

Asymptotic Analysis of a Matrix Latent Decomposition Model

Clément Mantoux, Stanley Durrleman, Stéphanie Allassonnière

▶ **To cite this version:**

Clément Mantoux, Stanley Durrleman, Stéphanie Allassonnière. Asymptotic Analysis of a Matrix Latent Decomposition Model. ESAIM: Probability and Statistics, 2022, 26, pp.208-242. 10.1051/ps/2022004 . hal-03674722

HAL Id: hal-03674722
https://hal.science/hal-03674722

Submitted on 20 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESAIM: PS 26 (2022) 208–242
https://doi.org/10.1051/ps/2022004

ESAIM: Probability and Statistics
www.esaim-ps.org

ASYMPTOTIC ANALYSIS OF A MATRIX LATENT DECOMPOSITION MODEL

Clément Mantoux[1,2,3,*], Stanley Durrleman[1,2] and Stéphanie Allassonnière[4,5]

Abstract. Matrix data sets arise in network analysis for medical applications, where each network belongs to a subject and represents a measurable phenotype. These large dimensional data are often modeled using lower-dimensional latent variables, which explain most of the observed variability and can be used for predictive purposes. In this paper, we provide asymptotic convergence guarantees for the estimation of a hierarchical statistical model for matrix data sets. It captures the variability of matrices by modeling a truncation of their eigendecomposition. We show that this model is identifiable, and that consistent Maximum A Posteriori (MAP) estimation can be performed to estimate the distribution of eigenvalues and eigenvectors. The MAP estimator is shown to be asymptotically normal for a restricted version of the model.

Mathematics Subject Classification. 62F12, 62H21.

Received February 7, 2022. Accepted April 7, 2022.

1. Introduction

Latent variable models are powerful tools to capture the complexity of high-dimensional data. Their hierarchical structure decouples this complexity into a low-dimensional distribution of latent variables and a mechanism to generate observations from latent variables. Over the last decades, they have proven relevant to perform regression and classification tasks as well as to provide interpretable representations of the data. In this paper, we are interested more specifically in the analysis of matrix data sets: in this context, an observation is a matrix which represents the interactions between a given number of entities. The main case of interest is network data set analysis, where matrices represent the evolution of a given network across time, or the same network structure measured in different situations.

Recently, the analysis of network data sets has received increased attention in the literature, in particular for medical applications, where each network represents a different patient, typically its brain connectivity

Keywords and phrases: Hierarchical model, matrix data sets, low rank, stiefel manifold, identifiability, strong consistency, asymptotic normality.

[1] ARAMIS Project Team, Inria, Paris 75013, France.
[2] ARAMIS Lab, Brain and Spine Institute, ICM, INSERM UMR 1127, CNRS UMR 7225, Sorbonne Université, Hôpital de la Pitié-Salpêtrière, Paris 75013, France.
[3] CMAP, École polytechnique, Palaiseau 91120, France.
[4] Centre de Recherche des Cordeliers, Université de Paris, INSERM UMR 1138, Sorbonne Université, Paris 75006, France.
[5] HEKA Project Team, Inria, Paris 75006, France.

* Corresponding author: clement.mantoux@inria.fr

© The authors. Published by EDP Sciences, SMAI 2022

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

network. The need to understand the complex structure of the interactions within networks has brought the development of low-dimensional representations of these networks, with methods like sparse dictionary learning or graph auto-encoders [18, 39]. In many cases, the core modeling assumption relies on the low rank of the observed matrices [10]. In that regard, such models can be interpreted as constraints on the distribution of the eigenvalues and the eigenvectors. However, although these recent works have achieved great performance on practical tasks, little has been done in the literature so far to analyze their theoretical soundness.

In this paper, we provide an asymptotic analysis for a recently proposed network data set analysis model [43] which, in terms of generative modeling, can be considered a generalization of several current similar models relying on graph auto-encoders [34] and dictionary learning [19]. The model quantifies the variability in the spectral decomposition of network adjacency matrices: the leading eigenvectors, taking values in the Stiefel manifold, and the related eigenvalues are considered as latent variables in a hierarchical generative model. It relies on the classical assumption that the relevant information in a matrix of interaction coefficients can be captured by a low-rank approximation [49]. The model structure introduced in [43] was shown to be able to account for the complex variability of functional brain networks using a restricted number of parameters, and provides an interpretable representation of this variability.

We first show that the model is identifiable, and consider the parameter estimation problem. We show that, although the Maximum Likelihood Estimator may not be defined, the Maximum A Posteriori estimator exists for wide classes of prior distributions. Finally, we show the almost sure consistency of the estimator and its asymptotic normality as the number of samples goes to infinity. The technical difficulties arise from the hierarchical structure of the model: only a few specific such cases have received attention in the literature. For instance, the identifiability of latent variable models remains an open question for most latent variable network analysis models. Although our results take stock on the model structure, we believe that they can be transposed without hurdle to many similar models.

Notations

In the next sections, we use the following notations:

- $A^\top$ denotes matrix transposition, $\mathrm{Tr}(A)$ the trace and $\det(A)$ the determinant,
- $\|x\|$ denotes the canonical Euclidean norm for vectors, and the related operator norm for matrices,
- $\|A\|_F$ denotes the Frobenius norm and $\langle A, B\rangle_F = \mathrm{Tr}(A^\top B)$ the related inner product for matrices,
- If X is a $n \times p$ matrix, $x_i \in \mathbb{R}^n$ denotes its i-th column, so that $X = (x_1, \dots, x_p)$,
- $\mathcal{V}_{np}$ is the Stiefel manifold of $n \times p$ matrices X such that $X^\top X = I_p$.
- $O_n(\mathbb{R})$ is the orthogonal group $\mathcal{V}_{nn}$,
- For λ a vector and X a matrix, we define $\lambda \cdot X = X^\top \mathrm{Diag}(\lambda) X$,
- For A a $n \times n$ matrix and X a $n \times p$ matrix, we define $A * X = (x_i^\top A x_i)_{i=1}^p$.

2. A statistical model for spectral decomposition

2.1. Model definition

2.1.1. Observations distribution

We study the generative model for sets of weighted graph adjacency matrices $A_1, \dots, A_N \in \mathbb{R}^{n\times n}$ proposed in [43]. It draws symmetric low rank adjacency matrices A by generating their eigenvectors $X = (x_1, \dots, x_p) \in \mathbb{R}^{n\times p}$ and eigenvalues $\lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$, and combining them with an additive noise $\varepsilon \in \mathbb{R}^{n\times n}$.

$$A = X \mathrm{Diag}(\lambda) X^\top + \varepsilon \tag{2.1}$$

In practice, the adjacency matrix A represents a network. n corresponds to the number of nodes (*e.g.* in the case of brain connectivity, the number of brain regions), and $p \ll n$ is chosen such that the residual term ε is small. The eigenvectors take values in the Stiefel manifold $\mathcal{V}_{np}$ of matrices such that $X^\top X = I_p$. Their

probability distribution will be described in the next section. The eigenvalues follow a multivariate Gaussian distribution $\lambda \sim \mathcal{N}(\mu, \sigma_\lambda^2 I_p)$. The noise ε is a symmetric matrix whose coefficients above the diagonal also follow a Gaussian distribution $\mathcal{N}(0, \sigma_\varepsilon^2 I_{n\times(n+1)/2})$. We assume that the variables λ, X, ε are independent. This assumption is strong: it might not be satisfied in practice, as the variation of a pattern x_i should be naturally correlated to a variation of the related λ_i. However, it also allows keeping a small number of parameters, which allows for robust estimation in practice when the number of observed matrices is low. Their interpretations will be given in Section 2.2.2 on simpler alternative models.

