we get the desired result:

$$\sup_{\theta \in K} \int_z \frac{1}{p_\theta(z)} \left(\frac{p_\theta(z)^{\frac{1}{T_n}}}{\int_{z'} p_\theta(z')^{\frac{1}{T_n}}} - p_\theta(z) \right)^2 dz \underset{n \to \infty}{\longrightarrow} 0 \,.$$

5.4.4 Examples of models that verify the conditions

In this section we illustrate that the conditions of theorem 5.4.1 are easily met by common models. We take two examples, first the Mixture of Gaussian (GMM) where the hidden variables belong to a finite space, then the Poisson count with random effect, where the hidden variables live in a continuous space.
In order to apply theorem 5.4.1, we need to verify the conditions

- $M1$, $M2$ and $M3$
- for any compact $K \in \Theta, \exists \epsilon \in]0,1[, \forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$,

$$\sup_{\theta \in K} \int_z p_\theta^\alpha(z) dz < \infty \quad ,$$
$$\forall i, \quad \sup_{\theta \in K} \int_z S_i^2(z) p_\theta^\alpha(z) dz < \infty \quad .$$

As previously stated, in both examples, we will actually verify the much stronger conditions: for any compact $K \in \Theta, \forall \alpha \in \mathbb{R}_+^*$:

$$\sup_{\theta \in K} \int_z p_\theta^\alpha(z) dz < \infty \text{ and } \forall i, \quad \sup_{\theta \in K} \int_z S_i^2(z) p_\theta^\alpha(z) dz < \infty \,.$$

Gaussian Mixture Model

Despite being one of the most common models the EM is applied to, the GMM have many known irregularities and pathological behaviours, see [153]. As a consequence none of the convergence results of the EM and their variants [33, 49, 90, 169] apply to the GMM. The hypothesis that the GMM fail to verify is the condition that the level lines have to be compact (called M2 (d) in this Chapter and [49]). In all the previously mentioned papers this hypothesis is used to prove that the EM sequence stays within a compact. All is not lost however for the GMM, indeed they verify all the other hypothesis of the convergence theorem (including the new tempering hypothesis introduced in theorem 5.4.1 of this paper). As a result, if an EM sequence applied to a GMM were to stay within a compact, then the convergence theorems would apply (including our theorem 5.4.1 for a tempered EM sequence) and the sequence would be guaranteed to converge towards a critical point of the likelihood function. Hence all that is needed in practice to ensure that there is convergence is to observe that the EM sequence remains in a compact. The GMM belongs to the curved exponential family, the complete likelihood is

$$\begin{aligned} h(z;\theta) = exp\Bigg(\sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\mathbb{1}_{z_i=k}}{2} \Big(& -(x_i - \mu_k)^T \Theta_k (x_i - \mu_k) + ln\left(|\Theta_k|\right) \\ & + 2ln\left(\pi_k\right) - pln\left(2\pi\right) \Big) \Bigg) \,. \end{aligned} \tag{5.11}$$

and the observed likelihood:

$$g(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} exp\left(\frac{1}{2}\left(-(x_i-\mu_k)^T\Theta_k(x_i-\mu_k) + ln\left(|\Theta_k|\right) + 2ln\left(\pi_k\right) - pln\left(2\pi\right)\right)\right). \tag{5.12}$$

This is an exponential model with

$$\theta := \left(\{\pi_k\}_{k=1}^K, \{\mu_k\}_{k=1}^K, \{\Theta_k\}_{k=1}^K\right) \in \left\{ \{\pi_k\}_k \in [0,1]^K \,\middle|\, \sum_k \pi_k = 1 \right\} \otimes \mathbb{R}^{p\times K} \otimes S_p^{++K}.$$

The verification of conditions M1-3 for the GMM (except M2 (d) of course) is a classical exercise since these are the conditions our theorem shares with any other EM convergence result on the exponential family. We focus here on the hypothesis specific to our Deterministic Approximate EM.

Condition on $\int_z p_\theta^\alpha(z)dz$ Let $\alpha \in \mathbb{R}_+^*$, in the finite mixture case, the integrals on z are finite sums:

$$\int_z p_\theta^\alpha(z)dz = \sum_k p_\theta^\alpha(z=k),$$

which is continuous in θ since $\theta \mapsto p_\theta(z=k) = h(z=k;\theta)/g(\theta)$ is continuous. Hence

$$\forall \alpha \in \mathbb{R}_+^*, \quad \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty.$$

Condition on $\int_z S_i^2(z)p_\theta^\alpha(z)dz$ The previous continuity argument is still valid.

Poisson count with random effect

This model is discussed in [49], the authors prove, among other things, that this model verifies the conditions M1-3.
The complete likelihood of the model, not accounting for irrelevant constants, is:

$$h(z;\theta) = e^{\theta\sum_k Y_k}.exp\left(-e^\theta \sum_k e^{z_k}\right). \tag{5.13}$$

$g(\theta) = \int_z h(z;\theta)dz$ can be computed analytically up to a constant:

$$\begin{aligned}
g(\theta) &= \int_{z\in\mathbb{R}^d} h(z;\theta)dz \\
&= e^{\theta\sum_k Y_k} \int_{z\in\mathbb{R}^d} exp\left(-e^\theta \sum_k e^{z_k}\right)dz \\
&= e^{\theta\sum_k Y_k} \prod_{k=1}^{d} \int_{z_k\in\mathbb{R}} exp\left(-e^\theta e^{z_k}\right)dz_k \\
&= e^{\theta\sum_k Y_k} \left(\int_{u\in\mathbb{R}_+} \frac{exp\left(-u\right)}{u}du\right)^d \\
&= e^{\theta\sum_k Y_k} E_1(0)^d,
\end{aligned} \tag{5.14}$$

where $E_1(0)$ is a finite, non zero, constant, called "exponential integral", in particular independent of α and θ.

Condition on $\int_z p_\theta^\alpha(z)dz$ Let K be a compact in Θ.
We have $p_\theta(z) = \frac{h(z;\theta)}{g(\theta)}$. Let us compute $\int_z h(z;\theta)^\alpha$ for any positive α. The calculations work as in Equation (5.14):

$$\int_{z\in\mathbb{R}^d} h(z;\theta)^\alpha = e^{\alpha\theta\sum_k Y_k} \prod_{k=1}^{d} \int_{z_k\in\mathbb{R}} exp\left(-\alpha e^\theta e^{z_k}\right) dz_k = e^{\alpha\theta\sum_k Y_k} E_1(0)^d \,.$$

Hence:

$$\int_z p_\theta^\alpha(z)dz = E_1(0)^{(1-\alpha)d} \,.$$

Since $E_1(0)$ is finite, non zero, and independent of θ, we easily have:

$$\forall\alpha\in\mathbb{R}_+^*, \quad \sup_{\theta\in K} \int_z p_\theta^\alpha(z)dz < \infty \,.$$

θ does not even have to be restricted to a compact.

Condition on $\int_z S_i^2(z)p_\theta^\alpha(z)dz$ Let K be a compact in Θ and α a positive real number.
In this Poisson count model, $S(z) = \sum_k e^{z_k} \in \mathbb{R}$. We have:

$$S^2(z)p_\theta^\alpha(z) = \left(\sum_k e^{z_k}\right)^2 \frac{exp\left(-\alpha e^\theta \sum_k e^{z_k}\right)}{E_1(0)^{\alpha d}} \,. \tag{5.15}$$

First, let us prove that the integral is finite for any θ. We introduce the variables $u_k := \sum_{l=1}^k e^{z_l}$. The Jacobi matrix is triangular and its determinant is $\prod_k e^{z_k} = \prod_k u_k$.

$$\begin{aligned}\int_z S^2(z)p_\theta^\alpha(z)dz &= \frac{1}{E_1(0)^{\alpha d}} \int_{z\in\mathbb{R}^d} \left(\sum_k e^{z_k}\right)^2 exp\left(-\alpha e^\theta \sum_k e^{z_k}\right) dz \\ &\propto \int_{u_1=0}^{+\infty} u_1 \int_{u_2=u_1}^{+\infty} u_2 ... \int_{u_d=u_{d-1}}^{+\infty} u_d^3\, exp\left(-\alpha e^\theta u_d\right) du_d ... du_2 du_1 \,.\end{aligned}$$

Where we removed the finite constant $\frac{1}{E_1(0)^{\alpha d}}$ for clarity. This integral is finite for any θ because the exponential is the dominant term around $+\infty$. Let us now prove that $\theta \mapsto \int_z S^2(z)p_\theta^\alpha(z)dz$ is continuous. From Equation (5.15), we have that

- $z \mapsto S^2(z)p_\theta^\alpha(z)$ is measurable on $\mathbb{R}^d$.
- $\theta \mapsto S^2(z)p_\theta^\alpha(z)$ is continuous on K (and on $\Theta = \mathbb{R}$).
- With $\theta_M := \min_{\theta\in K} \theta$, then $\forall\theta\in K,\ 0 \le S^2(z)p_\theta^\alpha(z) \le S^2(z)p_{\theta_M}^\alpha(z)$

Since we have proven that $S^2(z)p_{\theta_M}^\alpha(z) < \infty$, then we can apply the intervertion theorem and state that $\theta \mapsto \int_z S^2(z)p_\theta^\alpha(z)dz$ is continuous.
It directly follows that:

$$\forall\alpha\in\mathbb{R}_+^*, \quad \sup_{\theta\in K} \int_z S^2(z)p_\theta^\alpha(z)dz < \infty \,.$$

Note that after the change of variable, the integral could be computed explicitly, but involves d successive integration of polynomial $\times$ exponential function products of the form $P(x)e^{-\alpha e^\theta x}$. This would get tedious, especially since after each successful integration, the product with the next integration variable u_{k-1} increases by one the degree of the polynomial, i.e. starting from 3, the degree ends up being $d+2$. We chose a faster path.

5.4.5 Experiments with Mixtures of Gaussian

Context and experimental protocol

In this section, we will assess the capacity of tmp-EM to escape from deceptive local maxima, on a very well know toy example: likelihood maximisation within the Gaussian Mixture Model. We confront the algorithm to situations where the true classes have increasingly more ambiguous positions, combined with initialisations designed to be hard to escape from. Although the EM is an optimisation procedure, and the log-likelihood reached is a critical metric, in this example, we put more emphasis on the correct positioning of the cluster centroids, that is to say on the recovery of the μ_k. The other usual metrics are also in favour of tmp-EM, and can be found in section 5.7.
For the sake of comparison, the experimental design is similar to the one in [2] on the tmp-SAEM. It is as follows: we have three clusters of similar shape and same weight. One is isolated and easily identifiable. The other two are next to one another, in a more ambiguous configuration. fig. 5.2 represents the three, gradually more ambiguous configurations. Each configuration is called a "parameter family".
We use two different initialisation types to reveal the behaviours of the two EMs. The first - which we call "barycenter" - puts all three initial centroids at the centre of mass of all the observed data points. However, none of the EM procedures would move from this initial state if the three GMM centroids were at the exact same position, hence we actually apply a tiny perturbation to make them all slightly distinct. The blue crosses on Figure 5.3 represent a typical barycenter initialisation. With this initialisation method, we assess whether the EM procedures are able to correctly estimate the positions of the three clusters, despite the ambiguity, when starting from a fairly neutral position, providing neither direction nor misdirection. On the other hand, the second initialisation type - which we call "2v1" - is voluntarily misguiding the algorithm by positioning two centroids on the isolated right cluster and only one centroid on the side of the two ambiguous left clusters. The blue crosses on Figure 5.4 represent a typical 2v1 initialisation. This initialisation is intended to assess whether the methods are able to escape the potential well in which they start and make theirs centroids traverse the empty space between the left and right clusters to reach their rightful position.
For each of the three parameter families represented on fig. 5.2, 1000 datasets with 500 observations each are simulated, and the two EMs are ran with both the barycenter and the 2v1 initialisation.
Regarding the temperature profile of tmp-EM, the only constraint is that $T_n \longrightarrow 1$ and $T_n > 0$. We use an oscillating profile inspired from [2]: $T_n = th(\frac{n}{2r}) + (T_0 - b\frac{2\sqrt{2}}{3\pi})\, a^{n/r} + b\, sinc(\frac{3\pi}{4} + \frac{n}{r})$. These oscillations are meant to momentarily increase the convergence speed (when the temperature reaches low values) to "lock-in" some of the most obviously good decisions of the algorithm, before re-increasing the temperature and continuing the exploration on the other, more ambiguous parameters. Those two regimes are alternated in succession with gradually smaller oscillations, resulting in a multi-scale procedure that "locks-in" gradually harder decisions. The hyper-parameters are chosen by grid-search. The used parameters are $T_0 = 5$, $r = 2$, $a = 0.6$, $b = 20$ for the experiments with the barycenter initialisation, and $T_0 = 100$, $r = 1.5$, $a = 0.02$, $b = 20$ for the 2v1 initialisation. Although, we observe that in the case of 2v1, the oscillations are not critical, and a simple decreasing exponential profile: $T_n = 1 + (T_0 - 1)\, exp(-r.n)$, with $T_0 = 100$ and $r = 1.5$, works as well. We have two different sets of tempering hyper-parameters values, one for each of the two very different initialisation types. However, these values then remain the same for the three different parameter families and for every data generation within them. Underlining that the method is not excessively sensitive to the tempering parameters. Likewise, a simple experiment with 6 clusters, in section 5.7, demonstrates that the same hyper-parameters can be kept over different initialisation (and different data generations as well) when they were made in a non-adversarial way, by drawing random initial centroids uniformly among the data points.

Quantitative analysis

In this section, we quantify the performances of EM and tmp-EM over all the simulations.
Figure 5.3 and 5.4 depict the results of one typical simulation for each of the three ambiguity level (the three parameter families) starting from the barycenter and 2v1 initialisation respectively. The

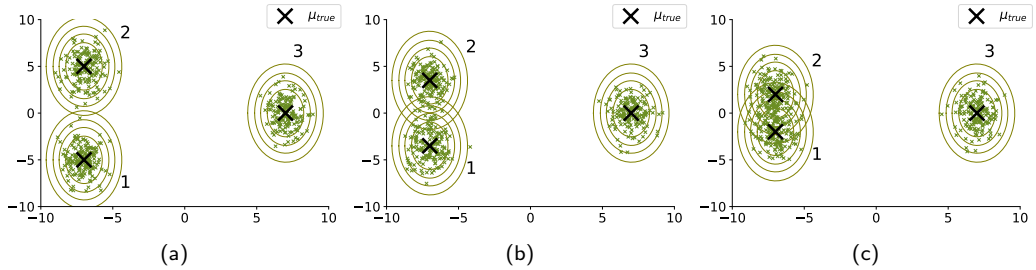


Figure 5.2: 500 sample points from a Mixture of Gaussians with 3 classes. The true centroid of each Gaussian are depicted by black crosses, and their true covariance matrices are represented by the confidence ellipses of level 0.8, 0.99 and 0.999 around the centre. There are three different versions of the true parameters. From left to right: the true μ_k of the two left clusters (mu_1 and mu_2) are getting closer while everything else stays identical.

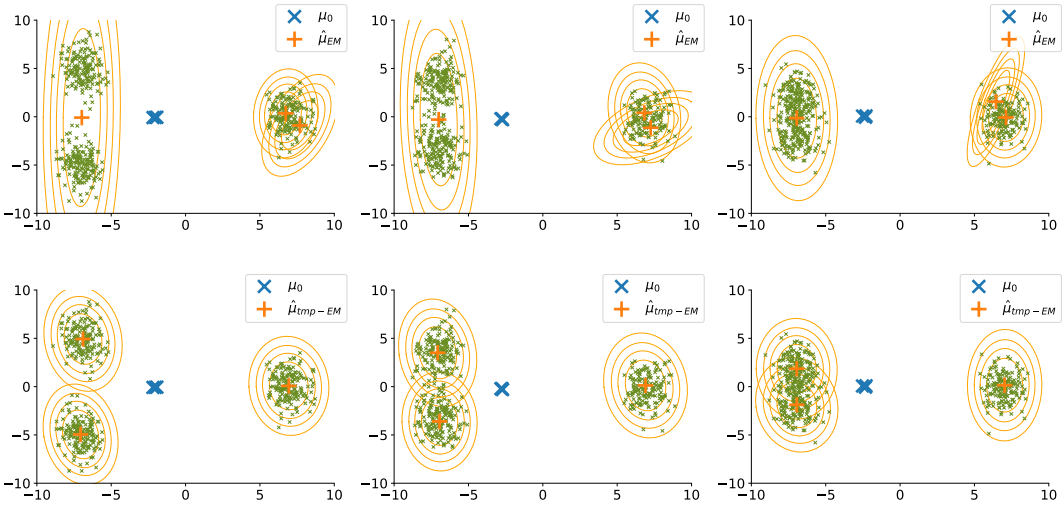


Figure 5.3: Typical final positioning of the centroids by EM (first row) and tmp-EM (second row) **when the initialisation is made at the barycenter of all data points** (blue crosses). The three columns represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those example, tmp-EM managed to correctly identify the position of the three real centroids.

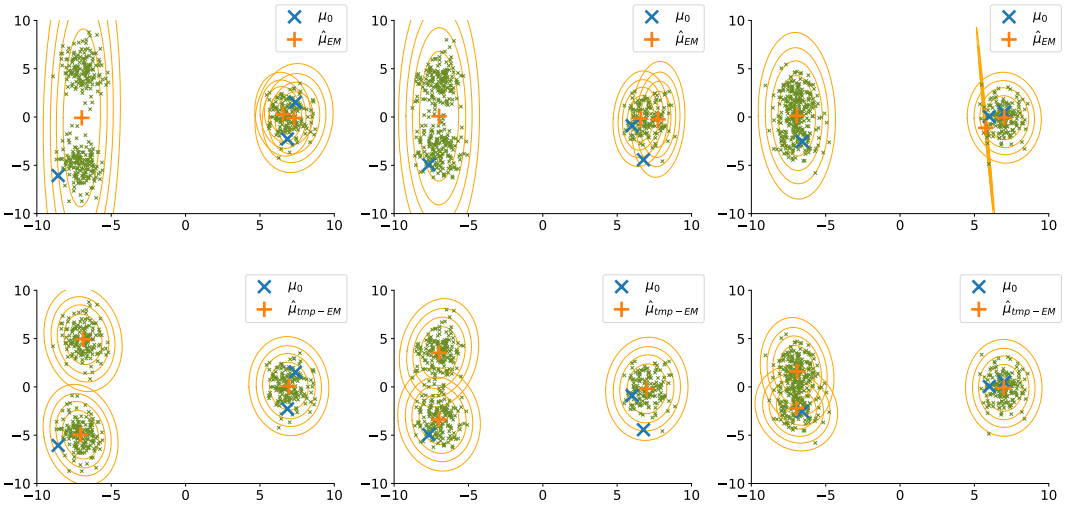


Figure 5.4: Typical final positioning of the centroids by EM (first row) and tmp-EM (second row) **when the initialisation is made by selecting two points in the isolated cluster and one in the lower ambiguous cluster** (blue crosses). The three columns represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those examples, although EM kept two centroids on the isolated cluster, tmp-EM managed to correctly identify the position of the three real centroids.

simulated data is represented by the green crosses. The initial centroids are in blue. The orange cross represents the estimated centroids positions $\hat{\mu}_k$, and the orange confidence ellipses are visual representations of the estimated covariance matrices $\hat{\Sigma}_k$. In section 5.7, we show step by step the path taken by the estimated parameters of tmp-EM before convergence, providing much more detail on the method's behaviours.
On these examples, we note that tmp-EM is more correct than EM. The results over all simulations are aggregated in table 5.1, and confirm this observation.
table 5.1 presents the average and the standard deviation of the relative l_2 error on μ_k of the EMs. For each category, the better result over EM and tmp-EM is highlighted in bold. The recovery of the true class averages μ_k is spotlighted as it is the essential success metric for this experiment.
First we focus on the effect of the different initialisations and placement of (μ_1, μ_2) on the performance of the classical EM. In the first parameter family of table 5.1, μ_1 and μ_2 are still far from one another. The relative error on these two positions is around 0.50 when the initialisation is a the neutral position at the barycenter of the dataset, and 1.50 when the initialisation is made by placing two centroids in the right cluster ("2v1"), a much more adversarial initialisation. In the second parameter family, μ_1 and μ_2 are getting closer. The relative error with the barycenter initialisation has doubled to reach 1.00, and, with the adversarial 2v1, it has increased to 1.70. Finally, in the third parameter family, where μ_1 and μ_2 are so close that their distributions are hard to distinguish with the naked eye, the relative error with the barycenter initialisation has gained another 0.50 points to reach over 1.50, which was the initial error level with the 2v1 initialisation when μ_1 and μ_2 were well separated (parameter family 1). In this very ambiguous setting however, the relative error with 2v1 initialisation has gone up to around 1.80-1.90. As expected, we see that the performances are always hindered in average by the 2v1 initialisation, and that they also worsen when the relative positions of μ_1 and μ_2 become more ambiguous, regardless of the initialisation. The barycenter initialisation however is the one that suffers the most from the increasing ambiguity, gaining 0.5 points of relative error at every transition, whereas 2v1 gain "only" around 0.2 points.
We compare these results and their progression with the ones of tmp-EM in table 5.1. In the first

parameter family - the least ambiguous situation - the relative errors on μ_1 and μ_2 are around 0.05 with the barycenter initialisation and 0.30 with 2v1. In other words, with the tempered E step, we divide by 10 and 5 respectively the relative errors with the barycenter and 2v1 initialisation. In the next position of μ_1 and μ_2, in the second parameter family, the relative error with the barycenter initialisation is now around 0.10, staying 10 times smaller than without tempering. With 2v1, the relative error stayed fairly stable, reaching now 0.35 in average, and remaining approximately 5 times smaller than without tempering. We underline that up until this point (parameter families 1 and 2), the standard deviation of these errors was 3 times smaller with tempering in the case of the barycenter initialisation, and around 2 times smaller in the case of the 2v1 initialisation. In the final configuration, parameter family 3, the relative errors with tempering are 0.30 with the barycenter initialisation (5 times smaller than without tempering) and 0.40 with the 2v1 initialisation (more than 4.5 times smaller than without tempering). Moreover, the standards deviations are at least 1.8 times smaller with tempering. We note that, in similar fashion to EM, the errors on μ_1 and μ_2 with the barycenter initialisation reached, in the most ambiguous configuration, the level of error seen with the 2v1 initialisation in the least ambiguous situation: 0.30. Which, as stated, remains 5 times smaller than the corresponding level of error without tempering: 1.50.
In the end, the progression of errors when μ_1 and μ_2 get closer is alike between EM and tmp-EM: the barycenter initialisation is the most affected, the 2v1 initialisation error being higher but fairly stable. However the level of error is much smaller with tmp-EM, being 5 to 10 times smaller in the case of the barycenter initialisation, and 4.5 to 5 times smaller for the 2v1 initialisation. Similarly, the standard deviation around those average levels is 1.8 to 2 times smaller with tmp-EM.
These quantitative results on the reconstruction error of μ_1 and μ_2 confirm exactly what was observed on the illustrative examples: with tempering, the EM procedure is much more likely to discern the true position of the three clusters regardless of the initialisation, and able to reach a very low error rate even with the most adversarial initialisations. To bolster this last point, we underline that even in the worst case scenario, 2v1 initialisation and very close μ_1 and μ_2, tmp-EM still outperforms EM in the best scenario, barycenter initialisation and well separated clusters, with an error rate of 0.40 versus 0.50.

Table 5.1: Average and standard deviation of the relative error on μ_k, $\frac{\|\hat{\mu}_k - \mu_k\|^2}{\|\mu_k\|^2}$, made by EM and tmp-EM over 1000 simulated dataset with two different initialisations. The three different parameter families, described in fig. 5.2, correspond to increasingly ambiguous positions of classes 1 and 2. For both initialisations type, the identification of these two clusters is drastically improved by the tempering.

		EM		tmp-EM	
Parameter family	cl.	barycenter	2v1	barycenter	2v1
1	1	0.52 (1.01)	1.52 (1.24)	**0.04 (0.26)**	**0.29 (0.64)**
	2	0.55 (1.05)	1.53 (1.25)	**0.05 (0.31)**	**0.30 (0.64)**
	3	**0.01 (0.06)**	**0.01 (0.03)**	0.03 (0.17)	0.03 (0.19)
2	1	1.00 (1.42)	1.69 (1.51)	**0.09 (0.47)**	**0.37 (0.86)**
	2	1.03 (1.44)	1.71 (1.52)	**0.12 (0.57)**	**0.32 (0.79)**
	3	0.01 **(0.05)**	**0.02 (0.03)**	**5.10^{-3}** (0.05)	0.04 (0.22)
3	1	1.56 (1.75)	1.79 (1.77)	**0.31 (0.97)**	**0.39 (0.98)**
	2	1.51 (1.74)	1.88 (1.76)	**0.30 (0.93)**	**0.39 (0.97)**
	3	0.02 (0.04)	**0.02 (0.04)**	**0.01 (0.04)**	0.07 (0.30)

5.5 Tempered Riemann approximation EM

5.5.1 Context, Theorem and proof

The Riemann approximation of section 5.3 makes the EM computations possible in hard cases, when the conditional distribution has no analytical form for instance. It is an alternative to the many stochastic approximation methods (SAEM, MCMC-SAEM...) that are commonly used in those cases. The tempering approximation of section 5.4 is used to escape the initialisation by allowing the procedure to explore more the likelihood profile before committing to convergence. We showed that both these approximation are particular cases of the wider class of Deterministic Approximate EM, introduced in section 5.2. However, since they fulfil different purposes, it is natural to use them in coordination and not as alternatives of one another. In this section, we introduce another instance of the Approximate EM: a combination of the tempered and Riemann sum approximations. This "tempered Riemann approximation EM" (tmp-Riemann approximation) can compute EM steps when there is no closed form thanks to the Riemann sums as well as escape the initialisation thanks to the tempering. For a bounded hidden variable $z \in [0,1]$, we define the approximation as: $\tilde{p}_{n,\theta}(z) := h(\lfloor nz \rfloor/n;\theta)^{\frac{1}{T_n}} / \int_{z'} h(\lfloor nz' \rfloor/n;\theta)^{\frac{1}{T_n}} dz'$, for a sequence $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$.
In the following theorem, we prove that the tempered Riemann approximation EM verifies the applicability conditions of theorem 5.2.1 with no additional hypothesis from the regular Riemann approximation EM covered by theorem 5.3.1.

Theorem 5.5.1. *Under conditions $M1-3$ of theorem 5.2.1, and when z is bounded, the (Stable) Approximate EM with $\tilde{p}_{n,\theta}(z) := \frac{h(\lfloor nz \rfloor/n;\theta)^{\frac{1}{T_n}}}{\int_{z'} h(\lfloor nz' \rfloor/n;\theta)^{\frac{1}{T_n}} dz'}$, which we call "tempered Riemann approximation EM", verifies the remaining conditions of applicability of theorem 5.2.1 as long as $z \mapsto S(z)$ is continuous and $\{T_n\}_n \in (\mathbb{R}_+^*)^{\mathbb{N}}$, $T_n \underset{n\infty}{\longrightarrow} 1$.*

Proof. This proof of theorem 5.5.1 is very similar to the proof of theorem 5.3.1 for the regular Riemann approximation EM. The first common element is that for the tempered Riemann approximation EM, the only remaining applicability condition of the general theorem 5.2.1 to prove is also:

$$\forall \text{compact } K \subseteq \Theta, \quad \sup_{\theta \in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\infty}{\longrightarrow} 0\,.$$

In the proof of theorem 5.3.1, we proved that having the uniform convergence of the approximated complete likelihood $\{\tilde{h}_n\}_n$ towards the real h - with both $\tilde{h}_n(z;\theta)$ and $h(z;\theta)$ uniformly bounded - was sufficient to fulfil this condition. Hence, we prove in this section that these sufficient properties still hold, even with the tempered Riemann approximation, where $\tilde{h}_n(z;\theta) := h\left(\lfloor nz \rfloor/n;\theta\right)^{\frac{1}{T_n}}$.
We recall that $h(z;\theta)$ hence uniformly continuous on the compact set $[0,1] \times K$, and verifies:

$$0 < m \leq h(z;\theta) \leq M < \infty\,.$$

Where m and M are constants independent of z and θ.
Since $T_n > 0$, $T_n \underset{n\infty}{\longrightarrow} 1$, then the sequence $\{1/T_n\}_n$ is bounded. Since $\tilde{h}_n(z;\theta) = h\left(\lfloor nz \rfloor/n;\theta\right)^{\frac{1}{T_n}}$, with $0 < m \leq h\left(\lfloor nz \rfloor/n;\theta\right) \leq M < \infty$ for any z, θ and n, then we also have:

$$0 < m' \leq \tilde{h}_n(z;\theta) \leq M' < \infty\,,$$

with m' and M' constants independent of z, θ and n.
We have seen in the proof of theorem 5.3.1, that:

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall (z,\theta) \in [0,1] \times K, \quad \left|h(z;\theta) - h\left(\lfloor nz \rfloor/n;\theta\right)\right| \leq \epsilon\,.$$

To complete the proof, we control in a similar way the difference $h\left(\lfloor nz \rfloor/n;\theta\right) - h\left(\lfloor nz \rfloor/n;\theta\right)^{\frac{1}{T_n}}$.
The function $(h,T) \in [m,M] \times [T_{min}, T_{max}] \mapsto h^{\frac{1}{T}} \in \mathbb{R}$ is continuous on a compact, hence uniformly continuous in (h,T). As a consequence: $\forall \epsilon > 0, \exists \delta > 0, \forall (h,h') \in [m,M]^2, (T,T') \in [T_{min}, T_{max}]^2$,

$$|h - h'| \leq \delta \text{ and } |T - T'| \leq \delta \implies \left|h^{\frac{1}{T}} - (h')^{\frac{1}{T'}})\right| \leq \epsilon\,.$$

Hence, with $N \in \mathbb{N}$ such that $\forall n \geq N, |T_n - 1| \leq \delta$, we have:

$$\forall n \geq N, \forall (z, \theta) \in [0,1] \times K, \quad \left| h\left(\lfloor nz \rfloor / n; \theta\right) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \leq \epsilon \,.$$

In the end, $\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall (z, \theta) \in [0,1] \times K$:

$$\begin{aligned} \left| h(z;\theta) - \tilde{h}_n\left(z;\theta\right) \right| &= \left| h(z;\theta) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \\ &\leq \left| h(z;\theta) - h\left(\lfloor nz \rfloor / n; \theta\right) \right| + \left| h\left(\lfloor nz \rfloor / n; \theta\right) - h\left(\lfloor nz \rfloor / n; \theta\right)^{\frac{1}{T_n}} \right| \\ &\leq 2\epsilon \,. \end{aligned}$$

In other words, we have the uniform convergence of $\{\tilde{h}_n\}$ towards h. From there, we conclude following the same steps as in the proof of theorem 5.3.1. □

5.5.2 Application to a Gaussian model with the Beta prior

We illustrate the method with the model of section 5.3.3:

$$h(z;\theta) = \frac{\alpha z^{\alpha - 1}}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y - \lambda z)^2}{2\sigma^2} \right) .$$

We apply the tempered Riemann approximation. As in section 5.3.3, the resulting conditional probability density is a step function defined by the n different values it takes on $[0,1]$. For the observation x_i, $\forall k \in [\![0, n-1]\!]$:

$$\tilde{p}^{(i)}_{\theta,n}\left(\frac{k}{n}\right) = \frac{h^{(i)}\left(\frac{k}{n};\theta\right)^{\frac{1}{T_n}}}{\frac{1}{n}\sum_{l=0}^{n-1} h^{(i)}\left(\frac{l}{n};\theta\right)^{\frac{1}{T_n}}} \,.$$

The M step, seen in Equation (5.9), is unchanged. We compare the tempered Riemann EM to the simple Riemann EM on a case where the parameters are ambiguous. With real parameters $\alpha = 0.1, \lambda = 10, \sigma = 0.8$, for each of the 100 simulations, the algorithms are initialised at $\alpha_0 = 10, \lambda_0 = 1, \sigma_0 = 7$. The initialisation is somewhat adversarial, since the mean and variance of the marginal distribution of y are approximately the same with the real of the initialisation parameter, even though the distribution is different. fig. 5.5 shows that the tempered Riemann EM better escapes the initialisation than the regular Riemann EM, and reaches errors on the parameters orders of magnitude below. The tempering parameters are here $T_0 = 150, r = 3, a = 0.02, b = 40$.

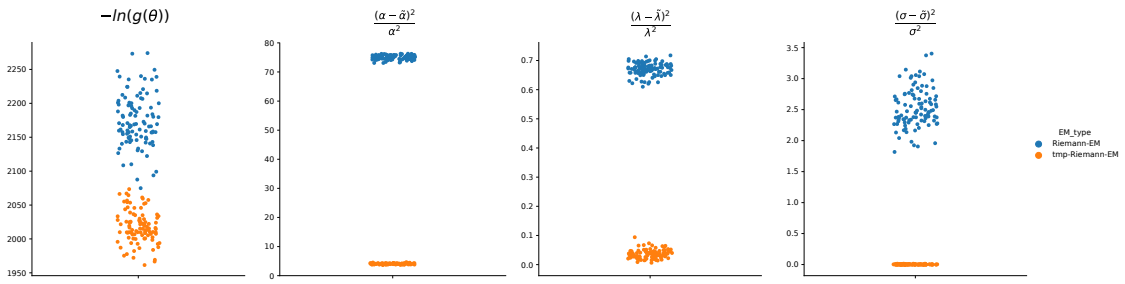


Figure 5.5: Results over many simulations of the Riemann EM and tmp-Riemann EM on the Beta-Gaussian model. The tempered Riemann EM reaches relative errors on the real parameters that are orders of magnitude below the Riemann EM with no temperature. The likelihood reached is also lower with the tempering.

5.6 Proofs of the two main Theorems

5.6.1 Proof of the general theorem

In this Section, we prove theorem 5.2.1 of the main paper, for the convergence of the Deterministic Approximate EM algorithm.
We follow the proof of [49] with very few variations. We use two intermediary results of [49]: their "Proposition 9" and "Proposition 11", which we recall here:

Proposition 8 ("Proposition 9"). *Let $\Theta \subseteq \mathbb{R}^l, K$ compact $\subset \Theta, \mathcal{L} \subseteq \Theta$ such that $\mathcal{L} \cap K$ compact. Let us assume*

- *WC^0 Lyapunov function with regards to $(T, \mathcal{L})$.*
- *$\exists u_n \in K^{\mathbb{N}}$ such that $|W(u_{n+1}) - W \circ T(u_n)| \underset{n\,\infty}{\longrightarrow} 0$*

Then

- *$\{W(u_n)\}_{n\in\mathbb{N}}$ converges towards a connected component of $W(\mathcal{L} \cap K)$*
- *If $W(\mathcal{L} \cap K)$ has an empty interior, then $\{W(u_n)\}_n$ converges towards w^* and $\{u_n\}_n$ converges towards the set $\mathcal{L}_{w^*} \cap K$*

$$\mathcal{L}_{w^*} = \{\theta \in \mathcal{L} | W(\theta) = w^*\}$$

Proposition 9 ("Proposition 11"). *Let $\Theta \subseteq \mathbb{R}^l, T$ and $\{F_n\}_n$ point to point maps on Θ. Let $\{\theta_n\}_n$ be the sequence defined by the stable approximate EM with likelihood f and approximate maps sequence $\{F_n\}_n$. Let $\mathcal{L} \subset \Theta$. We assume*

- *the $A1 - 2$ conditions of Proposition 10 of [49].*
  - ***(A1)** There exists W, a C^0 Lyapunov function with regards to $(T, \mathcal{L})$ such that $\forall M > 0$, $\{\theta \in \Theta, W(\theta) > M\}$ is compact, and:*

    $$\Theta = \cup_{n\in\mathbb{N}} \left\{\theta \in \Theta | W(\theta) > n^{-1}\right\} .$$

  - ***(A2)** $W(\mathcal{L})$ is compact OR **(A2')** $W(\mathcal{L} \cap K)$ is finite for all compact $K \subseteq \Theta$.*
- *$\forall u \in K_0, \quad \lim\limits_{n\,\infty} |W \circ F_n - W \circ T|(u) = 0$*
- *$\forall$ compact $K \subseteq \Theta, \quad \lim\limits_{n\,\infty} |W \circ F_n(u_n) - W \circ T(u_n)| \mathbb{1}_{u_n \in K} = 0$*

Then

$$\textit{With probability 1, } \limsup\limits_{n\,\infty} p_n < \infty \textit{ and } \{u_n\}_n \textit{ compact sequence}$$

Remark. In [49], condition **(A1)** is mistakenly written as:

$$\Theta = \cup_{n\in\mathbb{N}} \{\theta \in \Theta | W(\theta) > n\} .$$

This is a typo that we have corrected here.

We need to prove that, under the conditions of theorem 5.2.1, we verify the conditions of Proposition proposition 8 and proposition 9. Then we will have the results announced in theorem 5.2.1.

Verifying the conditions of 9

f is the likelihood function of a model of the curved exponential family. $\mathcal{L}$ the set of its critical points: $\mathcal{L} := \{\theta \in \Theta | \nabla f(\theta) = 0\}$. Let T be the point to point map describing the transition between θ_n and θ_{n+1} in the exact EM algorithm. The general properties of the EM tell us that its stationary points are the critical points of f: $\mathcal{L} = \{\theta \in \Theta | T(\theta) = \theta\}$. Additionally, f is a C^0 Lyapunov function associated to $(T, \mathcal{L})$. Let $\{\theta_n\}_n$ be the sequence defined by the stable approximate EM with $\{F_n\}_{n\in\mathbb{N}}$ our sequence of point to point maps.

We verify that under this framework - and with the assumptions of theorem 5.2.1 - we check the conditions of proposition 9.

As in [49], $M1-3$ implies $A1-2$.

Let us show that we have the last two conditions for proposition 9:

$$\forall \theta \in K_0, \quad \lim_{n\,\infty} |f \circ F_n - f \circ T|(\theta) = 0\,, \tag{5.16}$$

and

$$\forall \text{ compact } K \subseteq \Theta, \quad \lim_{n\,\infty} |f \circ F_n(\theta_n) - f \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0\,. \tag{5.17}$$

We focus on eq. (5.17), since eq. (5.16) is easier to verify and will come from the same reasoning. To that end, we reproduce the first steps of the reasoning of [49].

Equivalent formulation of the convergence We write Equation (5.17) under an equivalent form. First note that $F_n(\theta_n) = \hat{\theta}(\tilde{S}_n(\theta_n))$ and $T(\theta_n) = \hat{\theta}(\bar{S}(\theta_n))$. Hence $|f \circ F_n(u_n) - f \circ T(u_n)| = |f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))|$. To show Equation (5.17):

$$|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \underset{n\,\infty}{\to} 0\,,$$

it is sufficient and necessary to have:

$$\forall \epsilon > 0,\, \exists N \in \mathbb{N},\, \forall n \geq N,\, |f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

An other equivalent formulation is that there are a finite number of integers n such that $|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon$, in other words:

$$\forall \epsilon > 0,\, \sum_{n=1}^{\infty} \mathbb{1}_{|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon} < \infty\,.$$

Use the uniform continuity We aim to relate the proximity between the images $f \circ \hat{\theta}$ of to the proximity between the antecedents of $f \circ \hat{\theta}$. The function $f \circ \hat{\theta} : \mathbb{R}^q \to \mathbb{R}$ is continuous, but not necessarily uniformly continuous on $\mathbb{R}^q$. As a consequence, we will need to restrict ourselves to a compact to get uniform continuity properties. We already have a given compact K. $\tilde{S} : \Theta \to \mathbb{R}^l$ is continuous, hence $S(K)$ is a compact as well. Let δ be a strictly positive real number. Let $\bar{S}(K, \delta) := \left\{ s \in \mathbb{R}^q \middle| \inf_{t \in K} ||\bar{S}(t) - s|| \leq \delta \right\}$. Where we use any norm $||.||$ on $\mathbb{R}^q$ since they are all equivalent. $\bar{S}(K, \delta)$ is a compact set as well. As a consequence $f \circ \theta$ is uniformly continuous on $\bar{S}(K, \delta)$, which means that:

$$\forall \epsilon > 0,\, \exists \eta(\epsilon, \delta) > 0,\, \forall x, y \in \bar{S}(K, \delta),\, \|x - y\| \leq \eta(\epsilon, \delta) \implies |f \circ \hat{\theta}(x) - f \circ \hat{\theta}(y)| \leq \epsilon\,. \tag{5.18}$$

Let us show that, with $\alpha := min(\delta, \eta(\epsilon, \delta))$, $\forall n$,

$$|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon \implies \left\| \tilde{S}_n(\theta_n) - \bar{S}(\theta_n) \right\| \mathbb{1}_{\theta_n \in K} > \alpha\,. \tag{5.19}$$

To that end, we show that:

$$\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \implies |f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

Let us assume that $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha$.
If $\theta_n \notin K$, then $|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} = 0 \leq \epsilon$.
If, in contrary, $\theta_n \in K$, then $\bar{S}(\theta_n) \in \bar{S}(K) \subset \bar{S}(K, \delta)$. Since $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| = \left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \leq \delta$, then $\tilde{S}_n(\theta_n) \in \bar{S}(K, \delta)$. Since $(\bar{S}(\theta_n), \tilde{S}_n(\theta_n)) \in \bar{S}(K, \delta)^2$ and $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \leq \alpha \leq \eta(\epsilon, \delta)$, then we get from Equation (5.18)

$$|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \leq \epsilon\,.$$

In both cases, we get that:

$$\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \alpha \implies |f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \leq \epsilon\,,$$

which proves Equation (5.19).

Sufficient condition for convergence We use Equation (5.19) to find a sufficient condition for eq. (5.17). Equation (5.19) is equivalent to

$$\mathbb{1}_{|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon} \leq \mathbb{1}_{\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} > \alpha}\,.$$

From that, we get

$$\forall \epsilon > 0,\, \exists \alpha > 0 \sum_{n=1}^{\infty} \mathbb{1}_{|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon} \leq \sum_{n=1}^{\infty} \mathbb{1}_{\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} > \alpha}\,.$$

As a consequence, if

$$\forall \alpha > 0,\, \sum_{n=1}^{\infty} \mathbb{1}_{\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} > \alpha} < \infty$$

Then

$$\forall \epsilon > 0,\, \sum_{n=1}^{\infty} \mathbb{1}_{|f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} > \epsilon} < \infty$$

In other, equivalent, words:

$$\begin{aligned} \text{If} \quad & \left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \underset{n \infty}{\longrightarrow} 0 \\ \text{Then} \quad & |f \circ \hat{\theta}(\tilde{S}_n(\theta_n)) - f \circ \hat{\theta}(\bar{S}(\theta_n))| \mathbb{1}_{\theta_n \in K} \underset{n \infty}{\longrightarrow} 0\,. \end{aligned} \tag{5.20}$$

Hence, having for all compact sets $K \subset \Theta$, $\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \underset{n \infty}{\longrightarrow} 0$ is sufficient to have the desired condition eq. (5.17). Similarly, we find that $\forall \theta \in K_0$:

$$\begin{aligned} & \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\| \underset{n \infty}{\longrightarrow} 0 \\ \implies & |f \circ \hat{\theta}(\tilde{S}_n(\theta)) - f \circ \hat{\theta}(\bar{S}(\theta))| \underset{n \infty}{\longrightarrow} 0\,, \end{aligned} \tag{5.21}$$

which gives us a sufficient condition for eq. (5.16).

Further simplifications of the desired result with successive sufficient conditions We find another, simpler, sufficient condition for eq. (5.17) from Equation (5.20). We first remove the dependency on the terms $\{\theta_n\}_n$ of the EM sequence:

$$\left\|\tilde{S}_n(\theta_n) - \bar{S}(\theta_n)\right\| \mathbb{1}_{\theta_n \in K} \leq \sup_{\theta \in K} \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\|. \tag{5.22}$$

From Equation (5.20), eq. (5.21) and eq. (5.22) we get that:

$$\forall \text{ compact } K \subset \Theta, \quad \sup_{\theta \in K} \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\| \underset{n \infty}{\longrightarrow} 0\,,$$

is a sufficient condition to have both Equation (5.16) and eq. (5.17).

To show that the hypotheses of theorem 5.2.1 imply this sufficient condition, we express it in integral form. Let $S = \{S_i\}_{i=1,\dots,q}$. We recall that $\tilde{S}_n(\theta) = \left\{\int_z S_i(z)\tilde{p}_{\theta,n}(z)dz\right\}_i$ and $\bar{S}(\theta) = \left\{\int_z S_i(z)p_\theta(z)dz\right\}_i$. Hence:

$$\tilde{S}_n(\theta) - \bar{S}(\theta) = \left\{\int_z S_i(z)\left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)dz\right\}_i.$$

These q terms can be upper bounded by two different terms depending on the existence of the involved quantities:

$$\int_z S_i(z)\left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)dz \leq \left(\int_z S_i(z)^2 dz\right)^{\frac{1}{2}} \left(\int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz\right)^{\frac{1}{2}},$$

and

$$\int_z S_i(z)\left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)dz \leq \left(\int_z S_i(z)^2 p_\theta(z)dz\right)^{\frac{1}{2}} \left(\int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz\right)^{\frac{1}{2}}.$$

As a consequence, if $\int_z S_i(z)^2 dz$ exists, then it is sufficient to show have:

$$\sup_{\theta \in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n \infty}{\longrightarrow} 0\,,$$

and if $\int_z S_i(z)^2 p_\theta(z)dz$ exists, then it is sufficient to show have:

$$\sup_{\theta \in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n \infty}{\longrightarrow} 0\,.$$

Among the assumptions of theorem 5.2.1 is one that states that for all compacts $K \subseteq \Theta$, one of those scenarios has to be true. Hence our sufficient condition is met.

Conclusion With the hypothesis of theorem 5.2.1, we have

$$\forall \text{ compact } K \subseteq \Theta, \quad \sup_{\theta \in K} \left\|\tilde{S}_n(\theta) - \bar{S}(\theta)\right\| \underset{n \infty}{\longrightarrow} 0\,,$$

which is a sufficient condition to verify both Equation (5.16) and eq. (5.17). With these two conditions, we can apply proposition 9.

Applying 9

Since we verify all the conditions of proposition 9, we can apply its conclusions:

$$\text{With probability 1, } \underset{n \infty}{limsup}\, p_n < \infty \text{ and } \{\theta_n\}_n \text{ compact sequence },$$

which is specifically the result $(i)(a)$ of theorem 5.2.1.

Verifying the conditions of 8

With proposition 8, we prove the remaining points of theorem 5.2.1: $(i)(b)$ and (ii).

For the application of proposition 8:

- $Cl\left(\{\theta_n\}_n\right)$ plays the part of the compact K
- $\{\theta \in \Theta | \nabla f(\theta) = 0\} = \{\theta \in \Theta | T(\theta) = \theta\}$ plays the part of the set $\mathcal{L}$
- The likelihood f is the C^0 Lyapunov function with regards to $(T, \mathcal{L})$
- $\{\theta_n\}_n$ is the K valued sequence (since K is $Cl\left(\{\theta_n\}_n\right)$).

The last condition that remains to be shown to apply proposition 8 is that:

$$\lim_{n \infty} |f(\theta_{n+1}) - f \circ T(\theta_n)| = 0\,.$$

We have more or less already proven that, in the previous section of the Proof, with $F_n(\theta_n)$ in place of θ_{n+1}. The only indices where $F_n(\theta_n) \neq \theta_{n+1}$ are when the value of the sequence p_n experiences an increment of 1. We have proven with proposition 9 that there is only a finite number of such increments.

$$|f(\theta_{n+1}) - f \circ T(\theta_n)| = |f(\theta_0) - f \circ T(\theta_n)| \mathbb{1}_{p_{n+1}=p_n+1} \; + \; |f \circ F_n(\theta_n) - f \circ T(\theta_n)| \mathbb{1}_{p_{n+1}=p_n}\,.$$

Since there is only a finite number of increments of the value of p_n, then $\exists N \in \mathbb{N}$, $\forall n \geq N$, $\mathbb{1}_{p_{n+1}=p_n+1} = 0$ and $\mathbb{1}_{p_{n+1}=p_n} = 1$. In other words:

$$\begin{aligned}
&\exists N \in \mathbb{N}, \forall n \geq N, \;\; |f(\theta_{n+1}) - f \circ T(\theta_n)| = |f \circ F_n(\theta_n) - f \circ T(\theta_n)| \\
&\exists N \in \mathbb{N}, \forall n \geq N, \;\; |f(\theta_{n+1}) - f \circ T(\theta_n)| = |f \circ F_n(\theta_n) - f \circ T(\theta_n)| \mathbb{1}_{\theta_n \in Cl(\{\theta_k\}_k)}\,.
\end{aligned}$$

Since θ_n is always in $Cl(\{\theta_k\}_k)$ by definition. Additionally proposition 9 tells us that $Cl(\{\theta_k\}_k)$ is a compact. Moreover, in order to use proposition 9 in the first place, we had proven that:

$$\forall \text{ compact } K \subseteq \Theta, \quad \lim_{n \infty} |f \circ F_n(\theta_n) - f \circ T(\theta_n)| \mathbb{1}_{\theta_n \in K} = 0.$$

We can apply this directly with $K = Cl(\{\theta_k\}_k)$ to conclude the desired result:

$$\lim_{n \infty} |f(\theta_{n+1}) - f \circ T(\theta_n)| = 0$$

Hence we verify all the conditions to apply proposition 8.

Applying 8

Since we verify all we need, we have the conclusions of proposition 8:

- $\{f(\theta_n)\}_{n \in \mathbb{N}}$ converges towards a connected component of $f(\mathcal{L} \cap Cl(\{\theta_n\}_n)) \subset f(\mathcal{L})$
- If $f(\mathcal{L} \cap Cl(\{\theta_n\}_n))$ has an empty interior, then $\{f(\theta_n)\}_{n \in \mathbb{N}}$ converges towards a $f^* \in \mathbb{R}$ and $\{\theta_n\}_n$ converges towards $\mathcal{L}_{f^*} \cap Cl(\{\theta_n\}_n)$. Where $\mathcal{L}_{f^*} := \{\theta \in \mathcal{L} | f(\theta) = f^*\}$

Both points are respectively the statements $(i)(b)$ and (ii) of theorem 5.2.1.
Which concludes the proof of the theorem.

5.6.2 Proof of the tempering theorem

In this Section, we prove theorem 5.4.1 of the main paper, the convergence of the tempered EM algorithm. For that, we need to show that we verify each of the hypothesis of the more general theorem 5.2.1.
We already assumed the conditions M1, M2 and M3 in the hypothesis of theorem 5.4.1. To apply theorem 5.2.1, we need to show that when $\tilde{p}_{\theta,n}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'}$, then $\forall$ compact $K \subseteq \Theta$, one of the two following configurations holds:

$$\int_z S(z)^2 dz < \infty \text{ and } \sup_{\theta\in K} \int_z \left(\tilde{p}_{\theta,n}(z) - p_\theta(z)\right)^2 dz \underset{n\,\infty}{\longrightarrow} 0\,,$$

or

$$\sup_{\theta\in K} \int_z S(z)^2 p_\theta(z)dz < \infty \text{ and } \sup_{\theta\in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0\,.$$

Since we have assumed:

$$\forall \text{ compact } K \in \Theta,\ \forall \alpha \in \overline{\mathcal{B}}(1,\epsilon), \forall i, \quad \sup_{\theta\in K} \int_z S_i^2(z) p_\theta^\alpha(z)dz < \infty\,,$$

then we already verify the first half of the second configuration for all the compacts K. Hence it is sufficient to prove that:

$$\forall \text{ compact } K \in \Theta,\ \sup_{\theta\in K} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0\,, \tag{5.23}$$

to have the desired result. The rest of the proof is dedicated to this goal.

Taylor development

We use the Taylor's formula of the first order with the mean-value form of the reminder. For a derivable function g:

$$g(x) = g(0) + g'(a)x, \quad a \in [0,x]\,, \tag{5.24}$$

where the interval $[0,x]$ has a flexible meaning since x could be negative.
We apply it to:

$$g(x) = e^x, \quad g'(x) = e^x, \quad g(x) = 1 + xe^a,\ a \in [0,x]\ ,$$

and:

$$g(x) = \frac{1}{1+x}, \quad g'(x) = -\frac{1}{(1+x)^2}, \quad g(x) = 1 - \frac{x}{(1+a)^2},\ a \in [0,x]\ .$$

To make the upcoming calculation more readable, we momentarily replace $p_\theta(z)$ by simply p and T_n by T.

$$\begin{aligned} p^{\frac{1}{T}} &= p\left(p^{\frac{1}{T}-1}\right) \\ &= pe^{(\frac{1}{T}-1)ln\,p} \\ &= p + \left(\frac{1}{T}-1\right) p\,ln\,p\,e^a, \quad a \in \left[0, \left(\frac{1}{T}-1\right) ln\,p\right]\ , \end{aligned}$$

where $a = a(z,\theta,T_n)$ since it depends on the value of $p_\theta(z)$ and T_n. Provided that the following quantities are defined, we have:

$$\int_z p^{\frac{1}{T}} = 1 + \left(\frac{1}{T}-1\right) \int_z p\,ln\,p\,e^a\,,$$

Hence:

$$\frac{1}{\int_z p^{\frac{1}{T}}} = 1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p \ln p \, e^a}{(1+b)^2}, \quad b \in \left[0, \left(\frac{1}{T} - 1\right) \int_z p \ln p \, e^a\right] ,$$

where $b = b(\theta, T_n)$ since it depends on the value of T_n the integral over z of a function of z and θ. In the end, we have:

$$\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} = p + \left(\frac{1}{T} - 1\right) p \ln p \, e^a \left(1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p \ln p \, e^a}{(1+b)^2}\right) - \left(\frac{1}{T} - 1\right) p \frac{\int_z p \ln p \, e^a}{(1+b)^2} . \tag{5.25}$$

Since for any real numbers $(x+y)^2 \leq 2(x^2+y^2)$, then:

$$\begin{aligned}\left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 &\leq 2\left(\frac{1}{T} - 1\right)^2 p^2 \left((\ln p \, e^a)^2 \left(1 - \left(\frac{1}{T} - 1\right) \frac{\int_z p \ln p \, e^a}{(1+b)^2}\right)^2 + \left(\frac{\int_z p \ln p \, e^a}{(1+b)^2}\right)^2\right) \\ &= 2\left(\frac{1}{T} - 1\right)^2 p^2 \left((\ln p \, e^a)^2 A + B\right) .\end{aligned}$$

where $A = A(\theta, T_n)$ and $B = B(\theta, T_n)$. So far the only condition that has to be verified for all the involved quantities to be defined is that $\int_z p \ln p \, e^a$ exists. With this Taylor development on hand, we state, prove and apply two lemmas which allow us to get eq. (5.23) and conclude the proof of the theorem.

Two intermediary lemmas

The two following lemmas provides every result we need to finish the proof.

Lemma 5.6.1. *With*

$$p_\theta(z) = exp\left(\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right) ,$$

then

$$\int_z p_\theta^\alpha(z) \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2 \left\|\phi(\theta)\right\|^2 . \sum_i \int_z S_i^2(z) p_\theta^\alpha(z) .$$

and

$$\int_z p_\theta^\alpha(z) \left|\ln p_\theta(z)\right| dz \leq \left|\psi(\theta)\right| \int_z p_\theta^\alpha(z) dz + \left\|\phi(\theta)\right\| . \left(\sum_i \int_z S_i^2(z) p_\theta^\alpha(z) \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}} .$$

Proof. For the first inequality, using the fact that $(a+b)^2 \leq 2(a^2+b^2)$, we have:

$$\int_z p_\theta^\alpha(z) \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2 \int_z p_\theta^\alpha(z) \left\langle S(z), \phi(\theta)\right\rangle^2 ,$$

We use Cauchy-Schwartz:

$$\left\langle S(z), \phi(\theta)\right\rangle^2 \leq \left\|\phi\right\|^2 \left\|S(z)\right\|^2 = \left\|\phi\right\|^2 \sum_i S_i(z)^2 ,$$

to get the desired result:

$$\int_z p_\theta^\alpha(z) \ln^2 p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2 \left\|\phi(\theta)\right\|^2 . \sum_i \int_z S_i^2(z) p_\theta^\alpha(z) .$$

For the second inequality, we start with Cauchy-Schwartz on $\left\langle \int_z S(z) p_\theta^\alpha(z), \phi(\theta)\right\rangle$:

$$\int_z p_\theta^\alpha(z) \left|\ln p_\theta(z)\right| dz \leq \left|\psi(\theta)\right| \int_z p_\theta^\alpha(z) dz + \left\|\phi(\theta)\right\| . \left\| \int_z S(z) p_\theta^\alpha(z) \right\| .$$

Moreover, since:

$$\int_z S_i(z)p_\theta^\alpha(z)dz \leq \left(\int_z S_i^2(z)p_\theta^\alpha(z)dz\right)^{\frac{1}{2}} \left(\int_z p_\theta^\alpha(z)dz\right)^{\frac{1}{2}} ,$$

then

$$\int_z p_\theta^\alpha(z)\ |ln\, p_\theta(z)|\, dz \leq |\psi(\theta)| \int_z p_\theta^\alpha(z)dz + \|\phi(\theta)\| \cdot \left(\sum_i \int_z S_i^2(z)p_\theta^\alpha(z) \int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}} .$$

□

Lemma 5.6.2. *With K compact and $\epsilon \in \mathbb{R}_+^*$,*

$$p_\theta(z) = exp\left(\psi(\theta) + \langle S(z), \phi(\theta)\rangle\right),$$

and

$$\tilde{p}_{\theta,n}(z) := \frac{p_\theta^{\frac{1}{T_n}}(z)}{\int_{z'} p_\theta^{\frac{1}{T_n}}(z')dz'},$$

if

(i) $T_n \in \mathbb{R}_+^* \underset{n\,\infty}{\longrightarrow} 1$,

(ii) $\underset{\theta\in K}{sup}\, \psi(\theta) < \infty$,

(iii) $\underset{\theta\in K}{sup}\, \|\phi(\theta)\| < \infty$,

(iv) $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon), \quad \underset{\theta\in K}{sup} \int_z p_\theta^\alpha(z)dz < \infty$,

(v) $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon),\ \forall i, \quad \underset{\theta\in K}{sup} \int_z S_i^2(z)p_\theta^\alpha(z)dz < \infty$.

then

$$\underset{\theta\in K}{sup} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1\right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0 .$$

Proof. Provided that the following integrals exist, we have, thanks to the Taylor development:

$$\begin{aligned}\int_z \frac{1}{p}\left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 &\leq 2\int_z \left(\frac{1}{T} - 1\right)^2 p\left(\left(ln\, p\, e^a\right)^2 A + B\right) \\ &= 2\left(\frac{1}{T} - 1\right)^2 A \int_z pe^{2a} ln^2\, p + 2\left(\frac{1}{T} - 1\right)^2 B .\end{aligned} \tag{5.26}$$

In this proof, we find finite upper bounds independent of θ and T_n for $A(\theta, T_n)$, $B(\theta, T_n)$ and $\int_z pe^{2a} ln^2\, p$, then - since $\left(\frac{1}{T_n} - 1\right) \longrightarrow 0$ - we have the desired result.

We start by studying $A(\theta, T_n) = \left(1 - \left(\frac{1}{T} - 1\right)\frac{\int_z p\, ln\, p\, e^a}{(1+b)^2}\right)^2$. The first term of interest here is $\int_z p\, ln\, p\, e^a$. We have:

$$\begin{aligned} a &\in \left[0, \left(\frac{1}{T} - 1\right) ln\, p\right] , \\ e^a &\in \left[1, p^{\frac{1}{T}-1}\right] , \\ p\, ln\, p\, e^a &\in \left[p\, ln\, p, p^{\frac{1}{T}} ln\, p\right] . \end{aligned}$$

where we recall that the interval is to be taken in a flexible sense, since we do not now a priory which bound is the largest and which is the smallest. What we have without doubt though is:

$$|p\,ln\,p\,e^a| \leq max\left(|p\,ln\,p|\,,\left|p^{\frac{1}{T}}ln\,p\right|\right)\,.$$

We find an upper bound on both those term. Let $\alpha \in \overline{\mathcal{B}}(1,\epsilon)$, the second result of lemma 5.6.1 gives us:

$$\int_z p_\theta^\alpha(z)\;|ln\,p_\theta(z)|\,dz \leq |\psi(\theta)|\int_z p_\theta^\alpha(z)dz + \|\phi(\theta)\|\,.\left(\sum_i \int_z S_i^2(z)p_\theta^\alpha(z)\int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}\,.$$

Thanks to the hypothesises (ii), (iii), (iv) and (v), we have:

$$\begin{aligned}\int_z p_\theta^\alpha(z)\;|ln\,p_\theta(z)|\,dz &\leq \sup_{\theta\in K}|\psi(\theta)|\,.\sup_{\theta\in K}\int_z p_\theta^\alpha(z)dz\\ &\quad+\sup_{\theta\in K}\|\phi(\theta)\|\,.\sum_i\left(\sup_{\theta\in K}\int_z S_i^2(z)p_\theta^\alpha(z)\right)^{\frac{1}{2}}.\left(\sup_{\theta\in K}\int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}\\ &=: C(\alpha)\\ &<\infty\,.\end{aligned}$$

The upper bound $C(\alpha)$ in the previous inequality is independent of θ and z but still dependant of the exponent α. However, since $\overline{\mathcal{B}}(1,\epsilon)$ is closed ball, hypothesises (iv) and (v) can be rephrased as:

$$\begin{aligned}&(iv)\quad \sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)}\sup_{\theta\in K}\int_z p_\theta^\alpha(z)dz<\infty\,,\\ &(v)\;\forall i,\quad \sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)}\sup_{\theta\in K}\int_z S_i^2(z)p_\theta^\alpha(z)dz<\infty\,.\end{aligned}$$

Hence we can actually take the supremum in α as well:

$$\begin{aligned}\int_z p_\theta^\alpha(z)\;|ln\,p_\theta(z)|\,dz &\leq \sup_{\theta\in K}|\psi(\theta)|\,.\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)}\sup_{\theta\in K}\int_z p_\theta^\alpha(z)dz\\ &\quad+\sup_{\theta\in K}\|\phi(\theta)\|\,.\sum_i\left(\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)}\sup_{\theta\in K}\int_z S_i^2(z)p_\theta^\alpha(z)\right)^{\frac{1}{2}}.\left(\sup_{\alpha\in\overline{\mathcal{B}}(1,\epsilon)}\sup_{\theta\in K}\int_z p_\theta^\alpha(z)\right)^{\frac{1}{2}}\\ &=: C'\\ &<\infty\,.\end{aligned}$$

This new upper bound C' is independent of α.
Since $T_n \mapsto 1$, then $\exists N\in\mathbb{N},\forall n\geq N,\;\frac{1}{T_n}\in\overline{\mathcal{B}}(1,\epsilon)$. Hence for $n\geq N$, we can apply the previous inequation to either $\alpha=1$ or $\alpha=\frac{1}{T_n}$. Which gives us that $\int_z p_\theta(z)\;|ln\,p_\theta(z)|$, $\int_z p_\theta^{\frac{1}{T_n}}(z)\;|ln\,p_\theta(z)|$ and their supremum in θ are all finite, all of them upper bounded by C'.

In the end, when $n\geq N$, we have the control $\sup_{\theta\in K}\left|\int_z p\,ln\,p\,e^a\right|<C'$.

The next term to control is $\frac{1}{(1+b)^2}$. Since $b\in\left[0,\left(\frac{1}{T}-1\right)\int_z p\,ln\,p\,e^a\right]$, then $|b|\leq\left(\frac{1}{T}-1\right)\sup_{\theta\in K}\int_z p\,ln\,p\,e^a$.
We already established that for all $n\geq N$, $\sup_{\theta\in K}\left|\int_z p\,ln\,p\,e^a\right|\leq C'<\infty$, hence $\sup_{\theta\in K}|b(\theta,T_n)|\underset{T_n\longrightarrow 1}{\longrightarrow}0$.
In particular, $\exists N'\in\mathbb{N},\forall n\geq N',\forall\theta\in K$ we have $|b(\theta,T_n)|\leq\frac{1}{2}$. In that case:

$$\begin{aligned}(1+b)^2 &> (1-|b|)^2 \geq \frac{1}{4}\\ \frac{1}{(1+b)^2} &< \frac{1}{(1-|b|)^2}\leq 4\,.\end{aligned}$$

In the end, when $n \geq max(N, N')$, for any $\theta \in K$:

$$\begin{aligned} A(\theta, T_n) &\leq 2 + 2\left(\frac{1}{T_n} - 1\right)^2 \left(\frac{\int_z p\, ln\, p\, e^a}{(1+b)^2}\right)^2 \\ &\leq 2 + 32\left(\frac{1}{T_n} - 1\right)^2 \left(\sup_{\theta \in K} \int_z p\, ln\, p\, e^a\right)^2 \\ &\leq 2 + 32\left(\frac{1}{T_n} - 1\right)^2 C'^2 \\ &\leq 2 + 32\epsilon^2 C'^2 \\ &=: C_1\,. \end{aligned}$$

This upper bound does not depend en θ anymore and the part in T_n simply converges towards 0 when $T_n \longrightarrow 1$.

Treating the term $B(\theta, T_n) = \left(\frac{\int_z p\, ln\, p\, e^a}{(1+b)^2}\right)^2 \leq 16\left(\sup_{\theta \in K} \int_z p\, ln\, p\, e^a\right)^2 \leq 16C'^2 =: C_2$ is immediate after having dealt with $A(\theta, T_n)$.

We now treat the term $\int_z p\, e^{2a} ln^2\, p$ in the exact same fashion as we did $A(\theta, T_n)$:

$$\begin{aligned} & p\, ln\, p\, e^a \in \left[p\, ln\, p, p^{\frac{1}{T}} ln\, p\right] \\ \implies & p\, (ln\, p\, e^a)^2 \in \left[p\, ln^2\, p, p^{\frac{2}{T}-1} ln^2\, p\right] \\ \implies & p\, (ln\, p\, e^a)^2 \leq max(p\, ln^2\, p, p^{\frac{2}{T}-1} ln^2\, p)\,. \end{aligned}$$

We control those two terms as previously. First we apply lemma 5.6.1 (its first result this time) with $\alpha \in \overline{\mathcal{B}}(1, \epsilon)$.

$$\int_z p_\theta^\alpha(z)\, ln^2\, p_\theta(z) dz \leq 2\psi(\theta)^2 \int_z p_\theta^\alpha(z) dz + 2\left\|\phi(\theta)\right\|^2 . \sum_i \int_z S_i^2(z) p_\theta^\alpha(z)\,.$$

Thanks to the hypothesis (ii), (iii), (iv) and (v), we can once again take the supremum of the bound over $\theta \in K$, then over $\alpha \in \overline{\mathcal{B}}(1, \epsilon)$ and conserve finite quantities:

$$\begin{aligned} \int_z p_\theta^\alpha(z)\, ln^2\, p_\theta(z) dz \leq\ & 2 \sup_{\theta \in K} \psi(\theta)^2 . \sup_{\alpha \in \overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta \in K} \int_z p_\theta^\alpha(z) dz \\ & + 2 \sup_{\theta \in K} \left\|\phi(\theta)\right\|^2 . \sum_i \sup_{\alpha \in \overline{\mathcal{B}}(1,\epsilon)} \sup_{\theta \in K} \int_z S_i^2(z) p_\theta^\alpha(z) \\ =:\ & C_3 \\ <\ & \infty\,. \end{aligned}$$

The previous result is true for $\alpha = 1$, and since once again $\exists N''$, $\forall n \geq N''$, $\frac{2}{T_n} - 1 \in \overline{\mathcal{B}}(1, \epsilon) \cap \mathbb{R}_+^*$, it is also true for $\alpha = \frac{2}{T_n} - 1$ when n is large enough. C_3 is independent of z, θ and T_n.
In the end $\forall n \geq N''$, $\int_z p\, e^{2a} ln^2\, p \leq C_3 < \infty$.

We replace the three terms $A(\theta, T_n)$, $B(\theta, T_n)$ and $\int_z p\, e^{2a} ln^2\, p$ by their upper bounds in the inequality eq. (5.26). When $n \geq max(N, N', N'')$:

$$\int_z \frac{1}{p}\left(\frac{p^{\frac{1}{T}}}{\int_z p^{\frac{1}{T}}} - p\right)^2 \leq 2\left(\frac{1}{T_n} - 1\right)^2 C_1 C_3 + 2\left(\frac{1}{T_n} - 1\right)^2 C_2\,.$$

Which converges towards 0 when $T_n \longrightarrow 1$, i.e. when $n \longrightarrow \infty$. This concludes the proof of the lemma.

□

Verifying the conditions of Lemma 5.6.2

Now that the lemmas are proven, all that remains is to apply lemma 5.6.2.

(i) We have $T_n \in \mathbb{R}_+^* \underset{n\,\infty}{\longrightarrow} 1$ by hypothesis.

(ii)* and *(iii) $\underset{\theta\in K}{sup}\, \psi(\theta) < \infty$ and $\underset{\theta\in K}{sup}\, \|\phi(\theta)\| < \infty$ are implied by the fact that $\psi(\theta) = \psi'(\theta) - log\, g(\theta)$ and $\phi(\theta)$ are continuous

(iv)* and *(v) Are also hypothesis of the theorem.
Hence we can apply lemma 5.6.2. This means that:

$$\underset{\theta\in K}{sup} \int_z \left(\frac{\tilde{p}_{\theta,n}(z)}{p_\theta(z)} - 1 \right)^2 p_\theta(z)dz \underset{n\,\infty}{\longrightarrow} 0.$$

With this last condition verified, we can apply theorem 5.2.1. Which concludes the proof.

5.7 Additional experiments on tmp-EM with Mixtures of Gaussian

In this section, we present more detailed experiments analysing the tempered EM and comparing it to the regular EM. As in section 5.4.5, we focus on likelihood maximisation within the Gaussian Mixture Model. From the optimisation point of view, we demonstrate that tmp-EM does not fall in the first local maximum like EM does but instead consistently finds better one. From the machine learning point of view, we illustrate how tmp-EM is able to better identify the real GMM parameters even when they are ambiguous and when the initialisation is voluntarily tricky.
The only constraints on the temperature profile is that $T_n \longrightarrow 1$ and $T_n > 0$. We use two different temperature profiles. First, a decreasing exponential: $T_n = 1 + (T_0 - 1)\, exp(-r.n)$. We call it the "simple" profile, it works most of the time. Second, we examine the capabilities of a profile with oscillations in addition to the main decreasing trend. These oscillations are meant to momentarily increase the convergence speed to "lock-in" some of the most obviously good decisions of the algorithm, before re-increasing the temperature and continuing the exploration on the other, more ambiguous parameters. Those two regimes are alternated in succession with gradually smaller oscillations, resulting in a multi-scale procedure that "locks-in" gradually harder decisions. The formula is taken from [2]: $T_n = th(\frac{n}{2r}) + (T_0 - b\frac{2\sqrt{2}}{3\pi})\, a^{n/r} + b\, sinc(\frac{3\pi}{4} + \frac{n}{r})$. The profile used, as well as the values of the hyper-parameters are specified for each experiment. The hyper parameters are chosen by grid-search.
For the sake of comparison, the following Experiment 1 and 2 are similar to the experiments of [2] on the tmp-SAEM.

5.7.1 Experiment 1: 6 clusters

We start by demonstrating the superior performance of the tempered EM algorithm on an example mixture of $K = 6$ Gaussians in dimension $p = 2$. The real parameters can be visualised on fig. 5.6, where the real centroids are represented by black crosses and confidence ellipses help visualise the real covariance matrices. In addition, 500 points were simulated in order to illustrate, among other things, the weights of each class. To quantify the ability of each EM method to increase the likelihood and recover the true parameters, we generate from this model 20 different datasets with $n = 500$ observations. For each of these datasets, we make 200 EM runs, all of them starting from a different random initialisation. To initialise the mixture parameters, we select uniformly 6 data points to act as centroids. In each run, EM and tmp-EM start with the same initialisation. The number K of clusters is known by the algorithms. For this experiment, the simple tempering profile is used with parameters $T_0 = 50$ and $r = 2$.

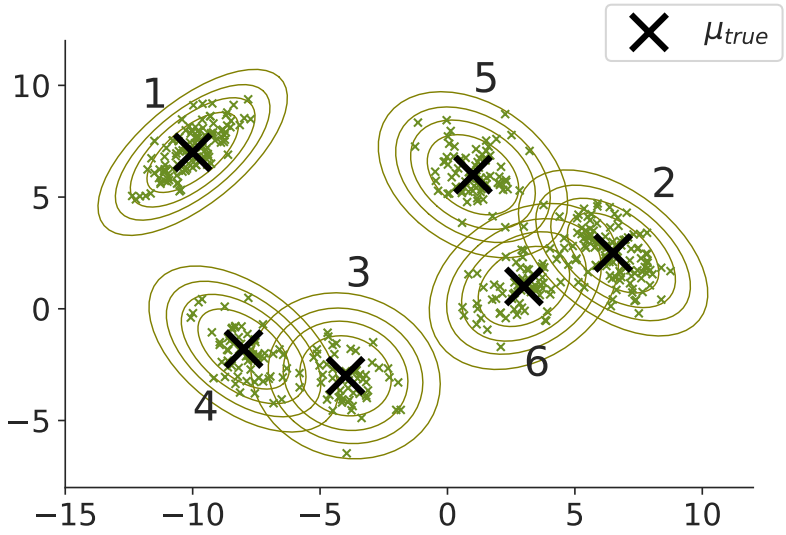


Figure 5.6: 500 sample points from a Mixture of Gaussians with 6 classes. The true centroid of each Gaussian are depicted by black crosses, and their true covariance matrices are represented by the confidence ellipses of level 0.8, 0.99 and 0.999 around the centre.

Illustrative

First, we observe on the left of fig. 5.7, one example of the final states of the EM algorithm. The observations can be seen in green, the initial centroids are represented by blue crosses, and the parameters $\{\hat{\mu}_k\}_{k=1}^K$ and $\left\{\widehat{\Sigma}_k\right\}_{k=1}^K$ estimated by the EM are represented in orange. In this EM run, one of the estimated clusters became degenerated and, as counterpart, two different real clusters were fused as one by the method. On the right of fig. 5.7, we observe the final state of the tmp-EM on the same dataset, from the same initialisation. This time all the clusters were properly identified.

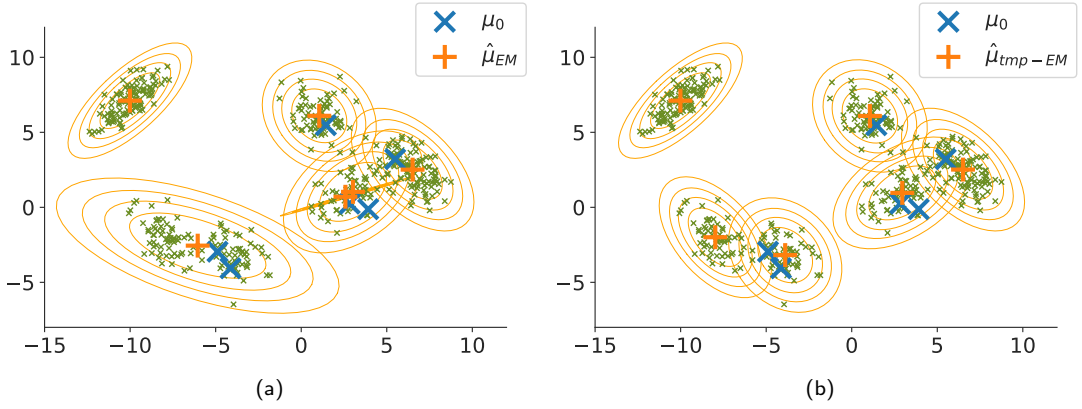
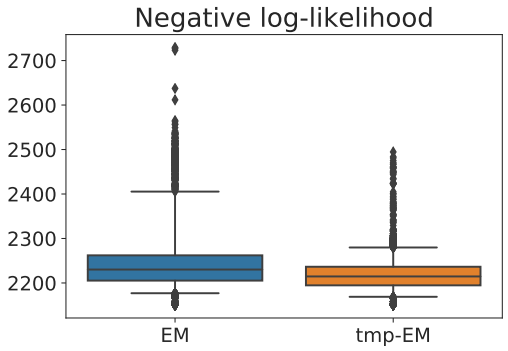


Figure 5.7: EM and tmp-EM final states on the same simulation with the same initialisation. tmp-EM positioned correctly the estimated centroides, whereas the regular EM made no distinction between the two bottom classes and ended up with a degenerate class instead.

Quantitative

To demonstrate the improvements made by tempering, we present aggregated quantitative results over all the simulated datasets and random initialisations.

Likelihood maximisation EM and tmp-EM are optimisation methods whose target function is the likelihood of the estimated mixture parameters. We represent on fig. 5.8 the empirical distribution of the negative log-likelihoods reached at the end of the two methods, EM in blue, tmp-EM in orange. On those boxplots, the coloured "box" at the centre contains 50% of the distribution, hence it is delimited by the 0.25 and 0.75 quantiles. The median of the distribution is represented by an horizontal black line inside the box. The space between the whiskers on the other end, contain 90% of the distribution, its limits are the 0.05 and 0.95 quantiles. The table provides the numeric values of these statistics.



(a)

	EM	tmp-EM
median	2230	**2215**
mean	2247	**2219**
std	70	**35**
Q5	2177	**2169**
Q25	2205	**2195**
Q75	2262	**2236**
Q95	2405	**2280**

(b)

Figure 5.8: Empirical distribution of the negative log-likelihood reached by the EM algorithms. EM is blue and tmp-EM in orange. The boxplot allow us to identify the quantiles 0.05, 0.25, 0.5, 0.75 and 0.95 of each distribution, as well as the outliers. Their numeric values can be found in the table, the better ones being in **bold**. tmp-EM is better overall.

We note that the negative log-likelihood reached by tmp-EM is lower on average (higher likelihood) than what EM obtains. Moreover, tmp-EM also has a lower variance, its standard deviation being approximately half of the std of EM. More generally, we observe that the distribution of the final loss of tmp-EM is both shifted towards the lower values and less variable. In particular, each of the followed quantiles are lower for tmp-EM, and both the difference Q95-Q5 (space between whiskers) and Q75-Q25 (size of the box) are lower for tmp-EM. This illustrates that it obtains better, more consistent results on our synthetic example.

Parameter recovery The EM algorithm is an optimisation procedure. Stricto sensu, the optimised metric - the likelihood - should be the only criterion for success. However, in the case of the Mixture of Gaussians, the underlying Machine Learning stakes are always very visible. Hence we dedicate time to assess the relative success parameter recovery of EM and tmp-EM.
The quality of parameter recovery is always dependent on the number of observation. The larger n, the more the likelihood will describe an actual ad-equation with the real parameters behind the simulation. Additionally, as n grows, the situation becomes less and less ambiguous, until all methods yield either the exact same, or at least very similar solutions, with all of them being fairly close to the truth. All of our simulation are done with $n = 500$ data points. Not a very large number, but since the lowest weight of our $K = 6$ classes is around 0.09, it is sufficient for all the classes to be guaranteed to contain several points. The three families of parameters in a GMM are the weights $\{\pi_k\}_{k=1}^K$, the averages (centroids positions) $\{\mu_k\}_{k=1}^K$ and the covariance matrices $\{\Sigma_k\}_{k=1}^K$ of the K classes. We evaluate the error made on μ with the relative different in squared norm 2: $\frac{\|\hat{\mu}_k - \mu_k\|_2^2}{\|\mu_k\|_2^2}$. For Σ, we compute the KL divergence between the real matrices and the estimates

$KL(\Sigma_k, \widehat{\Sigma}_k) = \frac{1}{2}\left(ln\,\frac{|\Theta_k|}{|\widehat{\Theta}_k|} + tr(\Sigma_k\widehat{\Theta}_k) - p\right)$, with $\Theta := \Sigma^{-1}$ for all those matrices. Finally, the analysis on π is harder to interpret and less interesting, but reveals the same trend, with lower errors for the tempering.

The error on the averages μ_k is usually the most informative and easy to interpret metric, quantifying how well each methods position the class centres. fig. 5.9 and table 5.2 represent the distribution of the relative error $\frac{\|\hat{\mu}_k - \mu_k\|_2^2}{\|\mu_k\|_2^2}$. The results of tmp-EM are much better with average and median errors often being orders of magnitude below the errors of EM, with similar or lower variance. The other quantiles of the tmp-EM distribution are also either equivalent to or order of magnitudes below the corresponding EM quantiles. The largest errors happen on Class 3 and 6, two of the ambiguous ones, but are always noticeably smaller and less variable with the tempering.

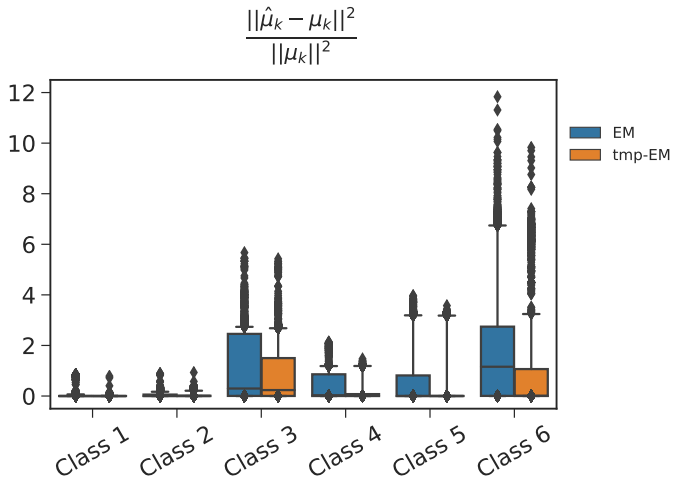


Figure 5.9: Empirical distribution of the relative error in squared norm 2 $\frac{\|\hat{\mu}_k - \mu_k\|_2^2}{\|\mu_k\|_2^2}$ between the real centroid positions in μ and the estimations by the EM algorithms.

Table 5.2: Quantiles and other statistics describing the empirical distribution of the relative error in squared norm 2 $\frac{\|\hat{\mu}_k - \mu_k\|_2^2}{\|\mu_k\|_2^2}$ between the real centroid positions in μ and the estimations by the EM algorithms. The error of tmp-EM is always closer to 0 with lower variance (with the exception of class 2 where the variance is similar).

Cl.		mean	std	Q5	Q25	Q50	Q75	Q95
1	EM	0.024	0.119	6.10^{-6}	6.10^{-5}	2.10^{-4}	0.002	0.065
	tmp-EM	**0.002**	**0.014**	$\mathbf{6.10^{-6}}$	$\mathbf{4.10^{-5}}$	$\mathbf{1.10^{-4}}$	$\mathbf{4.10^{-4}}$	**0.005**
2	EM	0.038	**0.066**	5.10^{-5}	2.10^{-4}	0.001	0.057	**0.169**
	tmp-EM	**0.032**	0.070	$\mathbf{5.10^{-5}}$	$\mathbf{2.10^{-4}}$	$\mathbf{5.10^{-4}}$	**0.013**	0.210
3	EM	0.971	1.153	4.10^{-4}	0.004	0.297	2.467	2.736
	tmp-EM	**0.743**	**1.072**	$\mathbf{3.10^{-4}}$	**0.003**	**0.235**	**1.500**	**2.681**
4	EM	0.310	0.487	7.10^{-5}	8.10^{-4}	0.031	0.859	**1.158**
	tmp-EM	**0.287**	**0.476**	$\mathbf{3.10^{-5}}$	$\mathbf{5.10^{-4}}$	**0.025**	**0.076**	1.188
5	EM	0.735	1.248	8.10^{-5}	5.10^{-4}	0.002	0.814	3.191
	tmp-EM	**0.432**	**1.054**	$\mathbf{6.10^{-5}}$	$\mathbf{4.10^{-4}}$	$\mathbf{7.10^{-4}}$	**0.002**	**3.180**
6	EM	1.940	2.828	7.10^{-4}	0.005	1.158	2.743	6.744
	tmp-EM	**0.807**	**1.735**	$\mathbf{4.10^{-4}}$	**0.002**	**0.010**	**1.066**	**3.243**

The KL divergences $KL(\Sigma_k, \widehat{\Sigma}_k)$ assess whether each the covariances Σ_k of each class are properly

replicated. Note that since the computation of the KL divergence involves the matrix inverse $\widehat{\Theta}_k = \widehat{\Sigma}_k^{-1}$, the outliers cases where a class vanishes in an EM have to be removed: they correspond to pathological, non invertible matrices. fig. 5.10 and table 5.3 describe the distribution of the KL divergence. The Figure is cropped and does not show some of the very rare, most upper outliers (less than 1%). Overall, the results are similar to what we get on μ: in terms of average KL and median KL, tmp-EM is better than EM, being either similar on some classes and much better on others. Its standard deviation is also lower - sometimes by one order of magnitude - on all classes except Class 4. The other quantiles are also overall better, with one exception on Q95 of class 4.

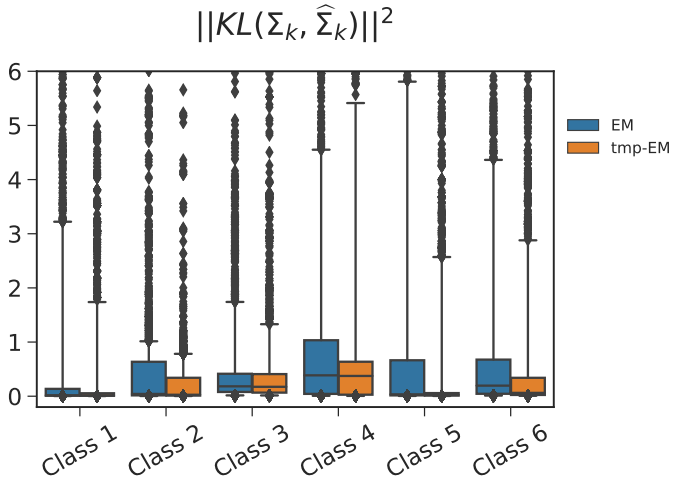


Figure 5.10: Empirical distribution of the KL divergence $KL(\Sigma_k, \widehat{\Sigma}_k)$ between each covariance matrix estimated by the EMs and the real covariance matrices Σ.

Table 5.3: Quantiles and other statistics describing the empirical distribution of the KL divergence $KL(\Sigma_k, \widehat{\Sigma}_k)$ between each covariance matrix estimated by the EMs and the real covariance matrices Σ. On every class but the 4th, the deviation of tmp-EM is closer to 0 with lower or similar variance.

Cl.		mean	std	Q5	Q25	Q50	Q75	Q95
1	EM	2.741	39.879	0.003	0.009	0.017	0.136	3.222
	tmp-EM	**0.845**	**8.683**	**0.003**	**0.008**	**0.013**	**0.055**	**1.745**
2	EM	0.852	**9.006**	0.004	0.015	0.042	0.636	1.015
	tmp-EM	**0.412**	9.072	0.004	**0.011**	**0.027**	**0.34**	**0.782**
3	EM	1.185	14.636	0.015	0.078	0.183	0.414	1.742
	tmp-EM	**0.648**	**4.435**	**0.014**	**0.066**	**0.174**	**0.408**	**1.331**
4	EM	**2.008**	**13.156**	0.008	0.043	0.386	1.034	**4.553**
	tmp-EM	2.998	20.1	**0.006**	**0.028**	**0.374**	**0.637**	5.468
5	EM	1.772	12.175	0.005	0.015	0.035	0.664	5.813
	tmp-EM	**0.791**	**7.088**	**0.005**	**0.011**	**0.026**	**0.058**	**2.57**
6	EM	2.909	59.913	0.012	0.045	0.195	0.676	4.371
	tmp-EM	**2.072**	**25.898**	**0.008**	**0.023**	**0.062**	**0.34**	**2.883**

Conclusion We saw that tmp-EM achieved better average and median results with lower variances both on likelihood maximisation and parameter recovery for every Class (with very rare exceptions). A more global look at the overall distributions confirms this trend: the errors of tmp-EM are more centred on 0 with less spread than EM. This indicates that the tempering allows the EM algorithm to avoid falling into the first local maximum available and consistently find better ones. From

Table 5.4: Synthetic table focusing solely on the average and standard deviation (in parenthesis) of the losses and parameter reconstruction errors made by EM and tmp-EM. We note that the likelihood reached is higher with lower variance, and similarly, the parameter metrics on almost every class are better with lower variance for tmp-EM.