2.1.2. Eigenvectors distribution

As an element of the Stiefel manifold $\mathcal{V}_{np}$, the eigenvector matrix X is described by a probability distribution over $\mathcal{V}_{np}$. The canonical framework for these distributions is exposed in [13], and consists in taking a measure with density with respect to the Haar measure over the Stiefel manifold. The Haar measure $[\mathrm{d}X]$ is defined, up to a constant, as the only measure invariant to orthogonal transformations, *i.e.*, for $S \subset \mathcal{V}_{np}$ and $O \in O_n(\mathbb{R})$, $\int_S [\mathrm{d}X] = \int_{OS} [\mathrm{d}X] = \int_{SO} [\mathrm{d}X]$. It can be rescaled by a constant factor to correspond to the Hausdorff measure over $\mathcal{V}_{np}$ [29].

The distribution considered for X is the von Mises-Fisher (vMF) distribution, also called Matrix Langevin distribution in the literature. It was first introduced by [32], who derived basic properties of the distribution and its Maximum Likelihood Estimator (MLE), and was further studied for both theoretical and algorithmic purposes [12, 30, 35, 46]. The von Mises-Fisher distribution over $\mathcal{V}_{np}$ is defined by its probability density function (p.d.f.) with respect to the Haar measure:

$$p(X) = \frac{1}{\mathcal{C}(F)} \exp(\mathrm{Tr}(X^\top F)) = \frac{1}{\mathcal{C}(F)} \exp\left(s_1 \langle x_1, m_1\rangle + \cdots + s_p \langle x_p, m_p\rangle\right), \tag{2.2}$$

with $\mathcal{C}(F)$ the normalizing constant and $F = (f_1, \ldots, f_p) = M\mathrm{Diag}(s) = (m_1, \ldots, m_p)\mathrm{Diag}(s_1, \ldots, s_p)$ the parameter of the distribution ($F \in \mathbb{R}^{n\times p}$). In the model considered here, $M \in \mathcal{V}_{np}$ and the s_i's are non-negative to ensure identifiability. By definition, the modal point M has maximal probability. The s_i's control the spread around the modal point, and are called the *concentration parameters* of the distribution.

The vMF distribution has a simple interpretation and requires few parameters. It imposes no dependency between the columns of X, except the orthogonality constraint. It forms an exponential family of distributions, and as such lends itself to efficient numerical estimation procedures. The normalizing constant $\mathcal{C}(F)$ has an analytic expression relying on the hypergeometric function of a matrix argument, and represents the main difficulty when analyzing the distribution, as it prevents from getting an explicit expression of its moments.

With this definition, we can write the full density of the model defined in the previous section. The likelihood of an observed matrix A writes:

$$\begin{aligned} p(A \mid \theta) &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} p(A \mid X, \lambda, \theta) p(X \mid \theta) p(\lambda \mid \theta)\, [\mathrm{d}X]\mathrm{d}\lambda \\ &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{\mathcal{C}(F)(2\pi)^{(n^2+p)/2}\sigma_\varepsilon^{n^2}\sigma_\lambda^p} \exp\left(\langle X, F\rangle_F - \frac{1}{2\sigma_\lambda^2}\|\lambda - \mu\|^2 - \frac{1}{2\sigma_\varepsilon^2}\|A - \lambda\cdot X\|_F^2\right) [\mathrm{d}X]\mathrm{d}\lambda\,, \end{aligned}$$

where we introduced the notation $\lambda \cdot X = X\mathrm{Diag}(\lambda)X^\top$ to lighten the formula, and $\theta = (F, \mu, \sigma_\lambda, \sigma_\varepsilon)$ regroups the model parameters.

Remark 2.1. The overall model structure (2.1) can be compared with equation (1.11) in [23], which states that, for any continuous probability distribution $p(A)$ over the space of symmetric matrices: for any bounded

continuous function h,

$$\int_{\mathbb{R}^{n\times n}} h(A)p(A)\,\mathrm{d}A = \iint_{O_n(\mathbb{R})\times\mathbb{R}^n} h(\lambda\cdot X)p(\lambda\cdot X)\prod_{i<j}(\lambda_i-\lambda_j)\,[\mathrm{d}X]\mathrm{d}\lambda\,,$$

with $[\mathrm{d}X]$ the normalized Haar measure over the group of orthogonal matrices $O_n(\mathbb{R})$. In other words, any matrix distribution is equivalently characterized by the joint distribution of its eigenvalues and eigenvectors. In that regard, our main hypotheses consist in constraining on the number of non-zero eigenvalues and imposing that the distributions of X and λ can be decoupled.

2.2. Motivation: network modeling

2.2.1. Beyond the graphon model

The graphon [42] is the standard reference model used in network theory to analyze large graphs from a probabilistic perspective. Many pieces of work in both the theoretical [28, 33, 55] and applied [36, 44, 50] literatures focus on the properties of the model it describes and its statistical estimation.

A graphon is a symmetric function $w : [0,1]^2 \to [0,1]$, which is to be understood as a continuous adjacency matrix with an infinite number of nodes. The graphon defines a distribution over $n \times n$ symmetric adjacency matrices by drawing n uniform numbers $U_1, \ldots, U_n \sim \mathcal{U}([0,1])$, and forming the matrix $A_{ij} = w(U_i, U_j)$, or $A_{ij} \sim \mathcal{B}(w(U_i, U_j))$ in the case of binary networks. The graphon inference problem thus consists, given one or several matrices A, in determining both the function w and the positions (U_i) of the nodes.

The main application of the graphon model is the Stochastic Block-Model (SBM), which assumes that w is block-wise constant. It amounts to dividing the set of nodes into clusters with given probabilities, and determining the connection between the nodes with the connection between their clusters. The SBM provides a well-studied [1, 45, 47] framework which is particularly relevant for a clustering analysis of networks, *i.e.* finding the most relevant partition among the nodes.

Both the graphon model and the SBM were conceived to analyze networks where nodes are drawn randomly and play interchangeable roles. They mostly focus on understanding the structure of the hidden graphon dynamic, which requires identifying the U_i's or the cluster labels.

Given a data set of matrices, both graphon and SBM would either (1) assume that the U_i's are drawn independently for each matrix or (2) take the same U_i's for each matrix in the data set. The first case yields a distribution whose expectation has constant coefficients: $\mathbb{E}[A_{ij}] = \mathbb{E}[w(U, U')]$ with $U, U' \sim \mathcal{U}([0,1])$. The second case results in a constant distribution with $A_{ij} = w(U_i, U_j)$ for every sample matrix A, or a matrix of independent Bernoulli variables $A_{ij} \sim \mathcal{B}(w(U_i, U_j))$ in the case of binary networks. Both options lead to simplistic distributions which are not relevant from a practical perspective.

In the context considered here, the nodes remain the same from one matrix to another (*e.g.* brain regions), and cannot be permuted. This allows to easily estimate the average interactions, which is the main difficulty for the graphon and the SBM. Modeling the matrices' spectral decomposition goes one step further than the SBM, and induces a dependency between the coefficients. It allows for instance computing the distribution of a set of matrix coefficients given other observed matrix coefficients.

2.2.2. Accounting for the full network variability

Two similar approaches currently co-exist in the literature to analyze sets of networks. On the one hand, Variational Graph Auto-Encoders (VGAE) [34] assume that each node i is represented by a low-dimensional vector $z_i \in \mathbb{R}^p$, and models the adjacency matrix as $A_{ij} = h(z_i^\top z_j)$, with h a non-linear function. The model thus characterizes A by a low-dimensional representation $Z \in \mathbb{R}^{n\times p}$, and retrieves $A = h(ZZ^\top) = h(\mathbf{1}_p \cdot Z)$. The matrices $\mathbf{1}_p \cdot Z$ are constrained to having positive eigenvalues. Additionally, the VGAE model considers all variables z_i as independent and identically distributed.