Metric	class	EM	tmp-EM
$-ln\, p_{\hat{\theta}}$		2 247.08 (69.62)	**2 218.80 (35.21)**
$\frac{ln\, p_{\theta_0} - ln\, p_{\hat{\theta}}}{ln\, p_{\theta_0}}$		0.12 (0.04)	**0.13 (0.04)**
$\frac{\hat{\pi}_k - \pi_k}{\pi_k}$	1	−0.19 (0.36)	**−0.17 (0.29)**
	2	0.11 (0.57)	**0.04 (0.33)**
	3	0.56 **(0.81)**	**0.45** (0.83)
	4	**0.10** (0.57)	0.10 **(0.43)**
	5	−0.08 (0.48)	**−0.02 (0.31)**
	6	−0.20 (0.43)	**−0.13 (0.40)**
$\frac{\|\hat{\mu}_k - \mu_k\|^2}{\|\mu_k\|^2}$	1	0.02 (0.12)	**2.10^{-3} (0.01)**
	2	0.04 **(0.07)**	**0.03** (0.07)
	3	0.97 (1.15)	**0.74 (1.07)**
	4	0.31 (0.49)	**0.29 (0.48)**
	5	0.73 (1.25)	**0.43 (1.05)**
	6	1.94 (2.83)	**0.81 (1.74)**
$KL(\Sigma, \widehat{\Sigma})$	1	2.74 (39.88)	**0.84 (8.68)**
	2	0.85 **(9.01)**	**0.41** (9.07)
	3	1.18 (14.64)	**0.65 (4.44)**
	4	**2.01 (13.16)**	3.00 (20.10)
	5	1.77 (12.17)	**0.79 (7.09)**
	6	2.91 (59.91)	**2.07 (25.90)**

the Machine Learning point of view, we highlighted that with our GMM parameters and $n = 500$ observations, it was able to better identify the different centroids, despite their ambiguity than the regular EM procedure. table 5.4 presents a comparative synthesis of the results of EM and tmp-EM.

5.7.2 Experiment 2: 3 clusters

In this section, we will assess the capacity of tmp-EM to escape from sub-optimal local maxima near the initialisation. The experimental protocol is the same as in the main paper. Let us recall it here. We confront the algorithm to situations where the true classes have increasingly more ambiguous positions, combined with initialisations designed to be hard to escape from. Even though we still follow the log-likelihood as a critical metric, for illustrative purposes we put more emphasis in this section on visualising whether the clusters were properly identify and following the paths in the 2D space of the estimated centroids towards their final values during the EM procedures.
The setup is the following: we have three clusters of similar shape and same weight. One is isolated and easily identifiable. The other two are next to one another, in a more ambiguous configuration. fig. 5.11 represents the three, gradually more ambiguous configurations.
We use two different initialisation types to reveal the behaviours of the two EMs. The first - which we call "barycenter" - puts all three initial centroids at the centre of mass of all the observed data points. However, none of the EM procedures would move from this initial state if the three GMM centroids were at the exact same position, hence we actually apply a tiny perturbation to make them all slightly distinct. The blue crosses on fig. 5.12 represent a typical barycenter initialisation. With this initialisation method, we assess whether the EM procedures are able to correctly estimate the positions of the three clusters, despite the ambiguity, when starting from a fairly neutral position, providing neither direction nor misdirection. On the other hand, the second initialisation type - which we call "2v1" - is voluntarily misguiding the algorithm by positioning two centroids on the isolated right cluster and only one centroid on the side of the two ambiguous left clusters. The blue crosses on fig. 5.13 represent a typical 2v1 initialisation. This initialisation is intended to assess whether the

methods are able to escape the potential well in which they start and make theirs centroids traverse the empty space between the left and right clusters to reach their rightful position. For each of the three parameter families represented on fig. 5.11, 1000 datasets with 500 observations each are simulated, and the two EMs are ran with both the barycenter and the 2v1 initialisation. In the case of tmp-EM, the oscillating temperature profile is used with parameters $T_0 = 5$, $r = 2$, $a = 0.6$, $b = 20$ for the barycenter initialisation, and $T_0 = 100$, $r = 1.5$, $a = 0.02$, $b = 20$ for the 2v1 initialisation. Although in the case of 2v1, the oscillations are not critical, and the simple temperature profile with $T_0 = 100$ and $r = 1.5$ works as well. We have two different sets of tempering hyper-parameters values, one for each of the two very different initialisation types. However, these values then remain the same for the three different parameter families and for every data generation within them. Underlining that the method is not excessively sensitive to the tempering parameters. The experiment with 6 clusters in section 5.7.1, already demonstrated that the same hyper parameters could be kept over different initialisation (and different data generations as well) when they were made in a non-adversarial way, by drawing random initial centroids uniformly among the data points.

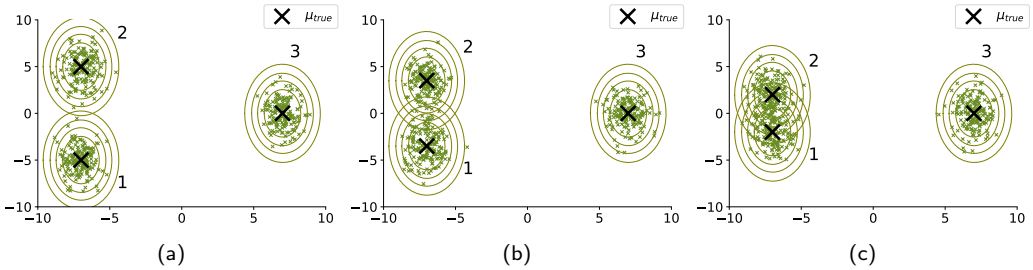


Figure 5.11: 500 sample points from a Mixture of Gaussians with 3 classes. The true centroid of each Gaussian are depicted by black crosses, and their true covariance matrices are represented by the confidence ellipses of level 0.8, 0.99 and 0.999 around the centre. There are three different versions of the true parameters. From left to right: the true μ_k of the two left clusters (mu_1 and mu_2) are getting closer while everything else stays identical.

Illustrative

First we illustrate on unique examples how tmp-EM is able to avoid falling for the tricky initialisations we set up.
As previously stated, the focus will be less on the likelihood optimisation for these illustrative examples. Indeed, they are meant to demonstrate that tmp-EM is able to cross the gaps and put the clusters in the right place even with the disadvantageous initialisation. The more relevant metric to assess success in this task is the error on μ (and in a lesser way, the error on Σ). One reason why the likelihood looses its ability to discriminate between failure and success in escaping the traps set by the initialisations is that there may not be a big likelihood gap between being completely wrong and mostly right. For instance placing two centroids (one of which is linked to an empty class) on the isolated left cluster and putting only one where the two ambiguously close clusters are could have a decent likelihood while being blatantly wrong.

On fig. 5.12, we represent the results of each EM after convergence for every of the three parameter set, when the start at the barycenter of all data points (blue crosses). The estimated means and covariance matrices of the GMM are represented by orange crosses and confidence ellipses respectively. In those examples, tmp-EM correctly identified the real clusters whereas EM put two centroids on the right, where only the isolated cluster stands, and only one on the left, where the two ambiguous clusters are. Figure 5.13 shows similar results, with the same conventions in the case of the "2v1" initialisation.

These different outcomes are exactly what one would expect: unlike the classical EM, tmp-EM

is by design supposed to avoid the local minima close to the initialisation by taking a more exploratory stance during its first steps. To demonstrate that point, we detail in fig. 5.14 to 5.17 the paths taken by the estimated centroids by tmp-EM in those simulations. The paths of the regular EM are straightforward convergences towards their final positions, and are not represented in these supplementary materials. fig. 5.14 represents the paths of the three cluster centroids during the iterations of tmp-EM. The parameter family is the least ambiguous (the two left cluster are well separated) with the "barycenter" initialisation. On fig. 5.15, the initialisation is "2v1" instead. The two following Figures, 5.16 and 5.17, also features the initialisations "barycenter" and "2v1" respectively, but with the most ambiguous parameter set, where the two left clusters are very close to one another.
These graphs are made of several rows of figures, each row representing a step in the EM procedure. In order to make the Figures informative, the number of steps between each row is not fixed, instead the most interesting steps are represented. Convergence is always achieved within 20 to 50 steps, so there are never big differences between the step gaps anyway. The first row is always the initial stage without any EM step, and the last one is the stage after convergence. Each of the three columns corresponds to one of the three centroids estimated by the EM procedure and represents its evolution in the 2D space, from initialisation to convergence. The corresponding estimated covariance matrix is represented by confidence ellipses. For each of the centroids, the observed data points are coloured accordingly to their (un-tempered) posterior probability of belonging to the associated class at this stage of the the algorithm. Plain blue being a low probability while bright green is a high probability.
We make the following observations on the steps taken by tmp-EM: with a "barycenter" initialisation (fig. 5.14 and 5.16), the three centroids gradually converge towards their final position (which correspond to true class centres in these cases) without too much hesitation. We also note that, since the three initial points are slightly distinct, there appears to be preferences at the very beginning, with each class having different high probability points right at the initialisation stage. However those preferences are not respected after a couple EM step, we generally see the centroids directing themselves towards different points than their initial favoured ones. This can be attributed to the tempering reshuffling the positions and preferences at the beginning. The "2v1" initialisation illustrates this phenomenon more clearly and in doing so, showcases the true power of the tempering. The very first steps after this very adversarial initialisation are not very remarkable: the single centroid on the left solidifies its position at the centre of the two ambiguous clusters, while the two centroids on the right try to share the single cluster they started in. However, very quickly this status quo is shattered and every estimated centroid jumps to a completely different position. On both fig. 5.15 and 5.17 we see the positions being completely reversed with the lonely centroid moving from the two left clusters to the isolated right one whereas the two close centroids make the inverse trip to reach the two clusters on the left. This jump is an indication that the tempering flattened the likelihood enough to allow each centroid to escape their potential wells. Effectively redoing the initialisation and allowing itself to start from more favourable positions. This behaviour is unattainable with the classical EM.

Quantitative

The quantitative analysis can be found in section 5.4.5.

5.7.3 Experiment on real data: Wine recognition dataset

To further validate tmp-EM, we compare it once more to the unmodified EM, this time on real observations from the scikit learn [128] classification data base "Wine" [42]. This dataset contains $p = 13$ chemical measurements of $n = 178$ wines each belonging to one of $K = 3$ families. Despite being in high dimension, this dataset is known as not very challenging (the classes are separable) and useful for testing new methods. We expect the unmodified EM to perform quite well already. For tmp-EM, we use the simple decreasing temperature profile, with no oscillations, the tempering parameters are $T_0 = 100$, $r = 4$. table 5.5 shows the result of 500 runs of the EMs from different random initial points. We focus on the likelihood and the error on μ_k, the other relevant metrics, not

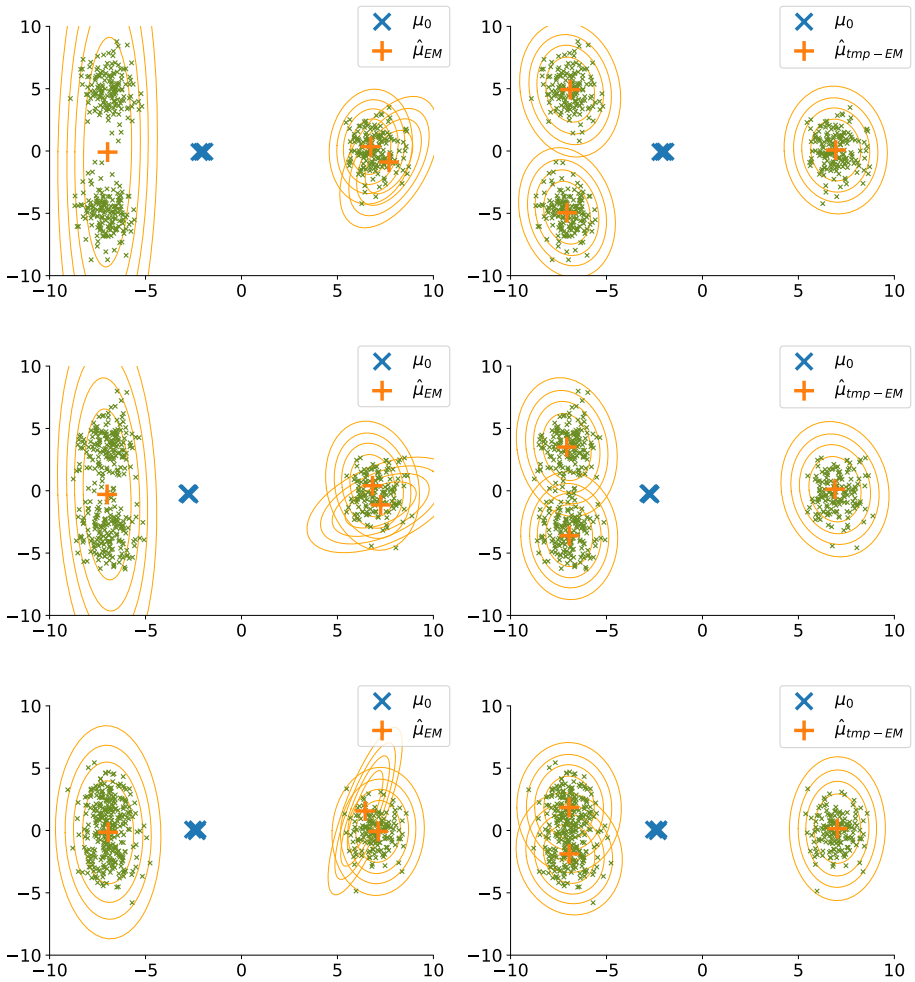


Figure 5.12: Typical final positioning of the centroids by EM (left column) and tmp-EM (right column) **when the initialisation is made at the barycenter of all data points** (blue crosses). The three rows represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those example, tmp-EM managed to correctly identify the position of the three real centroids.

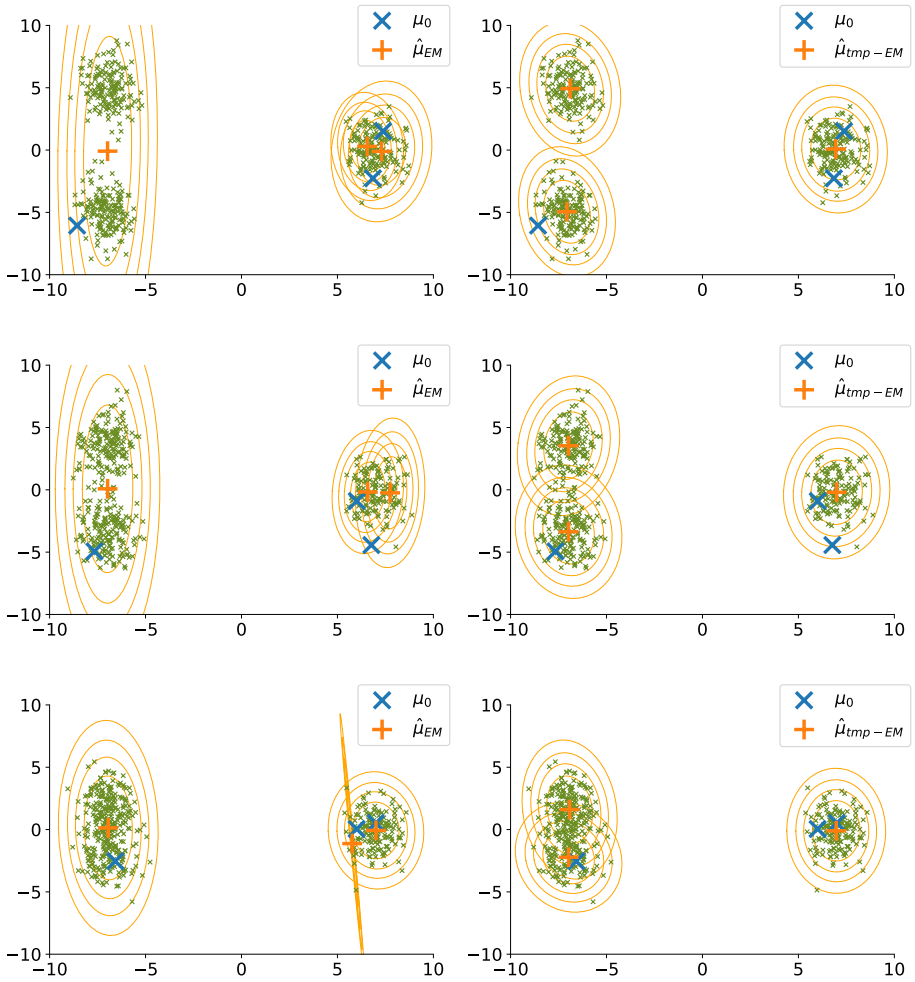


Figure 5.13: Typical final positioning of the centroids by EM (left column) and tmp-EM (right column) **when the initialisation is made by selecting two points in the isolated cluster and one in the lower ambiguous cluster** (blue crosses). The three rows represent the three gradually more ambiguous parameter sets. Each figure represents the positions of the estimated centroids after convergence of the EM algorithms (orange cross), with their estimated covariance matrices (orange confidence ellipses). In each simulation, 500 sample points were drawn from the real GMM (small green crosses). In those examples, although EM kept two centroids on the isolated cluster, tmp-EM managed to correctly identify the position of the three real centroids.

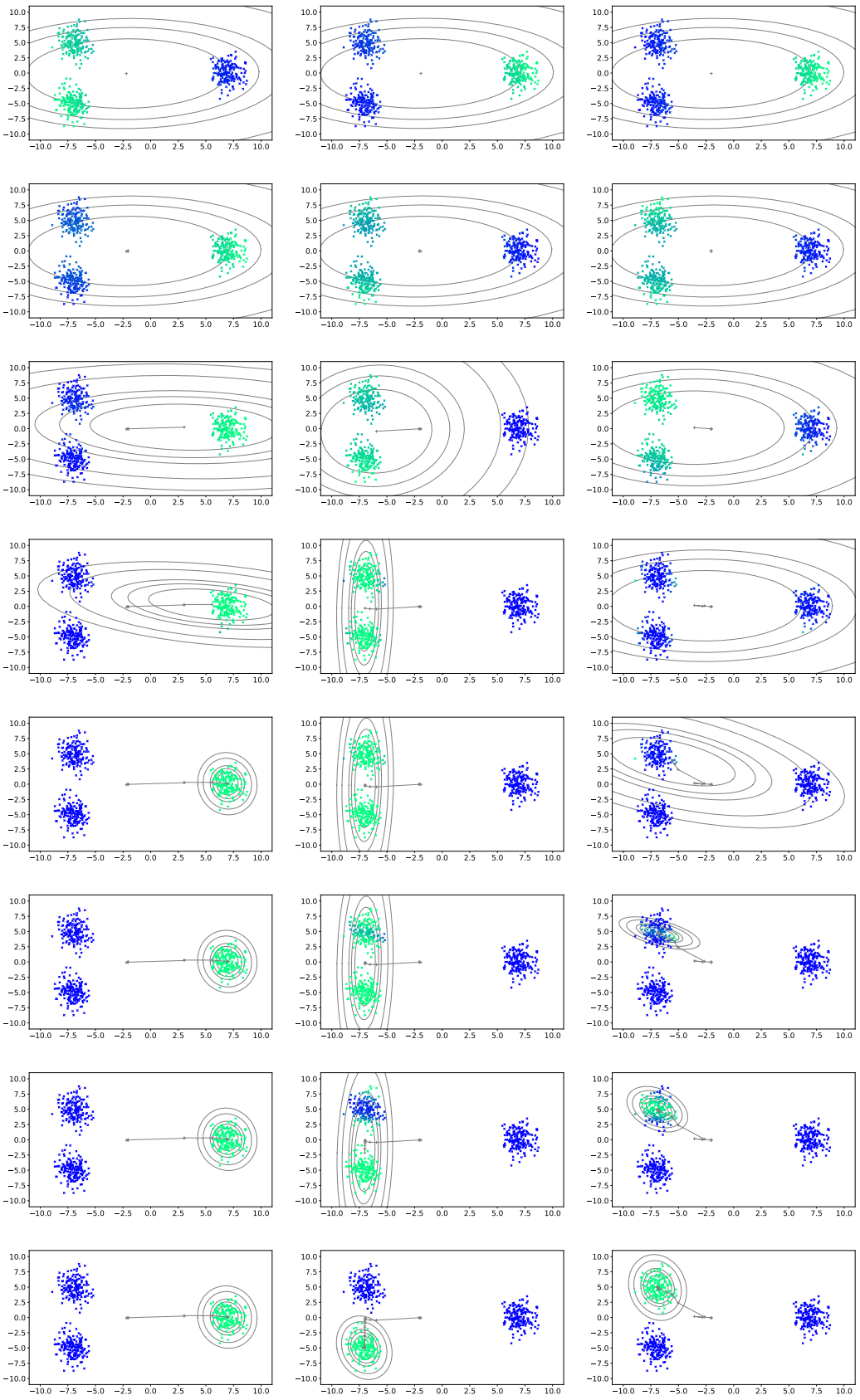


Figure 5.14: Paths of the centroids for tmp-EM with the "barycenter" initialisation. Parameter set 1 (least ambiguous).

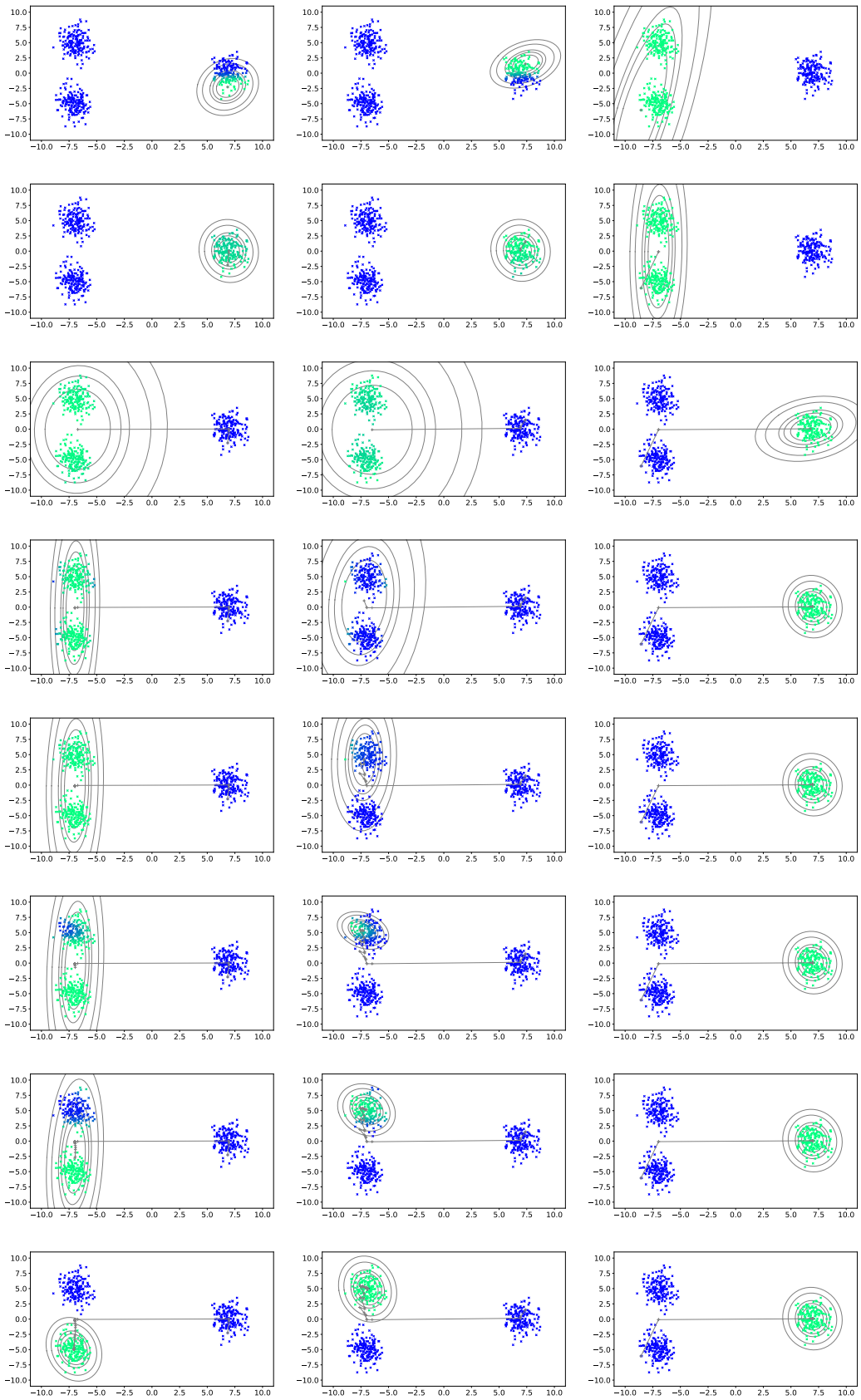


Figure 5.15: Paths of the centroids for tmp-EM with the "2v1" initialisation. Parameter set 1 (least ambiguous).

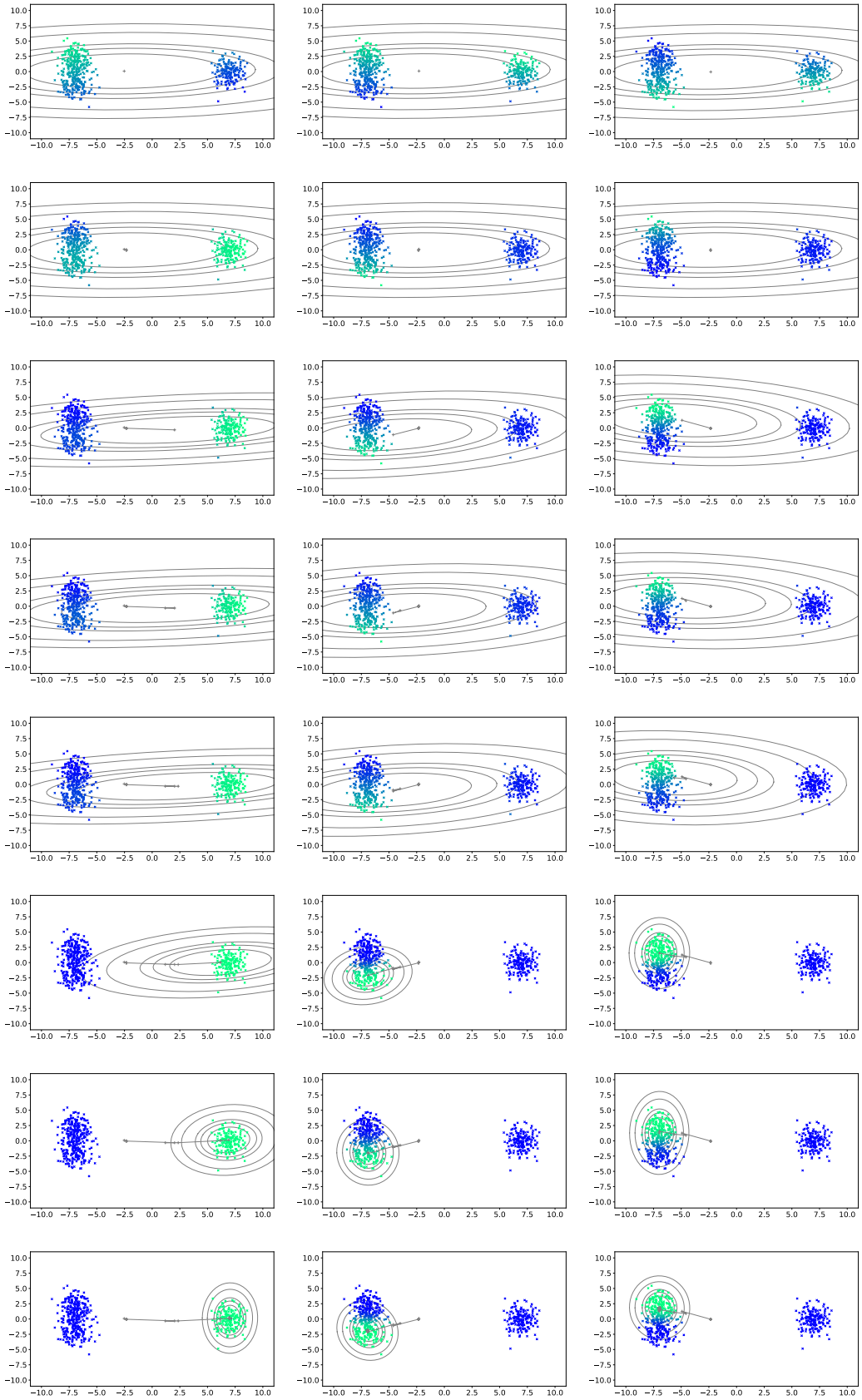


Figure 5.16: Paths of the centroids for tmp-EM with the "barycenter" initialisation. Parameter set 3 (most ambiguous).

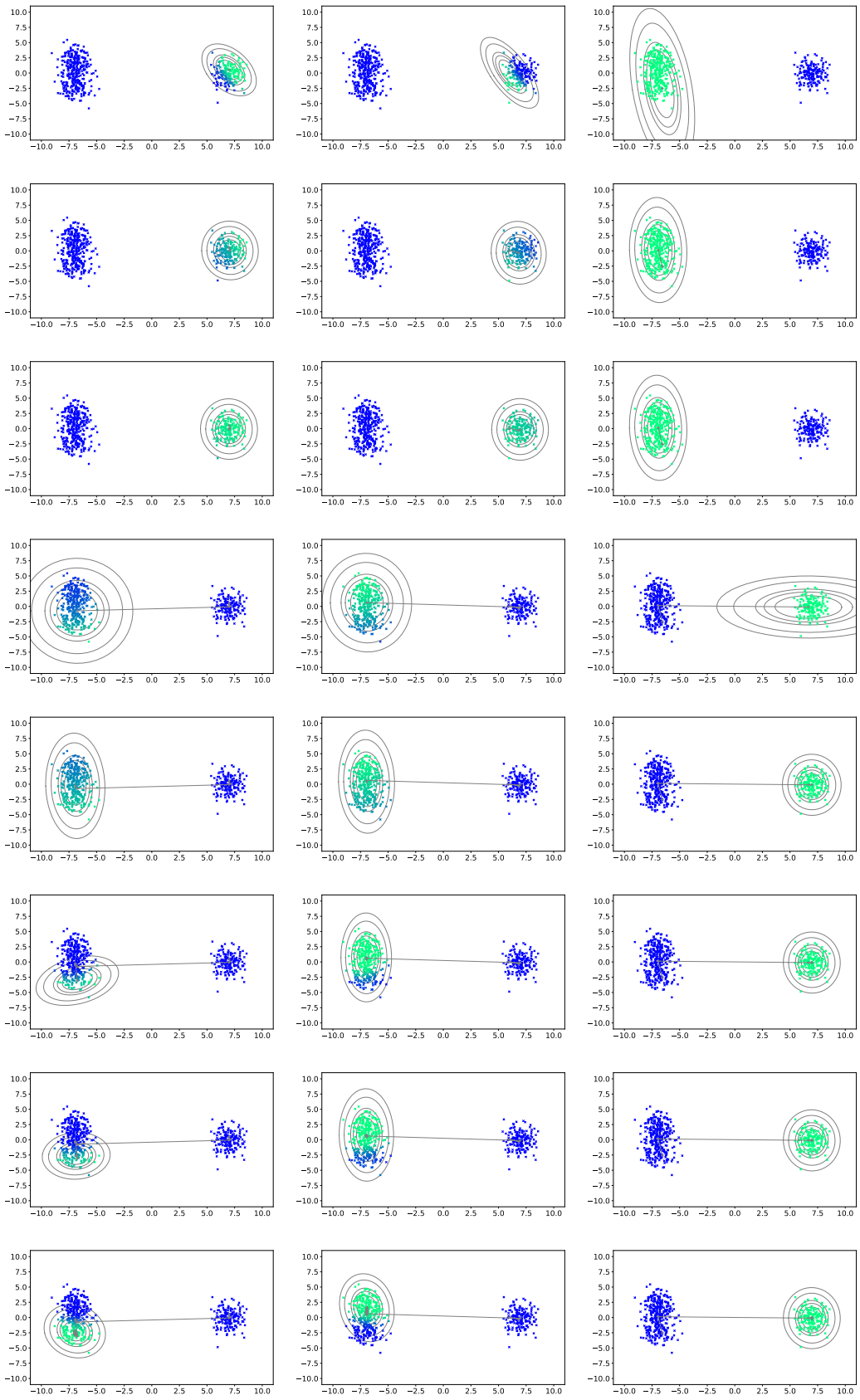


Figure 5.17: Paths of the centroids for tmp-EM with the "2v1" initialisation. Parameter set 3 (most ambiguous).

presented here, show the same tendencies. We observe, as usual, that tmp-EM reaches in average a lower negative log-likelihood with lower variance. The class centres are also better estimated. As expected, the errors made by the EM are already fairly small, however tmp-EM manages to go further and lower the errors on each class by approximately 17%, 18% and 11% respectively.
The results demonstrate that tmp-EM can improve the EM result on real data. Since this is an easy dataset, the difference is not as drastic as in the hard synthetic cases we ran the EMs by. Still, there was room to improve the EM results, and tmp-EM found those better solutions.

Table 5.5: Average and (standard deviation) of the EM and tmp-EM results over 500 random initialisation on the Wine recognition dataset. The classes on this dataset are easily identifiable hence the errors are low. Yet tmp-EM still improved upon the solutions of EM

metric	cl.	EM	tmp-EM
$-ln\, p_{\hat{\theta}}$		2923 (77)	**2905 (71)**
$\frac{\|\hat{\mu}_k - \mu_k\|^2}{\|\mu_k\|^2}$	1	0.017 (0.030)	**0.014 (0.028)**
	2	0.026 (0.034)	**0.021 (0.033)**
	3	0.089 (0.165)	**0.079 (0.156)**

5.8 Experiments on tmp-EM with Independent Factor Analysis

In this section, we present another application of the tmp-EM with Gaussian Mixture Models, but this time as part of a more complex model. The Independent Factor Analysis (IFA) model was introduced by [9] as an amalgam of Factor Analysis, Principal Component Analysis and Independent Component Analysis to identify and separate independent sources mixed into a single feature vector. From a practical standpoint, the mixing coefficient of each source is assumed to be drawn from a GMM, hence the EM. After estimation of the GMM parameters, the sources are recovered with an optimal non linear estimator. This is a complex model in which the EM plays a key part, works like [5] and [2] use it to assess new variants of the EM on a very practical application. The model is described as follows:

$$\forall i = 1, ..., L', \quad y_i = \sum_{j=1}^{L} H_{ij} x_j + u_i \,.$$

Where $y \in R^{L'}$ is one vector of observations, $H \in \mathbb{R}^{L'L}$ is the fixed matrix of the sources, $u \in R^{L'}$ the vector of noise, and $x \in R^L$ the random mixing coefficient. Each component x_j is assumed to be drawn from its own GMM.
An EM that converges too soon towards a local extremum has every chance to yield sub-optimal estimated sources. We demonstrate in this section that an IFA method with tmp-EM can recover sources closer to the original when they are known, and cleaner, more stable looking sources in general.

5.8.1 Synthetic IFA

We start with a toy example, where the true sources are two easily distinguishable images. As shown on fig. 5.19, one is a white square on a black background and the other is a white cross on a similar black background but positioned differently. However, once these two sources are mixed and noised, it becomes much harder to identify them with the naked eye - as illustrated by fig. 5.19 - and a quantitative method is required to properly separate them. To separate the sources, the identification model assumes that the coefficients used to mix the two sources are drawn from mixtures of Gaussian.

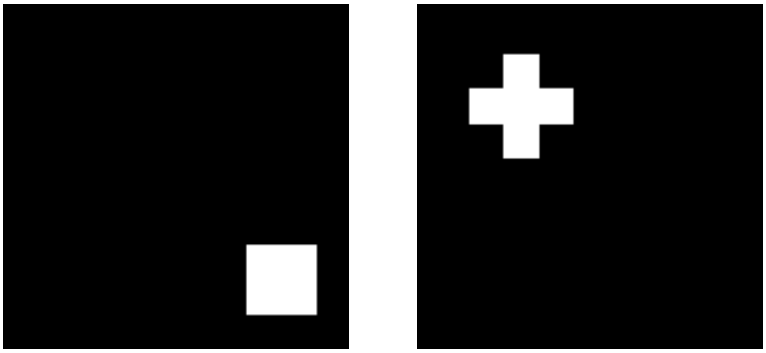


Figure 5.18: The two real sources of a synthetic source mixing model. They are images of size 20×20 made of a black background with a white symbol localised either on the bottom left or top right corner.

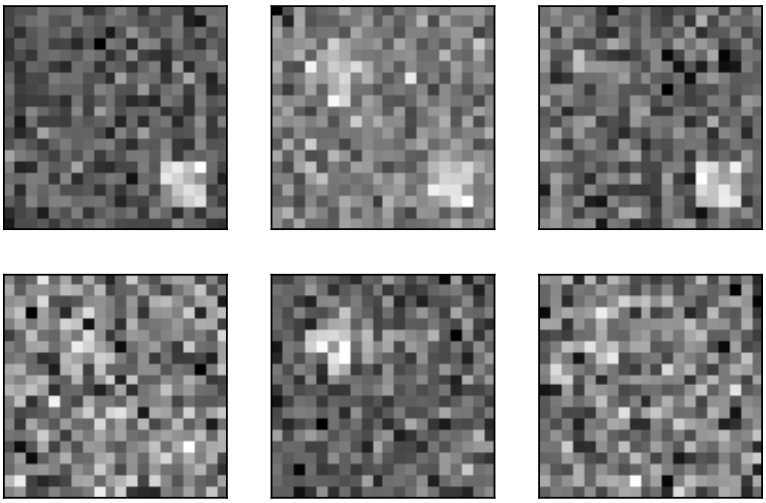


Figure 5.19: 6 typical observation obtained with the source mixing model. With the noise, the sources are harder to identify.

The outputs were voluntarily generated in a different way to show the generalisation capabilities of the mixture of Gaussian assumption. We run an EM and a tmp-EM algorithm to estimate the parameter of those mixtures, recovering in the process an estimation of the mixing matrix H. fig. 5.20 illustrates the sources typically estimated by each of the two procedures. Although there is noise, tmp-EM essentially identified and corrected the sources correctly. Whereas EM did not manage to completely turn off the square symbol in the estimated sources supposedly dedicated to the cross. fig. 5.21 displays the quantitative results of several runs over different simulated datasets. It represents the empirical distribution of l_2 errors made on the estimation of the source matrix H by the two EMs. As illustrated by the table in fig. 5.21, the solutions of tmp-EM have lower mean and median.

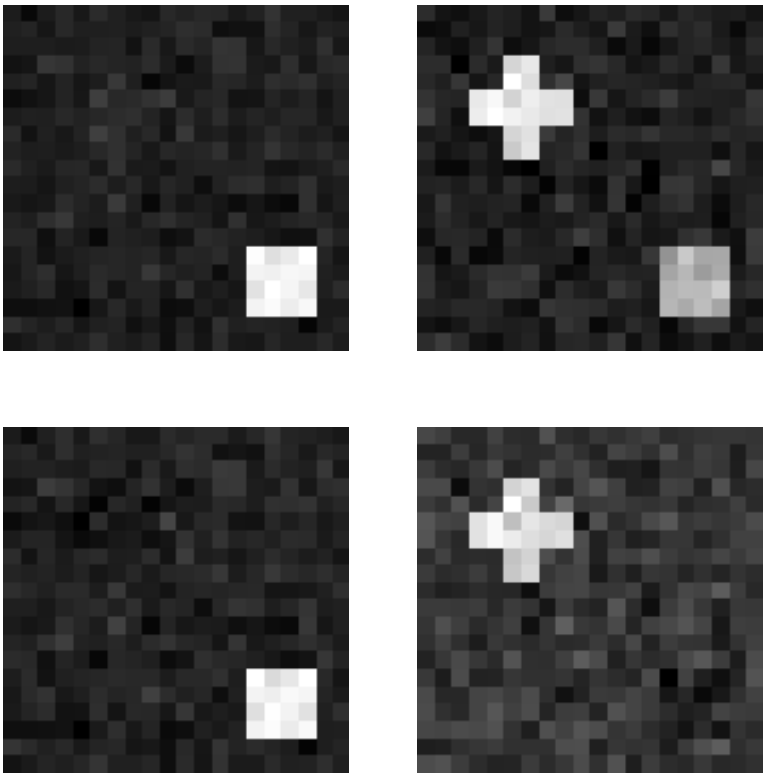
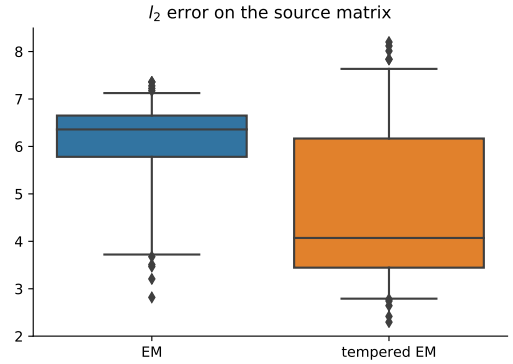


Figure 5.20: Estimated sources by EM (up) and tmp-EM (down). The two real sources were correctly identify by tmp-EM, but EM did not fully separate the cross and the square.

5.8.2 ZIP code

We apply this IFA algorithm to the ZIP code dataset from Elements of Statistical learning. This dataset contains handwritten digits between 0 and 9. In this study, we keep only the digits 0,3, 8 (all three being ambiguously similar) and 7 (very different from the three others). We make all classes even by removing half of the 0 which are originally more numerous. When applying Independent Factor Analysis to such data, one hopes that the distinct digits will be identified as the separable sources making up the signal. We run the IFA model with a Mixture of Gaussians model with a regular and a tempered EM. In the mixing model used, each mixture is composed of two classes. The tempering was made with the oscillating profile, with hyper-parameters: $T_0 = 50$, $b = 20$, $r = 3$, $a = 0.02$.
fig. 5.22 displays the estimated sources by the IFA procedure with either EM or tmp-EM at their core. EM did not really identify an "8" source. Instead, its "3" is a bit ambiguously close to and



(a)

	EM	tmp-EM
median	6.36	**4.07**
mean	6.01	**4.75**
std	**1.05**	1.65
Q5	3.72	**2.79**
Q25	5.78	**3.44**
Q75	6.65	**6.17**
Q95	**7.13**	7.64

(b)

Figure 5.21: Empirical distribution of the l_2 error on the source matrix H made by EM and tmp-EM. With tmp-EM, we shift the distribution towards the lower errors, with smaller average and median. The numeric values of the quantiles and other statistics can be found in the table, the better ones being in **bold**.

"8", and the rightmost source in fig. 5.22 seems like an amalgamation of the four digits. Moreover, the source "7" estimated by EM is actually a mix between a "7" and a "0". On the other hand, the sources estimated by tmp-EM each correspond clearly to a different digit. There is an "8", the "7" is not fused with a "0", the "3" is sharper and more distinct from an "8" then the corresponding EM source, and even the "0" is more symmetrical with tmp-EM than with EM. Tempering the EM within the IFA algorithm allowed for a cleaner separation of the sources. One can infer that tmp-EM was able to identify and reach a better local maximum of the loss function.

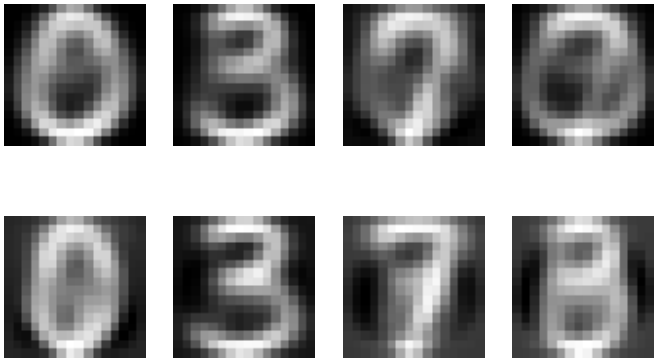


Figure 5.22: Estimated sources by EM (up) and tmp-EM (down). The "8" and the "7" in particular were much better identified by tmp-EM. Moreover, with tempering the "0" has a more symmetrical shape and the "3" is sharper and less ambiguous.

5.9 Conclusions

We proposed the Deterministic Approximate EM class to bring together the many possible deterministic approximations of the E step. We proved a unified theorem, with mild conditions on the approximation, which ensures the convergence of the algorithms in this class. Then, we showcased members of this class that solve the usual practical issues of the EM algorithm. For intractable E step, we introduced the Riemann approximation EM, a less parametric and deterministic alternative to the extensive family of MC-EM. We showed on an empirical intractable example how the

Riemann approximation EM was able to increase the likelihood and recover every parameter in a satisfactory manner with its simplest design, and no hyper parameter optimisation. For cases where one wants to improve the solution of the EM, we proved that the tempered EM, introduced under a different form in [156], is a specific case of the Deterministic Approximate EM. Moreover, we showed that the commonly used models benefit from the convergence property as long as the temperature profile converges towards 1. This justifies the use of many more temperature profiles than the ones tried in [156] and [120]. We ran an in-depth empirical comparison between tmp-EM and the regular EM. In particular, we showed how tmp-EM was able to escape from adversarial initial positions, a task that sometimes required complex non-monotonous temperature schemes, which are covered by our theorem. Finally, we added the Riemann approximation in order to apply the tempering in intractable cases. We were then able to show that the tmp-Riemann approximation massively improved the performances of the Riemann approximation, when the initialisation is ambiguous. Future works will improve both methods. The Riemann approximation will be generalised to be applicable even when the hidden variable is not bounded, and an intelligent slicing of the integration space will improve the computational performances in high dimension. For the tempered EM, tuning the temperature parameters in an adaptive way during the procedure will remove the necessity for preliminary hyper-parameter tuning by grid search.

Chapter 6

Mixture of Conditional Gaussian Graphical Models for unlabelled heterogeneous populations in the presence of co-factors

This Chapter has been submitted for review.

Conditional correlation networks, through Gaussian Graphical Modelling, are widely used to describe the direct interactions between the component of a random vector. In the case of a Heterogeneous population, Hierarchical Gaussian Graphical Models (GGM) are used to describe the different sub-populations with one graph each. When the sub-population labels are unknown, unsupervised methods must be implemented to estimate these labels in addition to the GGM parameters. Expectation Maximisation (EM) algorithms for Mixtures of GGM were proposed as a natural unsupervised extension of the Hierarchical GGM. However, we argue that, with most real data, class affiliation cannot be described with a Mixture of Gaussian, which mostly groups data points according to geometrical proximity in the feature space. In particular, there often exists external co-features whose values affect the features' average value, scattering across the feature space data points belonging to the same sub-population. Additionally, if the co-features' effect on the feature is Heterogeneous (sub-population dependent), then the estimation of this effect cannot be separated from the cluster identification. In this Chapter, we propose a Mixture of Conditional Graphical Models (CGGM) that integrates co-features with heterogeneous effects. Within this model, the position of each data point is readjusted in order to remove the effect of the co-features, and regroup the data points into sub-population corresponding clusters. We develop a penalised EM algorithm to estimate the model parameters with any desired sparsity structure for the graphs. We demonstrate on synthetic data how this method fulfils its goal and succeeds in identifying the sub-populations where the Mixtures of GGM are disrupted by the effect of the co-features. Likewise, on real Alzheimer's Disease data, we show that our method recovers patient clusters better correlated with the diagnostic.

6.1 Introduction

The conditional correlation networks are a popular tool to describe the co-variations between the components of a random vector. Within the Gaussian Graphical Model (GGM) framework, introduced in [35], the random vector of interest is modelled as a Gaussian vector $\mathcal{N}(\mu, \Sigma)$, and the

conditional correlation networks can be recovered from the sparsity of the inverse covariance matrix $\Lambda := \Sigma^{-1}$. In this Chapter, we consider the case of an unlabelled heterogeneous population, in which different sub-populations (or "classes") are described by different networks. Additionally, we take into account the presence of observed co-features (discrete and/or continuous) that have a heterogeneous (class-dependent) impact on the values of the features. The absence of known class labels turns the analysis of the population into an unsupervised problem. As a result, any inference method will have to tackle the problem of cluster discovery in addition to the parameter estimation. The former is a crucial task, especially since the relevance of the estimated parameters is entirely dependent on the clusters identified. The co-features, if their effects on the features are consequent, can greatly disrupt the clustering. Indeed, any unsupervised method will then be more likely to identify clusters correlated with the values of the co-features than with the hidden sub-population labels. This occurs frequently when analysing biological or medical features. To provide a simple illustration, if one runs an unsupervised method on an unlabelled population containing both healthy and obese patients, using the body fat percentage as a feature, then the unearthed clusters are very likely to be more correlated with the gender of the patients (a co-feature) rather than with the actual diagnostic (the hidden variable). Additionally, the fact that the effect of the gender on the average body fat is also dependent on the diagnostic (class-dependent effect) makes the situation even more complex.

Unsupervised GGM have received recent attention, with works such as [52] and [61] adapting the popular supervised joint Hierarchical GGM methods of [119] and [30] to the unsupervised case. When the labels are known in advance, these joint Hierarchical GGM are useful models to estimate several sparse conditional correlation matrices and are modular enough to allow for the recovery of many different forms of common structure between classes. However, we argue that they are not designed for efficient cluster identification in the unsupervised scenario, and will very likely miss the hidden variable and find clusters correlated to the most influential co-features instead. Which in turn will result in the estimation of irrelevant parameters. Even when there are no pre-existing hidden variables to recover, and the unsupervised method is run "blindly", it is uninteresting to recover clusters describing the values to already known co-features. Instead, one would rather provide beforehand the unsupervised method with the information of the co-features' values and encourage it to recover new information from the data.

In order to take into account the effect of co-features on features, [173] and [171] introduced the Conditional Gaussian Graphical Models (CGGM). Within this model, the average effect of the co-features is subtracted from the features, in order to leave only orthogonal effects. Both [173] and [143] worked with homogeneous populations, but the Hierarchical form of the CGGM was introduced by [27] to study labelled heterogeneous populations, with heterogeneous effects of the co-features on the features. Recent works such as [69] and [123] have adapted the state of the art supervised joint Hierarchical GGM methods for the CGGM. However, to the best of our knowledge, there has been no effort to make use of the CGGM in the unsupervised case.

In this article, we introduce a Mixture of Conditional GGM that models the class-dependent effect of the co-features on the features. We propose an Expectation-Maximisation (EM) procedure to estimate this model without prior knowledge of the class labels. This EM algorithm can be regularised with all the structure-inducing penalties introduced for the supervised joint Hierarchical CGGM. Hence, the recovered sparse conditional correlation graphs can present any of the desired form of common structure. Moreover, with an additional penalty, we can also enforce structure within the parameter describing the relation between co-features and features.

Thanks to the inclusion of the co-features within the model, our EM algorithm is able to avoid trivial clusters correlated with the co-features' values, and instead unearths clusters providing new information on the population. Additionally, since our model takes into account heterogeneous effects of the co-features, our EM can handle the more complex scenarios, where the co-features act differently on the features in each sub-population.

We demonstrate the performance of our method on synthetic and real data. First with a 2-dimensional toy example, where we show the importance of taking into consideration the (heterogeneous) effects of co-features for the clustering. Then, in higher dimension, we demonstrate that our EM with Mixture of CGGM consistently outperforms, both in terms of classification and

parameter reconstruction, the EM with a Mixture of GGM (used in [52] and [61]), as well as an improved Mixture of GGM EM, that takes into consideration a homogeneous co-feature effect. Finally, on real Alzheimer's Disease data, we show that our method is the better suited to recover clusters correlated with the diagnostic, from both MRI and Cognitive Score features.

6.2 Supervised Hierarchical GGM and CGGM

In this section, we summarise the whys and wherefores of Gaussian Graphical Modelling: the simple models for homogeneous populations, as well as the hierarchical models for heterogeneous populations. First, we explore the classical Gaussian Graphical Models techniques to describe a vector of features $Y \in \mathbb{R}^p$, then we discuss the Conditional Gaussian Graphical Models implemented in the presence of additional co-features $X \in \mathbb{R}^q$. For every parametric model, we call θ the full parameter, and p_θ the probability density function. Hence, in the example of a Gaussian model $\theta = (\mu, \Sigma)$. For hierarchical models with K classes, we will have K parameters $(\theta_1, ..., \theta_K)$.

6.2.1 Basics of Hierarchical Gaussian Graphical Models

In the classical GGM analysis introduced by [35], the studied features $Y \in \mathbb{R}^p$ are assumed to follow a Multivariate Normal distribution: $Y \sim \mathcal{N}(\mu, \Sigma)$. The average μ is often ignored and put to 0. With $\Lambda := \Sigma^{-1}$, the resulting distribution is:

$$p_\theta(Y) = (2\pi)^{-p/2} \left|\Lambda\right|^{1/2} exp\left(-\frac{1}{2} Y^T \Lambda Y\right) . \tag{6.1}$$

In this case $\theta = \Lambda$. Using the property that $corr(Y_u, Y_v | (Y_w)_{w \neq u,v}) = -\frac{(\Lambda)_{uv}}{\sqrt{(\Lambda)_{uu}(\Lambda)_{vv}}}$, the conditional correlation network is obtained using a sparse estimation of the precision (or "inverse-covariance") matrix Λ. Heterogeneous population, where different correlation networks may exist for each subpopulation (or "class"), can be described with the Hierarchical version of the GGM (6.1). With K classes, Let $\theta := (\theta_1, ..., \theta_k)$ be the parameter for each class and $z \in [\![1, K]\!]$ the categorical variable corresponding to the class label of the observation Y. With $\theta_k := \Lambda_k$ and z known, the Hierarchical density can be written:

$$\begin{aligned} p_\theta(Y|z) &= \sum_{k=1}^{K} \mathbb{1}_{z=k} p_{\theta_k}(Y) \\ &= \sum_{k=1}^{K} \mathbb{1}_{z=k} (2\pi)^{-p/2} \left|\Lambda_k\right|^{1/2} exp\left(-\frac{1}{2} Y^T \Lambda_k Y\right) . \end{aligned} \tag{6.2}$$

Mirroring the famous Graphical LASSO (GLASSO) approach introduced by [179] and [12] for homogeneous populations, many authors have chosen to estimate sparse $\widehat{\Lambda}_k$ as penalised Maximum Likelihood Estimator (MLE) of Λ_k. For $i = 1, ..., n$, let $Y^{(i)}$ be independent identically distributed (iid) feature vectors and $z^{(i)}$ their labels. These MLE are computed from the simple convex optimisation problem

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \; -\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathbb{1}_{z^{(i)}=k} \, ln \, p_{\theta_k}(Y^{(i)}) + pen(\theta) . \tag{6.3}$$

Where the convex penalty $pen(\theta)$ is usually designed to induce sparsity within each individual $\hat{\Lambda}_k$ as well as to enforce a certain common structure between the $\hat{\Lambda}_k$. This common structure is a desirable outcome when the different sub-populations are assumed to still retain core similarities. Following in the footsteps of [59], most authors propose such a joint estimation of the matrices Λ_k. In the case of the penalised MLE estimation (6.3), the form of the resulting common structure is dependent on the penalty. For instance, [30] propose the "Fused Graphical LASSO"" and "Group Graphical LASSO" penalties that encourage shared values and shared sparsity pattern across the different

Λ_k respectively. Likewise, [172] propose another fused penalty to incentivise common values across matrices. With their node based penalties, [119] can encourage the recovery of common hubs in the graphs.

Remark. Within a hierarchical model, one can also take $\theta_k := (\mu_k, \Lambda_k)$, and adapt $p_{\theta_k}(Y)$ accordingly, since it is natural to allow each sub-population to have different average levels μ_k.

6.2.2 Conditional GGM in the presence of co-features

In some frameworks, additional variables, noted $X \in R^q$ and called "co-features" or "cofactors" can be observed alongside the regular features within the Gaussian vector $Y \in \mathbb{R}^p$. In all generality, X can be a mix of finite, discrete and continuous random variables. In the GGM analysis, these co-features are not included as nodes of the estimated conditional correlation graph. Instead, they serve to enrich the conditioning defining each edge: in the new graph, there is an edge between the nodes Y_u and Y_v if and only if $cov(Y_u, Y_v|(Y_w)_{w\neq u,v}, X) \neq 0$. The Conditional Gaussian Graphical Models (CGGM) were introduced by [173] and [143] in order to properly take into account the effect of X on Y and easily identify the new conditional correlation network in-between the Y. They propose a linear effect, expressed by the conditional probability density function (pdf):

$$p_\theta(Y|X) = (2\pi)^{-p/2} \left|\Lambda\right|^{1/2} exp\left(-\frac{1}{2}(Y + \Lambda^{-1}\Theta^T X)^T \Lambda (Y + \Lambda^{-1}\Theta^T X)\right), \tag{6.4}$$

with $\Theta \in \mathbb{R}^{q\times p}$ and $\theta = \{\Lambda, \Theta\}$. In other words: $Y|X \sim \mathcal{N}\left(-\Lambda^{-1}\Theta^T X, \Lambda^{-1}\right)$. Two main branches of CGGM exist, depending on whether the pdf of X is also modelled. In this work, we chose to impose no model on X. The lack of assumption on the density of X provides far more freedom than the joint Gaussian assumption. In particular, X can have categorical and even deterministic components. This allows us to integrate any observed variables without restriction to the model.
To tackle heterogeneous populations, works such as [27] have introduced the Hierarchical version of the CGGM pdf:

$$p_\theta(Y|X, z) = \sum_{k=1}^{K} \mathbb{1}_{z=k} \left(\frac{|\Lambda_k|}{(2\pi)^p}\right)^{\frac{1}{2}} exp\left(-\frac{1}{2}(Y + \Lambda_k^{-1}\Theta_k^T X)^T \Lambda_k (Y + \Lambda_k^{-1}\Theta_k^T X)\right). \tag{6.5}$$

In particular, [69] have adapted the penalised MLE (6.3) to the Hierarchical CGGM density for some of the most popular GGM penalties. With a idd sample $(Y^{(i)}, X^{(i)}, z^{(i)})_{i=1}^n$, the corresponding penalised CGGM MLE can be written;

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \mathbb{1}_{z^{(i)}=k}\, ln\, p_{\theta_k}(Y^{(i)}|X^{(i)}) + pen(\theta)\,. \tag{6.6}$$

Remark. To include a regular average value for Y, independent of the values of X, one can simply add a constant component equal to "1" in X.

6.3 Mixtures of CGGM for unlabelled heterogeneous population

In this section, we tackle the problem of an unlabelled heterogeneous population. We introduce a Mixture of Conditional Gaussian Graphical Model to improve upon the state of the art unsupervised methods by taking into consideration the potent co-features that can drive the clustering. We develop a penalised EM algorithm to both identify data clusters and estimate sparse, structured, model parameters. We justify that our algorithm is usable with a wide array of penalties and provide detailed algorithmic for the Group Graphical LASSO (GGL) penalty.

6.3.1 Presentation and motivation of the model

When the labels of a heterogeneous population are missing, supervised parameter estimation methods like (6.3) have to be replaced by unsupervised approaches that also tackle the problem of cluster discovery. When z is unknown, the Hierarchical model (6.2) can easily be replaced by a Mixture model with observed likelihood:

$$p_{\theta,\pi}(Y) = \sum_{k=1}^{K} \pi_k p_{\theta_k}(Y)\,, \tag{6.7}$$

and complete likelihood:

$$p_{\theta,\pi}(Y,z) = \sum_{k=1}^{K} \mathbb{1}_{z=k}\pi_k p_{\theta_k}(Y)\,. \tag{6.8}$$

Where $\pi_k := \mathbb{P}(z=k)$ and $\pi := (\pi_1, ..., \pi_k)$. Then, the supervised penalised likelihood maximisation (6.3) can be adapted into the penalised observed likelihood optimisation:

$$\hat{\theta}, \hat{\pi} = \underset{\theta,\pi}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k\, p_{\theta_k}\left(Y^{(i)}\right)\right) + pen(\theta,\pi)\,. \tag{6.9}$$

This is a non-convex problem, and authors such as [187] and [86] have proposed EM algorithms to find local solutions to (6.9). They omit however the common structure inducing penalties that are the signature of the supervised joint GGM methods. The works of [52] and [61] correct this by proposing EM algorithms that solve (6.9) for some of the joint-GGM penalties, such as the Fused and Group Graphical LASSO penalties.
By design, the EM algorithm must handle the cluster identification jointly with the mixture parameters estimation. The underlying assumption is that the different sub-populations can be identified as different clusters in the feature space. With real data, and especially medical data, this is generally untrue, as many factors other than the class label can have a larger impact on the position of the data points in the feature space. Even when there are no specific sub-populations to recover, and the EM is ran "blindly" in order to observe which data points are more naturally grouped together by the method, the unearthed clusters have every chance to be very correlated with very influential but trivial external variables, such as the age group or the gender. In order to guide the cluster discovery of the EM algorithm, we propose a Mixture of Conditional Gaussian Graphical Models with which the overbearing effect of trivial external variables can be removed. By placing all external observed variable into X, we define the Mixture of CGGM with its observed likelihood:

$$\begin{aligned} p_{\theta,\pi}(Y|X) :=& \sum_{k=1}^{K} \pi_k p_{\theta_k}(Y|X) \\ =& \sum_{k=1}^{K} \pi_k \left(\frac{|\Lambda_k|}{(2\pi)^p}\right)^{\frac{1}{2}} exp\left(-\frac{1}{2}(Y+\Lambda_k^{-1}\Theta_k^T X)^T \Lambda_k (Y+\Lambda_k^{-1}\Theta_k^T X)\right)\,. \end{aligned} \tag{6.10}$$

Within this model, the position of each feature vector Y is corrected by its, class-dependent, linear prediction by the co-features X: $\mathbb{E}[Y|X,z=k] = -\Lambda_k^{-1}\Theta_k^T X$. In other words the "Mixture of Gaussians" type clustering is done on the residual vector $Y - \mathbb{E}[Y|X,z=k] = Y + \Lambda_k^{-1}\Theta_k^T X$. Hence, even if the co-features X have a class-dependent impact on the average level of the features Y, the Mixture of CGGM model is still able to regroup in the feature space the observations $Y^{(i)}$ that belong to the same class, $z^{(i)} = k$. We illustrate this dynamic in section 6.4.1.
Like the previous works on joint-GGM estimation, our goal is to estimate the parameters of model (6.10) with sparse inverse-covariance matrices Λ_k and common structure across classes. Sparsity in the matrices Θ_k is also desirable for the sake of interpretation. Hence, we define the following penalised Maximum Likelihood problem:

$$\hat{\theta}, \hat{\pi} = \underset{\theta,\pi}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k\, p_{\theta_k}\left(Y^{(i)}|X^{(i)}\right)\right) + pen(\theta,\pi)\,. \tag{6.11}$$

As with (6.9), this is a non-convex problem, and we define an EM algorithm to find local minima of the optimised function.

6.3.2 Penalised EM for the Mixture of CGGM

In this section, we provide the detailed steps of a penalised EM algorithm to find local solution of the non-convex penalised MLE (6.11) in order to estimate the parameters of the mixture model (6.10) with inverse-covariance sparsity as well as common structure. First we provide the different steps of the algorithm and justify that it can be run with a wide array of penalty functions. Then, we provide a detailed optimisation scheme for the Group Graphical Lasso (GGL) penalty specifically.

EM algorithm for Mixtures of CGGM. With n fixed $\left\{X^{(i)}\right\}_{i=1}^n$ and n iid observations $\left\{Y^{(i)}\right\}_{i=1}^n$ following the mixture density $p_{\theta,\pi}(Y|X)$ given in (6.10), the penalised observed negative log-likelihood to optimise is:

$$-\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K}\pi_k\, p_{\theta_k}\left(Y^{(i)}|X^{(i)}\right)\right) + pen(\theta,\pi)\,. \tag{6.12}$$

We will not redo here all the calculations for the EM applied to a mixture. In the end, we get an iterative procedure updating the current parameter $(\theta^{(t)},\pi^{(t)})$ with two steps. The Expectation (E) step is:

$$p_{i,k}^{(t)} := \mathbb{P}_{\theta^{(t)},\pi^{(t)}}(z^{(i)}=k|Y^{(i)},X^{(i)}) = \frac{p_{\theta_k^{(t)}}(Y^{(i)}|X^{(i)})\pi_k^{(t)}}{\sum_{l=1}^{K} p_{\theta_l^{(t)}}(Y^{(i)}|X^{(i)})\pi_l^{(t)}}\,.$$

More explicitly, by replacing $p_{\theta_k}(Y|X)$ by its formula (6.4):

$$(E)\quad p_{i,k}^{(t)} = \frac{|\Lambda_k|^{-\frac{1}{2}}\, exp\left(\frac{1}{2}(Y^{(i)}+\Lambda_k^{-1}\Theta_k^T X^{(i)})^T\Lambda_k(Y^{(i)}+\Lambda_k^{-1}\Theta_k^T X^{(i)})\right)}{\sum_{l=1}^{K}|\Lambda_l|^{-\frac{1}{2}}\, exp\left(\frac{1}{2}(Y^{(i)}+\Lambda_l^{-1}\Theta_l^T X^{(i)})^T\Lambda_l(Y^{(i)}+\Lambda_l^{-1}\Theta_l^T X^{(i)})\right)}\,. \tag{6.13}$$

The M step is:

$$\theta^{(t+1)},\pi^{(t+1)} = \underset{\theta,\pi}{\text{argmin}} -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)}\left(ln\, p_{\theta_k}(Y^{(i)}|X^{(i)}) + ln\,\pi_k\right) + pen(\theta,\pi)\,.$$

Assuming that there is no coupling between π and θ in the penalty, i.e. $pen(\pi,\theta) = pen_\pi(\pi)+pen_\theta(\theta)$, then the two optimisations can be separated:

$$\theta^{(t+1)} = \underset{\theta}{\text{argmin}} -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)}\, ln\, p_{\theta_k}(Y^{(i)}|X^{(i)}) + pen_\theta(\theta)\,,$$

$$\pi^{(t+1)} = \underset{\pi}{\text{argmin}} -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)}\, ln\,\pi_k + pen_\pi(\pi)\,.$$

Let us denote the sufficient statistics $n_k^{(t)} := \sum_{i=1}^n p_{i,k}^{(t)}$, $S_{YY}^{k,(t)} := \frac{1}{n}\sum_{i=1}^n p_{i,k}^{(t)} Y^{(i)}Y^{(i)\,T}$, $S_{YX}^{k,(t)} := \frac{1}{n}\sum_{i=1}^n p_{i,k}^{(t)} Y^{(i)}X^{(i)\,T}$ and $S_{XX}^{k,(t)} := \frac{1}{n}\sum_{i=1}^n p_{i,k}^{(t)} X^{(i)}X^{(i\,T)}$. Then, the M step can be formulated as:

$$\begin{aligned}\theta^{(t+1)} = \underset{\theta}{\text{argmin}}\ & \frac{1}{2}\sum_{k=1}^{K}\left(\left\langle\Lambda_k, S_{YY}^{k,(t)}\right\rangle + \left\langle 2\Theta_k, S_{YX}^{k,(t)}\right\rangle + \left\langle\Theta_k\Lambda_k^{-1}\Theta_k^T, S_{XX}^{k,(t)}\right\rangle\right)\\ & -\frac{1}{2}\sum_{k=1}^{K}\frac{n_k^{(t)}}{n} ln(|\Lambda_k|) + pen_\theta(\theta)\,,\end{aligned} \tag{6.14}$$

$$(M)$$

$$\pi^{(t+1)} = \underset{\pi}{\text{argmin}} -\sum_{k=1}^{K}\frac{n_k^{(t)}}{n}\, ln\,\pi_k + pen_\pi(\pi)\,.$$

The E step in Eq (6.13) is in closed form. With any reasonable penalty pen_π, the optimisation on the class weights π in Eq (6.14) will be trivial, and most likely in closed form as well. The update of θ in the M step (6.14) takes exactly the same form as the supervised penalised MLE of Eq (6.6), see [69] for the explicit supervised CGGM formulation. As a result, as long as the supervised case (6.6) is solved, then the M step is tractable as well. In their work on joint Hierarchical CGGM, [69] show that the supervised negative log-likelihood is a convex function of θ. As a consequence the problem (6.14) is solvable for a very wide array of penalties pen_θ, in particular the convex differentiable penalties.
In order to provide an algorithm with more specific and detailed steps, we consider in the rest of the section the special case of the GGL penalty. The GGL penalty was noticeably used in the supervised case by [69], who proposed a proximal gradient algorithm. Likewise, we can use a proximal gradient algorithm to compute the M step (6.14) of our EM algorithm.

Proximal gradient algorithm to solve the M step with the GGL penalty. The Group Graphical Lasso (GGL) penalty, introduced in [30] and adapted to the hierarchical CGGM by [69], can be written:

$$\begin{aligned} pen_\theta(\theta) := & \sum_{1\leq i\neq j\leq p} \left(\lambda_1^\Lambda \sum_{k=1}^K \left|\Lambda_k^{(ij)}\right| + \lambda_2^\Lambda \sqrt{\sum_{k=1}^K \left(\Lambda_k^{(ij)}\right)^2} \right) \\ & + \sum_{(i,j)\in[\![1,q]\!]\times[\![1,p]\!]} \left(\lambda_1^\Theta \sum_{k=1}^K \left|\Theta_k^{(ij)}\right| + \lambda_2^\Theta \sqrt{\sum_{k=1}^K \left(\Theta_k^{(ij)}\right)^2} \right) . \end{aligned} \tag{6.15}$$

Unlike in [69], where$\lambda_1^\Lambda = \lambda_1^\Theta$ and $\lambda_2^\Lambda = \lambda_2^\Theta$, we use different levels of penalisation for the parameters Λ and Θ, since both their scales and their desired sparsity level can be very different. This penalty borrows its design from the Group Lasso [178], where the l_1 norm induces individual sparsity of each coefficient, and the l_2 induces simultaneous sparsity of groups of coefficients. In Eq. (6.15), for each pair (i,j) belonging to the relevant space, $\left\{\Lambda_k^{(ij)}\right\}_{k=1}^K$ constitutes a group that can be entirely put to 0. This incites the algorithm to set a certain matrix coefficient to 0 over all K classes. These common zeros constitute the common structure sought after by the GGL approach. In our CGGM case, the same can be said for the group $\left\{\Theta_k^{(ij)}\right\}_{k=1}^K$. Regarding the theoretical analysis, we underline that the l_2 part of the penalty is not separable in a sum of K different penalties, which forces a joint optimisation problem to be solved, even in the supervised framework.
We detail here how to solve the M step (6.14) with $pen_\theta(\theta)$ defined as in Eq (6.15). We assume, as usual, that the optimisation in π is both independent from the optimisation in $\theta = \{\Lambda_k, \Theta_k\}_{k=1}^K$ and trivial. The function to minimise in θ at the M step is:

$$f(\theta) := \sum_{k=1}^K \left(-\frac{n_k^{(t)}}{n} ln(|\Lambda_k|) + \left\langle \Lambda_k, S_{YY}^{k,(t)} \right\rangle + \left\langle 2\Theta_k, S_{YX}^{k,(t)} \right\rangle + \left\langle \Theta_k \Lambda_k^{-1} \Theta_k^T, S_{XX}^{k,(t)} \right\rangle \right) + pen_\theta(\theta) .$$

As shown in [69], this function is convex and infinite on the border of its set of definition and as a unique global minimum. We note $f(\theta) =: g(\theta) + pen_\theta(\theta)$ for the sake of simplicity. The proximal gradient algorithm, see [28], is an iterative method based on a quadratic approximation on $g(\theta)$. If $\theta^{(s-1)}$ is the current state of the parameter within the proximal gradient iterations, then the next stage, $\theta^{(s)}$, is found by optimising the approximation:

$$\begin{aligned} f\left(\theta^{(s)}\right) &= f\left(\theta^{(s-1)} + \theta^{(s)} - \theta^{(s-1)}\right) \\ &\approx g\left(\theta^{(s-1)}\right) + \nabla g\left(\theta^{(s-1)}\right)^T . \left(\theta^{(s)} - \theta^{(s-1)}\right) + \frac{1}{2\alpha} \left\|\theta^{(s)} - \theta^{(s-1)}\right\|_2^2 + pen_\theta\left(\theta^{(s)}\right) \\ &\equiv \frac{1}{2\alpha} \left\|\theta^{(s)} - \left(\theta^{(s-1)} - \alpha \nabla g\left(\theta^{(s-1)}\right)\right)\right\|_2^2 + pen_\theta\left(\theta^{(s)}\right) . \end{aligned} \tag{6.16}$$

Where we removed in the last line the constants irrelevant to the optimisation in $\theta^{(s)}$ and α denotes the step size of the gradient descend. Note that we use the exponent (s) to indicate the current stage of the proximal gradient iteration, to avoid confusion with the exponent (t) used for the EM iterations (which are one level above). We underline that, in addition to $g(\theta)$ itself, the second order term in the Taylor development of $g(\theta)$ is also approximated. Using $\frac{1}{2\alpha}\left\|\theta^{(s)}-\theta^{(s-1)}\right\|_2^2$ instead of $\frac{1}{2}\left(\theta^{(s)}-\theta^{(s-1)}\right)^T.H_g\left(\theta^{(s-1)}\right).\left(\theta^{(s)}-\theta^{(s-1)}\right)$ spares us from computing the Hessian $H_g\left(\theta^{(s-1)}\right)$ and simplifies the calculations to come. The approximated formulation in Eq (6.16) leads to the definition of the proximal optimisation problem:

$$prox_\alpha(x) := \underset{\theta}{\operatorname{argmin}} \frac{1}{2\alpha}\|\theta - x\|_2^2 + pen_\theta(\theta)\,. \tag{6.17}$$

So that the proximal gradient step can be written:

$$\theta^{(s)} = prox_{\alpha_s}\left(\theta^{(s-1)} - \alpha_s \nabla g\left(\theta^{(s-1)}\right)\right)\,. \tag{6.18}$$

Where the step size α_s is determined by line search. The usual proximal gradient heuristic is to take a initial step size α^0, a coefficient $\beta \in]0,1[$, and to reduce the step size, $\alpha \longleftarrow \beta\alpha$, as long as:

$$g\left(\theta^{(s-1)} - \alpha G_\alpha\left(\theta^{(s-1)}\right)\right) > g\left(\theta^{(s-1)}\right) - \alpha\nabla g\left(\theta^{(s-1)}\right)^T.G_\alpha\left(\theta^{(s-1)}\right) + \frac{\alpha}{2}\left\|G_\alpha\left(\theta^{(s-1)}\right)\right\|_2^2\,,$$

with $G_\alpha\left(\theta^{(s-1)}\right) := \frac{\theta^{(s-1)} - prox_\alpha\left(\theta^{(s-1)} - \alpha\nabla g\left(\theta^{(s-1)}\right)\right)}{\alpha}$ the generalised gradient.
To apply the proximal gradient algorithm, we need to be able to solve the proximal (6.17) with the CGGM likelihood and the GGL penalty. Thankfully, [30] found an explicit solution to this problem in the GGM case, which [69] adapted to the CGGM. The proximal optimisation is separable in Λ and Θ, and the solutions $\Lambda^{(prox)}$ and $\Theta^{(prox)}$ share the same formula. As a result, we use D as a placeholder name for either Λ or Θ, i.e. depending on the context either $D_k^{ij} = \Lambda_k^{ij}$ or $D_k^{ij} = \Theta_k^{ij}$. Let S be the soft thresholding operator: $S(x,\lambda) := sign(x)\,max(|x|-\lambda, 0)$, and $\widetilde{D}_{k,\alpha}^{ij} := D_k^{ij,(s-1)} - \alpha\frac{\partial g}{\partial D_k^{ij}}\left(\theta^{(s-1)}\right)$. The solution of (6.17), with $x = \theta^{(s-1)} - \alpha\nabla g\left(\theta^{(s-1)}\right)$, is given coefficient-by-coefficient in Eq (6.19):

$$D_k^{ij,(prox)} = S\left(\widetilde{D}_{k,\alpha}^{ij}, \lambda_1^D\alpha\right) max\left(1 - \frac{\lambda_2^D\alpha}{\sqrt{\sum_k S(\widetilde{D}_{k,\alpha}^{ij}, \lambda_1^D\alpha)^2}}, 0\right)\,. \tag{6.19}$$

Note that the partial derivatives $\frac{\partial g}{\partial D_k^{ij}}\left(\theta^{(s-1)}\right)$, necessary to get $\widetilde{D}_{k,\alpha}^{ij}$, are easily calculated in closed form from the likelihood formula. With the proximal problem (6.17) and the line search easily solvable, the proximal gradient steps can be iterated until convergence to find the global minimum of $f(\theta)$. With $f(\theta)$ optimised, the M step (6.14) is solved.

6.4 Experiments

In this section, we demonstrate the performances of our EM with Mixture of CGGM. First on a visual toy example in 2 dimension, then on a higher dimensional synthetic example and finally on real Alzheimer's Disease data. We compare the Mixture of CGGM to the regular Mixture of GGM which ignores co-features and to a Mixture of GGM that assumes a uniform linear effect of the co-features on the features.

6.4.1 An illustration of co-features with class-dependent effect

In this section, we present a simple visual example to illustrate the importance of taking into account heterogeneous co-feature effects. We show that even with a single binary co-feature, and with low dimensional features, the state of the art unsupervised GGM techniques are greatly disrupted

by the co-features. Whereas our EM with Mixture of CGGM (which we call "Conditional EM" or "C-EM") achieves near perfect classification.
Under the Mixture of Gaussians (MoG) model, the observed data, $Y \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$, belongs to K classes which can directly be represented as K clusters in the feature space $\mathbb{R}^p$. Each cluster centred around a centroid at position μ_k and with an ellipsoid shape described by Σ_k. However, when there exists conditioning variables $X \in \mathbb{R}^q$ that have an effect on Y, this geometric description becomes more complex. Typically, the value of Y could depend linearly on the value of X, with $\mathbb{E}[Y|X, z = k] = \beta_k^T X$ for some $\beta_k \in \mathbb{R}^{p \times q}$. In this case, the average position in class k is not a fixed μ_k but a function of X. If X contains categorical variables, this creates as many different centroid positions as there are possible category combinations in X. The number of these *de facto* clusters geometrically increases with the dimension q, which deters from simply running a clustering method with an increased number of clusters K' to identify all of them. Moreover, if X contains continuous variables, there is a continuum of positions for the centroid, not a finite number of *de facto* clusters. If X mixes the two types of variables, the two effects coexist. This shatters any hope to run a traditional MoG-based EM clustering algorithm, since its success is heavily dependent on its ability to identify correctly the K distinct cluster centroids μ_k.
Since the X are observed, a possible solution is to run the linear regression $\hat{Y} = \hat{\beta} X$ beforehand, and run the EM algorithm on the residual $Y - \hat{Y}$ to remove the effect of X. This is what we call the "residual EM" or "residual Mixture of GGM". However this does not take into account the fact that this effect can be different for each class k, $\beta_1 \neq \beta_2 \neq ... \neq \beta_K$. Since the label is not known beforehand in the unsupervised context, the linear regression $\hat{Y} = \hat{\beta} X$ can only be run on all the data indiscriminately, hence is insufficient in general. On the other hand, the hierarchical CGMM (6.5), which verifies: $\mathbb{E}[Y|X, z = k] = -\Lambda_k^{-1} \Theta_k^T X$, is designed to capture heterogeneous co-feature effects. We design a simple experiment to substantiate this intuition.

In this example, $Y \in \mathbb{R}^2, X \in \{-1, 1\}$ and $z \in \{1, 2\}$. $Y|X, z$ follows the hierarchical conditional model of (6.5). In this simple case, this can be written as $Y = (\beta_1 X + \epsilon_1) \mathbb{1}_{z=1} + (\beta_2 X + \epsilon_2) \mathbb{1}_{z=2}$. With $\epsilon_1 \sim \mathcal{N}(0, \Lambda_1^{-1})$ and $\epsilon_2 \sim \mathcal{N}(0, \Lambda_2^{-1})$. A typical iid data sample $(Y_{i=1}^{(i)})^n$ is represented on the left sub-figure of Figure 6.1. The hidden variable z is represented by the colour (blue or orange). The observable co-feature X is represented by the shape of the data point (dot or cross). It is clear from the figure that a Mixture of Gaussians model with $K = 2$ cannot properly separate the blue and orange points in two clusters. Indeed, on the right sub-figure of Figure 6.1, we observe the final state of an EM that fits a Mixture of Gaussians on Y. The two recovered clusters are more correlated with the co-feature X than the hidden variable z. However, this method did not take advantage of the knowledge of the co-feature X. As previously mentioned, one could first subtract the effect of X from Y before running the EM. On the left sub-figure of Figure 6.2, we represent the residual data $\tilde{Y} := Y - \hat{\beta} X$. Where $\hat{\beta}$ is the Ordinary Least Square estimator of the linear regression between X and Y over all the dataset ($\hat{\beta} \approx \frac{\beta_1 + \beta_2}{2}$ if n is large enough). Since the linear effect between X and Y is not uniform over the dataset, but class dependent, the correction is imperfect, and the two class clusters remain hardly separable. This is why the residual EM, that fits a Mixture of GGM on $\tilde{Y}$ is also expected to fail to identify clusters related to the hidden variable. Which is shown by the right sub-figure of Figure 6.2, where we see a typical final state of the residual EM.
On the leftmost sub-figure of Figure 6.3, we display the proper correction for the co-features' effect $\tilde{Y}' = Y - \beta_1 X \mathbb{1}_{z=1} - \beta_2 X \mathbb{1}_{z=2} = \epsilon_1 \mathbb{1}_{z=1} + \epsilon_2 \mathbb{1}_{z=2}$. Under this form, a Mixture of Gaussian can separate the data by colour. This is precisely the kind of translation that each data point undergoes within a Hierarchical CGGM. Hence a Mixture of CGGM can succeed in identifying the hidden variable z, provided that it estimates correctly the model parameters. To illustrate this point, the two next sub-figures in Figure 6.3 represent the same final state of the EM fitting a Mixture of CGGM on Y. The middle sub-figure represents $\tilde{Y}'$ as well as the two estimated centered distributions $\mathcal{N}(0, \widehat{\Lambda}_k^{-1})$ for $k = 1, 2$. We can see the two formally identified clusters after removing the effect of X. The rightmost sub-figure represents the original data Y as well as the four estimated distributions $\mathcal{N}(\pm \widehat{\Sigma}_k \widehat{\Theta}_k^T, \widehat{\Lambda}_k^{-1})$ for $k = 1, 2$. The four *de facto* clusters present in the data Y before removing the effect of X are well estimated by the method.
We confirm these illustrative results by running several simulations. We generate 50 datasets with

$n = 500$ data points. For each simulation, we make 10 random initialisations from which we run the three EMs: with GGM, residualised GGM or CGGM. Table 6.4.1 summarises the results. We follow the errors made by the estimated class probabilities or "soft labels", $\widehat{\mathbb{P}}(z_i = k)$, which we call the "soft misclassification error", as well as the error made by the "hard labels", $\mathbb{1}_{\widehat{z}_i=k}$, which we call the "hard misclassification error". They can be expressed as $\frac{1}{2n}\sum_{i,k}\left|\mathbb{1}_{z_i=k} - \widehat{\mathbb{P}}(z_i = k)\right|$ and $\frac{1}{2n}\sum_{i,k}|\mathbb{1}_{z_i=k} - \mathbb{1}_{\widehat{z}_i=k}|$ respectively. We see that the Mixture of CGGM performs much better, with less than 10% of misclassification in average, while the two GGM methods are both above 40% of error, fairly close to the level of a random uniform classifier, 50%.

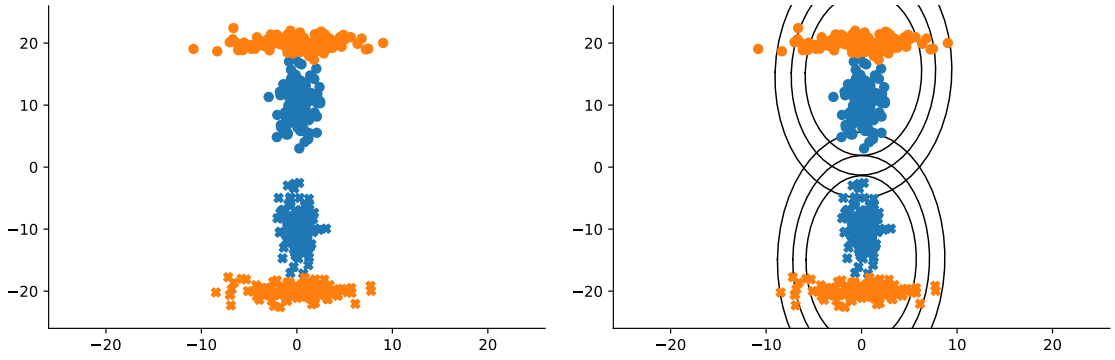


Figure 6.1: (Left) Observed data Y in the 2D space. The observed conditioning variable X is binary. Data points with $X = -1$ are represented as crosses, and the ones with $X = 1$ are represented as dots. In addition, there is an unknown "class" variable z. Class 1 is in blue, class 2 in orange. $Y|X, z$ follows the hierarchical conditional model. As a result, the two classes (orange and blue) are hard to separate in two clusters. (Right) Typical clusters estimated by an EM that fits a GGM mixture on Y

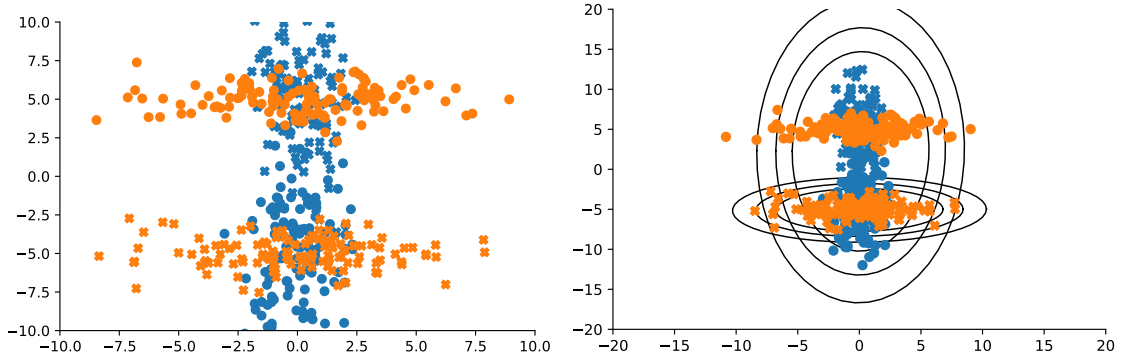


Figure 6.2: (Left) Residual $\tilde{Y} = Y - \hat{\beta}Y$ data after taking into account the estimated effect of X. Since the effect had different intensities on class 1 and 2, only the average effect was subtracted, and two classes are still not well separated. (Right) Typical clusters estimated by the "residual EM", that fits a GGM mixture on $\tilde{Y}$

6.4.2 Experiments in high dimension

In this section, we perform a quantitative analysis of the algorithms in a higher dimension framework, where the matrix parameters Λ and Θ are more naturally interpreted as sparse networks. We confirm

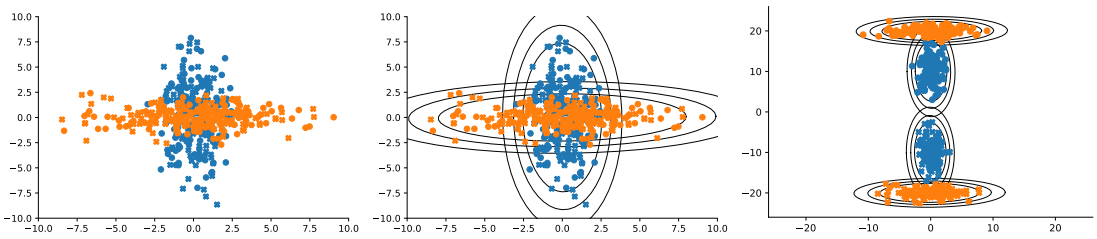


Figure 6.3: (Left) Observations $\tilde{Y}' = Y - \beta_1 X \mathbb{1}_{z=1} - \beta_2 X \mathbb{1}_{z=2}$ exactly corrected for the class-dependent effect of X. In this state the two classes appear as two distinct clusters. The Conditional-EM is designed to transform the data in this manner. (Middle) One possible representation of the CEM results. The corrected observations $\tilde{Y}'$ are displayed alongside centered normal distributions with the two estimated covariance matrices: $\mathcal{N}(0, \widehat{\Lambda}_k^{-1})$. (Right) Another possible representation of the same CEM results. The original observations Y are displayed, alongside with the four *de facto* estimated distributions $\mathcal{N}(\pm\widehat{\Sigma}_k\widehat{\Theta}_k^T, \widehat{\Lambda}_k^{-1})$.

Table 6.1: Average and standard deviation of the misclassification error achieved on the 2-dimensional example with the EMs on the Mixture of GGM, the Mixture of GGM with residualised data, and the Mixture of CGG. The two GGM methods are close to the threshold of random classification (0.50), while the Mixture of CGGM is in average below 10% of error.