On the other hand, a dictionary model was proposed by [19], and writes each adjacency matrix A as a weighted combination of fixed rank one matrices: $A = \lambda_1 x_1 x_1^\top + \cdots + \lambda_p x_p x_p^\top$, which rewrites as $\lambda \cdot X$. Here, the goal is to find the best λ for each matrix A, while the matrix X is the same for all networks. This model thus imposes a strong dimension constraint on the adjacency matrices.

Each of these two approaches capture one aspect of the variability: for the VGAE, only the "eigenvectors" vary, and for the dictionary model, only the "eigenvalues" depend on the network. The model we study here simultaneously accounts for these two sources of variability, and thus allows for a richer representation, while keeping a latent space dimension comparable to that of VGAE. From the VGAE perspective, the rows $(M_{ki})_{i=1}^p$ shape the distribution of z_k, and the parameters (μ, σ_λ^2) determine the (possibly non-positive) inner products between the z_i's. From the dictionary model perspective, the column $m_i = (M_{ki})_{k=1}^n$ gives the i-th dictionary element and s_i its concentration; the coefficient M_{ki} gives the strength of the contribution of pattern i to the interactions of node k in the network. The parameters (μ, σ_λ^2) give the distribution of the dictionary weights.

2.3. Conditional distribution

Summarizing the model definition in Section 2.1, we assume that an observed adjacency matrix A writes as $A = \lambda \cdot X + \varepsilon$, with $(\lambda, X) \in \mathbb{R}^p \times \mathcal{V}_{np}$ being independent latent variables and ε a symmetric matrix of Gaussian distributed noise coefficients. The full model p.d.f. writes:

$$p(A, X, \lambda \mid \theta) = \frac{1}{\mathcal{C}(F)(2\pi)^{(n^2+p)/2}\sigma_\varepsilon^{n^2}\sigma_\lambda^p} \exp\left(\langle X, F\rangle_F - \frac{1}{2\sigma_\lambda^2}\|\lambda - \mu\|^2 - \frac{1}{2\sigma_\varepsilon^2}\|A - \lambda \cdot X\|_F^2\right).$$

From this expression, we can express the conditional distribution of the latent variables (X, λ) given A as follows. In the remainder of the paper, we will denote

$$A * X = (x_k^\top A x_k)_{k=1}^p, \quad \frac{1}{\sigma_p^2} = \frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\lambda^2} \quad \text{and} \quad \mu_{AX} = \sigma_p^2\left[\frac{1}{\sigma_\varepsilon^2} A * X + \frac{1}{\sigma_\lambda^2}\mu\right]. \tag{2.3}$$

The expression of the conditional density $p(X, \lambda \mid A, \theta)$ of the latent variables given the observed variable A writes as:

$$\begin{cases} p(X \mid A, \theta) \propto \exp\left(\langle X, F\rangle_F + \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right) \\ p(\lambda \mid X, A, \theta) = \mathcal{N}(\mu_{AX}, \sigma_p^2). \end{cases}$$

The proof of this equation follows the same lines as in Lemma C.1 in Appendix C. We will be using this expression of the conditional distribution in Section 5 on asymptotic normality. The $\frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2$ term in the distribution of $(X \mid A)$ is typically much larger than $\langle X, F\rangle_F$ as long as $n \gg p$, and it thus determines the shape of the distribution. As shown in the following proposition, it is maximized by the eigenvectors of A.

Proposition 2.2. *For $A \in \mathbb{R}^{n\times n}$, $\|\mu_{AX}\|^2$ is maximized by taking X among the eigenvectors of A. Furthermore, if the eigenvalues of A all have multiplicity one, this maximization is strict.*

Proof. Let $A = U^\top D U$ be the eigendecomposition of A, with $U^\top U = I_n$. Without loss of generality, we take $\sigma_\lambda = \sigma_\varepsilon = 1$. We have:

$$\begin{aligned} \max_{X\in\mathcal{V}_{np}} 2\|\mu_{AX}\|^2 &= \max_{X\in\mathcal{V}_{np}} \sum_{i=1}^p (x_i^\top A x_i + \mu_i)^2 \\ &= \max_{Y\in\mathcal{V}_{np}} \sum_{i=1}^p (y_i^\top (D + \mu_i I_n) y_i)^2 \quad (\text{setting } Y = UX) \end{aligned}$$

$$
\begin{aligned}
&= \max_{Y \in \mathcal{V}_{np}} \sum_{i=1}^{p} \sum_{k=1}^{n} [(d_k + \mu_i) y_{ik}^2]^2 \\
&\leq \max_{Y \in \mathcal{V}_{np}} \sum_{i=1}^{p} \sum_{k=1}^{n} (d_k + \mu_i)^2 y_{ik}^2 \quad \text{(Jensen's inequality)} \\
&= \max_{Y \in \mathcal{V}_{np}} \langle K, Y \odot Y \rangle_F \, .
\end{aligned}
$$

With $K \in \mathbb{R}^{n \times p}$ defined by $K_{ij} = d_k + \mu_i$ and $Y \odot Y$ the Hadamard (entrywise) product. If we extend K to a $n \times n$ matrix K' by padding zeros, and extend Y to an orthogonal matrix Q by completing Y into a basis, the objective function remains unchanged: $\langle K, Y \odot Y \rangle_F = \langle K', Q \odot Q \rangle_F$.

Since Q is orthogonal, the matrix $S = Q \odot Q$ is doubly stochastic. Furthermore, the Birkhoff-von Neumann theorem states that the set of doubly stochastic matrices is the convex hull of the set of permutation matrices. As a consequence, the linear function $\langle K', S \rangle_F$ is maximized by taking for S a permutation matrix. Such matrices are orthogonal and verify $S \odot S = S$, and their only square roots for the Hadamard product are permutation matrices with negative coefficients allowed. Therefore, the optimal choice for Y has its columns in the canonical basis. Hence the optimal choice for $X = U^\top Y$ is to take its columns among the eigenvectors of A.[1]

When Y is a permutation matrix, Jensen's inequality becomes an equality, so that taking the related $X = U^\top Y$ is also an optimal choice for the original objective $\|\mu_{AX}\|^2$. Furthermore, if A has n distinct eigenvalues, Jensen's inequality is strict except when y_i is a vector of the canonical basis. Therefore, in that case, the optimal subset of eigenvectors of A (up to permutation and change of sign) is the only maximizer of $\|\mu_{AX}\|^2$. □

Remark 2.3. When taking $\mu = 0$, the result can be proved more simply by using Ky Fan's principle on eigenvectors [22]. A closely related, yet different result, was recently obtained by [40]. We believe that obtaining a closed-form formula for maximizing the complete conditional density $p(X \mid A, \theta)$ would require significantly more work. The eigenvectors of A are no longer optimal: the best value of X is obtained as a trade-off between M and the closest optimal eigenvalue combination of A, with the concentration and variance parameters determining the balance between both.

3. Model identifiability

Identifiability of statistical model $p(x \mid \theta)$ refers to the property that, if $\theta_1 \neq \theta_2$, then the distributions $p(\cdot \mid \theta_1)$ and $p(\cdot \mid \theta_2)$ must differ. It is a generally desirable property, as it ensures that the model is well-defined and behaves in an intuitive way. It also has an immediate theoretical interest, since it enables to prove that Maximum Likelihood Estimators converge to the correct value when the data is generated according to the model. It can be proved for instance by retrieving the parameter θ from a set of moments of $p(\cdot \mid \theta)$.

The identifiability of latent variable models is a general, long-standing question, which has been studied and proved for only few specific models. It relates to the question of identifying the parameters of graphical models where only a fraction of the variables is observed. Much work has been devoted to the identifiability of finite mixture models [27, 51, 52, 56]. In a similar spirit, classes of statistical models with discrete latent variables have also recently been proved to be identifiable [4, 25]. Partial results have been shown for mixed-effects models, in particular in a longitudinal setting [37, 51]. In a less closely related domain, identifiability results exist on time series model with latent variables [17]. Finally, general identifiability results are available for (possibly infinite) mixtures of exponential models [6, 7]. Although the latter result is related to the model we consider here, its necessary theoretical conditions turn out to be hard to verify in practice.

[1]The authors thank the `math.stackexchange.com` community member `user1551` for his helpful answer on the Birkhoff-von Neumann theorem.

The main difficulty with identifying latent variable models comes from the expression of the observations' likelihood:

$$p(A \mid \theta) = \iint_{\mathcal{V}_{np} \times \mathbb{R}^p} p(A, X, \lambda \mid \theta)\, [\mathrm{d}X]\mathrm{d}\lambda\,.$$

Even though our full model $p(A, X, \lambda \mid \theta)$ is identifiable, the marginalized model $p(A \mid \theta)$ may not be: permuting two eigenvalues μ_i, μ_j and the related eigenvector parameters f_i, f_j, or changing the sign of f_i does not change the distribution of A. This first obvious source of non-identifiability is easily overcome, by imposing that the normalized columns $(m_1, \dots, m_p)$ (denoting $m_i = f_i/|f_i|$) are sorted according to the lexicographical order and that each column has its first non-zero element positive. An additional constraint allows getting a provably identifiable marginal model: we shall assume that the f_i's are non-zero, *i.e.* that the concentration parameters $s_i = \|f_i\|$ are positive. These two constraints form the set of identifiable parameters Θ^{id}:

$$\Theta^{\mathrm{id}} = \{\theta \mid m_1 \prec \cdots \prec m_p \text{ and } \min_i s_i > 0\}\,.$$

With this definition, we have the following result:

Theorem 3.1. *If $p < n$, over Θ^{id}, different parameters $\theta_1 \neq \theta_2$ yield different marginal probability distributions $p(A \mid \theta_1)$ and $p(A \mid \theta_2)$.*

Proof. Given $\theta \in \Theta^{\mathrm{id}}$, we show that all parameters $(F, \mu, \sigma_\lambda, \sigma_\varepsilon)$ can be retrieved from the distribution $p(A \mid \theta)$. We first identify the noise variance. This allows identifying the eigenvalue parameters, and finally the eigenvector parameters.

Identifying σ_λ and σ_ε. Using Lemma C.1 and $\alpha I_n * X = \alpha \mathbf{1}_p$, we have, for all $\alpha \in \mathbb{R}$:

$$p(A = \alpha I_n \mid \theta) = \frac{1}{\sqrt{2\pi}^{n^2} \sigma_\varepsilon^{n^2}} \frac{\sigma_p^p}{\sigma_\lambda^p} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} n^2\alpha^2 + \frac{\sigma_p^2}{2\sigma_\varepsilon^4} p^2\alpha^2 + \|\mu\|^2 \left(\frac{\sigma_p^2}{2\sigma_\lambda^4} - \frac{1}{2\sigma_\lambda^2} \right) + \alpha \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \langle \mu, \mathbf{1}_p \rangle \right),$$

with $\sigma_p^{-2} = \sigma_\varepsilon^{-2} + \sigma_\lambda^{-2}$. The function $\alpha \mapsto \log p(A = \alpha I_n \mid \theta)$ is a second-order polynomial, its coefficients (a_0, a_1, a_2) can thus be identified. In particular, the degree two coefficient gives the value of

$$a_2 = -\frac{n^2}{2\sigma_\varepsilon^2} + \frac{p^2\sigma_p^2}{2\sigma_\varepsilon^4}\,. \tag{3.1}$$

Similarly, the computation in Lemma C.1 can be used to derive the gradient $\nabla_A p(A \mid \theta)$. It writes, for $A = \alpha I$:

$$\nabla_A p(A \mid \theta) = \frac{1}{\sigma_\varepsilon^2} p(A = \alpha I \mid \theta)\, (-\alpha I + \mathbb{E}[B])\,,$$

where B is the random variable given by $B = \lambda_p \cdot X$, with $\lambda_p \sim \mathcal{N}\left(\frac{\sigma_p^2}{\sigma_\lambda^2}\mu + \frac{\sigma_p^2}{\sigma_\varepsilon^2}\alpha \mathbf{1}_p, \sigma_p^2 \right)$. Furthermore, since we have

$$\mathbb{E}[B] = \mathbb{E}[X^\top \mathrm{Diag}(\lambda_p) X] = \sum_{i=1}^{p} \mathbb{E}[\lambda_{p,i}] \mathbb{E}[x_i x_i^\top]\,,$$

we deduce:

$$\frac{\nabla_A p(A=\alpha I \mid \theta)}{p(A=\alpha I \mid \theta)} = \frac{1}{\sigma_\varepsilon^2}\left(-\alpha I + \sum_{i=1}^{p}\left[\frac{\sigma_p^2}{\sigma_\lambda^2}\mu_i + \frac{\sigma_p^2}{\sigma_\varepsilon^2}\alpha\right]\mathbb{E}[x_i x_i^\top]\right).$$

Finally, since $\operatorname{Tr}\mathbb{E}[x_i x_i^\top] = \left\langle \mathbb{E}[x_i x_i^\top], I\right\rangle_F = \mathbb{E}[x_i^\top I x_i] = 1$, we have:

$$\operatorname{Tr}\left(\frac{\nabla_A p(A=\alpha I \mid \theta)}{p(A=\alpha I \mid \theta)}\right) = \frac{1}{\sigma_\varepsilon^2}\left(-\alpha n + \frac{\sigma_p^2}{\sigma_\lambda^2}\langle \mu, \mathbf{1}_p\rangle + p\frac{\sigma_p^2}{\sigma_\varepsilon^2}\alpha\right).$$

As a consequence, the α-linear function above can be deduced from the distribution of A, hence we know its coefficients. In particular, the leading coefficient a_3 writes:

$$a_3 = -\frac{n}{\sigma_\varepsilon^2} + p\frac{\sigma_p^2}{\sigma_\varepsilon^4}.$$

The formulas of a_3 and a_2 in equation (3.1) can be combined to obtain $-\frac{1}{2\sigma_\varepsilon^2}(n^2 - np) = a_2 - pa_3/2$. Therefore, since $p \neq n$, σ_ε can be identified, along with σ_λ.

Identifying μ.

The moment generating function of A writes as:

$$G_A(T) = \mathbb{E}[e^{\langle T, A\rangle_F}] = \mathbb{E}[e^{\langle T, \lambda\cdot X + \varepsilon\rangle_F}] = G_{\lambda\cdot X}(T) \times G_\varepsilon(T).$$

Since the distribution of ε has been characterized, $G_\varepsilon(T)$ is known, and hence $G_{\lambda\cdot X}$ can be deduced as $G_A(T)/G_\varepsilon(T)$. As the moment generating function characterizes the probability distribution, if the distribution $\lambda\cdot X$ is identifiable then the distribution of A is identifiable. We thus turn on the problem of identifying μ given the distribution of $\lambda\cdot X$ (and proceed similarly for the eigenvector parameters in the next paragraph). We have for $t \in \mathbb{R}$:

$$\begin{aligned}\mathbb{E}[e^{t\lambda\cdot X}] &= \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{1}{k!}t^k X^\top \operatorname{Diag}(\lambda)^k X\right] = \mathbb{E}\left[X^\top \operatorname{Diag}((e^{t\lambda_i})_{i=1}^p) X\right] \\ &= \sum_{i=1}^{p}\mathbb{E}[e^{t\lambda_i} x_i x_i^\top] = \sum_{i=1}^{p}\mathbb{E}[e^{t\lambda_i}]\mathbb{E}[x_i x_i^\top] = \sum_{i=1}^{p} e^{t\mu_i + \frac{1}{2}\sigma_\lambda^2 t^2}\mathbb{E}[x_i x_i^\top],\end{aligned}$$

which in particular gives $\operatorname{Tr}(\mathbb{E}[e^{t\lambda\cdot X}]) = \sum_{i=1}^{p} e^{t\mu_i + \frac{1}{2}\sigma_\lambda^2 t^2}$.