	EM GGM	EM resid. GGM	EM CGGM
$\frac{1}{2n}\sum_{i,k}\left\lvert\mathbb{1}_{z_i=k} - \widehat{\mathbb{P}}(z_i=k)\right\rvert$	0.41 (0.11)	0.47 **(0.05)**	**0.08** (0.17)
$\frac{1}{2n}\sum_{i,k}\lvert\mathbb{1}_{z_i=k} - \mathbb{1}_{\widehat{z}_i=k}\rvert$	0.41 (0.12)	0.46 **(0.06)**	**0.07** (0.17)

that the Mixture of Conditional Gaussian Graphical Models is better suited to take into account the heterogeneous effects of co-features on the graph.
For this experiment, the observed data follows a mixture model with $K = 3$ classes. Each class k has the probability weight $\pi_k = \frac{1}{3}$. An observation $(Y, X) \in \mathbb{R}^p \times \mathbb{R}^q$ belonging to the class k is described by the distribution: $Y|X \sim \mathcal{N}\left(-\Lambda_k^{-1}\Theta_k^T X, \Lambda_k^{-1}\right)$. No model assumption are made on X. In this example, X contains two binary variables, two continuous variables, and a constant variable always equal to 1. The inverse-covariance matrix $\Lambda_k \in \mathbb{R}^{p\times p}$ and the transition matrix $\Theta \in \mathbb{R}^{q\times p}$ are both sparse, with $p = 10$ and $q = 5$. We run 20 simulations. A simulation consists of $n = 300$ generated data points. On these data points, we run the compared methods, all initialised with the same random parameters. For all simulations, we make 10 of these runs, each with a different random initialisation. We compared the same three algorithms as in section 6.4.1: the EM for the Mixture of GGM, the EM for the Mixture of GGM with average effect of X subtracted, and the EM applied to the Mixture of CGGM. Additionally, we also run the tempered version of these three EM algorithms.
We follow four metrics to assess the method's success in terms of cluster recovery and fit with the data. The classification error (both soft and hard labels versions), the recovery of the network matrix Λ and an "ABC-like" metric. The "ABC-like" metric is meant to assess how well each of the estimated solutions is able to replicate the observed data. Since each solution is the parameter of a probability distribution, at the end of each EM, we generate new data following this proposed distribution. Then, for each synthetic data point, we compute the distance to the closest neighbour among the real data points. These minimal distances constitute our "ABC-like" metric. Finally, we also compute the execution time of each EM, knowing that they all have the same stopping criteria. We represent on Figure 6.4 the empirical distribution of these four metrics and we quantify with Table 6.2 the key statistics (mean, standard deviation, median) that characterise them. With $K = 3$ and balanced classes, a uniform random classifier would guess the wrong label 66.7% of the time. We observe that the two Mixture of GGM method are dangerously close to this threshold, with more than 50% hard misclassification. The EM on the Mixture of CGGM (C-EM) on the other hand, achieves a much better classification with less than 15% hard misclassifcation. This demonstrate that, even when faced with a more complex situation, in higher dimension, the Mixture of CGGM is better suited to correct for the effect of the co-features and discover the right clusters of data points. This also underlines once more the importance of allowing different values of the effect of X on Y for each class. Indeed, the residual Mixture of GGM - which took into account the average effect of X on Y over the entire population - was unable to achieve better performances than the EM that did not even use the co-features X. In terms of reconstruction of the observed data by the estimated model (ABC-like metric), we see that the synthetic data points generated from the estimated Mixture of CGGM model have closer nearest neighbours than the data points generated by the other estimated models. In addition to all these observations, the C-EM is also faster than the other two methods, reaching the convergence threshold faster.
In addition to the cluster recovery, we can also assess the parameter reconstruction of each method.

Table 6.2: Average, standard deviation and median (below) of the four followed performance metrics over the 30×5 simulations. The best values are in **bold**. We can see that the classification performances with the Mixture of CGGM are much better than the two methods with Mixtures of GGM, and with faster computation times.

	soft misclassif.	hard misclassif.	ABC-like metric	runtimes
GGM	0.56 **(0.03)** 0.57	0.55 (0.04) 0.56	5.57 **(0.09)** 5.58	115 **(61)** 93
GGM resid.	0.51 **(0.03)** 0.51	0.50 **(0.03)** 0.49	4.64 (0.22). 4.64	253 (137) 256
CGGM	**0.17** (0.05) **0.16**	**0.14** (0.06) **0.13**	**4.13** (0.14) **4.14**	**58** (91) **16**

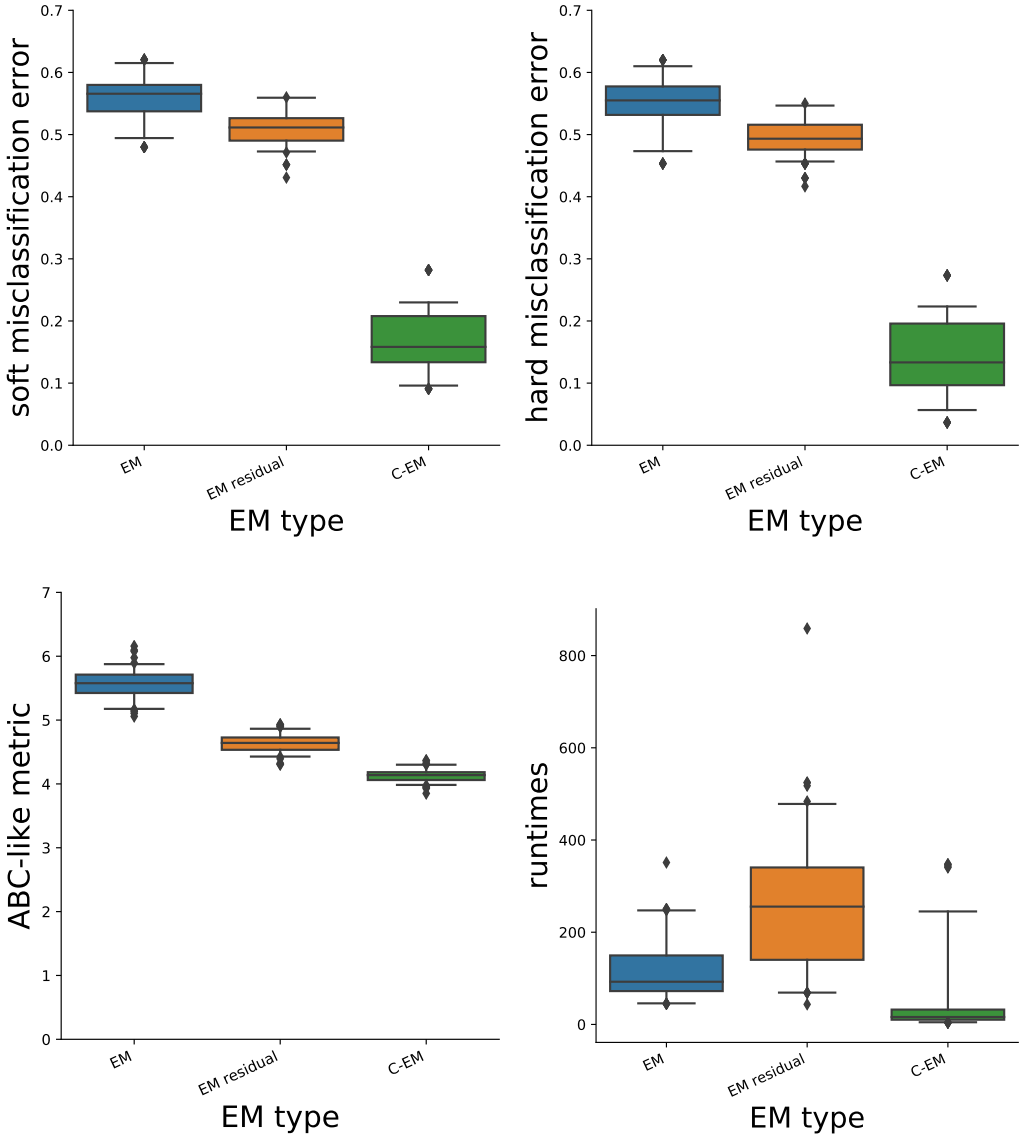


Figure 6.4: Empirical distribution of several performance metrics measured over many simulations. The sample is made of 30 simulations with 5 different initialisations each. Three methods are compared. The EM and EM residual algorithms estimate a Mixture of GGM. The C-EM algorithm estimates a Mixture of CGGM. The C-EM is much better performing and faster. (Upper left) Soft mis-classification error $\left|\mathbb{1}_{z_i=k} - \widehat{\mathbb{P}}(z_i = k)\right|$. (Upper right) Hard mis-classification error $|\mathbb{1}_{z_i=k} - \mathbb{1}_{\widehat{z}_i=k}|$. (Bottom left) ABC-like metric. (Bottom right) Run time.

Since the three clustering methods estimate different parametric models over the data, they do not actually try to estimate the same parameters. Regardless, all the methods still estimate a certain inverse covariance matrix Λ_k (conditional or not on the X depending on the model) of each sub-population that they identify. In Table 6.3, we can check that the $\widehat{\Lambda}_k$ estimated by with the Mixture of CGGM are indeed a much better fit for the real Λ_k then the estimated matrices from the other models. This is expected, since the real Λ_k actually correspond to the CGGM model. The two metrics followed are the Kullback–Leibler (KL) divergence between the Gaussian distribution $f_{\Lambda_k} \sim \mathcal{N}(0, \Lambda_k^{-1}))$ and $f_{\widehat{\Lambda}_k} \sim \mathcal{N}(0, \widehat{\Lambda}_k^{-1}))$, and the l_2 difference given by the Froebenius norm: $\left\|\Lambda_k - \widehat{\Lambda}_k\right\|_F^2$.
To illustrate the different level of success concerning the conditional correlation graph recovery, we display on Figure 6.5 the conditional correlation matrix (i.e. the conditional correlation graph with weighted edges) estimated by each method. The three columns of figures correspond to the three sub-populations. The first two rows of figures are the matrices estimated by the two Mixture of GGM methods, with and without residualisation with the co-features. The third row of figures correspond to the matrices estimated by the Mixture of CGGM. The final row displays the real conditional correlation matrices. We observe that the two Mixtures of GGM recover way too many edges, with no particular fit with the real matrix. By contrast, the matrices from the CGGM Mixture exhibit the proper edge patterns. Not all the true edges are recovered, and some of the recovered ones have a lower intensity than the real ones, but there are very few False Positives. This level of fidelity is very impressive since the method was run from a random initialisation on a totally unsupervised dataset, with heavily translated data points all over the 10 dimensional space. Moreover, the matrices in Figure 6.5 all result from the inversion of the empirical covariance matrix, which is neither a very geometrical nor a very stable operation. The figures in this example were obtained without fine tuning of the penalty intensity. A lower penalty intensity should allow the CGGM Mixture graph to resemble the true graph even more. Note that the matrices estimated by the Mixtures of GGM EMs are less sparse than the CGGM ones, despite being estimated with the same penalisation intensity. This is because the estimated clusters by the GGM Mixtures contain data points so far apart that the corresponding empirical covariance coefficients are very large. Hence, a penalty with the same intensity is not enough to put as many coefficients to 0 as in the case of the Mixture of CGGM.
In Figure 6.6, we represent the regression parameter $\widehat{\Theta}_k$ estimated by with the Mixture of CGGM alongside the real Θ_k. Once again, we see that the sparsity pattern is very well identified, with no False Positive. Moreover, in this case, there are also almost no False Negative, and all the edge intensities are correct. This is not a surprise. Indeed, the parameter Θ plays a huge role in the correct classification of the data, since it serves to define the expected position of each data point in the feature space (playing the part of the "average" parameter in Mixtures of GGM). Hence, a good estimation of Θ is mandatory to reach a good classification. Since the EM with Mixture of CGGM achieved very good classification results, it was expected that Θ would be well estimated.

Table 6.3: Average and standard deviation of the metrics describing the reconstruction of each inverse-covariance matrix Λ_k. The matrices are consistently better reconstructed with the mixture of CGGM.

metric	class	EM GGM	EM res. GGM	EM CGGM
$KL(f_\Lambda, f_{\hat{\Lambda}})$	1	11.0 (3.0)	7.5 (6.8)	**0.8 (0.2)**
	2	10.3 (2.2)	8.5 (5.0)	**1.9 (0.3)**
	3	13.6 (2.5)	5.2 (2.3)	**3.4 (1.1)**
$\left\|\Lambda - \hat{\Lambda}\right\|_F^2$	1	39.2 (48.4)	44.2 (114)	**2.2 (0.8)**
	2	15.1 (12.2)	102 (73.9)	**6.6 (0.9)**
	3	14.2 (13.8)	15.1 (25.7)	**5.8 (4.0)**

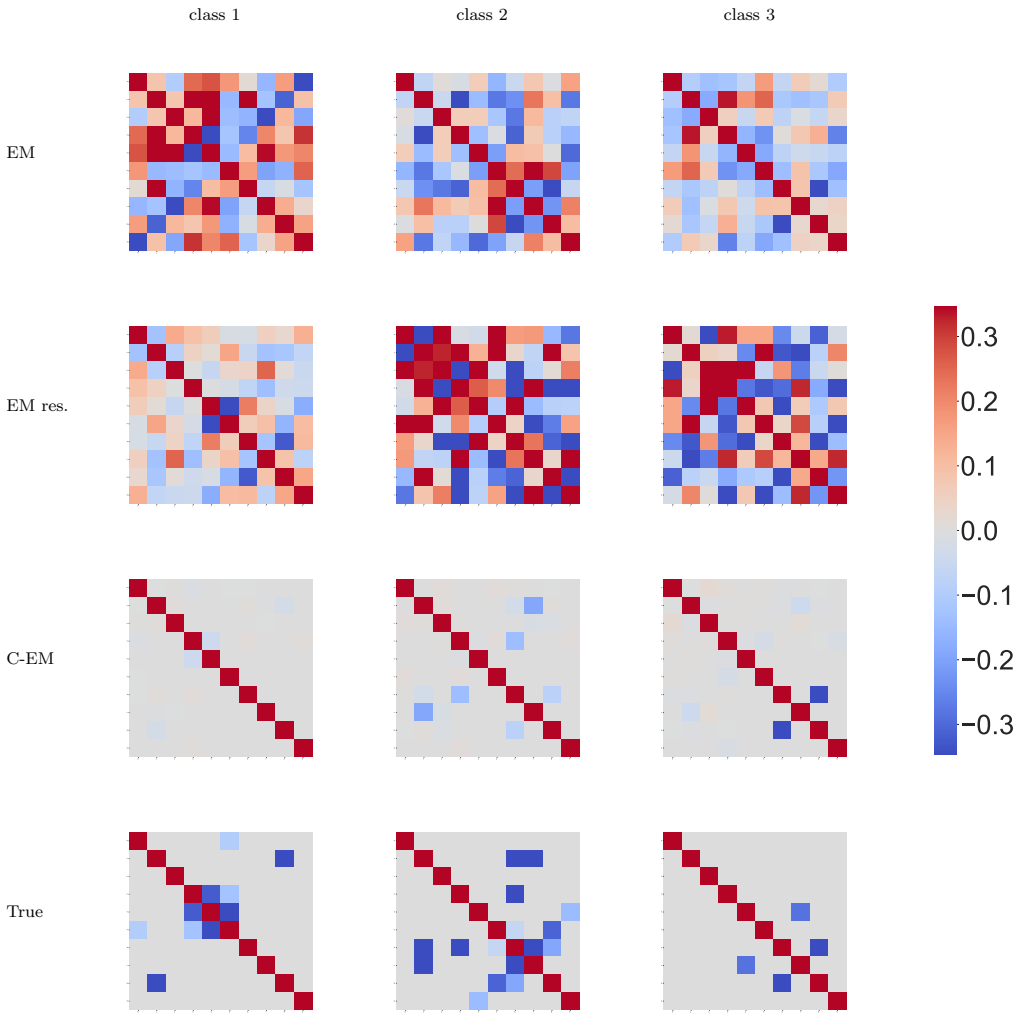


Figure 6.5: Comparison between the several estimated and the true conditional correlation matrices for each sub-population. The three columns of figures correspond to the three sub-populations. The first two rows of figures are the matrices estimated by the two Mixture of GGM methods, with and without residualisation with the co-features. The third row of figures correspond to the matrices estimated by the Mixture of CGGM. The final row displays the real conditional correlation matrices. Unlike the two GGM-based methods, the Mixture of CGGM recovers correct edges with very few False Positives.

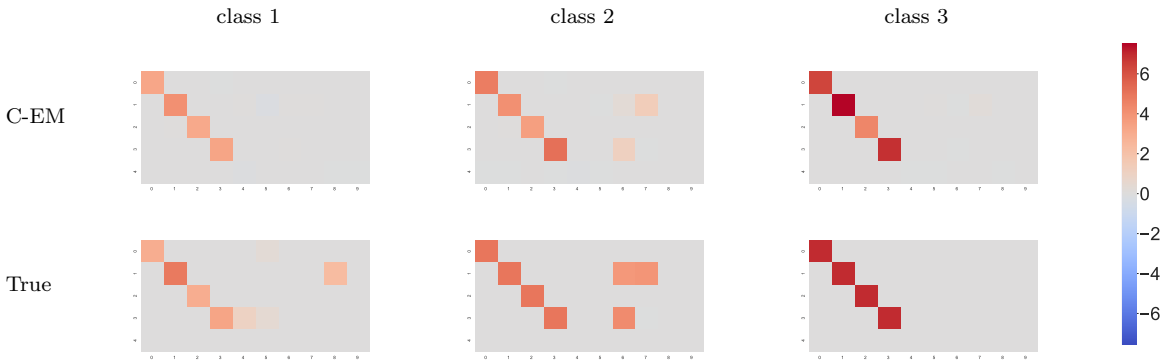


Figure 6.6: Reconstruction of the Θ_k by the EM on the Mixture of CGGM. The three columns of figures correspond to the three sub-populations. Almost all the edges are right, with no False Positive and almost no False Negative. Moreover, the intensities are also mostly correct.

6.4.3 Experiments on real data

In this section, we confirm our experimental observations with a real, high dimensional, Alzheimer's Disease dataset. We illustrate that the EM with Mixture of CGGM is better suited to identify clusters correlated with the diagnostic than the Mixture of GGM methods. We bring to light the effect of co-features such as the gender and age on the medical features.

Our dataset is composed of the parameters $\xi, \tau, (w_i)_{i=1,...,30}$ of longitudinal models estimated on real Alzheimer's Disease patients, see [141]. In summary, the evolution of several features are followed over time for each patients. The features $i = 1, ..., 10$ correspond to MRI measures of atrophy in different region of the brain. The features $i = 11, ..., 30$ correspond to cognitive scores obtained through tests. A longitudinal model estimates a geodesic trajectory within a Riemannian manifold of the parameter space that fits with the patient's own evolution. More specifically, the longitudinal model parameters describe how each patient's trajectory deviates from a specific reference geodesic trajectory. The parameter ξ is the time acceleration of the patient with regards to the reference. The parameter τ is the time shift, so that a smaller τ corresponds a disease which starts early. Each w_i describes the space shift of the trajectory with regards to its corresponding feature. With $Y := \xi, \tau, (w_i)_{i=1,...,30}$ the vector of features, we have $p = 32$. We add three co-features to describe each patient: the gender, the age baseline, and the number of years of education. With the addition of the constant co-feature = 1, the vector of co-features is 4-dimensional, $X \in \mathbb{R}^4$. The dataset contains 1400 patients, with half being healthy ("Control" patients), and the other half being diagnosed with the Alzheimer's Disease, either from the start or after a few visits ("AD" patients). The data is centered and normalised over the entire population.

We run the three algorithms: EM, EM residual and C-EM on this dataset. In order to check the stability of the results over several different runs, we implement a bootstrap procedure that only uses 70% of the data each time. We generate 10 such bootstrapped dataset. We initialise the algorithms with a KMeans on the $Y^{(i)}$ data points. Since KMeans is not deterministic, we make 5 different runs for each bootstrapped dataset, starting from 5 different possible KMeans initialisation each. Like previously, for the sake of fairness, the EM and C-EM are always provided with the same initialisation, and the residual EM is initialised with a KMeans on the residual of Y after subtracting the prediction by the X, a more relevant initialisation for this method. We make all these runs with four different feature sets. First with no space shift variable $Y = \{\xi, \tau\}, p = 2$, then we add only the MRI space shifts $Y = \{\xi, \tau, (w_i)_{i=1,...,10}\}, p = 12$, then only the Cognitive Scores space shifts $Y = \{\xi, \tau, (w_i)_{i=1,...,10}\}, p = 22$, and finally, with all the features $Y = \{\xi, \tau, (w_i)_{i=1,...,30}\}, p = 22$. The classification results are summarised in Table 6.4. With two balanced classes, the classification error of a uniform random classifier is 50%. On the smallest dataset, $p = 2$, we can see that the discovered cluster are somewhat correlated with the diagnostic, with classification errors below 30%. The Mixture of GGM on the uniformly residualised data and the Mixture of CGGM achieve similar levels of error, they are both better than the regular Mixture of GGM. When the MRI features are

added, all the discovered cluster become more correlated with the diagnostic. The regular Mixture of GGM achieves in average 16% of hard classification error, the residualised Mixture of GGM is at 11% of error, and the Mixture of CGGM even below, at 7%. The results with only the Cognitive Scores are very similar, simply a bit worse for every method. However, when both the MRI and Cognitive Scores feature are included, the performance of both GGM mixtures decrease, with both higher average error and higher variance. On the other hand, the Mixture of CGGM achieves here its best level of performance. This stability of the Mixture of CGGM's performance as the size of the feature set increases indicates that our model is the best suited to properly identify clusters correlated with the diagnostic in high dimension.

We analyse the estimated Mixture of CGGM parameters on the full feature set $p = 32$. First, since $\mathbb{E}[Y|X, z = k] = -\Lambda_k \Theta_k^T X$ in the CGGM, we display on Figure 6.7 the two estimated $\hat{\beta}_k := -\widehat{\Lambda}_k \widehat{\Theta}_k$ (averaged over the bootstrap). They play the role of linear regression coefficients in the model. The last column is the constant coefficient, while the first three are the gender, age baseline and years of education coefficients respectively. Since the data is centered, negative and positive values correspond to below average and above average values respectively. The cluster $k = 1$ is the one very correlated with the Control patients sub-population. Similarly, the cluster $k = 2$ is the one very correlated with the AD patients.

The most noticeable difference between the two $\hat{\beta}_k$ are the constant vectors, who have opposite effects on all features. In particular, the "AD cluster" is very correlated with high ξ and low τ, as well as high LL Delay and LM IMM. The opposite being true for the "Control cluster". These are the expected effects: a high ξ corresponds to a quickly progressing disease, and a low τ to an early starting disease. For some reason, the "AD cluster" is also negatively correlated with the atrophy of the ventricles, which may not be an actual trait of AD patients, but an accidental characteristic of this cluster. However, this behaviour is consistent over all the bootstraped datasets, which makes us question whether it might actually be a relevant description of the disease. Further testing with new datasets is required.

The non-constant linear regression coefficients are also different between the clusters, although these differences are often in intensity and not in sign. In order to visualise more clearly the differences in intensity, we represent on Figure 6.8, with the same conventions, the difference $\hat{\beta}_2 - \hat{\beta}_1$. In particular, within the AD cluster, we observe stronger positive effect of the Age baseline on the MRI Amygdala, Enthorinal, Hyppocampus and Parahip atrophies. On the contrary, there is a stronger positive effect of the education level on all the MRI attrophies for the Control patients. The age bl has a stronger negative impact on the scores ECOG SELF mem, lang and dispat for the AD patients, and a stronger negative impact on the scores LL Delay and LM IMM for the control patients.

Finally, we display on Figure 6.9 and 6.10 the average conditional correlation graphs estimated for these two clusters by the Mixture of CGGM. Their only noticeable difference is the negative conditional correlation between ξ and τ in the "Control cluster", which is reversed in the "AD cluster". For the AD patients, this means that a disease that appears later tends to also progress faster, which is in line with medical observations. Apart from this edge, the rest of the connections are almost identical in-between clusters. This suggests that the, cluster dependent, prediction $\mathbb{E}_{\hat{\theta}_k}[Y^{(i)}|X^{(i)}, z = k] = -\widehat{\Sigma}_k \widehat{\Theta}_k^T X^{(i)}$ takes into account enough of the cluster-specific effects so that the remaining unexplained variance has almost the same form in both clusters. Hence, the conditional correlations pictured in these graphs correspond to very general effects, such as the positive correlations between related cognitive tests or areas of the cortex.

6.5 Conclusion

We introduced the Mixture of Conditional Gaussian Graphical Models in order to guide the cluster discovery when estimating different Gaussian Graphical Models for an unlabelled heterogeneous population in the presence of co-features. We motivated its usage to deal with the potential inhomogeneous and class-dependent effect of the co-features on the observed data that would otherwise disrupt the clustering effort. To estimate our Mixture model, we proposed a penalised EM algorithm ("Conditional EM" or "C-EM") compatible with a wide array of penalties. Moreover, we provided detailed algorithmic steps in the specific case of the popular Group Graphical LASSO penalty.

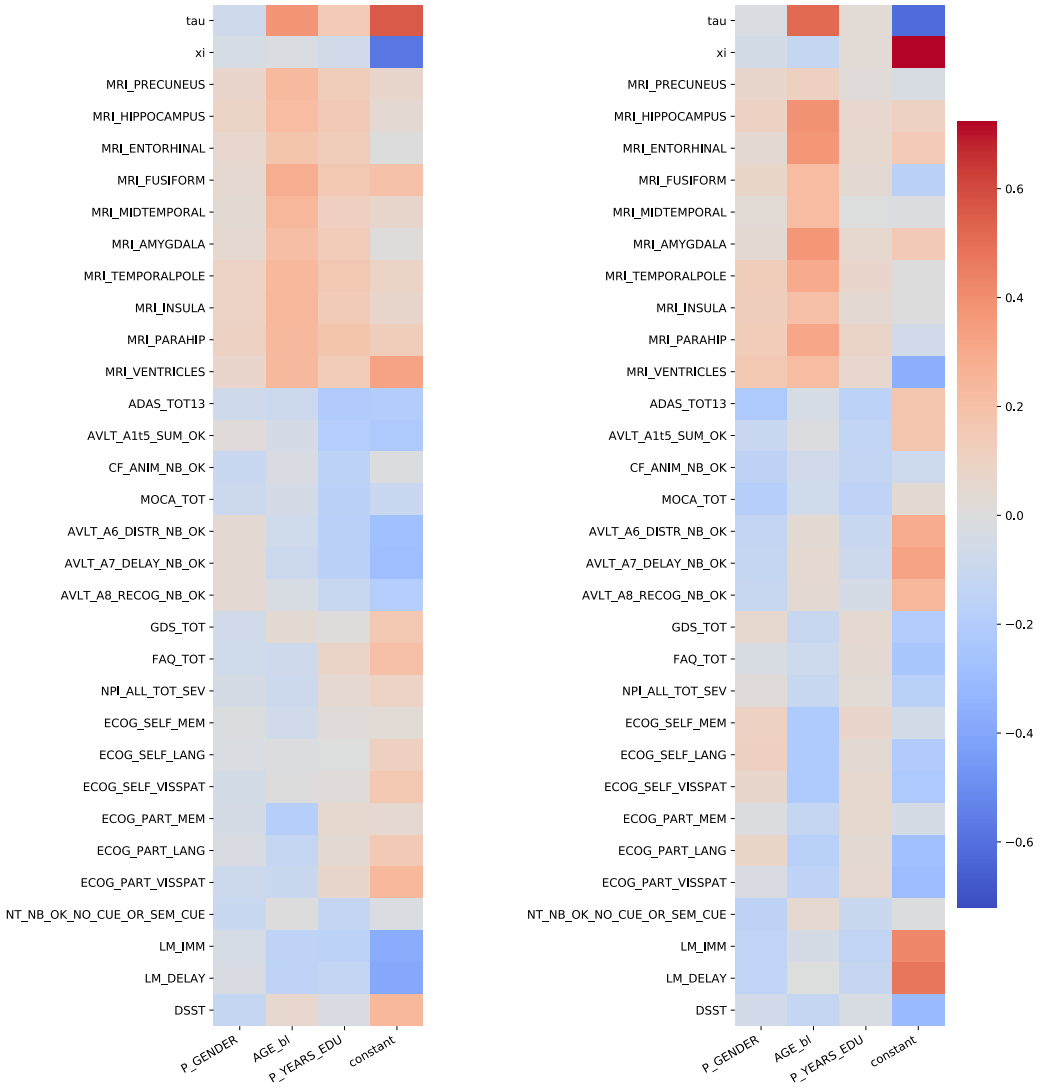


Figure 6.7: Average $\widehat{\beta}_k := -\widehat{\Sigma}_k \widehat{\Theta}_k^T$ over 10 bootstrap sampling of the data. For each bootstrapped dataset, 3 different runs of the C-EM are made, each with a different KMmeans initialisation of the labels. (Left) $\widehat{\beta}_1$, the cluster $k = 1$ is always very correlated with the **Control patients** subpopulation (less than 10% deviation). (Right) $\widehat{\beta}_2$, the cluster $k = 2$ is likewise very correlated with the **AD patients**. In each figure, the last column is the constant coefficient. The largest inter-cluster differences are between the two constant terms. However there are some noticeable difference on the other regression coefficients as well. Figure 6.8 makes these difference more explicit.

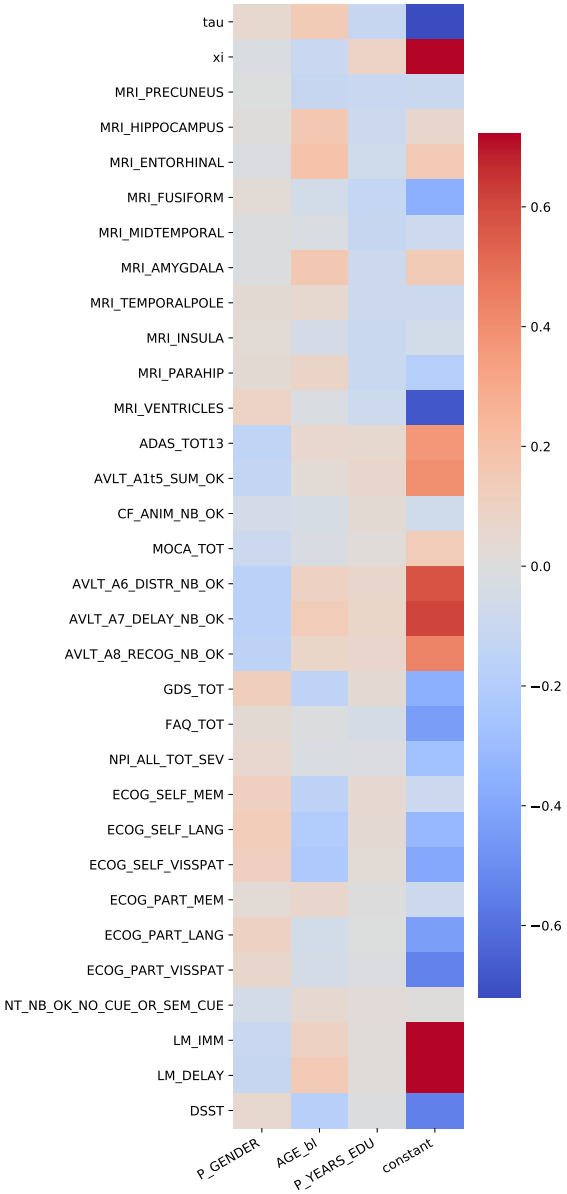


Figure 6.8: Average $\widehat{\beta}_2 - \widehat{\beta}_1$ over the 30 bootstrap runs of the C-EM. Here, the differences in intensity between AD ($k = 2$) and Control ($k = 1$) patients are more explicit.

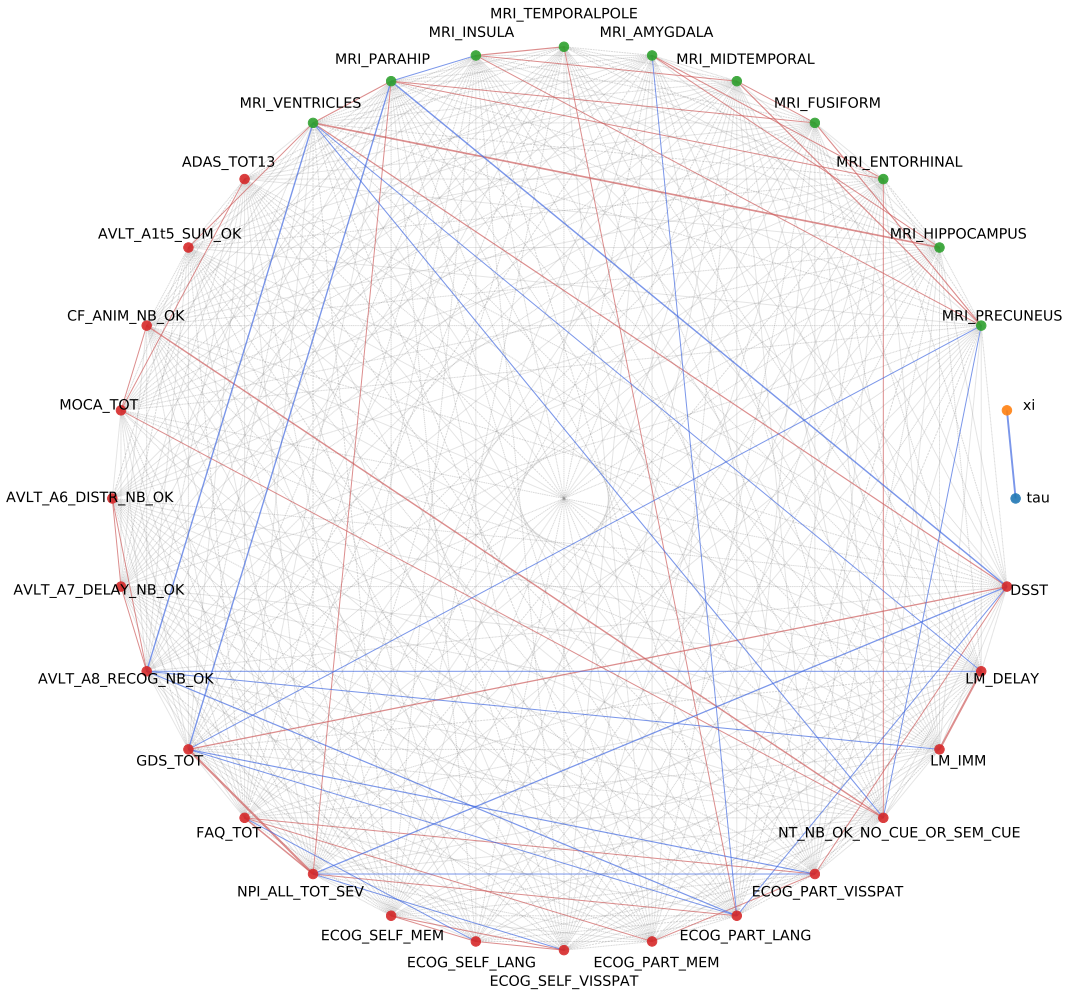


Figure 6.9: Graph Control.

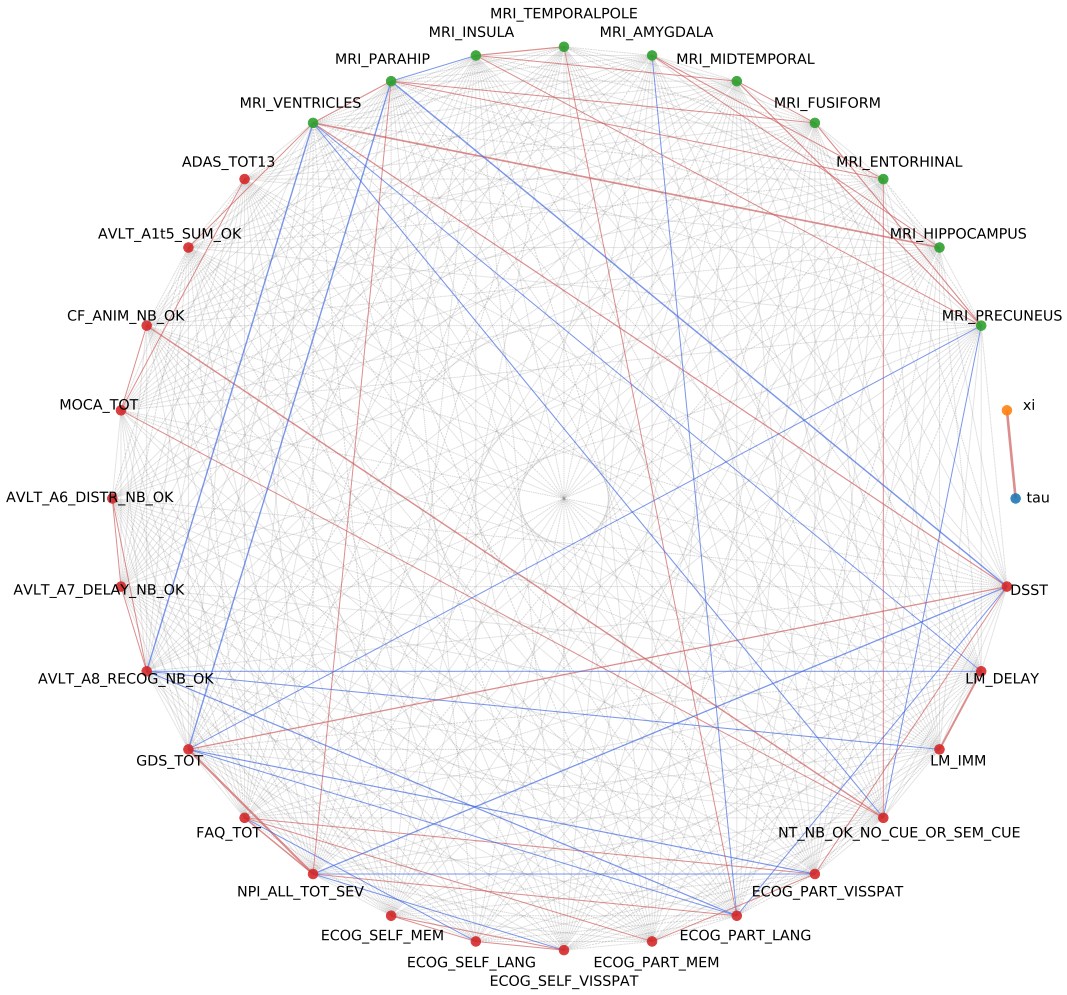


Figure 6.10: Graph AD.

Table 6.4: Recovery of the diagnostic labels (AD or control) with unsupervised methods on real longitudinal data. The three compared methods are the EM, EM residual (both GGM) and the C-EM (CGGM). Four different feature sets are tried: only $\{\tau, \xi\}$, adding the MRI space shift coefficients w_i, adding the Cognitive Score (CS) space shift coefficients w_i, and adding both the MRI and CS space shift coefficients. The table presents the average and standard deviation of the misclassification error over 10 bootstrap iteration, with 5 different KMeans initialisation each. The best results are in **bold**.

	metric	EM	EM resid.	C-EM
no CS, no MRI $p = 2$	soft misclassif.	0.31 (0.02)	0.22 (0.03)	**0.21 (0.01)**
	hard misclassif.	0.31 (0.03)	**0.18** (0.05)	0.19 **(0.01)**
only MRI $p = 12$	soft misclassif.	0.15 (0.01)	0.13 (0.01)	**0.08 (0.01)**
	hard misclassif.	0.12 (0.01)	0.10 (0.01)	**0.07 (0.01)**
only CS $p = 22$	soft misclassif.	0.17 (0.02)	0.15 (0.02)	**0.09 (0.01)**
	hard misclassif.	0.14 (0.03)	0.13 (0.03)	**0.08 (0.01)**
CS and MRI $p = 32$	soft misclassif.	0.24 (0.09)	0.17 (0.04)	**0.08 (0.01)**
	hard misclassif.	0.21 (0.10)	0.15 (0.05)	**0.07 (0.01)**

Additionally, we proved a theorem providing conditions on the penalty under which the C-EM benefits from state of the art convergence guarantees. Then, we demonstrated the interest of the method with experiments on synthetic and real data. First, we showed on a toy example - with a 2-dimensional feature space and a 1-dimensional co-feature - that the regular Mixture of GGM methods were inadequate to deal with even the most simple in-homogeneous co-feature. We confirmed on a more complex simulation, in higher dimension, that Mixtures of CGGM could identify much better the clusters in the feature space, and recover the actual GGM structure of the data. Finally, we tested all the methods on a real data set, with longitudinal model parameters describing the evolution of several Alzheimer's Disease patients. We demonstrated that our method was the best at identifying the diagnostic with an unlabelled dataset. We unearthed some in-homogeneous effect of co-features on the longitudinal parameter and recovered the conditional correlation graphs by cluster.

6.6 Appendix

In these appendices, we explore other aspects of our EM for Mixtures of CGGM. First, we prove a convergence result on our algorithm which provides condition on the regularisation in order to benefit from the convergence guarantees. Then, we propose a tempered version of our algorithm that fits within the framework of Chapter 5. We prove that the theorem of convergence for the tempered EM 5.4.1 of section 5 applies to the tempered EM for the Mixture of CGGM with conditions on the regularisation. Afterwards, we include the tempered version of each EM in the high dimensional experiments with synthetic data of section 6.4.2 to demonstrate how the tempering improves even further the Mixture of CGGM EM. Finally, we discuss how the method can be extended beyond the CGGM and to any model member of the curved exponential family.

6.6.1 Convergence of the EM for Mixtures of CGGM

In section 6.3, we proposed a tractable EM algorithm to estimate a Mixture of CGGM with joint structure. In this section, we prove that our algorithm verifies the state of the art convergence guarantees for EM algorithms on the exponential family [33]: convergence towards a critical point of the likelihood function.

Convergence Theorem

The authors of [33] propose in their Theorem 1 a convergence result for the EM algorithm applied to a distribution in the exponential family. In the formalism of [33], the different realisations of the

observed variable Y are considered fixed, hence Y is ignored in all notations. Hence, with $z \in \mathbb{R}^l$ the hidden variable and $\xi \in \Xi \subseteq \mathbb{R}^d$ the parameter, the parametric complete likelihood is simply noted $\mathbb{P}_\xi(Y, z) =: f(z; \xi)$. Note that in the particular case where this formalism is applied to a mixture model, the parameter called "ξ" here includes both the hierarchical model parameter θ and the class weight parameter π. Considering μ a $\sigma-$finite positive Borel measure on $\mathbb{R}^l$, they also define:

$$\begin{aligned} g(\xi) &:= \int_z f(z;\xi)\mu(dz) \\ l(\xi) &:= ln\, g(\xi) \\ p(z;\xi) &:= \begin{cases} f(z;\xi)/g(\xi) & \text{if } g(\xi) \neq 0 \\ 0 & \text{if } g(\xi) = 0\,, \end{cases} \end{aligned} \tag{6.20}$$

the observed likelihood (implicitly in Y), observed log-likelihood (likewise, in Y) and conditional likelihood in z respectively. The same notations are used for a single data point ($z = z^{(1)} \in \mathbb{R}^l$) or several independent identically distributed (iid) data points ($z \equiv \left(z^{(i)}\right)_{i=1}^n \in \mathbb{R}^{l \times n}$), since all the probabilities are simply decomposed into products (and the log-probabilities into sums) in the latter case. The authors of [33] then define the following conditions:

- **M1.** The complete likelihood belongs to the exponential density. That is to say: $f(z;\xi) = exp(\psi(\xi) + \left\langle \tilde{S}(z), \phi(\xi) \right\rangle)$. Where $\tilde{S}$ is a Borel function from $\mathbb{R}^l$ to $\mathcal{S} \subseteq \mathbb{R}^m$ such that the convex hull of $S(\mathbb{R}^l) \subseteq \mathcal{S}$. Additionally, $\forall \xi \in \Xi$:
$$\int_{z \in \mathbb{R}^l} \left|\tilde{S}(z)\right| p(z;\xi)\mu(dz) < \infty\,.$$
- **M2.** ϕ and ψ are twice continuously differentiable on Ξ.
- **M3.** $\bar{s} : \Xi \longrightarrow \mathcal{S}$ defined as:
$$\int_{z \in \mathbb{R}^l} \tilde{S}(z) p(z;\xi)\mu(dz)\,,$$
is continuously differentiable in ξ.
- **M4.** The function l is continuously differentiable on Ξ and
$$\partial \int_{z \in \mathbb{R}^l} f(z;\xi)\mu(dz) = \int_{z \in \mathbb{R}^l} \partial f(z;\xi)\mu(dz)$$
- **M5.** Let $L(s, \xi) := \psi(\xi) + \langle s, \phi(\xi) \rangle$ There exists a function $\hat{\xi} : \mathcal{S} \longrightarrow \Xi$ continuously differentiable such that:
$$\forall \xi \in \Xi, \quad \forall s \in \mathcal{S}, L(s, \hat{\xi}(s)) \geq L(s, \xi)\,.$$

Let us call T the map $\Xi \longrightarrow \Xi$ describing one step of the EM algorithm run on the observed likelihood $g(\xi)$: $\xi^{(t+1)} =: T(\xi^{(t)})$. With the formalism of [33] and under their assumptions, this step can be written: $T(\xi) = \hat{\xi}(\bar{s}(\xi))$. Moreover, the set of fixed points of T is equal to the set of critical points of $l(\xi)$: $\mathcal{L} := \{\xi | \partial_\xi l(\xi) = 0\} = \{\xi | T(\xi) = \xi\}$. Hence, under $M1-5$, the EM sequence $(\xi^{(t)})_t$ converges towards the set of critical points $\mathcal{L}$ of the log-likelihood $l(\xi)$ provided that the sequence $\xi^{(t)} := T^t(\xi_0)$ remains within a compact of Ξ. Using this remark, the authors of [33] call $\mathcal{L}(\xi)$ the set of limit points of the sequence $T^t(\xi)$ and state the following Theorem:

Theorem 6.6.1 (Theorem 1 of [33])**.** *Under $M1-5$, if additionally the closure $clos(\mathcal{L}(\xi))$ is compact of Ξ, then for any initial point ξ_0, the sequence $l(\xi^{(t)})$ is increasing and $\lim_{t\to\infty} d(\xi^{(t)}, \mathcal{L}) \longrightarrow 0$.*

Application to the Mixture of CGGM

In this section, we prove a theorem providing conditions under which the proposed EM for the Mixture of CGGM verifies the hypothesis of Theorem 6.6.1.