The functions of the form $t \mapsto e^{t\mu_i + \frac{1}{2}\sigma_\lambda^2 t^2}$ are linearly independent for distinct μ_i's: this allows retrieving both the μ_i's and the multiplicity count of each eigenvalue.

Identifying F. From there, we could use the matrices $\mathbb{E}[x_i x_i^\top]$ to identify the modal directions m_i. Indeed, as shown in [32] (Eqs. (2.9)–(2.11)), each m_k is an eigenvector of each $\mathbb{E}[x_i x_i^\top]$. However, the related eigenvalues and remaining $n-p$ eigenvectors are unknown, and the relevant eigenvectors cannot be identified easily. In the limit of large concentration parameters, $\mathbb{E}[x_i x_i^\top] \simeq m_i m_i^\top$, so that the largest eigenvalue is the one corresponding to m_i. Yet this argument cannot be quantified, as the eigenvalues involve partial derivatives of $\log \mathcal{C}(F)$ which are hard to manipulate.

Instead, we get a better result by expressing the density of the distribution of $B = \lambda\cdot X = \mathscr{D}(\lambda, X)$, with support on the set $\operatorname{Im}(\mathscr{D})$ of $n\times n$ square matrices with rank p. The distribution of B is characterized by the

expectations $\mathbb{E}[h(B)]$ with h continuous bounded. We have:

$$\mathbb{E}[h(B)] = \iint h(\mathscr{D}(\lambda, X)) \cdot p(\lambda \mid \theta) p(X \mid \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \,. \tag{3.2}$$

We want to perform a change of variable so as to express the expectation as an integral over $\mathrm{Im}(\mathscr{D})$. However this cannot be performed directly. First, the mapping $\mathscr{D}$ is not injective. Next, the most relevant change of variable formula for this problem is, to the best of our knowledge, the main result of [53], which gives a formula for mappings taking inputs in vector spaces (which is not the case here as $X \in \mathcal{V}_{np}$).

The first problem can be solved by splitting the integral over domains where $\mathscr{D}$ is injective, which means preventing permutation and change of signs in the columns of X. To that end, for $\pi \in S_p$ a permutation and $f \in \{\pm 1\}^p$, we denote $X_{\pi,f} = (f_1 x_{\pi(1)}, \dots, f_p x_{\pi(p)})$, and by $\lambda_\pi = (\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)})$. We also define the sets

$$\begin{cases} \Delta_0 = \{X \in \mathcal{V}_{np} \mid x_1 \prec \cdots \prec x_p \text{ and the first non-zero coefficient of each column is } > 0\} \\ \Delta_{\pi,f} = \{X_{\pi,f} \mid X \in \Delta_0\} \,, \end{cases}$$

where $\prec$ denotes the lexicographical order over $\mathbb{R}^n$. By construction, we have $\mathcal{V}_{np} = \cup_{\pi,f} \Delta_{\pi,f} \cup \mathcal{O}$ with $\mathcal{O}$ a set with measure zero. We get

$$\mathbb{E}[h(B)] = \sum_{\pi,f} \iint_{\mathbb{R}^p \times \Delta_{\pi,f}} h(\mathscr{D}(\lambda, X)) \cdot p(\lambda \mid \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \,.$$

Furthermore, the map $X \mapsto X_{\pi,f}$ corresponds to multiplying X by an orthogonal matrix. By construction, the Haar measure over $\mathcal{V}_{np}$ is invariant to this transformation [13]. Moreover, the map $\lambda \mapsto \lambda_\pi$ is also a linear orthogonal transformation, and as such has Jacobian determinant one. Hence we can perform the change of variable $(\lambda, X) \mapsto (\lambda_\pi, X_{\pi,f})$, and we get:

$$\begin{aligned} \mathbb{E}[h(B)] &= \sum_{\pi,f} \iint_{\mathbb{R}^p \times \Delta_0} h(\mathscr{D}(\lambda_\pi, X_{\pi,f})) \cdot p(\lambda_\pi \mid \theta) p(X_{\pi,f} \mid \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \\ &= \iint_{\mathbb{R}^p \times \Delta_0} h(\mathscr{D}(\lambda, X)) \cdot \sum_{\pi,f} p(\lambda_\pi \mid \theta) p(X_{\pi,f} \mid \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \\ &= \iint_{\mathbb{R}^p \times \Delta_0} h(\mathscr{D}(\lambda, X)) \cdot \sum_{\pi,f} p(\lambda \mid \theta_{\pi,f}) p(X \mid \theta_{\pi,f}) \, [\mathrm{d}X] \mathrm{d}\lambda \,, \end{aligned}$$

with $\theta_{\pi,f} = (F_{\pi,f}, \mu_\pi, \sigma_\lambda, \sigma_\varepsilon)$.

The first problem is now solved, as $\mathscr{D}$ is injective over $\mathbb{R}^p \times \Delta_0$. We now need to get to an integral formulation over a vector space. To that end, we consider the inverse of the Cayley transform of X: $D = C^{-1}(X)$. We refer the reader to Appendix A for a definition of the Cayley transform C. It is a smooth injective map from the tangent space at identity $T_{I_{np}} \mathcal{V}_{np} = \{\binom{A}{B} \mid A^\top = -A\}$ to the manifold $\mathcal{V}_{np}$, which covers the entire manifold apart from a set with measure zero. As explained in [29] (Thm. 4.1), a change of variable from D to X can be performed, and amounts to adding a multiplicative factor $J_1(D)$, with J_1 is a generalized Jacobian determinant. It follows that we can rewrite:

$$\mathbb{E}[h(B)] = \iint_{\mathbb{R}^p \times C^{-1}(\Delta_0)} h(\mathscr{D}(\lambda, C(D))) \cdot \sum_{\pi,f} p(\lambda \mid \theta_{\pi,f}) p(X \mid \theta_{\pi,f}) \cdot J_1(D) \, \mathrm{d}D \mathrm{d}\lambda \,.$$

Since the map $D \mapsto C(D)$ is injective on $T_{I_{np}}\mathcal{V}_{np}$ (Eqs. (1)–(3) in [29]), the map $(\lambda, D) \mapsto B = \mathscr{D}(\lambda, C(D))$ is injective over $\mathbb{R}^p \times C^{-1}(\Delta_0)$. Given B, we denote by λ_B, X_B and D_B its pre-images by $\mathscr{D}$ and C. Since the considered mapping is smooth, the main theorem of [53] applies. Letting $J_2(\lambda, D)$ be the generalized Jacobian determinant involved in the formula, it writes as:

$$\mathbb{E}[h(B)] = \int_{\mathrm{Im}(\mathscr{D})} h(B) \cdot \sum_{\pi,f} p(\lambda_B \mid \theta_{\pi,f}) p(X_B \mid \theta_{\pi,f}) \cdot \frac{J_1(D_B)}{J_2(\lambda_B, D_B)} \, \mathrm{d}B \,.$$

where $\mathrm{d}B$ denotes the Hausdorff measure over $\mathrm{Im}(\mathscr{D})$. Since both maps C and $\mathscr{D}$ are diffeomorphic, the generalized Jacobian determinants involved are non-zero.

As a consequence, the random variable B has density

$$\sum_{\pi,f} p(\lambda_B \mid \theta_{\pi,f}) p(X_B \mid \theta_{\pi,f}) \cdot \frac{J_1(D_B)}{J_2(\lambda_B, D_B)}$$

over its support w.r.t. the Hausdorff measure. Therefore, if the distribution of B is known, we can deduce the value of the function $B \mapsto \sum_{\pi,f} p(\lambda_B \mid \theta_{\pi,f}) p(X_B \mid \theta_{\pi,f})$. For $X \in \Delta_0$ and $\lambda \in \mathbb{R}^p$, it comes that we know the value of

$$f_\lambda(X) = \sum_{\pi,f} p(\lambda \mid \theta_{\pi,f}) p(X \mid \theta_{\pi,f}) \,.$$

Since the sum above is invariant by any permutation π and change of sign f, it follows that the value of this expression is known not only for $X \in \Delta_0$, but over the whole manifold $\mathcal{V}_{np}$. Now, we consider the specific case $\lambda = \mu$. Up to a normalizing constant, $f_\mu(X)$ is a probability distribution over $\mathcal{V}_{np}$: it is a mixture of von Mises-Fisher distributions with parameters $(F_{\pi,f})$ and mixture weights proportional to $p(\mu \mid \theta_{\pi,f})$. This structure allows using the main result of [31], which grants that the von Mises-Fisher densities given by the $F_{\pi,f}$ are linearly independent. This result can be combined with the main theorem of [56], which states that a family of finite mixtures is identifiable if and only if the mixture components form a linearly independent set.

As a consequence, we identify the parameter F up to a column permutation and change of sign. Moreover, in the sum above, the probabilities $p(\mu \mid \theta_{\pi,f})$ with maximal amplitude are given by $\pi = \mathrm{Id}$, and all the other permutations such that for all i, $\mu_{\sigma(i)} = \mu_i$ (which encompasses eigenvalue multiplicity). Since we assumed that all concentration parameters are positive, all $(F_{\pi,f})$ are distinct and hence the maximal mixture weights correspond to the matrices $(F_{\pi,f})$ with π as just described. This finally allows matching eigenvalues with eigenvectors, completing the identification of θ.

□

4. Existence and consistency of the MAP estimator

4.1. Maximum a posteriori *versus* maximum likelihood

We turn to the problem of estimating θ from samples $A_1, \ldots, A_N$ when the number of samples N grows large. In this section, we assume that the samples are distributed according to a distribution P, which may not be of the form $p(A \mid \theta)$.

However, the MLE may not be defined: the optimal value for F may theoretically be infinite, as the model likelihood does not necessarily decrease at infinity. For instance, if the samples $A_1, \ldots, A_N$ are drawn from a Gaussian distribution with i.i.d. coefficients and mean equal to a rank p matrix $\lambda_0 \cdot X_0$, the parameters σ_λ and s_i tend to take extreme values (σ_λ being very small and s_i being very large), and the distribution of latent variables is highly concentrated around (λ_0, X_0). This phenomenon occurs because the estimated data distribution asymptotically converges to the true data distribution, which lies at the boundary of the model

family (in the sense that taking very large s_i's and a very small σ_λ yields a distribution close to the true data distribution).

This problem is overcome numerically by adding a prior distribution $p(\theta)$ and considering the Maximum A Posteriori (MAP) estimator over the set Θ of all parameters:

$$\hat{\theta}_N \in \operatorname{argmax}_\Theta p(\theta \mid A_1, \dots A_N) = \operatorname{argmax}_\Theta p(A_1, \dots A_N \mid \theta) p(\theta) \,.$$

In this section, we want to account for the possible convergence of latent variable distributions to constant values. For this purpose, instead of the parameterization $\theta = (F, \mu, \sigma_\lambda, \sigma_\varepsilon)$, we will be defining the parameter set by $\Theta = \{\theta = (M, s, \mu, \sigma_\lambda, \sigma_\varepsilon) \mid \sigma_\lambda > 0, s_i < +\infty\}$, with the equivalence given by $F = M\mathrm{Diag}(s)$. In the next section, this representation will allow us to formally consider an extension of the set Θ accounting for the case where $s_i = +\infty$ and $\sigma_\lambda = 0$.

We consider inverse Gamma distributions for the prior $p(\sigma_\lambda, \sigma_\varepsilon)$, the uniform distribution over $\mathcal{V}_{np}$ for M, and any p.d.f. decreasing at infinity for $p(s)$ and $p(\mu)$. Unlike the MLE, with this prior specification the MAP estimator is guaranteed to exist.

Theorem 4.1. *Given the proposed model, with parameters following the prior distribution described above, for any set of matrices $(A_i)_{i=1}^N$, there exists $\hat{\theta}_N \in \operatorname{argmax}_{\theta\in\Theta} p(\theta \mid A_1, \dots, A_N)$.*

Proof. The bound obtained in Lemma C.2 gives with Bayes' formula:

$$\log p(\theta \mid A) \leq -\frac{n^2}{2}\log(2\pi) - (n^2 - p)\log\sigma_\varepsilon - p\log\sigma_\lambda + \log p(\theta) - \log p(A) \,.$$

Since

$$\begin{cases} p(\sigma_\lambda) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)}(1/\sigma_\lambda)^{\alpha_\lambda+1}\exp(-\beta_\lambda/\sigma_\lambda) \\ p(\sigma_\varepsilon) = \frac{\beta_\varepsilon^{\alpha_\varepsilon}}{\Gamma(\alpha_\varepsilon)}(1/\sigma_\varepsilon)^{\alpha_\varepsilon+1}\exp(-\beta_\varepsilon/\sigma_\varepsilon) \,, \end{cases}$$

and given the other assumptions on the prior distribution, we have $\log p(\theta \mid A) \to -\infty$ as any of the model variables reaches an open boundary of its domain. Furthermore, the function $\log p(\theta \mid A)$ is smooth: the integral representation given by Lemma C.1 writes as

$$p(A \mid \theta) = \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}}\frac{1}{C(F)}\frac{\sigma_p^p}{\sigma_\lambda^p}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2\right]\int_{\mathcal{V}_{np}}\exp\left[\langle F, X\rangle + \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right]\mathrm{d}X \,.$$

Since the manifold $\mathcal{V}_{np}$ is compact and the integrand $f(\theta, X) = \exp(\langle F, X\rangle + \|\mu_{AX}\|^2/2\sigma_p^2)$ is smooth on $\Theta \times \mathcal{V}_{np}$, classical integration theorems grant that $\log p(\theta \mid A)$ is smooth over every compact subset of Θ: given a compact set K, the domination function $g(X) = \max_{\theta\in K} f(X, \theta)$ is smooth over K. Hence $\log p(\theta \mid A)$ is smooth over Θ. In particular, the function $\log p(A \mid \theta)$ is coercive and continuous, and it thus admits a maximizer over Θ. □

4.2. MAP consistency

The above result motivates the study of the MAP estimator over the MLE. However, although adding a prior distribution grants the existence of a maximizer within Θ, the weight of the prior term decreases as the number of samples grows large, and we should expect the MAP estimator to diverge to the boundary of Θ for some empirical data distributions P. This phenomenon is accounted for by considering an extended set of parameters Θ^∞ allowing null eigenvector variance (*i.e.* λ constant) and infinite von Mises-Fisher concentrations (*i.e.* x_i constant for some i's):

$$\Theta^\infty = \{(M, s, \mu, \sigma_\lambda, \sigma_\varepsilon) \mid \sigma_\lambda \in [0, +\infty), s_i \in [0, +\infty]\} \,.$$

We prove in Lemma C.3 that the likelihood $p(A \mid \theta)$ extends continuously to Θ^∞. The extension essentially amounts to considering eigenvalue and eigenvector distributions restricted to a conditional subspace. With this convention, the objective function ℓ to be asymptotically maximized can be defined over Θ^∞ as the almost sure (a.s.) limit of the empirical objective function $\frac{1}{N}\sum_{i=1}^N \log p(A_i \mid \theta) + \frac{1}{N}p(\theta)$ defined over Θ:

$$\ell(\theta) = \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta)]\,.$$

If P has a density with respect to the Lebesgue measure, the function ℓ is equal, up to a constant term which depends only on P, to the opposite of the Kullback-Leibler divergence between P and $p(A \mid \theta)$. The MAP estimator is said to be *consistent* if it converges to the set Θ_* of maximizers of $\ell(\theta)$. In the case where P corresponds to some $p(A \mid \theta^*)$ for $\theta^* \in \Theta^{\mathrm{id}}$, ℓ only has one maximizer, which is the true model parameter θ^*. For large classes of sufficiently regular families of statistical models, the MLE and the MAP can be proved to be consistent and, in probability, to minimize the KL divergence to the optimal point [54].