CGGM belongs to the exponential family. We did not specify a statistical model for X. However we can formally consider a joint density of the form $p_\theta(Y,X) = p_\theta(Y|X) \times p(X)$ with the improper prior $p(X) = 1$. Then, since the CGGM density (6.4) can be put under the form:

$$p_\theta(Y|X) = exp\left(\frac{1}{2}\left(-pln(2\pi) + ln\,|\Lambda| - tr\left(\Lambda YY^T + 2\Theta YX^T + \Theta\Lambda^{-1}\Theta^T XX^T\right)\right)\right), \tag{6.21}$$

then the vector $(Y,X) \in \mathbb{R}^{p+q}$ belongs to the exponential family, with

- $\psi(\theta) := \frac{1}{2}\left(-pln(2\pi) + ln\,|\Lambda|\right)$,
- $S(Y,X) := vec(YY^T, 2YX^T, XX^T)$, and
- $\phi(\theta) := \frac{1}{2}vec(\Lambda, \Theta, \Theta\Lambda^{-1}\Theta^T)$

Preliminary results. Before stating our own theorem, we prove some general results. With a Mixture of CGGM, we have $\xi := (\theta, \pi)$. The observed likelihood $g(\xi)$ of [33] is $\prod_{i=1}^n p_{\theta,\pi}(Y^{(i)}|X^{(i)})$, where $p_{\theta,\pi}(Y|X)$ is the observed density of a single data point, defined by Eq (6.10). Our EM algorithm, defined by the E step (6.13) and the M step (6.14), actually optimises a penalised version of this likelihood. The target function of the algorithm is defined in Eq (6.24), and takes the form of a penalised negative log-likelihood: $-\frac{1}{n}ln\,g(\xi) + pen(\xi)$. An equivalent formulation is to consider the new likelihood: $\tilde{g}(\xi) := g(\xi)e^{-n\,pen(\xi)}$. In the following, we show that the corresponding complete likelihood $\tilde{f}(z;\xi) = f(z;\xi)e^{-n\,pen(\xi)}$ belongs to the exponential family as per hypothesis M1 of [33]. First, recall that we already showed with Eq (6.21) that the simple CGGM density belong to the exponential family, i.e. $p_\theta(Y|X) = exp\left(\psi(\theta) + \langle S(Y,X), \phi(\theta)\rangle\right)$. With $\psi(\theta) = \frac{1}{2}\left(-pln(2\pi) + ln\,|\Lambda|\right)$, $S(Y,X) = vec(YY^T, 2YX^T, XX^T)$, and $\phi(\theta) = \frac{1}{2}vec(\Lambda, \Theta, \Theta\Lambda^{-1}\Theta^T)$. Thanks to this observation, the mixture of CGGM density can easily be expressed as a member of the exponential family as well. With a single data point (X,Y,z), we have:

$$\begin{aligned}
f(z;\xi) &= p_{\theta,\pi}(Y,z|X) \\
&= \sum_{k=1}^K \mathbb{1}_{z=k}\pi_k p_{\theta_k}(Y|X) \\
&= \sum_{k=1}^K \mathbb{1}_{z=k}\pi_k exp\left(\psi(\theta_k) + \langle S(Y,X), \phi(\theta_k)\rangle\right) \\
&= exp\left(\sum_{k=1}^K \left(\mathbb{1}_{z=k}\left(ln(\pi_k) + \psi(\theta_k)\right) + \langle \mathbb{1}_{z=k}S(Y,X), \phi(\theta_k)\rangle\right)\right) \\
&=: exp\left(\left\langle \underline{S(Y,X,z)}, \underline{\phi(\theta,\pi)} \right\rangle\right),
\end{aligned} \tag{6.22}$$

with the sufficient statistics and natural parameters

$$\begin{aligned}
\underline{S(Y,X,z)} &:= (\mathbb{1}_{z=k},\ \mathbb{1}_{z=k}S(Y,X))_{k=1}^K \\
\underline{\phi(\theta,\pi)} &:= ((\log\pi_k + \psi(\theta_k)),\ \phi(\theta_k))_{k=1}^K .
\end{aligned}$$

With several iid observations $\left(Y^{(i)}, X^{(i)}, z^{(i)}\right)_{i=1}^n$, let $z := (z^{(1)}, ..., z^{(n)})$, the complete likelihood is also written under the exponential form:

$$f(z;\xi) = exp\left(\left\langle \sum_{i=1}^n \underline{S\left(Y^{(i)}, X^{(i)}, z^{(i)}\right)}, \underline{\phi(\theta,\pi)} \right\rangle\right). \tag{6.23}$$

Finally we get the penalised complete likelihood:

$$\tilde{f}(z;\xi) = exp\left(-n\,pen(\theta,\pi) + \left\langle \sum_{i=1}^n \underline{S\left(Y^{(i)}, X^{(i)}, z^{(i)}\right)}, \underline{\phi(\theta,\pi)} \right\rangle\right), \tag{6.24}$$

which indeed belongs to the exponential family. In the formalism of [33], the sufficient statistics and natural parameters are:

$$
\begin{aligned}
S(z) &:= \sum_{i=1}^{n} \underline{S\left(Y^{(i)}, X^{(i)}, z^{(i)}\right)} = \left(\sum_{i=1}^{n} \mathbb{1}_{z^{(i)}=k}, \sum_{i=1}^{n} \mathbb{1}_{z^{(i)}=k} vec(Y^{(i)}Y^{(i)T}, 2Y^{(i)}X^{(i)T}, X^{(i)}X^{(i)T}) \right)_{k=1}^{K} \\
\phi(\xi) &:= \underline{\phi(\theta, \pi)} = \left(\log \pi_k + \frac{1}{2}\left(-pln(2\pi) + ln\left|\Lambda_k\right|\right), \frac{1}{2} vec(\Lambda_k, \Theta_k, \Theta_k \Lambda_k^{-1} \Theta_k^T) \right)_{k=1}^{K} \\
\psi(\xi) &:= -n\, pen(\theta, \pi)\,.
\end{aligned}
\tag{6.25}
$$

Where $S(z) \in \mathcal{S} := \mathbf{S}_K \times {S_p^+}^K \times \mathcal{M}_{qp}^K \times {S_q^+}^K$. Under this form, we see that $\xi \mapsto \phi(\xi)$ is C^∞ on Ξ. Moreover, the hidden variable $z = (z^{(1)}, ..., z^{(n)}) \in [\![1, K]\!]^n$ belongs to a finite space, hence all the integrals expressed in the conditions of [33] are actually finite sums. Thanks to these two properties, the conditions M1-4 are quickly verified.

Statement of the Theorem

Theorem 6.6.2 (Convergence of the penalised Mixture of CGGM EM)**.** *Assume that* $\left\{Y^{(i)}, X^{(i)}\right\}_{i=1}^{n}$ *are observed data points. Let* $\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}} := \left\{\theta^{(t)}, \pi^{(t)}\right\}_{t\in\mathbb{N}}$ *be the EM sequence defined by the E step* (6.13) *and the M step* (6.14)*, with initial points* $\pi^{(0)} \in \mathbf{S}_K := \left\{(\pi_1, ..., \pi_K) \in]0,1[^K \,\middle|\, \sum_k \pi_k = 1\right\}$ *and* $\theta^{(0)} \in \left(S_p^{++} \times \mathcal{M}_{q,p}\right)^K$*.*

- ***Condition on*** pen_θ***.*** *Assume that* pen_θ *is* C^2 *in* θ*. Assume in addition that there exists* $\alpha, \beta, \gamma \in \mathbb{R}_+^*$ *such that:*

$$
pen_\theta(\theta) \geq \sum_k \left(\alpha \left\|\Lambda_k\right\| + \beta \left\|\Theta_k\right\|^2 + \gamma \left\|\Lambda_k^{-1}\right\| \right)\,, \tag{6.26}
$$

  and that the Hessian matrix $Hpen_\theta$ *of this penalty is definite positive for all values of the parameters:*

$$
\forall \theta, \quad (Hpen_\theta)(\theta) \succ 0\,. \tag{6.27}
$$

- ***Condition on*** pen_π***.*** *Assume that* pen_π *is convex,* C^2 *in* π*. Assume in addition that the solution* $\pi^{(t+1)}$ *to* (6.14) *is* C^1 *in the sufficient statistics* $\left(\frac{n_k^{(t)}}{n}\right)_{k=1}^{n}$*.*

Then the Theorem 1 of [33] applies: the penalised negative log-likelihood $-\frac{1}{n} ln\, g(\xi^{(t)}) + pen(\xi^{(t)})$ *is decreasing and* $\lim_{t\to\infty} d(\xi^{(t)}, \mathcal{L}) \longrightarrow 0$*. Where* $\mathcal{L}$ *is the set of critical points of* $\xi \mapsto -\frac{1}{n} ln\, g(\xi) + pen(\xi)$*.*

Remark.

- All the requirements on $pen_\pi(\pi)$ are verified by $pen_\pi(\pi) := 0$, since we have then $\pi_k^{(t+1)} = \frac{n_k^{(t)}}{n}$. A typical non-zero penalty that also verifies of all the requirements is $pen_\pi(\pi) := -\delta \sum_{k=1}^{K} ln(\pi_k)$ with $\delta > 0$. Most noticeably, the solution to (6.14) is $\pi_k^{(t+1)} = \frac{n_k^{(t)}/n + \delta}{1 + K\delta}$, which is indeed C^∞ in the sufficient statistic. In practice, most EM algorithms on mixture models include, explicitly or not, this penalty with a very small δ, in order to avoid vanishing clusters.
- The lower bound of $pen_\theta(\theta)$ by $\sum_k \gamma \left\|\Lambda_k^{-1}\right\|$ is only needed as a sufficient condition to verify that the EM sequence $\xi^{(t)}$ remains within a compact. If this lower bound condition is not verified ($\gamma = 0$), then as long as the EM sequence is still observed to remain within a compact of Ξ, then the theorem still applies, with all convergence guarantees.

Proof First, the $M1-4$ conditions. We already proved that the complete likelihood could be expressed as a member of the exponential family in (6.24). The sufficient statistics space is $\mathcal{S} := \mathbf{S}_K \times {S_p^+}^K \times \mathcal{M}_{qp}^K \times {S_q^+}^K$ in our case. It is a convex space, and $S(\mathbb{R}^l) \subseteq \mathcal{S}$, with S defined as in Eq (6.25). As a result, the convex hull of $S(\mathbb{R}^l)$ is as well included in $\mathcal{S}$. This part of $M1$ is verified as well. Since $z = (z^{(1)}, ..., z^{(n)}) \in [\![1, K]\!]^n$ belongs to a finite space, all the integrals expressed in $M1-4$ are actually finite sums. Additionally, the natural parameters $\phi(\xi)$ emerging from the Mixture of CGGM are all C^∞ in $\xi = (\theta, \pi)$ on their space of definition Ξ. The natural parameter $\psi(\xi)$, coming from the penalty $pen(\theta, \pi)$ is C^2 in ξ by hypothesis. Thanks to this combination of finite sums over z with the C^2 natural parameters $\phi(\xi)$ and $\psi(\xi)$, the conditions M1-4 are immediately verified.

Condition $M5$ requires the solution $\hat{\xi}(s) = (\hat{\pi}(s), \hat{\theta}(s))$ of the M step (6.14) to be continuously differentiable on the value of the sufficient statistics s. In our formulation of the M step (6.14), the sufficient statistic is $s = \overline{s}(\xi^{(t)})$, and it defines the next step $\xi^{(t+1)}$. However, $\overline{s}(\xi^{(t)})$ can be replaced by any element of $\mathcal{S}$ in the formula to define $\hat{\xi}(s)$ for any $s \in \mathcal{S}$. First, note that $\hat{\pi}(s)$ is already continuously differentiable in s by hypothesis. We underline that this property is not hard to have in practice, since we showed that it is achieved with one of the most used penalties on π. As a result, we just need to prove that, with our model and under the assumptions of the theorem, the solution $\hat{\theta}(s)$ exists, is unique and is continuously differentiable in s. This property can be shown using the theorem of the implicit function. In order to apply this theorem, we prove that for any $s \in \mathcal{S}$, the solution to the M step (6.14) $\hat{\theta}(s)$ is the global minimum of a convex function $\theta \mapsto L(s, \theta)$, and is reached on the inside of its set of definition, and not on the border. To do so, we prove that for any $s \in \mathcal{S}$, $\theta \mapsto L(s, \theta)$ is infinite on its border. For $s = \left(\frac{n_k}{n}, S_{YY}^k, S_{YX}^k, S_{XX}^k\right)_{k=1}^K \in S$, let:

$$\begin{aligned} L(s, \theta) :=& \frac{1}{2} \sum_{k=1}^K \left(\left\langle \Lambda_k, S_{YY}^k \right\rangle + \left\langle 2\Theta_k, S_{YX}^k \right\rangle + \left\langle \Theta_k \Lambda_k^{-1} \Theta_k^T, S_{XX}^k \right\rangle \right) \\ & - \frac{1}{2} \sum_{k=1}^K \frac{n_k}{n} ln(|\Lambda_k|) + pen_\theta(\theta) \,. \end{aligned} \tag{6.28}$$

Then, $\hat{\theta}(s)$, defined as in our M step (6.14), is written:

$$\hat{\theta}(s) := \underset{\Lambda, \Theta \in {S_p^+}^K \times \mathcal{M}_{qp}^K}{\operatorname{argmin}} L(s, \theta) \,. \tag{6.29}$$

We show that this minimisation is properly defined. First, let us show that the function $\theta \mapsto L(s, \theta)$ takes infinitely large values on its border. We lower bound or rewrite each term defining $L(s, \theta)$. First, notice that $\left\langle \Lambda_k, S_{YY}^k \right\rangle = tr(\Lambda_k S_{YY}^k) = tr(\Lambda_k^{\frac{1}{2}} S_{YY}^k \Lambda_k^{\frac{1}{2}}) \geq 0$ since $\Lambda_k^{\frac{1}{2}} S_{YY}^k \Lambda_k^{\frac{1}{2}} \in S_p^+$. Likewise, we have $\left\langle \Theta_k \Lambda_k^{-1} \Theta_k^T, S_{XX}^k \right\rangle = tr(\Theta_k \Lambda_k^{-1} \Theta_k^T S_{XX}^k) = vec(\Theta_k)^T (S_{XX}^k \otimes \Lambda_k^{-1}) vec(\Theta_k) \geq 0$ since $S_{XX}^k \otimes \Lambda_k^{-1} \in S_{pq}^+$, with $\otimes$ the Kronecker product. We also have $pen_\theta(\theta) \geq \sum_k \left(\alpha \left\| \Lambda_k \right\|_* + \beta \left\| \Theta_k \right\|_F^2 \right)$ by hypothesis. Since all norms are equivalent in finite dimension, we chose to express this inequality with the nuclear norm $\left\| \Lambda_k \right\|_* = tr(\Lambda_k) = \sum_{j=1}^p \lambda_k^j$ for Λ_k, and the Frobenius norm $\left\| \Theta_k \right\|_F = tr(\Theta_k \Theta_k^T)^{\frac{1}{2}}$ for Θ_k. The $(\lambda_k^j)_j$ are the eigenvalues of Λ_k. With them, we can rewrite $ln(|\Lambda_k|) = \sum_j ln(\lambda_k^j)$. Finally, the Cauchy-Schwartz inequality gives $\left\langle \Theta_k, S_{YX}^k \right\rangle \geq - \left\| \Theta_k \right\|_F \left\| S_{YX}^k \right\|_F$. Combining all those results, we get:

$$\begin{aligned} L(s, \theta) &\geq \sum_{k=1}^K \left(\left\| \Theta_k \right\|_F \left(\beta \left\| \Theta_k \right\|_F - \left\| S_{YX}^k \right\|_F \right) + \sum_{j=1}^p \left(\alpha \lambda_k^j - \frac{1}{2} \frac{n_k}{n} ln(\lambda_k^j) \right) \right) \\ &= \sum_{k=1}^K \left(h_1^k \left(\left\| \Theta_k \right\|_F \right) + \sum_{j=1}^p h_2^k \left(\lambda_k^j \right) \right) \,, \end{aligned} \tag{6.30}$$

With $h_1^k : x \mapsto \beta x \left(x - \frac{\|S_{YX}^k\|_F}{\beta} \right)$ and $h_2^k : x \mapsto \alpha \left(x - \frac{1}{2\alpha} \frac{n_k}{n} ln(x) \right)$. An analysis on $x \in \mathbb{R}_+$ shows that these functions can all be lower bounded by a fixed constant that we call c. Moreover $h_1^k(x) \underset{x \longrightarrow +\infty}{\longrightarrow} +\infty$, $h_2^k(x) \underset{x \longrightarrow +\infty}{\longrightarrow} +\infty$ and $h_2^k(x) \underset{x \longrightarrow 0}{\longrightarrow} +\infty$. As a result, for any $M > 0$, there exists $m > 0$ such that $\forall k \leq K$:

$$\begin{aligned} \forall x \geq m, \quad & h_1^k(x) \geq M \\ \forall x \geq m, \quad & h_2^k(x) \geq M \\ \forall x \leq \frac{1}{m}, \quad & h_2^k(x) \geq M \,. \end{aligned}$$

We define

$$V_m := \left\{ (\Lambda, \Theta) \in {S_p^{++}}^K \times \mathcal{M}_{qp}^K | \forall k, \lambda_{max}(\Lambda_k) \leq m, \lambda_{min}(\Lambda_k) \geq \frac{1}{m}, \|\Theta_k\|_F \leq m \right\} . \tag{6.31}$$

V_m is a compact set within ${S_p^{++}}^K \times \mathcal{M}_{qp}^K$. If $(\Lambda, \Theta) \in \left({S_p^{++}}^K \times \mathcal{M}_{qp}^K \right) \setminus V_m$, then by definition there is at least one $k \in [\![1, K]\!]$ such that either

- $\lambda_{max}(\Lambda_k) \geq m$, in which case $h_2^k(\lambda_{max}(\Lambda_k)) \geq M$,
- $\lambda_{min}(\Lambda_k) \leq \frac{1}{m}$, in which case $h_2^k(\lambda_{min}(\Lambda_k)) \geq M$,
- $\|\Theta_k\|_F \geq m$, in which case $h_1^k(\|\Theta_k\|_F) \geq M$.

These three options are of course not exclusive. Regardless, this results in the lower bound:

$$L(s, \theta) \geq M + (K(p-1) - 1)c \,.$$

Where $(K(p-1) - 1)c$ is a fixed constant depending only on s and the dimension of the problem. Hence, we have proven that for any $s \in \mathcal{S}$, for any $M > 0$, there exists $m > 0$ such that if $(\Lambda, \Theta) \in \left({S_p^{++}}^K \times \mathcal{M}_{qp}^K \right) \setminus V_m$, then

$$L(s, \theta) \geq M \,.$$

This means that $L(s, \theta)$ grows infinitely large outside of the sets V_m as $m \longrightarrow +\infty$. Moreover, for any $s \in \mathcal{S}$, there exists at least one $\theta \in {S_p^{++}}^K \times \mathcal{M}_{qp}^K$ such that $L(s, \theta)$ is finite ($L(s, \theta)$ is finite for every $\theta \in {S_p^{++}}^K \times \mathcal{M}_{qp}^K$ actually). As a consequence, there must exist a set V_m such that the minimum of $\theta \mapsto L(s, \theta)$ is reached inside V_m.
Now, let us prove that this minimum is reached on a unique point $\hat{\theta}(s)$, and that $\nabla_\theta L(s, \hat{\theta}(s)) = 0$. According to the formula (6.28), the function $\theta \mapsto L(s, \theta)$ is the sum of two terms: $L(s, \theta) =: \tilde{L}(s, \theta) + pen_\theta(\theta)$. The term $\theta \mapsto \tilde{L}(s, \theta)$ is the negative log-likelihood of the hierarchical CGGM. It is C^∞ and convex in θ. In particular, this means that the hessian in θ is semi-positive definite: $\forall s, \theta,\ (H_\theta \tilde{L})(s, \theta) \succeq 0$. By hypothesis, the penalty $pen_\theta(\theta)$ is C^2, convex in θ and $\forall \theta,\ (Hpen_\theta)(\theta) \succ 0$. As a consequence, $\theta \mapsto L(s, \theta)$ is itself C^2, convex in θ and its Hessian is always positive definite $\forall s \in \mathcal{S}, \forall \theta,\ (H_\theta L)(s, \theta) = (H_\theta \tilde{L})(s, \theta) + (Hpen_\theta)(\theta) \succ 0$. Which implies that $\theta \mapsto L(s, \theta)$ is even strictly convex in θ. Since we have shown that the minimum of $\theta \mapsto L(s, \theta)$ is reached on the inside of its definition set, this means that this minimum is reached for a unique value $\hat{\theta}(s)$ which verifies $\nabla_\theta L(s, \hat{\theta}(s)) = 0$. Since the Hessian $(H_\theta L)$ is always positive definite, we also have $(H_\theta L)(s, \hat{\theta}) \succ 0$. Among other things, we have proven that the optimisation problem (6.29) is well defined and has a unique solution.
Now we apply the theorem of the implicit function to the function $(s, \theta) \mapsto \nabla_\theta L(s, \theta)$, which is C^2 in θ and C^∞ in s. We have $\nabla_\theta L(s, \hat{\theta}(s)) = 0$. The Hessian $H_\theta L$ is the Jacobian of the gradient $\nabla_\theta L$. We have proven that $(H_\theta L)(s, \hat{\theta})$ is invertible at the point $\hat{\theta}(s)$. Hence, the theorem of implicit function applies to $\theta \mapsto \nabla_\theta L(s, \theta)$. It states that there exists an open set V_s containing s and a

unique function, continuously differentiable function $\hat{\theta}_s$ defined on V_s such that $\hat{\theta}_s(s) = \hat{\theta}(s)$ and for any $s' \in V_s$, $\nabla_\theta L(s', \hat{\theta}_s(s')) = 0$. Since $\theta \mapsto L(s', \theta)$ is strictly convex, this means that:

$$\hat{\theta}_s(s') = \underset{\Lambda,\Theta \in S_p^{++K} \times \mathcal{M}_{qp}^K}{\operatorname{argmin}} L(s', \theta)\,.$$

By uniqueness of the argmin, we have: $\forall s' \in V_s$, $\hat{\theta}_s(s') = \hat{\theta}(s')$. Hence $s \mapsto \hat{\theta}(s)$ is continuously differentiable on the open set V_s. Since this reasoning can be applied on all points $s \in \mathcal{S}$, and $\mathcal{S} \subseteq \cup_{s \in \mathcal{S}} V_s$, where the V_s are all open sets, we actually have that $\hat{\theta}(s)$ is the unique, continuously differentiable on $\mathcal{S}$ function such that $\forall s \in \mathcal{S}$, $\hat{\theta}(s) = \underset{\Lambda,\Theta \in S_p^{++K} \times \mathcal{M}_{qp}^K}{\operatorname{argmin}} L(s, \theta)$. With this, condition $M5$ is verified.

The final condition to apply Theorem 6.6.1 is that the sequence $\left\{ln(\tilde{g}(\xi^{(t)}))\right\}_t$ remains within a compact of Ξ. The auhtors of [33] provide a sufficient condition to have that:

$$L_C := \left\{\xi \in \Xi \middle| -ln(\tilde{g}(\xi)) \leq C\right\}\,, \tag{6.32}$$

is compact for any $C \geq 0$. We prove this property by obtaining a lower bound on $-ln(\tilde{g}(\xi))$ similar to (6.30). First, we express $-\frac{1}{n}ln(\tilde{g}(\xi))$ starting from its formula in Eq (6.12):

$$\begin{aligned}
-\frac{1}{n}ln(\tilde{g}(\xi)) &= -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k\, p_{\theta_k}\left(Y^{(i)}|X^{(i)}\right)\right) + pen(\theta, \pi) \\
&= -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k p_{\theta_k}\left(Y^{(i)}|X^{(i)}\right) e^{-pen_\theta(\theta)}\right) + pen_\pi(\pi) \\
&= -\frac{1}{n}\sum_{i=1}^{n} ln\Bigg(\sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{\frac{p}{2}}} exp\Bigg(\frac{1}{2}ln\,|\Lambda_k| - tr(\Theta_k Y^{(i)} X^{(i)\,T}) - pen_\theta(\theta) \\
&\qquad\qquad - \frac{1}{2}tr(\Lambda_k Y^{(i)} Y^{(i)\,T}) \\
&\qquad\qquad - \frac{1}{2}tr(\Sigma_k \Theta_k^T X^{(i)} X^{(i)\,T} \Theta_k)\Bigg)\Bigg) + pen_\pi(\pi) \\
&\geq -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \frac{1}{(2\pi)^{\frac{p}{2}}} exp\left(\frac{1}{2}ln\,|\Lambda_k| - tr(\Theta_k Y^{(i)} X^{(i)\,T}) - pen_\theta(\theta)\right)\right) + cst\,.
\end{aligned}$$

Where we used $\frac{1}{2}tr(\Lambda_k Y^{(i)} Y^{(i)\,T}) \geq 0$ and $\frac{1}{2}tr(\Sigma_k \Theta_k^T X^{(i)} X^{(i)\,T} \Theta_k) \geq 0$, which are true for the same reasons as before. We also used $\pi_k \leq 1$ as well as the fact that pen_π is lower bounded, since it is continuous, and π lives in a compact. In the following, we will assume this lower bound to be 0 and remove the "cst" term at the end. For the next steps, we re-use the hypothesis on pen_θ:

$$pen_\theta(\theta) \geq \sum_k \left(\alpha \left\|\Lambda_k\right\|_* + \beta \left\|\Theta_k\right\|_F + \gamma \left\|\Lambda_k^{-1}\right\|_*\right)\,.$$

Where we chose the nuclear norm $\|.\|_*$ on Λ_k^{-1} as we did for Λ_k. We write $\|\Lambda_k\|_* = \sum_{j=1}^{p} \lambda_k^j$, $\left\|\Lambda_k^{-1}\right\|_* = \sum_{j=1}^{p} (\lambda_k^j)^{-1}$ and $ln\,|\Lambda_k| = \sum_{j=1}^{p} ln(\lambda_k^j)$. We also re-use $tr(\Theta_k Y^{(i)} X^{(i)\,T}) \geq -\left\|\Theta_k\right\|_F \left\|Y^{(i)} X^{(i)\,T}\right\|_F$.

We get:

$$
\begin{aligned}
-\frac{1}{n}ln(\tilde{g}(\xi)) \geq -\frac{1}{n}\sum_{i=1}^{n} ln\Bigg(\sum_{k=1}^{K}\frac{1}{(2\pi)^{\frac{p}{2}}}exp\Bigg(\sum_{j=1}^{p}\frac{1}{2}ln(\lambda_k^j) - \alpha\lambda_k^j - \gamma(\lambda_k^j)^{-1} \\
+ \left\|\Theta_k\right\|_F \left\|Y^{(i)}X^{(i)\,T}\right\|_F - \beta\left\|\Theta_k\right\|_F^2 \\
+ \sum_{l\neq k,l=1}^{K}\sum_{j=1}^{p} -\alpha\lambda_l^j - \gamma(\lambda_l^j)^{-1} \\
+ \sum_{l\neq k,l=1}^{K} -\beta\left\|\Theta_l\right\|_F^2\Bigg)\Bigg) \\
= -\frac{1}{n}\sum_{i=1}^{n} ln\Bigg(\sum_{k=1}^{K}\frac{1}{(2\pi)^{\frac{p}{2}}}exp\Bigg(-\sum_{j=1}^{p} h_2^k\left(\lambda_k^j\right) \\
- h_1^{k,i}\left(\left\|\Theta_k\right\|_F\right) \\
- \sum_{l\neq k,l=1}^{K}\sum_{j=1}^{p} u_2(\lambda_l^j) \\
- \sum_{l\neq k,l=1}^{K} u_1\left(\left\|\Theta_k\right\|_F\right)\Bigg)\Bigg).
\end{aligned}
\tag{6.33}
$$

Where $h_1^{k,i} : x \mapsto \beta x\left(x - \frac{\left\|Y^{(i)}X^{(i)\,T}\right\|_F}{\beta}\right)$, $h_2^k : x \mapsto \alpha x + \gamma\frac{1}{x} - \frac{1}{2}ln(x)$, $u_1 : x \mapsto \beta x^2$ and $u_2 : x \mapsto \alpha x + \gamma\frac{1}{x}$. These functions are slightly different from the h_1 and h_2 previously defined in the bound on $L(s,\theta)$, but are functionally identical. First, for $x \in \mathbb{R}_+$, they are all lower bounded by a finite constant that we call c. Second, they have the same limits as before: $h_1^{k,i}(x) \underset{x\longrightarrow+\infty}{\longrightarrow} +\infty$, $h_2^k(x) \underset{x\longrightarrow+\infty}{\longrightarrow} +\infty$, $h_2^k(x) \underset{x\longrightarrow 0}{\longrightarrow} +\infty$ and $u_1(x) \underset{x\longrightarrow+\infty}{\longrightarrow} +\infty$, $u_2(x) \underset{x\longrightarrow+\infty}{\longrightarrow} +\infty$, $u_2(x) \underset{x\longrightarrow 0}{\longrightarrow} +\infty$. With the lower bound (6.33) established, let us prove that L_C is compact for any $C \geq 0$. Let us assume that $(\theta,\pi) \in L_C$. By definition, $\pi \in \mathbf{S}_K$, the space of stochastic vectors of size K, which is already a compact. To conclude the proof, we show that there exists $m \geq 0$ such that θ belongs to the compact V_m defined in Eq (6.31). We have that, for any $M > 0$, there exists $m > 0$ such that $\forall(k,i) \in [\![1,K]\!] \times [\![1,n]\!]$:

$$
\begin{aligned}
\forall x \geq m, \quad & h_1^{k,i}(x) \geq M \\
\forall x \geq m, \quad & h_2^k(x) \geq M \\
\forall x \leq \frac{1}{m}, \quad & h_2^k(x) \geq M \\
\forall x \geq m, \quad & u_1(x) \geq M \\
\forall x \geq m, \quad & u_2(x) \geq M \\
\forall x \leq \frac{1}{m}, \quad & u_2(x) \geq M\,.
\end{aligned}
$$

If $\theta = (\Lambda,\Theta) \in \left(S_p^{++\,K} \times \mathcal{M}_{qp}^K\right) \setminus V_m$, then by definition there is at least one $k \in [\![1,K]\!]$ such that either

- $\lambda_{max}(\Lambda_k) \geq m$, in which case $h_2^k(\lambda_{max}(\Lambda_k)) \geq M$ and $u_2(\lambda_{max}(\Lambda_k)) \geq M$,
- $\lambda_{min}(\Lambda_k) \leq \frac{1}{m}$, in which case $h_2^k(\lambda_{min}(\Lambda_k)) \geq M$ and $u_2(\lambda_{max}(\Lambda_k)) \geq M$,
- $\left\|\Theta_k\right\|_F \geq m$, in which case $\forall i,\ h_1^{k,i}(\left\|\Theta_k\right\|_F) \geq M$ and $u_1(\left\|\Theta_k\right\|_F) \geq M$.

These three options are of course not exclusive. Regardless, this means that for all $k \in [\![1, K]\!]$, at least one of the terms in the sum:

$$\sum_{j=1}^{p} h_2^k\left(\lambda_k^j\right) + h_1^{k,i}\left(\|\Theta_k\|_F\right) + \sum_{l\neq k, l=1}^{K}\sum_{j=1}^{p} u_2(\lambda_l^j) + \sum_{l\neq k, l=1}^{K} u_1\left(\|\Theta_k\|_F\right),$$

is larger than M. The rest are all at least larger than c, the common lower bound. Hence $\forall k \in [\![1, K]\!]$:

$$\sum_{j=1}^{p} h_2^k\left(\lambda_k^j\right) + h_1^{k,i}\left(\|\Theta_k\|_F\right) + \sum_{l\neq k, l=1}^{K}\sum_{j=1}^{p} u_2(\lambda_l^j) + \sum_{l\neq k, l=1}^{K} u_1\left(\|\Theta_k\|_F\right) \geq M + (K(p-1)-1)c\,.$$

Where $(K(p-1)-1)c$ is a fixed constant depending only on the observed data and the dimension of the problem. Hence, we have proven that for any $M > 0$, there exists $m > 0$ such that if $\theta \in \left(S_p^{++K} \times \mathcal{M}_{qp}^K\right) \setminus V_m$, then

$$-\frac{1}{n} ln(\tilde{g}(\xi)) \geq -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \frac{1}{(2\pi)^{\frac{p}{2}}} exp\left(-M\right)\right) = M - ln\left(\frac{K}{(2\pi)^{\frac{p}{2}}}\right).$$

Since $-ln(\tilde{g}(\xi)) \leq C$, then there must exists a $m > 0$ such that $\theta \in V_m$. We have proven that $L_C = \left\{(\theta, \pi) \in \Xi \middle| - ln(\tilde{g}(\xi)) \leq C\right\}$ is included in the compact set $V_m \times \mathbf{S}_K$. L_C is also a closed set as the reciprocal image of closed set by a continuous function, hence L_C is a compact set. This ensures that the sequence $\left\{ln(\tilde{g}(\xi^{(t)}))\right\}_t$ is bounded, and concludes the proof of the Theorem.

6.6.2 Convergence of the tempered EM for Mixtures of CGGM

Since the likelihood function of a mixture is non-convex, the critical point towards which the EM converges will most probably be one of the local maxima closest to the initial point. This is a consequent problem if the initialisation is not very good, and is a systematic hurdle in high dimension, where the procedure converges without having explored at all the parameter space.
Hence, in this section, we propose an approximated variant of our algorithm with tempering on the E step, in accordance with the framework of Chapter 5. We recall that the tempering approximation is borrowed from the Simulated Annealing [82] optimisation technique. It weakens the gradients and potential wells of the likelihood profile, which allows the procedure to escape the initialisation and explore more of the parameter space before convergence. We prove that despite the approximation, the tempered version of our algorithm still verifies the same theoretical convergence guarantees, and demonstrate on the data that in practice, the end solution of the algorithm in consistently better.
We recall here the formalism used in Chapter 5 to define the tempered EM. With $p(z, \xi^{(t)}) = \frac{f(z,\xi^{(t)})}{g(\xi^{(t)})}$ the posterior probability in z, the two steps E (6.13) and M (6.14) of our EM algorithm can be expressed compactly as:

$$\xi^{(t+1)} := \underset{\xi\in\Xi}{\operatorname{argmax}}\, \mathbb{E}_{z\sim p(z,\xi^{(t)})}\left[ln\, f(z;\xi) - npen(\xi)\right]. \tag{6.34}$$

Let $(T_t)_{t\in\mathbb{N}}$ be a sequence of positive temperatures, The tempered E step replaces $p(z, \xi^{(t)})$ by the approximated posterior density:

$$\tilde{p}^{(t)}(z, \xi^{(t)}) = \frac{p(z, \xi^{(t)})^{\frac{1}{T_t}}}{\int_z p(z, \xi^{(t)})^{\frac{1}{T_t}} \mu(dz)},$$

With our mixture model, this means replacing the posterior probability $p_{i,k}^{(t)}$, defined for the exact E step (6.13), by the tempered version:

$$\tilde{p}_{i,k}^{(t)} := \frac{\left(p_{i,k}^{(t)}\right)^{\frac{1}{T_t}}}{\sum_{l=1}^{K}\left(p_{i,l}^{(t)}\right)^{\frac{1}{T_t}}}. \tag{6.35}$$

Afterwards, the M step is still defined using the same formula as the exact EM (6.14), where $p_{i,k}^{(t)}$ is replaced by $\tilde{p}_{i,k}^{(t)}$. In the end, one step of this approximated EM can be written as $\xi^{(t+1)} := F^{(t)}(\xi^{(t)})$, with:

$$F^{(t)}(\xi^{(t)}) := \underset{\xi \in \Xi}{\operatorname{argmax}} \mathbb{E}_{z \sim \tilde{p}^{(t)}(z,\xi^{(t)})} \left[ln\, f(z;\xi) - npen(\xi) \right] . \tag{6.36}$$

Here, $\left\{F^{(t)}\right\}_{t\in\mathbb{N}}$ is the sequence of point to point maps in Θ describing the tempered EM step. It is dependent on the sequence of approximations $\left\{\tilde{p}^{(t)}(z,\xi^{(t)})\right\}_{t\in\mathbb{N}}$. Fort and Moulines [49], as well as our Chapter 5, both propose convergence theorem for similarly approximated E steps. They also include a slight modification of the EM dynamic. Assume that you dispose of an increasing sequence of compacts $\{K_t\}_{t\in\mathbb{N}}$ such that $\cup_{t\in\mathbb{N}} K_t = \Theta$ and $\xi^{(0)} \in K_0$. Define $j_0 := 0$. Then, the transition $\xi^{(t+1)} = F^{(t)}(\theta^{(t)})$ is accepted only if $F^{(t)}(\theta^{(t)})$ belongs to the current compact K_{j_t}, otherwise the sequence is reinitialised at $\theta^{(0)}$. This algorithm is called a "Stable Approximate EM". One step can be written as:

$$\begin{cases} if F^{(t)}(\xi^{(t)}) \in K_{j_t}, \text{ then} & \xi^{(t+1)} = F^{(t)}(\xi^{(t)}), \text{ and } j_{t+1} := j_t \\ if F^{(t)}(\xi^{(t)}) \notin K_{j_t}, \text{ then} & \xi^{(t+1)} = \xi^{(0)}, \text{ and } j_{t+1} := j_t + 1\,. \end{cases} \tag{6.37}$$

In Chapter 5, we provide a convergence theorems for Stable Approximate EM belonging to the exponential family, in particular for the tempering approximation. The convergence guarantees are functionally the same as for the exact EM. In the following, we recall Theorem 5.4.1 of Chapter 5, then we state a new theorem, providing conditions under which the Stable Approximate EM applied to the penalised Mixture of CGGM benefits from these convergence guarantees.

General result for the tempered EM Let us recall quickly our convergence theorem from Chapter 5 for the tempered EM. Theorem 5.4.1 describes the convergence of the Stable Approximate EM (6.37) for any model of the exponential family, with the tempering approximation (6.35) and no penalty $pen(\xi)$, that is to say $F^{(t)}(\xi^{(t)})$ is defined from Eq (6.36) with $pen(\xi) = 0$. The Theorem makes use of the following assumptions:

C1. $f(z;\xi)$ can be written $f(z;\xi) = exp(\psi(\xi) + \left\langle \tilde{S}(z), \phi(\xi) \right\rangle)$.

C2.

(a*) ψ and ϕ are continuous on Ξ;

(b) for all $\xi \in \Xi$, $\bar{s}(\xi) := \int_z S(z) p(z;\xi) \mu(dz)$ is finite and continuous on Ξ;

(c) with $L(s;\hat{\xi}(s)) := \psi(\xi) + \langle s, \phi(\xi) \rangle$, there exists a continuous function $\hat{\xi} : \mathcal{S} \to \Xi$ such that for all $s \in \mathcal{S}$, $L(s;\hat{\xi}(s)) = \underset{\xi\in\Xi}{sup}\, L(s;\xi)$;

(d) g is positive, finite and continuous on Ξ and, for any $M > 0$, the level set $\{\xi \in \Xi, g(\xi) \geq M\}$ is compact.

C3. Assume either that:

(a) The set $g(\mathcal{L})$ is compact or

(a') for all compact sets $K \subseteq \Xi$, $g\left(K \cap \mathcal{L}\right)$ is finite.

Then, our theorem can be stated as:

Theorem (Theorem 5.4.1 of Chapter 5)**.** *Under C1−3, if additionally* $T_n \underset{n\,\infty}{\longrightarrow} 1$ *and for any compact* $K \in \Xi$, $\exists \epsilon \in]0,1[$, $\forall \alpha \in \overline{\mathcal{B}}(1,\epsilon)$*:*

- $\underset{\xi\in K}{sup} \int_z p^\alpha(z;\xi) dz < \infty$,
- $\forall i \in [\![1,m]\!]$, $\underset{\xi\in K}{sup} \int_z S_i^2(z) p^\alpha(z;\xi) dz < \infty$.

Where $\overline{\mathcal{B}}(1,\epsilon)$ is the closed ball centered in 1 and with radius ϵ in $\mathbb{R}$, and the index i of $S_i(z)$ indicates each of the real component of the vector $S(z) \in \mathcal{S}$.

Then, with $\mathcal{L} := \{\xi \in \Xi | \nabla g(\xi) = 0\}$ and $\mathcal{L}_{g^*} := \{\xi \in \mathcal{L} | g(\xi) = g^*\}$, the Stable Approximate EM sequence defined in Eq (6.37) benefits from the convergence guarantees:

(i) (a) With probability 1, $\lim_{n\infty} j_t < \infty$ and $\sup_{t\in\mathbb{N}} \left\|\xi^{(t)}\right\| < \infty$;

(b) $g(\xi^{(t)})$ converges towards a connected component of $g(\mathcal{L})$.

(ii) If, additionally, $g\left(\mathcal{L} \cap Cl\left(\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}}\right)\right)$ has an empty interior, then:

$$g(\xi^{(t)}) \underset{t\infty}{\longrightarrow} g^*\,,$$
$$d(\xi^{(t)}, \mathcal{L}_{g^*}) \underset{t\infty}{\longrightarrow} 0\,.$$

Convergence theorem for the tempered EM applied to the Mixture of CGGM We prove a new theorem, which provides conditions under which the tempered EM algorithm applied to the Mixture of CGGM - as defined by equations (6.35), (6.36) and (6.37) - falls under the convergence Theorem 5.4.1 of Chapter 5.

Theorem 6.6.3 (Convergence of the Stable Approximate EM for penalised Mixtures of CGGM)**.** *Assume that $\left\{Y^{(i)}, X^{(i)}\right\}_{i=1}^{n}$ are observed data points. Let $\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}} := \left\{\theta^{(t)}, \pi^{(t)}\right\}_{t\in\mathbb{N}}$ be the EM sequence defined by the Stable Approximate EM (6.37), with initial points $\pi^{(0)} \in \mathbf{S}_K$ and $\theta^{(0)} \in \left(S_p^{++} \times \mathcal{M}_{q,p}\right)^K$. Where $\mathbf{S}_K := \left\{(\pi_1, ..., \pi_K) \in]0,1[^K \,\middle|\, \sum_k \pi_k = 1\right\}$. Call $\mathcal{L} := \{\xi \in \Xi | \nabla \tilde{g}(\xi) = 0\}$ and $\mathcal{L}_{\tilde{g}^*} := \{\xi \in \mathcal{L} | \tilde{g}(\xi) = \tilde{g}^*\}$. Assume that $T_t \underset{t\longrightarrow\infty}{\longrightarrow} 1$ and:*

- ***Condition on pen_θ.*** *Assume that pen_θ analytic in θ. Assume in addition that there exists $\alpha, \beta \in \mathbb{R}_+^*$ such that:*

$$pen_\theta(\theta) \geq \sum_k \left(\alpha \left\|\Lambda_k\right\| + \beta \left\|\Theta_k\right\|\right)\,, \tag{6.38}$$

and that the Hessian matrix $Hpen_\theta$ of this penalty is definite positive for all values of the parameters:

$$\forall \theta, \quad (Hpen_\theta)(\theta) \succ 0\,.$$

- ***Condition on pen_π.*** *Assume that pen_π is convex and analytic in π. Assume in addition that the solution $\pi^{(t+1)}$ to (6.14) is C^1 in the sufficient statistics $\left(\frac{n_k^{(t)}}{n}\right)_{k=1}^{n}$.*

Then $\tilde{g}\left(\mathcal{L} \cap Cl\left(\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}}\right)\right)$ has an empty interior and all the results of the Theorem 5.4.1 of Chapter 5 apply. That is to say:

(i) (a) With probability 1, $\lim_{n\infty} j_t < \infty$ and $\sup_{t\in\mathbb{N}} \left\|\xi^{(t)}\right\| < \infty$;

(b) $\tilde{g}(\xi^{(t)})$ converges towards a connected component of $\tilde{g}(\mathcal{L})$.