The consistency of MLE for latent variable models has been studied for several classes of models, like Hidden Markov Models [16], Independent Component Analysis [9] or longitudinal mixed effects models [3, 11]. Along these results, we obtain the almost sure (a.s.) consistency of the MAP estimator. We study two particular cases: in the first case, we assume that the parameters which may diverge stay bounded, and obtain a.s. convergence to the set of maximizers over the constrained set. In the second case, we show that the unconstrained MAP estimator converges a.s. to the set of maximizers over Θ^∞.

The convergence to the set of maximizers of $\ell(\theta)$ is quantified by the distance $d(\hat{\theta}_N, \Theta_*)$. However, the set Θ_*^∞ of maximizers of ℓ over Θ^∞ may have some elements with infinite coordinates, which prevents from quantifying distances. To overcome this issue, we consider the reparameterization $\xi(\theta) = (M, h(s), \mu, \sigma_\lambda, \sigma_\varepsilon)$, with $h : [0, +\infty]^p \to [0, 1]^p$ applying the same continuous increasing transformation to each s_i, for instance $h(s)_i = \operatorname{atanh}(s_i)$. Over the new parameter space $\Xi^\infty = \xi(\Theta^\infty)$, we also obtain the almost sure consistency of the MAP $\hat{\xi}_N = \xi(\hat{\theta}_N)$.

Theorem 4.2. *Let Θ^η be the set of parameters with each s_i and σ_λ^{-1} upper bounded by η, and let Θ_*^η be the the set of maximizers of ℓ over Θ^η. Consider the following hypotheses:*

- ***H1*** *The number of latent patterns is strictly lower than the number of nodes: $p < n$.*
- ***H2*** *The samples $(A_i)_{i=1}^N$ are independent and identically distributed.*
- ***H3*** *The true data distribution $P(\mathrm{d}A)$ has a density w.r.t. the Lebesgue measure and exponentially decaying tails beyond a compact set: there exist $a, b > 0$, such that for x large enough, $\sup_{\|A\|_F \geq x} P(A) \leq a\exp(-bx)$.*

*Then, assuming **H1**, **H2** and **H3**:*

1. *For all $\eta > 0$, $\Theta_*^\eta \neq \varnothing$ and the MAP estimator $\hat{\theta}_N^\eta$ on Θ^η is consistent: for every continuous metric δ, almost surely,*

$$\delta(\hat{\theta}_N^\eta, \Theta_*^\eta) \xrightarrow{N\to+\infty} 0\,.$$

2. *The extended set of maximizers is non empty: $\Theta_*^\infty \neq \varnothing$. Denoting $\Xi_*^\infty = \xi(\Theta_*^\infty)$, for every continuous metric δ, almost surely,*

$$\delta(\hat{\xi}_N, \Xi_*^\infty) \xrightarrow{N\to+\infty} 0\,.$$

Remark 4.3. As a consequence, if all the elements of Θ^∞ are equal on a coordinate, the corresponding coordinate of $\hat{\theta}_N$ converges to this value. In particular, for some distributions P we may have $s_i \to +\infty$ or $\sigma_\lambda \to 0$ almost surely. This explains the phenomenon observed in the previous section on Gaussian empirical data distributions.

The proof follows the architecture of [11, 54]. The main difficulties and specificities lie in the proofs of the required lemmas which are specific to the model, and the possibility of having partially constant latent variable distributions. We thus only present here the structure of the main proof, and refer the reader to Appendix B for the detailed argument. As the proof for the first assertion is a strictly simpler version of the proof of the second assertion, we omit it for the sake of brevity.

Sketch of the proof. The proof is divided into four parts. We define

$$\mathbb{E}^* = \sup_{\theta\in\Theta^\infty} \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta)] \quad \text{and} \quad K_\varepsilon = \{\theta \in \overline{\Theta^\infty} \mid \delta(\xi(\theta), \Xi_*^\infty) \geq \varepsilon\},$$

with $\overline{\Theta^\infty}$ the Alexandrov compactification of Θ^∞, as detailed in Appendix B.

A) We prove that, for all $\theta_\infty \in \overline{\Theta^\infty}$ such that $\delta(\xi(\theta_\infty), \Xi_*^\infty) \geq \varepsilon$, there exists an open neighborhood $\mathcal{U} \subset \overline{\Theta^\infty}$ of θ_∞ such that

$$\mathbb{E}_{P(\mathrm{d}A)}\left[\sup_{\theta\in\mathcal{U}\cap\Theta^\infty} \log p(A \mid \theta)\right] < \mathbb{E}^*.$$

B) The set K_ε described above is compact, and therefore among all the sets $\mathcal{U}$ defined in part A we can extract a finite cover of K_ε. This allows proving that

$$\limsup_{N\to+\infty} \sup_{\theta\in K_\varepsilon\cap\Theta^\infty} \frac{1}{N}\sum_{i=1}^N \log p(A_i \mid \theta) < \mathbb{E}^*.$$

C) Using the definition of $\hat{\theta}_N$ and the law of large numbers, we show that

$$\liminf_{N\to+\infty} \frac{1}{N}\sum_{i=1}^N \log p(A_i \mid \hat{\theta}_N) \geq \mathbb{E}^*.$$

D) Finally, combining the two arguments above allows getting a contradiction if $\hat{\theta}_N \in K_\varepsilon$ for an infinite number of N. As a consequence, for all $\varepsilon > 0$, $\hat{\theta}_N \notin K_\varepsilon$ almost surely as $N \to +\infty$, which gives precisely $\delta(\hat{\xi}_N, \Xi_*^\infty) \to 0$.

□

5. Asymptotic normality of the MAP estimator

A consequence of Theorem 3.1 is that, if the empirical data distribution P corresponds to $p(A \mid \theta_0)$ for some $\theta_0 \in \Theta^{\mathrm{id}}$, we have $\Theta_*^{\mathrm{id}} = \{\theta_0\}$: thus, by Theorem 4.2, the MAP estimator over Θ^{id} converges almost surely to θ_0. A classical question is then to establish the rate of convergence of $\hat{\theta}_N$ toward θ_0, as well as the limiting asymptotic distribution. An answer for the more general case of M and Z-estimators is provided in chapter 5 of [54], which we restate with adapted notations:

Theorem 5.1 (Thems. 5.23 in [54])**.** *Let $m_\theta(A) = \log p(A \mid \theta)$. Assume that m_θ is a measurable function such that $\theta \mapsto m_\theta(A)$ is differentiable at θ_0 for P-almost every A with derivative $\nabla_A m_{\theta_0}(A)$. Assume that there exists a function $\overline{m}$ with $\mathbb{E}_{P(\mathrm{d}A)}[\overline{m}(A)^2] < +\infty$, such that, for every θ_1 and θ_2 in a neighborhood of θ_0:*

$$|m_{\theta_1}(A) - m_{\theta_2}(A)| \leq \overline{m}(A)\,\|\theta_1 - \theta_2\|\,. \tag{5.1}$$

Furthermore, assume that the map $\ell(\theta) = \mathbb{E}_{P(\mathrm{d}A)}[m_\theta(A)]$ *admits a second-order Taylor expansion at a point of maximum* θ_0 *with nonsingular symmetric second derivative* $V = \nabla^2\ell(\theta_0)$. *If*

$$\frac{1}{N}\sum_{i=1}^{N} m_{\hat{\theta}_N}(A_i) \geq \sup_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} m_\theta(A_i) - o_{\mathbb{P}}(1/N) \tag{5.2}$$

and $\hat{\theta}_N \to \theta_0$ *in probability, then*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = -V^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \nabla_\theta m_{\theta_0}(A_i) + o_{\mathbb{P}}(1)\,.$$

In particular, the sequence $\sqrt{N}(\hat{\theta}_N - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $V^{-1}\mathbb{E}_{P(\mathrm{d}A)}[\nabla_\theta m_\theta(A)\nabla_\theta m_\theta(A)^\top]V^{-1}$.