(ii) $\tilde{g}(\xi^{(t)}) \underset{t\infty}{\longrightarrow} \tilde{g}^*$ *and*

$d(\xi^{(t)}, \mathcal{L}_{\tilde{g}^*}) \underset{t\infty}{\longrightarrow} 0\,.$

Proof We verify the hypothesis of the Theorem 5.4.1 of Chapter 5. Conditions $C1-2$ either are immediate or were proven in the proof of Theorem 6.6.2. Although the assumptions of Theorem 6.6.3 are slightly different from those of Theorem 6.6.2, the relevant parts to verify $C1-2$ are included.

Condition $C3$. Note that $\xi \mapsto g(\xi)$ is composed of exponential, scalar products, determinants (which are polynomial in the matrix entries) and logarithms, hence it is an analytic function. Then, by assumption on $pen(\xi)$, $\xi \mapsto g(\xi)e^{-n\,pen(\xi)} = \tilde{g}(\xi)$ is also an analytic function on Ξ. Analytic functions only have a finite number of zeros on any compact space. As a consequence, for all compact sets $K \subseteq \Xi$, $K \cap \mathcal{L}$ is finite, and $\tilde{g}(K \cap \mathcal{L})$ is also finite. $C3(a')$ is verified.

We now prove that $\tilde{g}\left(\mathcal{L} \cap Cl\left(\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}}\right)\right)$ has an empty interior. We have already shown that $\xi \mapsto \tilde{g}(\xi)$ is C^∞. As a result, Sard's theorem applies and $\tilde{g}(\mathcal{L})$ is of empty interior. Then, $\tilde{g}\left(\mathcal{L} \cap Cl\left(\left\{\xi^{(t)}\right\}_{t\in\mathbb{N}}\right)\right) \subseteq \tilde{g}(\mathcal{L})$ also has an empty interior.

Finally, we prove the two upper bounds on the integrals against $p^\alpha(z;\xi)dz$. Let α be any real number. Since this is a mixture model, $\int_z p^\alpha(z;\xi)dz$ and $\int_z S_i^2(z)p^\alpha(z;\xi)dz$ are actually finite sums over z. We also have

$$p^\alpha(z;\xi) = \left(\frac{\tilde{f}(z;\xi)}{\tilde{g}(\xi)}\right)^\alpha = \left(\frac{f(z;\xi)}{g(\xi)}\right)^\alpha .$$

Hence $\xi \mapsto p^\alpha(z;\xi)$ is a finite, continuous function for any α and z. As a consequence, the applications $\xi \mapsto \int_z p^\alpha(z;\xi)dz$ and $\xi \mapsto \int_z S_i^2(z)p^\alpha(z;\xi)dz$ are also continuous, and we have for any compact $K \subseteq \Xi$:

- $\sup_{\xi\in K} \int_z p^\alpha(z;\xi)dz < \infty$,
- $\forall i \in [\![1,m]\!], \quad \sup_{\xi\in K} \int_z S_i^2(z)p^\alpha(z;\xi)dz < \infty$.

This is true for any $\alpha \in \mathbb{R}$, a much stronger property than the required one. This concludes the proof.

6.6.3 High dimensional experiments with tempering

In this section, we present the results of the high dimensional synthetic experiments where each of the three EM is also run with its counterpart tempered version. We consider two initialisation procedures. A random one, where three initial centroids are picked at random from data points, and a "smart" one, which initialise the EM with a KMeans on the l_2 norm of Y. This initialisation is good because the heterogeneous effect of the co-features in this experiment have different average amplitudes for each sub-population. In this experiment, we do not penalise the likelihood optimised by the EM. There is no sparsity constraints on the estimated parameters.
In Figure 6.11, we represent with boxplots the empirical distribution of the followed metrics for each EM, tempered or not. Table 6.5 provides quantitative statistics on these distributions. Regardless of the tempering, the EM and EM residual that estimate Mixtures of GGM are at a level of classification error around 60%, very close to the threshold of 66.7%. On the other hand, the non-tempered EM for Mixtures of CGGM is at half this level of error, with around 30% of error in average. These performances are further improved by the tempering approximation, with an average classification error that goes down to 24%. The tempering approximation allows the EM to escape the random initialisation and find better values for the Mixture of CGGM parameters. Similar observations can be made about the ABC-like metric that computes the average distance to the nearest real data point of the virtual observations generated by the estimated statistical models. We note that according to this metric, the residual EM represents a certain improvement from the EM. The computation times are very low. This is a results of the absence of penalty, which makes each M step explicit, and the overall method much faster.
Figure 6.12 and Table 6.6 offer the same description for the experiments done with the smart

initialisation. In this case, the performances of the EM and residual EM are still poor, this is because the Mixture of GGM that they estimate cannot properly reproduce the form of the true data clusters. The EM with Mixture of CGGM on the other hand achieves very good performances, 6% of error, because this is a very favourable initialisation for the Mixture of CGGM parameters. This is an example of situation in which the tempering is not required and does not offer significant improvements. Looking back at the empirical distributions drawn on Figure 6.11 in the case of the random initialisation, we see that the Mixture of CGGM EM with tempering approximation can sometimes reach such levels of performances, but that this is rarer for the non-tempered version of the method.
With this study, we have confirmed the observations of Chapter 5 on the benefits of tempering in the case of the conditional EM with co-factors. We also delineated the cases where the tempering is most useful by studying different types of initialisations.

Table 6.5: Average, standard deviation and median (below) of the four followed performance metrics over the 30×5 simulations. Both the regular and tempered (tmp) version of each EM was run from the **random initialisation**. The best values are in **bold**. We can see that the classification performances with the Mixture of CGGM are better than the two methods with Mixtures of GGM. The tempered version of the EM for CGGM is even better performing.

		soft misclassif.	hard misclassif.	ABC-like metric	runtimes
GGM	not tmp	0.60 **(0.04)** 0.61	0.60 (0.05) 0.61	5.36 (0.23) 5.33	**4.1** (2.5) **3.5**
	tmp	0.59 (0.05) 0.60	0.59 (0.05) 0.59	5.37 (0.24) 5.36	4.5 (2.4) 4.2
GGM resid.	not tmp	0.59 **(0.04)** 0.60	0.58 (0.05) 0.59	4.66 **(0.11)** 4.65	4.8 (2.9) 3.9
	tmp	0.58 **(0.04)** 0.59	0.57 **(0.04)** 0.58	4.66 (0.12) 4.66	5.7 (2.8) 5.4
CGGM	not tmp	0.31 (0.17) 0.34	0.30 (0.17) 0.34	4.08 (0.20) 4.04	4.7 (2.9) 3.8
	tmp	**0.24** (0.18) **0.27**	**0.24** (0.18) **0.27**	**4.03** (0.21) **3.98**	4.9 **(2.4)** 4.4

6.6.4 Extension: EM for the exponential family

In this section, we prove that for any model $p_\theta(Y)$ of the exponential family, as long as the supervised penalised maximum likelihood estimation (6.3) is tractable, then a tractable EM can be designed to solve the corresponding unsupervised problem. Most of the works on unsupervised GGM [52, 61, 86, 187] adapt supervised MLE problems to the unsupervised scenario with an EM algorithm. Likewise, our EM for the Mixture of CGGM relies on the tractability of the supervised Hierarchical CGGM problem [69]. There is a more general result behind all of this, any supervised regularised MLE of a Hierarchical model of exponential family densities can be adapted, with the same regularisation, into an EM for the unsupervised scenario. This justifies the application of the EM algorithm for a wider range of models and penalties than those previously studied. In addition to the CGGM, this includes models such as the MNGM. This also allows the use of any of the structure-defining penalty functions previously introduced by the literature for the supervised maximum likelihood estimation.
First, we recall the definition of the exponential family and re-write the original optimisation problem (6.3) for a density of the exponential family. Let us assume that the random variable $Y \in \mathcal{X} \subseteq \mathbb{R}^p$ belongs to the curved exponential family of distributions with the parameter $\theta \in \Xi \subseteq \mathbb{R}^l$. Then there are three applications $\psi : \Xi \to \mathbb{R}$, $\phi : \Xi \to \mathbb{R}^d$ and $S : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}^d$ such that the density is:

$$p_\theta(Y) := exp\left(\psi(\theta) + \langle S(Y), \phi(\theta)\rangle\right) . \tag{6.39}$$

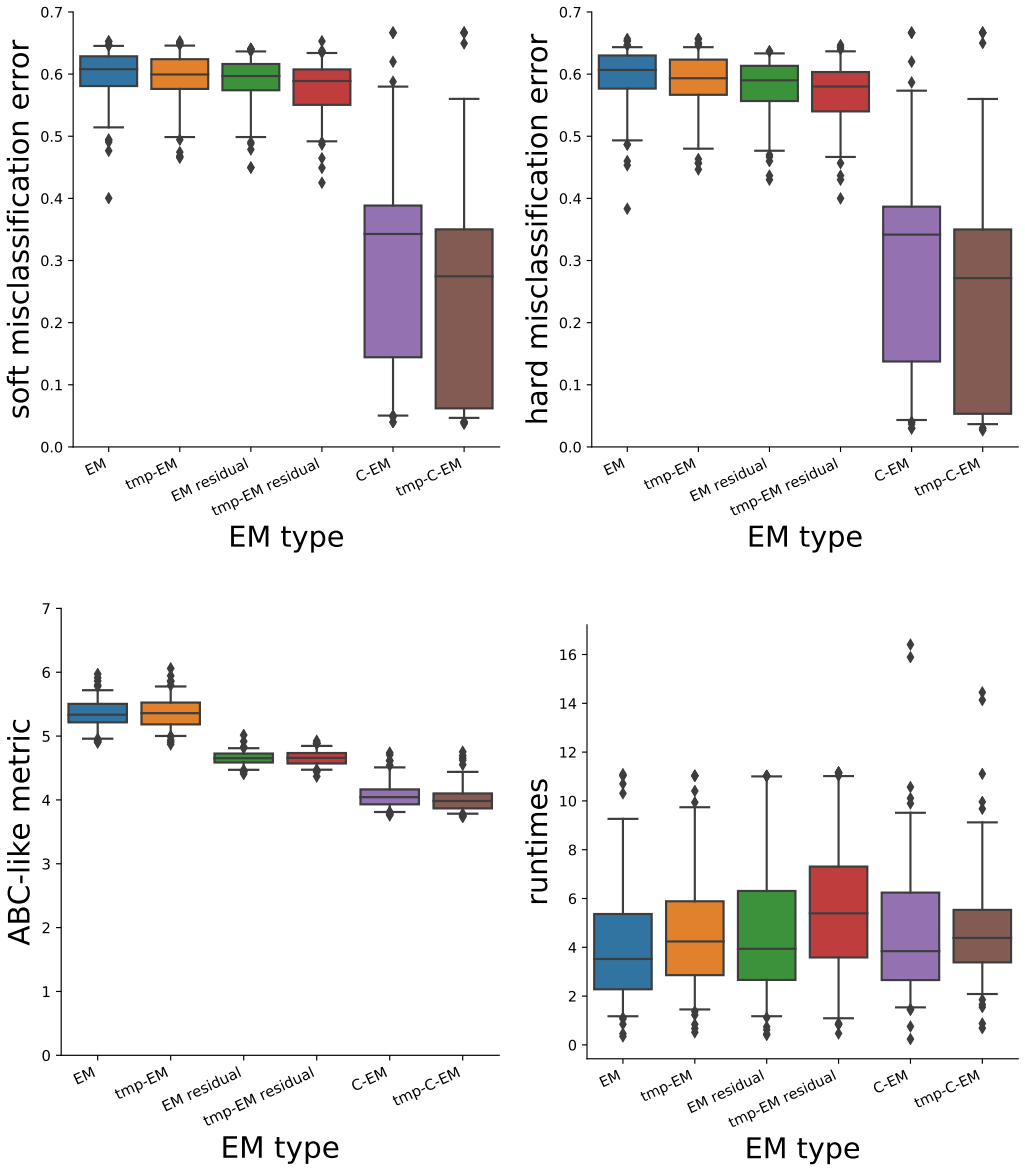


Figure 6.11: Empirical distribution of several performance metrics measured over many simulations with the **random initialisation**. The sample is made of 30 simulations with 5 different initialisations each. Three methods are compared. The EM and EM residual algorithms estimate a Mixture of GGM. The C-EM algorithm estimates a Mixture of CGGM. The C-EM is better performing and its performances are improved even further by the tempering. (Upper left) Soft mis-classification error $\left|\mathbb{1}_{z_i=k} - \widehat{\mathbb{P}}(z_i = k)\right|$. (Upper right) Hard mis-classification error $|\mathbb{1}_{z_i=k} - \mathbb{1}_{\widehat{z}_i=k}|$. (Bottom left) ABC-like metric. (Bottom right) Run time.

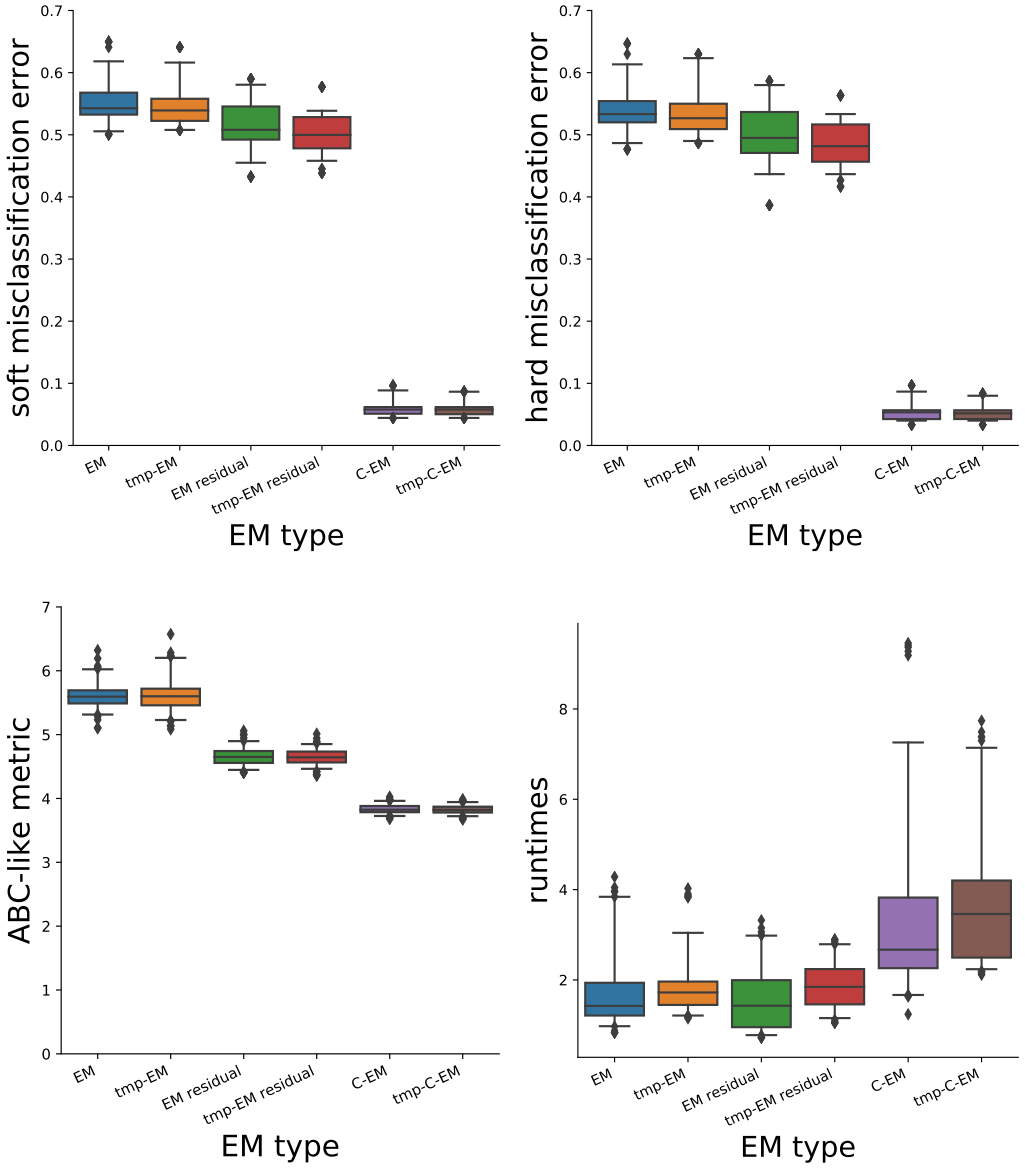


Figure 6.12: Empirical distribution of several performance metrics measured over many simulations with the **Smart initialisation**. The sample is made of 30 simulations with 5 different initialisations each. Three methods are compared. The EM and EM residual algorithms estimate a Mixture of GGM. The C-EM algorithm estimates a Mixture of CGGM. The C-EM is much better performing and faster. (Upper left) Soft mis-classification error $\left|\mathbb{1}_{z_i=k} - \widehat{\mathbb{P}}(z_i = k)\right|$. (Upper right) Hard mis-classification error $|\mathbb{1}_{z_i=k} - \mathbb{1}_{\widehat{z}_i=k}|$. (Bottom left) ABC-like metric. (Bottom right) Run time.

Table 6.6: Average, standard deviation and median (below) of the four followed performance metrics over the 30×5 simulations. Both the regular and tempered (tmp) version of each EM was run from the **smart initialisation**. The best values are in **bold**. We can see that the classification performances with the Mixture of CGGM are much better than the two methods with Mixtures of GGM, and with faster computation times.

		soft misclassif.	hard misclassif.	ABC-like metric	runtimes
GGM	not tmp	0.55 (0.03) 0.54	0.54 (0.04) 0.53	5.62 (0.22) 5.59	1.7 (0.8) **1.4**
	tmp	0.55 (0.03) 0.54	0.53 (0.04) 0.53	5.62 (0.27) 5.60	1.9 (0.6) 1.7
GGM resid.	not tmp	0.51 (0.04) 0.51	0.50 (0.05) 0.50	4.66 (0.14) 4.65	**1.6** (0.7) **1.4**
	tmp	0.50 (0.03) 0.50	0.49 (0.04) 0.48	4.65 (0.13) 4.64	1.9 **(0.5)** 1.8
CGGM	not tmp	**0.06 (0.01)** **0.06**	**0.05** (0.02) **0.05**	**3.83 (0.07)** **3.82**	3.5 (2.0) 2.7
	tmp	**0.06 (0.01)** **0.06**	**0.05 (0.01)** **0.05**	**3.83 (0.07)** **3.82**	3.7 (1.4) 3.5

We consider the case of a supervised Hierarchical model of densities of the form (6.39), let $(Y^{(i)})_{i=1}^n$ be an independent sample following this Hierarchical distribution with $z^{(i)}$ their known labels. Within the exponential family, the Hierarchical negative log-likelihood function minimised in the optimisation problem (6.3) can be re-written:

$$\begin{aligned}\sum_{k=1}^{K} -\frac{\sum_{i=1}^n \mathbb{1}_{z^{(i)}=k}}{n}\psi(\theta_k) + \sum_{k=1}^{K} -\left\langle \frac{\sum_{i=1}^n \mathbb{1}_{z^{(i)}=k} S(Y^{(i)})}{n}, \phi(\theta_k)\right\rangle + pen(\theta)\\ = -\left\langle \underline{P(z)}, \underline{\psi(\theta)}\right\rangle - \left\langle \underline{S(Y,z)}, \underline{\phi(\theta)}\right\rangle + pen(\theta)\,.\end{aligned}$$

Where $\underline{P(z)} := \left\{\frac{\sum_{i=1}^n \mathbb{1}_{z^{(i)}=k}}{n}\right\}_{k=1}^K \in [0,1]^K$ is a stochastic vector, $\underline{\psi(\theta)} := \{\psi(\theta_k)\}_{k=1}^K \in \mathbb{R}^K$, $\underline{S(Y,z)} := \left\{\frac{\sum_{i=1}^n \mathbb{1}_{z^{(i)}=k} S(Y^{(i)})}{n}\right\}_{k=1}^K \in \mathcal{S}^K \subseteq \mathbb{R}^{dK}$, $\underline{\phi(\theta)} := \{\phi(\theta_k)\}_{k=1}^K \in \mathbb{R}^{dK}$. Depending on the observed data, $\underline{P(z)}$ and $\underline{S(Y,z)}$ can take any value in their respective spaces. As a result, the optimisation problem (6.3) is solvable in θ if and only if, for any stochastic vector $\underline{P} \in [0,1]^K$ and real vector $\underline{S} \in \mathcal{S}^K$, the problem (6.40) has a solution $\hat{\theta}$.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; -\left\langle \underline{P}, \underline{\psi(\theta)}\right\rangle - \left\langle \underline{S}, \underline{\phi(\theta)}\right\rangle + pen(\theta)\,. \tag{6.40}$$

Now, we introduce an EM algorithm, solving the unsupervised version of (6.3) and put its Maximisation (M) step under the form of (6.40). We recall that in the unsupervised setting, the hierarchical model becomes a mixture model with class probabilities $\mathbb{P}\left(z^{(i)} = k\right) =: \pi_k$ and the following likelihood to optimise:

$$\hat{\theta}, \hat{\pi} = \underset{\theta,\pi}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k \, p_{\theta_k}\left(Y^{(i)}\right)\right) + pen(\theta, \pi)\,.$$

We consider in all generality the possibility that there could be a penalty on π in addition to the structure-inducing penalty on θ already present in the supervised problem. We recall that the

EM can be expressed as an iteration of two steps updating the current parameter $(\theta^{(t)}, \pi^{(t)})$. The Expectation (E) step:

$$p_{i,k}^{(t)} := \mathbb{P}_{\theta^{(t)},\pi^{(t)}}(z^{(i)} = k|Y^{(i)}) = \frac{p_{\theta_k^{(t)}}(Y^{(i)})\pi_k^{(t)}}{\sum_{l=1}^{K} p_{\theta_l^{(t)}}(Y^{(i)})\pi_l^{(t)}}\,. \tag{6.41}$$

And the M step:

$$\theta^{(t+1)}, \pi^{(t+1)} = \underset{\theta,\pi}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)} \left(ln\, p_{\theta_k}(Y^{(i)}) + ln\, \pi_k \right) + pen(\theta, \pi)\,. \tag{6.42}$$

Assuming once again that there is no coupling between π and θ in the penalty, i.e. $pen(\pi, \theta) = pen_\pi(\pi) + pen_\theta(\theta)$, then the M step (6.42) can be written as in (6.43).

$$\begin{aligned} \theta^{(t+1)} &= \underset{\theta}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)}\, ln\, p_{\theta_k}(Y^{(i)}) + pen_\theta(\theta)\,, \\ \pi^{(t+1)} &= \underset{\pi}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{i,k}^{(t)}\, ln\, \pi_k + pen_\pi(\pi)\,. \end{aligned} \tag{6.43}$$

The problem in π is easily solvable as long as pen_π is convex and differentiable. Using the same manipulations as on the supervised loss, the problem in θ can be re-written under the form of Eq (6.40):

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmin}} \; -\left\langle \underline{P(Y, \theta^{(t)})}, \underline{\psi(\theta)} \right\rangle - \left\langle \underline{S(Y, \theta^{(t)})}, \underline{\phi(\theta)} \right\rangle + pen(\theta)\,. \tag{6.44}$$

With the stochastic vector $\underline{P(Y, \theta^{(t)})} := \left\{ \frac{\sum_{i=1}^{n} p_{i,k}^{(t)}}{n} \right\}_{k=1}^{K} \in [0,1]^K$, and the real vector $\underline{S(Y, \theta^{(t)})} := \left\{ \frac{\sum_{i=1}^{n} p_{i,k}^{(t)} S(Y^{(i)})}{n} \right\}_{k=1}^{K} \in \mathcal{S}^K \subseteq \mathbb{R}^{dK}$. The parameter vectors $\underline{\psi(\theta)}$ and $\underline{\phi(\theta)}$ are the same as in Eq (6.40). We have already established that the supervised hierarchical problem (6.3) is solvable if and only if the problem (6.40) is solvable for all $\underline{P}$ and $\underline{S}$. As a consequence, when the supervised problem (6.3) is solvable, then the M step (6.44) is as well, and the EM algorithm can be run to completion. This concludes the proof.

Chapter 7

Conclusions and perspectives

7.1 Conclusions

The analysis of unlabelled heterogeneous population through the lens of Gaussian Graphical Models is still a fledgling field of research. Most of the currently state of the art works are supervised methods re-purposed for the unsupervised scenario. In this thesis, we set out to develop several aspects of the estimation of Gaussian Graphical Models that we considered crucial for the successful description of such unlabelled populations.

At first, we tackled the problem of model selection. These studies were made in the simple homogeneous case, but the results and conclusions have applicability to the heterogeneous case. In particular regarding the use of the Kullback–Leibler divergence between distribution to estimate how well, despite being a sparse approximation, a proposed graph can describe the multivariate relations between features. We notably highlighted how local edge-wise methods that carefully consider each edge to add to the graph are more efficient than global methods that estimate a full graph all at once when there is few data available. Hence, we introduced, for the simple case, a new algorithm that makes use of an edge-wise graph exploration procedure and a global KL divergence selection criterion. We demonstrated how this composite method is able to capitalise on the strengths of the local and global approaches and improve upon the state of the art.

Afterwards, we studied the Expectation-Maximisation algorithm that is a centrepiece of many of the blooming unsupervised GGM methods. Since graphs are often considered in high dimensional setting, the convergence of the EM algorithm is very tied to the initialisation. We proposed a stable tempered variant of the EM algorithm to navigate the high dimension non-convex likelihood functions. We proved that this approach is theoretically sound, and still benefits from the same convergence guarantees as the regular case. We demonstrated with an in-depth experimental analysis the capacities of this method, showing how it was able to escape even the most adversarial initialisations and identify the right clusters within thorny data sets. Moreover, we introduced this algorithm as part of a wider class of deterministic EM algorithms that all enjoy the convergence guarantees and can be implemented to fulfil many different roles.

Finally, we introduced a new unsupervised method with a Mixture of Conditional Gaussian Graphical Models that takes into accounts co-factors and their heterogeneous effect on the graph variables. Indeed, we argued that real data has every chance to be organised into either trivial or very ambiguous clusters and we showcased how this disrupts regular maximum likelihood approaches. Hence, we designed a model that takes into consideration the knowledge of observed co-factors, in order to ensure that the clustering procedure will recover new, non-trivial information. We demonstrated experimentally that this approach was able to correctly correct for the effect of co-factors and recover more subtle, hidden clusters. We justified the theoretical validity of our method with a convergence result, that we also extended to the tempered version of our EM algorithm. Subsequent experiments showed how the tempering was able to improve even further the performances of our method in high dimension.

In the following, we will widen the discussion with a few practical application of our work to medical problems. We will showcase a few additional GGM problematics, such as the search for long distance

correlations. In the end, we will discuss some of the future works that this PhD thesis leads to.

7.2 Recent clinical collaborations

In this section, we present some of our recent collaborations and our more applied work.

7.2.1 Ciliopathies

In the following, we present our contribution in collaboration with the *Imagine Institute* to the C'IL-LICO project dedicated to next generation medicine for renal ciliopathies.
We were asked to propose visual interpretation of their data. The database follows 130 medical concepts and is constituted of 75 ciliopathy patients, 30 “”Differential Diagnosis” (DD) patients who have similar renal disease that are not ciliopathies, and 30 control, healthy, patients. In Figures 7.1, 7.2 and 7.3, we display the conditional correlation graph obtained for each of these three groups with a supervised joint-GGM model, using the Group Graphical LASSO penalty for structure. The graphical conventions are the same as Chapter 6, red and blue edges represent positive and negative correlations respectively, and the edge width is proportional to the eight weight. The three structure are indeed somewhat different, with the ciliopathy patients - who are the more numerous - offering the most sensible graph. Most of the correlations however are completely obvious, if not semantically mandatory. All the different type of cyst are related, so are the dysplasia, deafness is related to hear loss, blindness to vision loss... This is to be expected with so many redundancies among the features. Our graphs only capture the most basic, trivial correlations. We would prefer them to capture the information of the less obvious correlations specific to each sub-population. Thankfully, we have at our disposal a prior correlation matrix Σ_0 established by the Institute that describes all the prior correlation between all the concepts. The corresponding conditional correlation graph can be seen on Figure 7.4. It is not sparse, hence the weaker edges were removed from the display. This is a vision of overall “average” correlations, for any random individual of the world population. Most of the obvious correlations are indeed present in this graph. We can remove this baseline correlation structure from the data by replacing Y by $\Sigma_0^{-\frac{1}{2}}Y$. An operation which effectively takes for each feature the residual of its linear prediction from the values of the others. With the new residualised data, we make the same supervised joint-GGM estimation on the three sub-populations. The resulting graphs are depicted in Figures 7.5, 7.6 and 7.7. We see that the obvious correlations of Figure 7.4 have been removed, and there only remains “long distance” correlations.

We also estimate an unsupervised Mixture of GGM with a regularised EM algorithm to find out which clusters are naturally identified by the method without knowledge of the labels. The three estimated graphs can be seen on Figures 7.8, 7.9 and 7.10. Interestingly, the three graphs are very distinct. In terms of sub-populations, the Ciliopathy patients are spread cross each clusters, with cluster 2 having the most. All the DD patients are in cluster 2, which is the largest of the three and contains a mix of all patient types. Cluster 3 has a similar constitution to cluster 1, but with more control patients. From an interpretation standpoint, since the unsupervised GGM was estimated from the raw data, the recovered correlations are of the obvious, trivial kind. In Figures 7.11, 7.12 and 7.13 we represent the result of the same method ran on the residualised data. We notice that the patient distribution by cluster is curiously very similar to the case with the raw data. The ciloptahies are spread out, cluster 2 contains the most. The DD patients are once again only found in cluster 2. And the cluster with the largest number of controls is cluster 3. The estimated correlations are “long distance” as expected, with all clusters being quite different. In particular, cluster 2, which is the largest and most diverse one, has a graph with many more strong correlations than any other. Whereas the two other clusters, who are penalised with the same intensity, have sparse graphs. This could be due to a high number of somewhat dissimilar individuals within this cluster that results in an empirical precision matrix with higher values that get less affected by the penalty.

These results were presented at the April 2020 RHU3 scientific meeting and are currently under scrutiny of the C'IL-LICO team members in order to provide a more versed interpretation.

7.2.2 Cushing's syndrome

In this section, we give an overview of our work with the *Cochin Institute* on Cushing's syndrome. In the following, you can find an abstract by Roberta Armignacco, submitted to the 37th congress of the French Endocrinology Society (SFE 2020). In this work, we contributed to the establishment of a proper statistical methodology to demonstrate the clinical interest of Genome methylation for the diagnosis of Cushing's syndrome.

Identification of a molecular signature for hypercortisolism by whole blood methylome analysis
Armignacco R, Lartigue T, Jouinot A, Septier S, Neou M, Gaspar C, Perlemoine K, Bouys L, Braun L, Riester A, Allassonnière S, Zennaro MC, Reincke M, Bertherat J, Beuschlein F, Assié G

The diagnosis of hypercortisolism is based on hormone assays. However, these assays do not reflect the individual risk for each complication of hypercortisolism, as inter-individual susceptibility varies, particularly in sub-clinical forms.
The objective of this work is to identify biomarkers reflecting individual glucocorticoid impregnation from the whole blood methylome. Methods: 47 patients with Cushing's disease, separated into two cohorts: training (clear cases, n=24) and validation (borderline cases, n=23). For each, a pair of blood samples were taken: before and > 3 months after correction of hypercortisolism. The generation of the whole blood methylmeter is made by Illumina chip (850K probes).
The results are the following: in the training cohort, 28% genome methylation varies according to cortisolic status. The neutrophil count inferred from the methylome is highly correlated with the CBC (r=0.81). This variation in CBC is, as expected, associated with cortisolic status. To predict the Cushing/non Cushing status from the methyloma, a penalized regression approach based solely on the part not predicted by the CBC identifies a combination of a few genome locations (accuracies of 1 and 0.8 on training and validation cohorts respectively), which improves the prediction on the validation cohort compared to the CBC alone (accuracy=0.65).
We conclude that genome methylation is a biomarker of hypercortisolism. Several questions are pending, such as the possibility of optimizing simple techniques for targeted methylation measurements, the performance of this biomarker in hypercortisolism at minimal levels, and the links of the methylation profile with the complications of hypercortisolism.

7.3 Perspectives

7.3.1 Neurology

Although the work in this thesis is mostly methodological, I supported each of our contributions with applications on neurological data, in particular Alzheimer's Disease data. Each of them gave rise to observations that mostly matched, but sometimes challenged the current knowledge in the field. From a clinical standpoint, these observations need to be verified and reproduced on new independent datasets. Then, it is necessary to bring in the perspective of medical experts such as neurolgist and neuroradiologist in order to properly validate and interpret each result. Only then can we draw conclusions that contribute to the medical knowledge pool and begin to set up clinical applications.

7.3.2 Statistical theory and methodology

In Chapter 3, I brought to light that edge-wise graph exploration could be more data-efficient than global, self-contained, optimisation problems when it comes to estimate graphs. Despite that, many of the unsupervised joint-GGM approaches propose Maximum Likelihood Estimators, including our own method in Chapter 6. In the future, it will be interesting to design edge-wise methods for the unsupervised hierarchical case, in particular within the low sample size setting.
The EM algorithm I proposed in Chapter 6 to optimise a regularised Mixture likelihood relies on

our ability to optimise the corresponding supervised regularised Hierarchical density. In this, our method is alike the previous unsupervised methods that use the EM algorithm to estimate Mixtures of GGM. This observation however, has wider implications. At the end of Chapter6, I discuss the fact that for any Hierarchical density that belongs to the exponential family, if the penalised MLE is tractable, then the M step of the EM algorithm that optimises the corresponding Mixture likelihood is tractable as well. Since these are finite Mixtures, the E step will also always be tractable. In short, any supervised penalised MLE built from a density within the exponential family can be adapted to the unsupervised case with an EM algorithm. This means that the extensive catalogue of Maximum Likelihood Estimators for Hierarchical GGM can be ported to the unsupervised case. This include any alternative model, such as the Matrix Normal Graphical Model, that also belongs to the exponential family. This also includes the alternative frameworks such as the differential networks, the censored data and so on...
I have laid the theoretical groundwork to propose a very large extension of many of the methods designed for the labelled heterogeneous populations to the unsupervised case. This will be explored in future works.

Acknowledgments

The research leading to these results has received funding from the European Research Council (ERC) under grant agreement No 678304, European Union's Horizon 2020 research and innovation program under grant agreement No 666992 (EuroPOND) and No 826421 (TVB-Cloud), and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (IHU-A-ICM). I would like to thank Pascal Houillier for his insightful comments on the nephrological experiments.

Figure 7.1: Supervised hierarchical GGM. Ciliopathies.

Figure 7.2: Supervised hierarchical GGM. Differential Diagnosis.

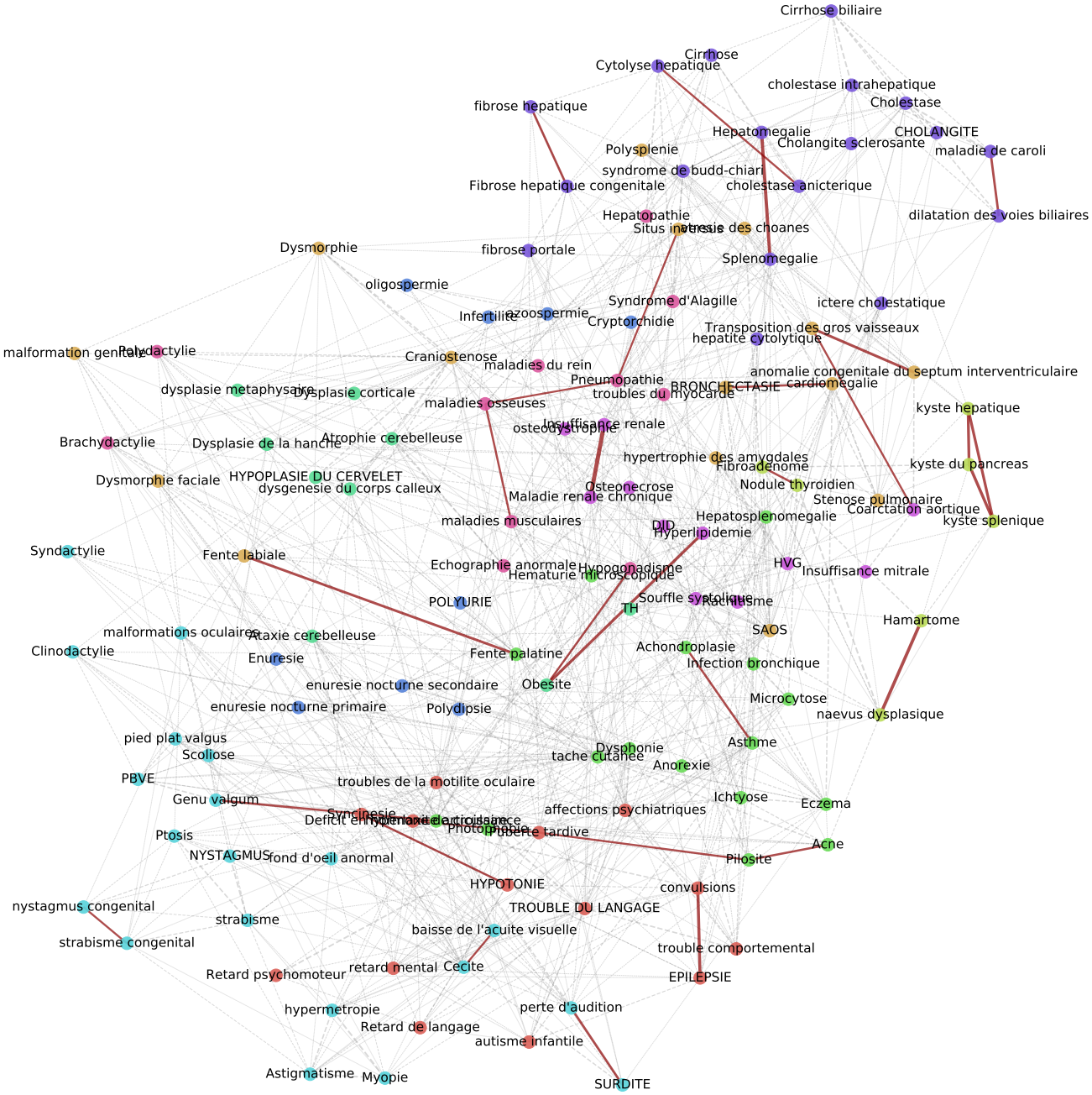


Figure 7.3: Supervised hierarchical GGM. Controls.

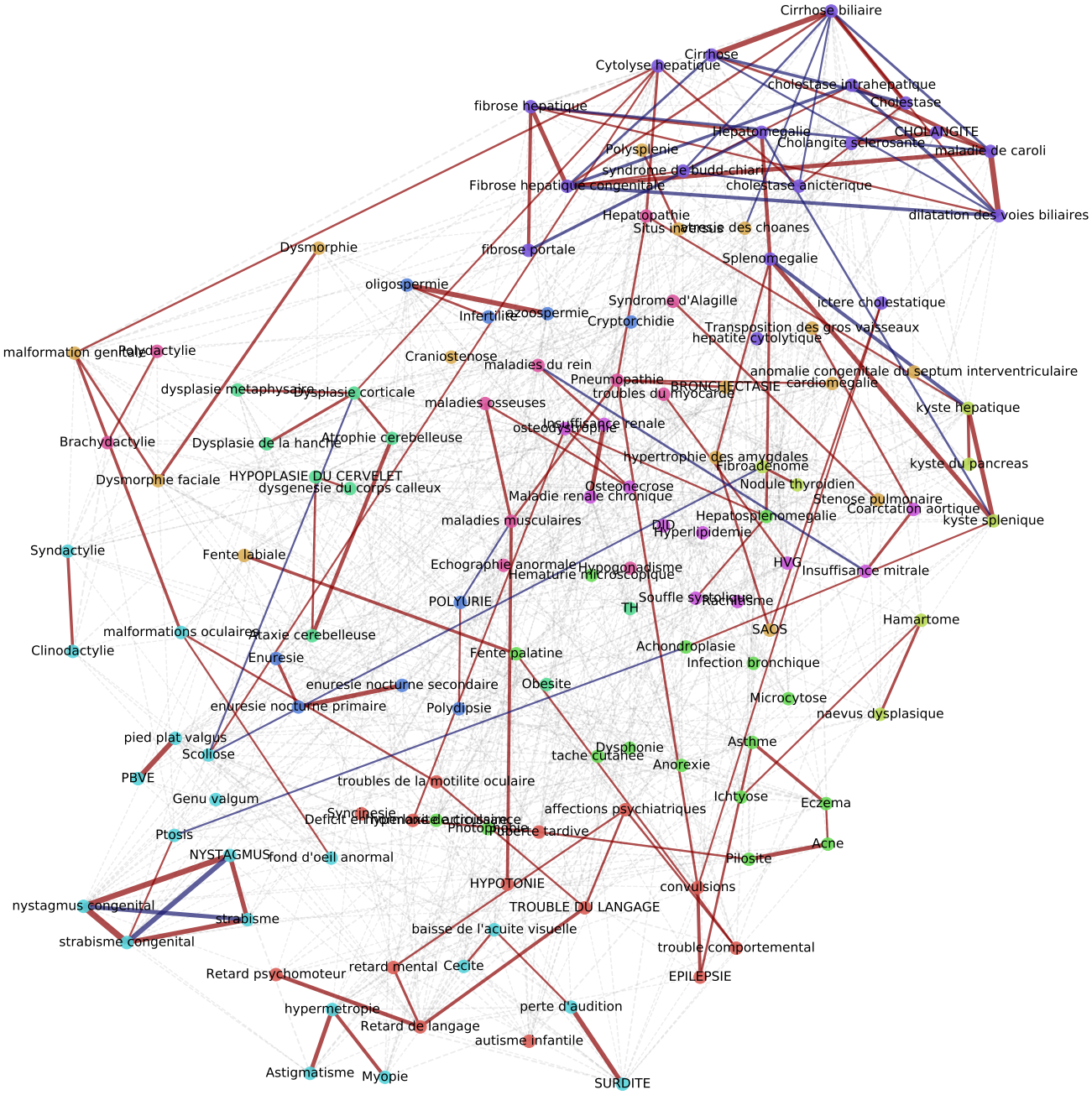


Figure 7.4: Graph prior.

Figure 7.5: Supervised hierarchical GGM. Ciliopathies minus prior.

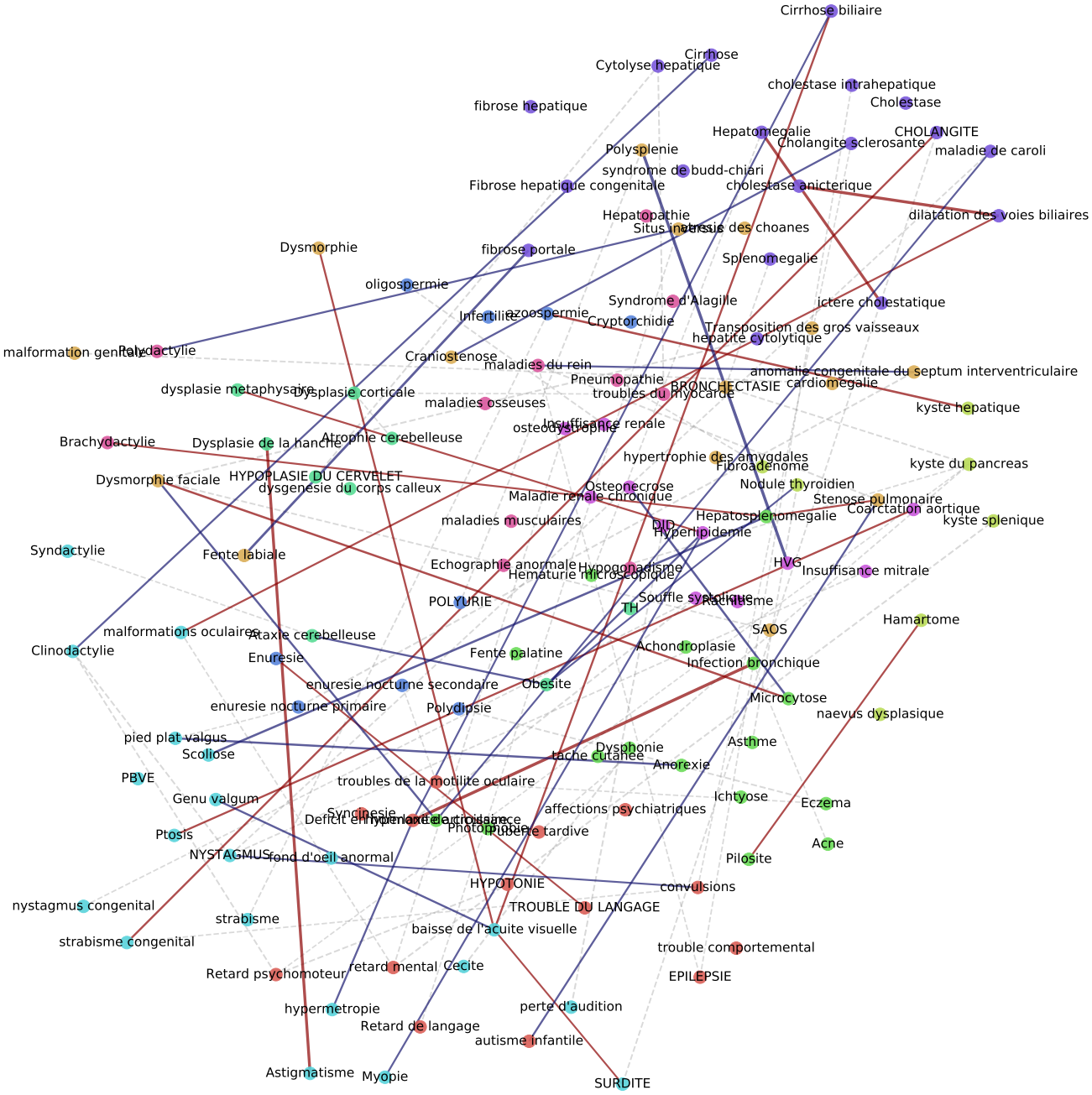


Figure 7.6: Supervised hierarchical GGM. Differential Diagnosis minus prior.

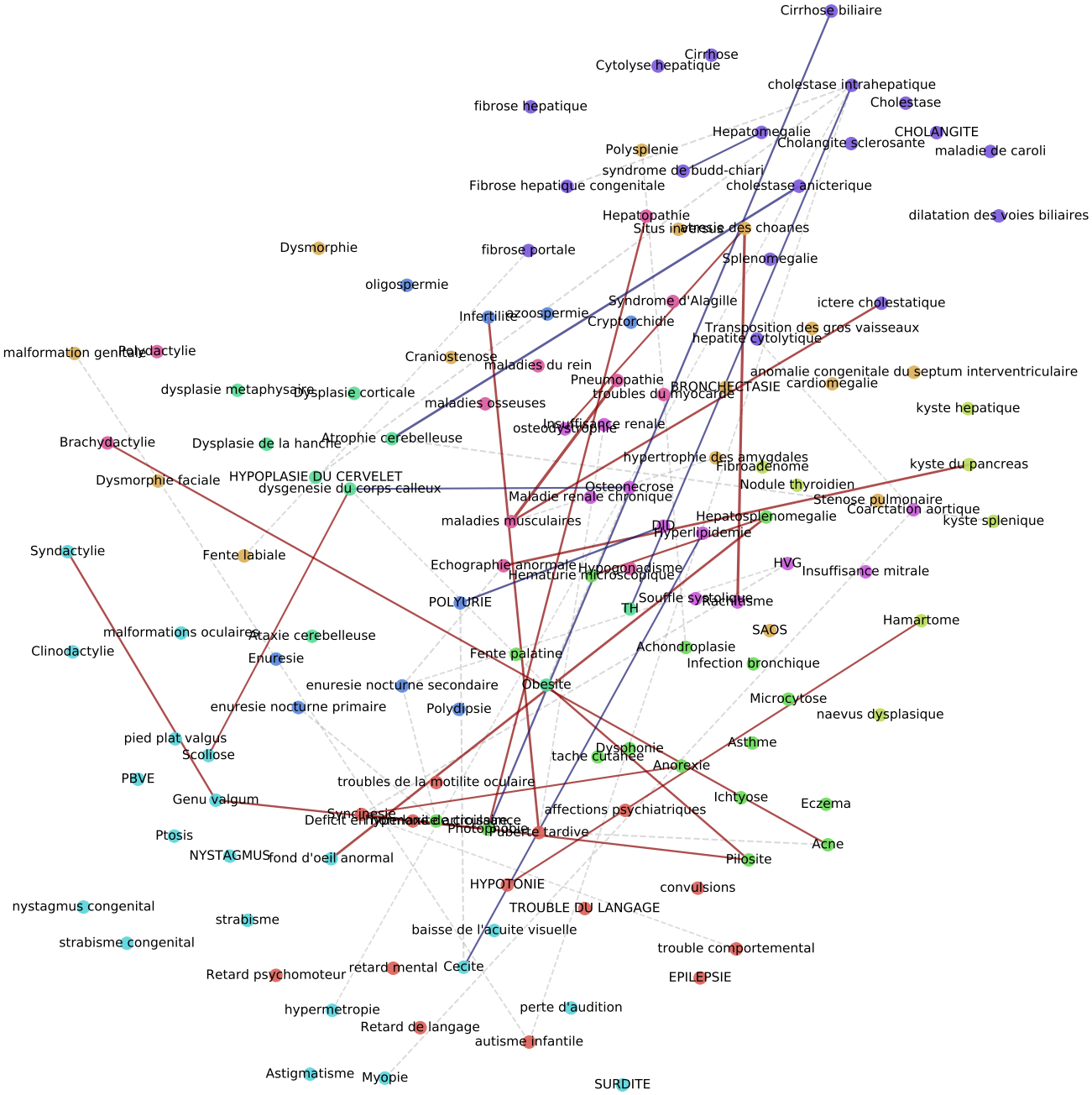


Figure 7.7: Supervised hierarchical GGM. Controls minus prior.

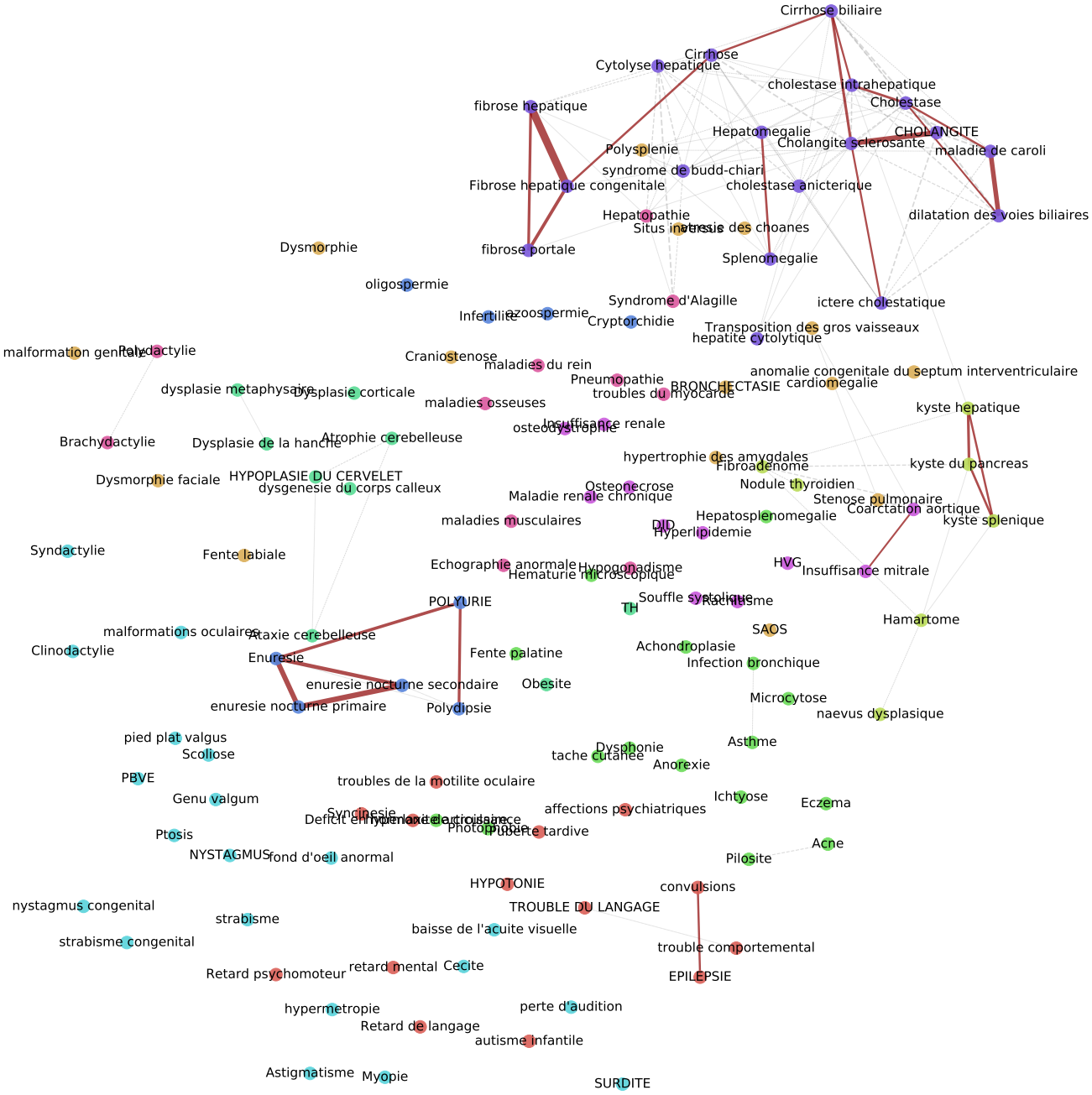


Figure 7.8: Unsupervised identified cluster 1. Contains 24 Ciliopathies and 7 Controls.

Figure 7.9: Unsupervised identified cluster 2. Contains 31 Ciliopathies, 30 Differential Diagnosis and 9 Controls.

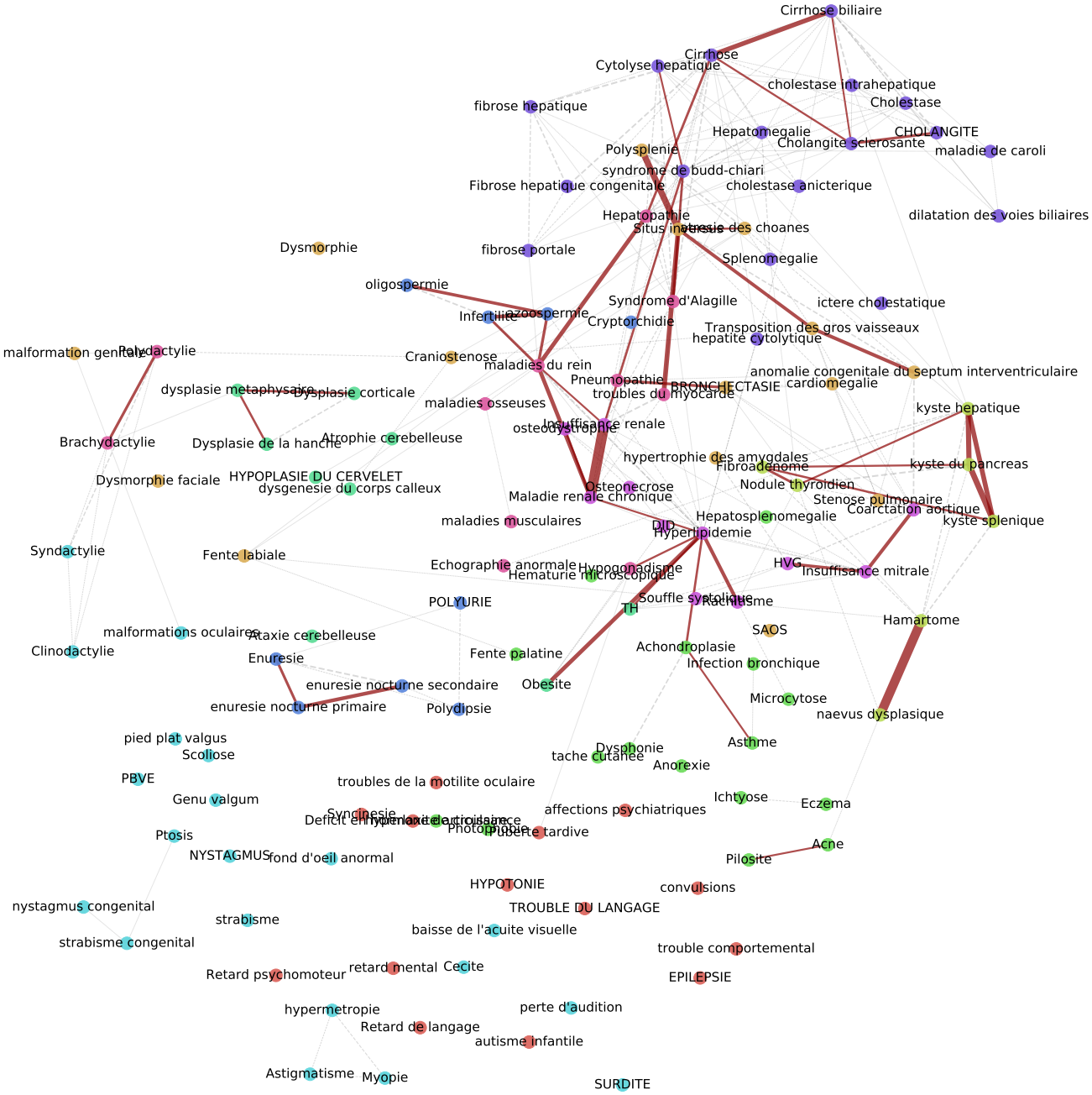


Figure 7.10: Unsupervised identified cluster 3. Contains 20 Ciliopathies and 14 Controls.

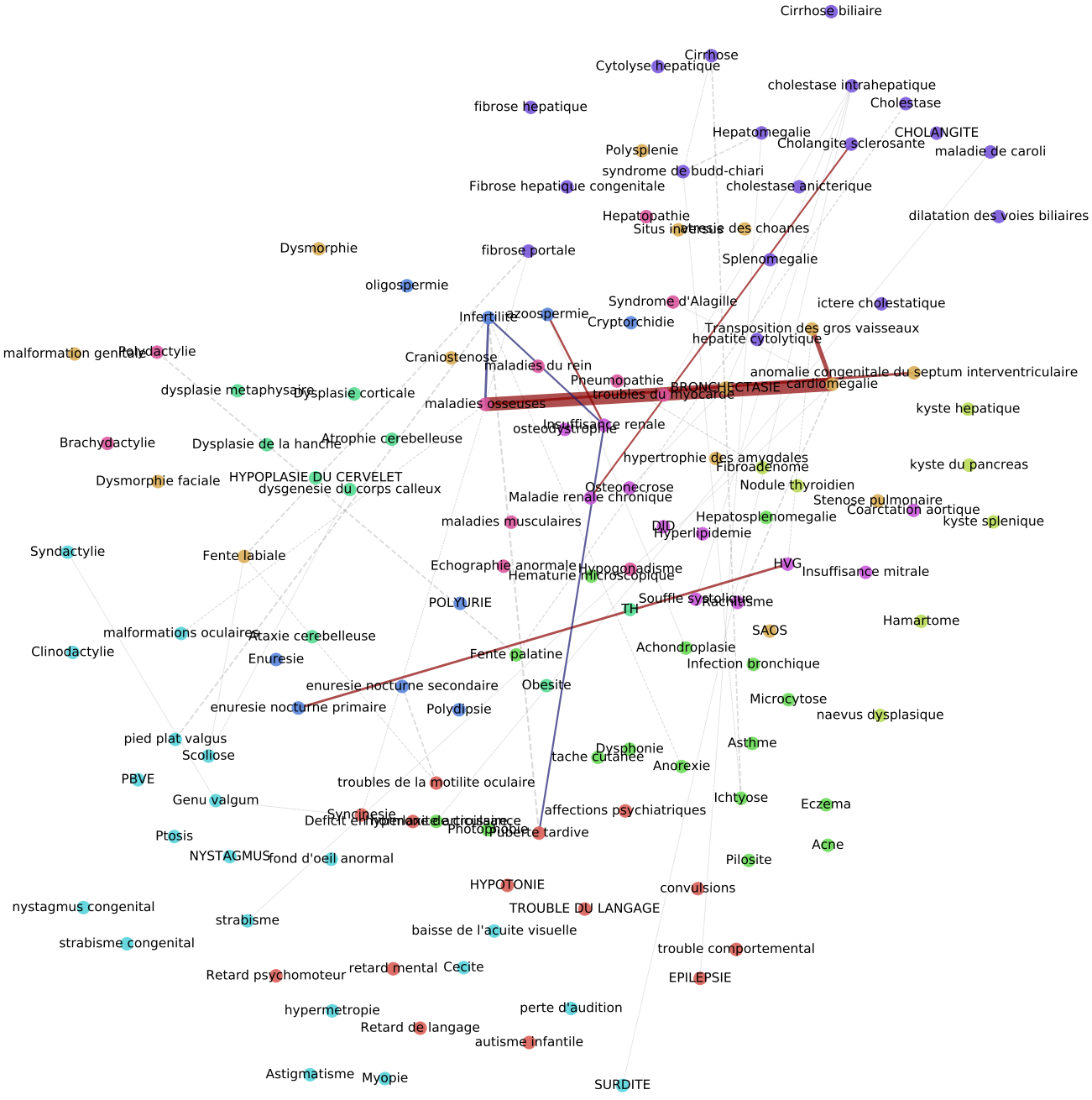


Figure 7.11: Unsupervised identified cluster 1 with residualised data. Contains 19 Ciliopathies and 6 Controls.

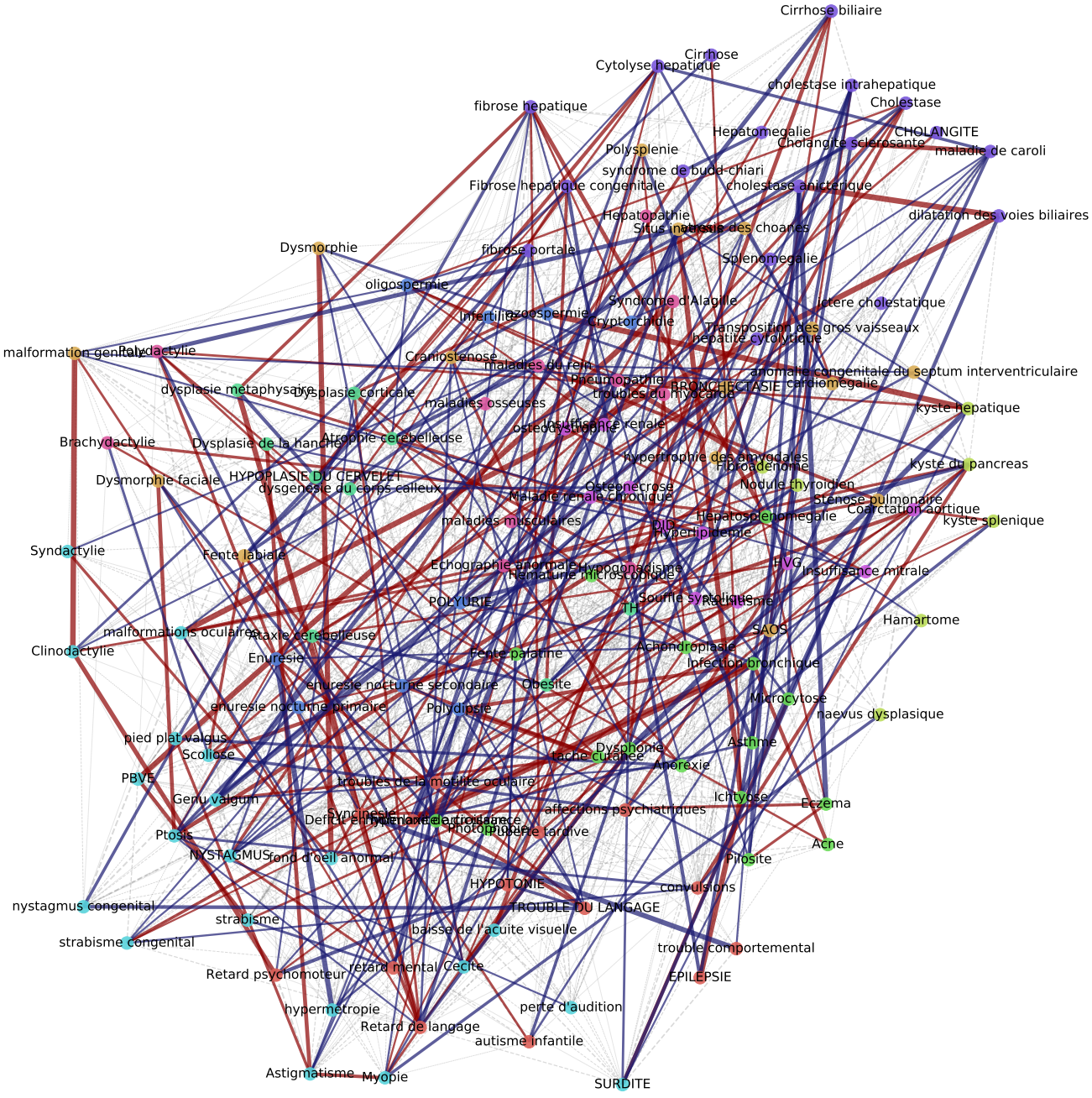


Figure 7.12: Unsupervised identified cluster 2 with residualised data. Contains 31 Ciliopathies, 30 Differential Diagnosis and 9 Controls.

Figure 7.13: Unsupervised identified cluster 3 with residualised data. Contains 25 Ciliopathies and 15 Controls.

Bibliography

[1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines.* New York, NY; John Wiley and Sons Inc., 1988.

[2] S. Allassonnière and J. Chevallier. A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling. working paper or preprint, Jun 2019.

[3] S. Allassonniere, P. Jolivet, and C. Giraud. Detecting long distance conditional correlations between anatomical regions using gaussian graphical models. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 111–122, 2011.

[4] S. Allassonnière, E. Kuhn, A. Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.

[5] S. Allassonniere, L. Younes, et al. A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160, 2012.

[6] G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764, 2010.

[7] C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.

[8] K. Ashurbekova, S. Achard, and F. Forbes. Robust structure learning using multivariate T-distributions. In *50e Journées de Statistique de la SFdS (JdS'2018)*, pages 1–6, Saclay, France, May 2018.

[9] H. Attias. Independent factor analysis. *Neural computation*, 11(4):803–851, 1999.

[10] L. Augugliaro, A. Abbruzzo, and V. Vinciotti. L1-penalized censored gaussian graphical model. *arXiv preprint arXiv:1801.07981*, 2018.

[11] L. Augugliaro, G. Sottile, and V. Vinciotti. The conditional censored graphical lasso estimator. *arXiv preprint arXiv:1910.12775*, 2019.

[12] O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd international conference on Machine learning*, pages 89–96. ACM, 2006.

[13] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[14] A. Berry, P. Heggernes, and Y. Villanger. A vertex incremental approach for maintaining chordality. *Discrete Mathematics*, 306(3):318–336, 2006.

[15] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

[16] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):47–50, 1983.

[17] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186, 2009.

[18] T. Cai, W. Liu, and X. Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[19] T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2012.

[20] T. T. Cai, H. Li, W. Liu, and J. Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445, 2016.

[21] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

[22] M. Chen, Z. Ren, H. Zhao, and H. Zhou. Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *Journal of the American Statistical Association*, 111(513):394–406, 2016.

[23] X. Chen and W. Liu. Graph estimation for matrix-variate gaussian data. *arXiv preprint arXiv:1509.05453*, 2015.

[24] Z. Chen and C. Leng. Dynamic covariance models. *Journal of the American Statistical Association*, 111(515):1196–1207, 2016.

[25] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

[26] J. Chiquet, T. Mary-Huard, and S. Robin. Structured regularization for conditional gaussian graphical models. *Statistics and Computing*, 27(3):789–804, 2017.

[27] H. Chun, M. Chen, B. Li, and H. Zhao. Joint conditional gaussian graphical models with multiple sources of genomic data. *Frontiers in genetics*, 4:294, 2013.

[28] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[29] J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software*, 23(4):501–520, 2008.

[30] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[31] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

[32] A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

[33] B. Delyon, M. Lavielle, E. Moulines, et al. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[34] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[35] A. P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[37] W. S. DeSarbo and W. L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2):249–282, 1988.

[38] A. Deshpande, M. Garofalakis, and M. I. Jordan. Efficient stepwise selection in decomposable models. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 128–135. Morgan Kaufmann Publishers Inc., 2001.

[39] S. K. Deshpande, V. Ročková, and E. I. George. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, pages 1–11, 2019.

[40] Q. T. Dinh, A. Kyrillidis, and V. Cevher. A proximal newton framework for composite minimization: Graph learning without cholesky decompositions and matrix inversions. In *International Conference on Machine Learning*, pages 271–279, 2013.

[41] M. Drton and M. D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

[42] D. Dua and C. Graff. UCI machine learning repository, 2017.

[43] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. *arXiv preprint arXiv:1206.3249*, 2012.

[44] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[45] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521, 2009.

[46] X. Fan, K. Fang, S. Ma, and Q. Zhang. Integrating approximate single factor graphical models. *Statistics in medicine*, 39(2):146–155, 2020.

[47] M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics*, pages 1057–1080, 2011.

[48] M. Fop, T. B. Murphy, and L. Scrucca. Model-based clustering with sparse covariance matrices. *arXiv preprint arXiv:1711.07748*, 2017.

[49] G. Fort, E. Moulines, et al. Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259, 2003.

[50] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[51] L. Gan, N. N. Narisetty, and F. Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114(527):1218–1231, 2019.

[52] C. Gao, Y. Zhu, X. Shen, and W. Pan. Estimation of multiple networks in gaussian mixture models. *Electronic journal of statistics*, 10:1133, 2016.

[53] C. J. Geyer and E. A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[54] A. J. Gibberd and J. D. Nelson. Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634, 2017.

[55] C. Giraud et al. Estimation of gaussian graphs by model selection. *Electronic Journal of Statistics*, 2:542–563, 2008.

[56] C. Giraud, S. Huet, and N. Verzelen. Graph selection with GGMselect. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

[57] P. Giudici and P. Green. Decomposable graphical gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[58] M. Grechkin, M. Fazel, D. Witten, and S.-I. Lee. Pathway graphical lasso. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[59] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[60] A. Gupta and D. Nagar. *Matrix Variate Distributions*, volume 104. CRC Press, 1999.

[61] B. Hao, W. W. Sun, Y. Liu, and G. Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *The Journal of Machine Learning Research*, 18(1):7981–8038, 2017.

[62] S. Hara and T. Washio. Learning a common substructure of multiple graphical gaussian models. *Neural Networks*, 38:23–38, 2013.

[63] Y. He, X. Zhang, J. Ji, and B. Liu. Joint estimation of multiple high-dimensional gaussian copula graphical models. *Australian & New Zealand Journal of Statistics*, 59(3):289–310, 2017.

[64] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans. Enhancement of mr images using registration for signal averaging. *Journal of computer assisted tomography*, 22(2):324–333, 1998.

[65] J. Honorio and D. Samaras. Multi-task learning of gaussian graphical models. In *ICML*, pages 447–454. Citeseer, 2010.

[66] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in neural information processing systems*, pages 2330–2338, 2011.

[67] F. Huang and S. Chen. Joint learning of multiple sparse matrix gaussian graphical models. *IEEE Trans. Neural Netw. Learning Syst.*, 26(11):2606–2620, 2015.

[68] F. Huang and S. Chen. Learning dynamic conditional gaussian graphical models. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):703–716, 2018.

[69] F. Huang, S. Chen, and S.-J. Huang. Joint estimation of multiple conditional gaussian graphical models. *IEEE transactions on neural networks and learning systems*, 29(7):3034–3046, 2018.

[70] J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[71] K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[72] L. Ibarra. Fully dynamic algorithms for chordal graphs and split graphs. *ACM Transactions on Algorithms (TALG)*, 4(4):40, 2008.

[73] W. Jank. Quasi-monte carlo sampling to improve the efficiency of monte carlo em. *Computational statistics & data analysis*, 48(4):685–701, 2005.

[74] J. Janková and S. van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

[75] J. Jankova and S. van de Geer. Inference in high-dimensional graphical models. *arXiv preprint arXiv:1801.08512*, 2018.

[76] J. Jankova, S. Van De Geer, et al. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[77] K. Jedidi, V. Ramaswamy, W. S. DeSarbo, and M. Wedel. On estimating finite mixtures of multivariate regression and simultaneous equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(3):266–289, 1996.

[78] J. Ji, Y. He, and L. Xie. Dynamic brain connectivity alternation detection via matrix-variate differential network model. *bioRxiv*, page 446237, 2018.

[79] B. Jiang, X. Wang, and C. Leng. A direct approach for sparse quadratic discriminant analysis. *The Journal of Machine Learning Research*, 19(1):1098–1134, 2018.

[80] A. Kalaitzis, J. Lafferty, N. Lawrence, and S. Zhou. The bigraphical lasso. In *International Conference on Machine Learning*, pages 1229–1237, 2013.

[81] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

[82] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[83] M. Kolar and E. Xing. On time varying undirected graphs. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 407–415, 2011.

[84] M. Kolar and E. P. Xing. Estimating networks with jumps. *Electronic journal of statistics*, 6:2069, 2012.

[85] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *arXiv preprint arXiv:1405.2468*, 2014.

[86] A. Krishnamurthy. High-dimensional clustering with sparse gaussian mixture models. *Unpublished paper*, pages 191–192, 2011.

[87] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038, 2005.

[88] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[89] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[90] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):425–437, 1995.

[91] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[92] W. Lee and Y. Liu. Joint estimation of multiple precision matrices with common structures. *The Journal of Machine Learning Research*, 16(1):1035–1062, 2015.

[93] C. Leng and C. Y. Tang. Sparse matrix graphical models. *Journal of the American Statistical Association*, 107(499):1187–1200, 2012.

[94] A. Lenkoski and A. Dobra. Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157, 2011.

[95] E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

[96] R. A. Levine and G. Casella. Implementations of the monte carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

[97] R. A. Levine and J. Fan. An automated (markov chain) monte carlo EM algorithm. *Journal of Statistical Computation and Simulation*, 74(5):349–360, 2004.

[98] B. Li, H. Chun, and H. Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107(497):152–167, 2012.

[99] L. Li and K.-C. Toh. An inexact interior point method for l_1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315, 2010.

[100] Y. Li, B. A. Craig, and A. Bhadra. The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757, 2019.

[101] Z. R. Li and T. H. McCormick. An expectation conditional maximization approach for gaussian graphical models. *Journal of Computational and Graphical Statistics*, pages 1–11, 2019.

[102] Z. R. Li, T. H. McCormick, and S. J. Clark. Bayesian joint spike-and-slab graphical lasso. *arXiv preprint arXiv:1805.07051*, 2018.

[103] Z. Lin, T. Wang, C. Yang, and H. Zhao. On joint estimation of gaussian graphical models for spatial and temporal data. *Biometrics*, 73(3):769–779, 2017.

[104] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[105] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

[106] H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1):241–294, 2017.

[107] W. Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

[108] W. Liu and X. Luo. High-dimensional sparse precision matrix estimation via sparse column inverse operator. *Preprint. Available at*, 2012.

[109] Y. Liu and A. Willsky. Learning gaussian graphical models with observed or latent fvss. In *Advances in Neural Information Processing Systems*, pages 1833–1841, 2013.

[110] Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

[111] J. Ma and G. Michailidis. Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, 17(1):5777–5824, 2016.

[112] S. Ma, L. Xue, and H. Zou. Alternating direction methods for latent variable gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

[113] H. P. Maretic and P. Frossard. Graph laplacian mixture model. *arXiv preprint arXiv:1810.10053*, 2018.

[114] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[115] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[116] Z. Meng, B. Eriksson, and A. Hero. Learning latent variable gaussian graphical models. In *International Conference on Machine Learning*, pages 1269–1277, 2014.

[117] M. I. Miller, C. E. Priebe, A. Qiu, B. Fischl, A. Kolasny, T. Brown, Y. Park, J. T. Ratnanather, E. Busa, J. Jovicich, et al. Collaborative computational anatomy: an mri morphometry study of the human brain via diffeomorphic metric mapping. *Human brain mapping*, 30(7):2132–2141, 2009.

[118] K. Mohan, M. Chung, S. Han, D. Witten, S.-I. Lee, and M. Fazel. Structured learning of gaussian graphical models. In *Advances in neural information processing systems*, pages 620–628, 2012.

[119] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

[120] I. Naim and D. Gildea. Convergence of the EM algorithm for gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427*, 2012.

[121] Y. Ni, P. Müller, Y. Zhu, and Y. Ji. Heterogeneous reciprocal graphical models. *Biometrics*, 74(2):606–615, 2018.

[122] Y. Ning and H. Liu. High-dimensional semiparametric bigraphical models. *Biometrika*, 100(3):655–670, 2013.

[123] L. Ou-Yang, X.-F. Zhang, X. Hu, and H. Yan. Differential network analysis via weighted fused conditional gaussian graphical model. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[124] L. Ou-Yang, X.-F. Zhang, X.-M. Zhao, D. D. Wang, F. L. Wang, B. Lei, and H. Yan. Joint learning of multiple differential networks with latent variables. *IEEE transactions on cybernetics*, PP(99):1–13, 2018.

[125] F. Oztoprak, J. Nocedal, S. Rennie, and P. A. Olsen. Newton-like methods for sparse inverse covariance estimation. In *Advances in neural information processing systems*, pages 755–763, 2012.

[126] J.-X. Pan and R. Thompson. Quasi-monte carlo EM algorithm for mles in generalized linear mixed models. In *COMPSTAT*, pages 419–424. Springer, 1998.

[127] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[128] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[129] K. Perrakis, F. Dondelinger, and S. Mukherjee. Latent group structure and regularized regression. *arXiv preprint arXiv:1908.07869*, 2019.

[130] C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[131] C. Peterson, M. Vannucci, C. Karakas, W. Choi, L. Ma, and M. MALETIĆ-SAVATIĆ. Inferring metabolic networks using the bayesian adaptive graphical lasso with informative priors. *Statistics and its Interface*, 6(4):547, 2013.

[132] H. Qiu, F. Han, H. Liu, and B. Caffo. Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):487–504, 2016.

[133] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, et al. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[134] Z. Ren, Y. Kang, Y. Fan, and J. Lv. Tuning-free heterogeneity pursuit in massive networks. *USC-INET Research Paper No 16-27*, 2016.

[135] Z. Ren, T. Sun, C.-H. Zhang, H. H. Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[136] G. V. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*, 2008.

[137] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[138] T. Saegusa and A. Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic journal of statistics*, 10(1):1341, 2016.

[139] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in neural information processing systems*, pages 2101–2109, 2010.

[140] K. Scheinberg and I. Rish. Learning sparse gaussian markov networks using a greedy coordinate ascent approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 196–212. Springer, 2010.

[141] J.-B. Schiratti, S. Allassonniere, A. Routier, O. Colliot, S. Durrleman, A. D. N. Initiative, et al. A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In *International Conference on Information Processing in Medical Imaging*, pages 564–575. Springer, 2015.

[142] E. Shaddox, C. B. Peterson, F. C. Stingo, N. A. Hanania, C. Cruickshank-Quinn, K. Kechris, R. Bowler, and M. Vannucci. Bayesian inference of networks across multiple sample groups and data types. *arXiv preprint arXiv:1909.02058*, 2019.

[143] K.-A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Artificial Intelligence and Statistics*, pages 1081–1089, 2012.

[144] G. Song, L. Han, and K. Xie. Overlapping decomposition for gaussian graphical modeling. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2217–2230, 2015.

[145] L. Song, M. Kolar, and E. P. Xing. Time-varying dynamic bayesian networks. In *Advances in neural information processing systems*, pages 1732–1740, 2009.

[146] T. P. Speed and H. T. Kiiveri. Gaussian markov distributions over finite graphs. *The Annals of Statistics*, pages 138–150, 1986.

[147] N. Städler and P. Bühlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235, 2012.

[148] T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

[149] R. H. Swendsen and J.-S. Wang. Replica monte carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

[150] R. Talluri, V. Baladandayuthapani, and B. K. Mallick. Bayesian sparse graphical models and their mixtures. *Stat*, 3(1):109–125, 2014.

[151] L. S. Tan, A. Jasra, M. De Iorio, and T. M. Ebbels. Bayesian inference for multiple gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11(4):2222–2251, 2017.

[152] Z. Tang, Z. Yu, and C. Wang. A fast iterative algorithm for high-dimensional differential network. *arXiv preprint arXiv:1901.07150*, 2019.

[153] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[154] J.-D. Tournier, F. Calamante, and A. Connelly. Mrtrix: diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, 22(1):53–66, 2012.

[155] T. Tsiligkaridis, A. O. Hero III, and S. Zhou. Convergence properties of kronecker graphical lasso algorithms. *arXiv preprint arXiv:1204.0585*, 2012.

[156] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural networks*, 11(2):271–282, 1998.

[157] C. Uhler. Gaussian graphical models: An algebraic and geometric perspective. *arXiv preprint arXiv:1707.04345*, 2017.

[158] C. Uhler et al. Geometry of maximum likelihood estimation in gaussian graphical models. *The Annals of Statistics*, 40(1):238–261, 2012.

[159] P. J. Van Laarhoven and E. H. Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[160] G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in neural information processing systems*, pages 2334–2342, 2010.

[161] C. Wang, D. Sun, and K.-C. Toh. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010.

[162] H. Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

[163] H. Wang et al. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

[164] J. Wang. Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, pages 831–851, 2015.

[165] J. Wang and M. Kolar. Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638*, 2014.

[166] G. C. Wei and M. A. Tanner. A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

[167] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):R92, 2004.

[168] A. Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

[169] C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1):95–103, 1983.

[170] N. Wu, J. Huang, X.-F. Zhang, L. Ou-Yang, S. He, Z. Zhu, and W. Xie. Weighted fused pathway graphical lasso for joint estimation of multiple gene networks. *Frontiers in genetics*, 10:623, 2019.

[171] M. Wytock and Z. Kolter. Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International conference on machine learning*, pages 1265–1273, 2013.

[172] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015.

[173] J. Yin and H. Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics*, 5(4):2630, 2011.

[174] J. Yin and H. Li. Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis*, 107:119–140, 2012.

[175] J. Yin and H. Li. Adjusting for high-dimensional covariates in sparse precision matrix estimation by $l1$-penalization. *Journal of multivariate analysis*, 116:365–381, 2013.

[176] H. Yuan, R. Xi, C. Chen, and M. Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

[177] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

[178] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[179] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[180] X. Yuan. Alternating direction methods for sparse covariance selection. *preprint*, 2(1), 2009.

[181] X.-T. Yuan and T. Zhang. Partial gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687, 2014.

[182] B. Zhang and Y. Wang. Learning structural changes of gaussian graphical models in controlled experiments. *arXiv preprint arXiv:1203.3532*, 2012.

[183] X.-F. Zhang, L. Ou-Yang, T. Yan, X. T. Hu, and H. Yan. A joint graphical model for inferring gene networks across multiple subpopulations and data types. *IEEE Transactions on Cybernetics*, 2019.

[184] Y. Zhang and J. G. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems*, pages 2550–2558, 2010.

[185] S. D. Zhao, T. T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

[186] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

[187] H. Zhou, W. Pan, and X. Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics*, 3:1473, 2009.

[188] S. Zhou et al. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

[189] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2):295–319, 2010.

[190] S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12(Oct):2975–3026, 2011.

[191] Y. Zhu and L. Li. Multiple matrix gaussian graphs estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

[192] Y. Zhu, X. Shen, and W. Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Mélanges de modèles graphiques gaussiens sous contraintes

Mots clés : Modèles graphiques, corrélations conditionnelles, estimation non-supervisée, algorithme EM

Résumé : La description des co-variations entre plusieurs variables aléatoires observées est un problème délicat. Les réseaux de dépendance sont des outils populaires qui décrivent les relations entre les variables par la présence ou l'absence d'arêtes entre les nœuds d'un graphe. En particulier, les graphes de corrélations conditionnelles sont utilisés pour représenter les corrélations "directes" entre les nœuds du graphe. Ils sont souvent étudiés sous l'hypothèse gaussienne et sont donc appelés "modèles graphiques gaussiens" (GGM).
Un seul réseau peut être utilisé pour représenter les tendances globales identifiées dans un échantillon de données. Toutefois, lorsque les données observées sont échantillonnées à partir d'une population hétérogène, il existe alors différentes sous-populations qui doivent toutes être décrites par leurs propres graphes. De plus, si les labels des sous-populations (ou "classes") ne sont pas disponibles, des approches non supervisées doivent être mises en œuvre afin d'identifier correctement les classes et de décrire chacune d'entre elles avec son propre graphe.
Dans ce travail, nous abordons le problème relativement nouveau de l'estimation hiérarchique des GGM pour des populations hétérogènes non labellisées. Nous explorons plusieurs axes clés pour améliorer l'estimation des paramètres du modèle ainsi que l'identification non supervisée des sous-populations. Notre objectif est de s'assurer que les graphes de corrélations conditionnelles inférés sont aussi pertinents et interprétables que possible.
Premièrement - dans le cas d'une population simple et homogène - nous développons une méthode composite qui combine les forces des deux principaux paradigmes de l'état de l'art afin d'en corriger les faiblesses. Pour le cas hétérogène non labellisé, nous proposons d'estimer un mélange de GGM avec un algorithme espérance-maximisation (EM). Afin d'améliorer les solutions de cet algorithme EM, et d'éviter de tomber dans des extrema locaux sous-optimaux quand les données sont en grande dimension, nous introduisons une version tempérée de cet algorithme EM, que nous étudions théoriquement et empiriquement. Enfin, nous améliorons le clustering de l'EM en prenant en compte l'effet que des co-facteurs externes peuvent avoir sur la position des données observées dans leur espace.

Title : Mixtures of Gaussian Graphical Models with constraints

Keywords : Graphical Models, Conditional Correlations, Unsupervised Estimation, EM algorithm

Abstract : Describing the co-variations between several observed random variables is a delicate problem. Dependency networks are popular tools that depict the relations between variables through the presence or absence of edges between the nodes of a graph. In particular, conditional correlation graphs are used to represent the "direct" correlations between nodes of the graph. They are often studied under the Gaussian assumption and consequently referred to as "Gaussian Graphical Models" (GGM).
A single network can be used to represent the overall tendencies identified within a data sample. However, when the observed data is sampled from a heterogeneous population, then there exist different sub-populations that all need to be described through their own graphs. What is more, if the sub-population (or "class") labels are not available, unsupervised approaches must be implemented in order to correctly identify the classes and describe each of them with its own graph.
In this work, we tackle the fairly new problem of Hierarchical GGM estimation for unlabelled heterogeneous populations. We explore several key axes to improve the estimation of the model parameters as well as the unsupervised identification of the sub-populations. Our goal is to ensure that the inferred conditional correlation graphs are as relevant and interpretable as possible.
First - in the simple, homogeneous population case - we develop a composite method that combines the strengths of the two main state of the art paradigms to correct their weaknesses. For the unlabelled heterogeneous case, we propose to estimate a Mixture of GGM with an Expectation Maximisation (EM) algorithm. In order to improve the solutions of this EM algorithm, and avoid falling for sub-optimal local extrema in high dimension, we introduce a tempered version of this EM algorithm, that we study theoretically and empirically. Finally, we improve the clustering of the EM by taking into consideration the effect of external co-features on the position in space of the observed data.

Institut Polytechnique de Paris
91120 Palaiseau, France