Remark 5.2. The notation $o_{\mathbb{P}}(1/N)$ designates a random variable Z_N such that $NZ_N \to 0$ in probability.

The most important condition in the theorem above is the non singularity of the Hessian matrix at θ_0. In general, $\nabla^2_\theta\ell(\theta)$ is impossible to compute for latent variable models, as it involves the Hessian of $\log p(A \mid \theta)$. However, the problem gets more tractable when the data distribution P corresponds to $p(A \mid \theta_0)$ for some $\theta_0 \in \Theta$. The Hessian matrix at θ_0 then classically rewrites as the Fisher information matrix $I(\theta_0)$ (see for instance Lemma 5.3 in [38]):

$$\nabla^2_\theta\ell(\theta_0) = \mathbb{E}_{p(A|\theta_0)}[\nabla^2_\theta \log p(A \mid \theta_0)] = -\mathbb{E}_{p(A|\theta_0)}\left[(\nabla_\theta \log p(A \mid \theta_0))(\nabla_\theta \log p(A \mid \theta_0))^\top\right] = -I(\theta_0)\,.$$

The non-singularity of the Fisher information matrix remains difficult to prove for general latent variable models. Some papers consider it as a base hypothesis to obtain the asymptotic normality, *e.g.* for Factor Analysis [5] or Hidden Markov Models [8]. In the latter case, the more recent work of [15] provided a condition to obtain the non-singularity of the Fisher information matrix. A recent result was obtained by [48] on the asymptotic normality of MLE for Gaussian graphical models and apply it to estimation from partial observations. In this specific case, the Fisher information has a simple closed form expression.

For the model considered here, no closed form expression can be expected, as the gradient of the log-likelihood writes with integrals on $\mathcal{V}_{np}$. Instead, we notice that, since the observation density $p(A \mid \theta_0)$ is continuous and $I(\theta_0)$ writes as the integral of $(\nabla_\theta \log p(A \mid \theta_0))(\nabla_\theta \log p(A \mid \theta_0))^\top$, the matrix will be non-singular if we can find $\dim(\theta_0)$ matrices A_i such that the related gradients $\nabla_\theta \log p(A_i \mid \theta_0)$ are linearly independent. This is formalized in the following lemma:

Lemma 5.3. *Let* $d = \dim(\theta_0)$. *If* $A_1, \ldots, A_d$ *matrices can be found such that the related* $\log p(A_i \mid \theta_0)$ *are independent, then* $I(\theta_0)$ *is positive definite.*

Proof. Let $x \in \mathbb{R}^d$. We have:

$$x^\top I(\theta_0)x = \mathbb{E}_{p(A|\theta_0)}\left[\langle x, \nabla_\theta \log p(A \mid \theta_0)\rangle^2\right] \geq 0\,.$$

If $x^\top I(\theta_0)x = 0$, then $\langle x, \nabla_\theta \log p(A \mid \theta_0)\rangle^2$ must be zero everywhere. Therefore, since $\theta \mapsto \log p(A \mid \theta)$ is infinitely smooth, x is orthogonal to all the gradients $\nabla_\theta \log p(A_i \mid \theta_0)$, and thus to their linear span, which covers the full space, which implies $x = 0$. As a consequence, $I(\theta_0)$ is positive definite. □

In the case of our model, it turns out that, although the expression of $\nabla_\theta \log p(A \mid \theta)$ is intractable, it simplifies as $\|A\|_F$ grows large. This simplification comes from the so-called Fisher identity:

$$\begin{aligned}\nabla_\theta \log p(A \mid \theta) &= \frac{1}{p(A \mid \theta)} \nabla_\theta p(A \mid \theta) \\ &= \frac{1}{p(A \mid \theta)} \iint_{\mathcal{V}_{np} \times \mathbb{R}^p} \nabla_\theta p(A, X, \lambda \mid \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \\ &= \iint_{\mathcal{V}_{np} \times \mathbb{R}^p} \nabla_\theta \log p(A, X, \lambda \mid \theta) \cdot p(X, \lambda \mid A, \theta) \, [\mathrm{d}X] \mathrm{d}\lambda \\ &= \mathbb{E}[\nabla_\theta \log p(A, X, \lambda) \mid A] \,,\end{aligned}$$

and the gradient rewrites as an expectation of the complete log-likelihood over the latent variables. Given the complete expression

$$p(A, X, \lambda \mid \theta) = \frac{1}{(2\pi)^{n^2/2} \sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2} \sigma_\lambda^p} \frac{1}{\mathcal{C}(F)} \exp\left[\langle F, X \rangle_F - \frac{1}{2\sigma_\varepsilon^2} \|A - \lambda \cdot X\|_F^2 - \frac{1}{2\sigma_\lambda^2} \|\lambda - \mu\|^2 \right] ,$$

this expectation yields for instance:

$$\nabla_F \log p(A \mid \theta) = -\nabla_F \log \mathcal{C}(F) + \mathbb{E}[X \mid A] \,.$$

As $\|A\|_F \to +\infty$, we show in the upcoming Proposition 5.4 that the eigenvector distribution of $(X \mid A)$ concentrates around the permutations of the p eigenvectors of A related to the p largest eigenvalues. As a consequence, $\nabla_F \log p(A \mid \theta)$ writes as the sum of $\nabla_F \log \mathcal{C}(F)$ and a linear combinations of all $(X_A)_{\pi,f}$, with X_A the $n \times p$ eigenvector matrix of A and $\pi \in S_p, f \in \{\pm 1\}^p$. However, although X_A can be chosen freely, the subsequent linear combination turns out to be hard to compute and manipulate, which ultimately prevents from getting an explicit expression for the gradient in F. The same phenomenon happens with the other gradients, which all rely on an expectation given A.

This observation motivates the main hypothesis for our normality result. We shall consider a **restricted variant of the main model** $\tilde{p}(A, X, \lambda)$, where the X variable is constrained to the set Δ_0 defined in equation (3): the density of X writes as

$$\tilde{p}(X \mid \theta) = \frac{\mathbf{1}_{X \in \Delta_0}}{\mathcal{C}'(F)} \exp\left(\langle X, F \rangle_F \right) , \tag{5.3}$$

with $\mathcal{C}'(F) = \int_{\Delta_0} \exp\left(\langle X, F \rangle_F \right) [\mathrm{d}X]$. This constraint does not fundamentally change the model in the limits where $s_i \to 0$ and $s_i \to +\infty$. For intermediate values, it truncates the other sections $\Delta_{\pi,f}$ of the vMF distribution, but does not change the support of the distribution of $\lambda \cdot X$, as it still covers the set of rank p matrices. The resemblance between p and $\tilde{p}$ is optimized when the maximum of $\langle X, F \rangle_F$ is reached in Δ_0, *i.e.* when choosing the normalized columns of F to be in Δ_0. We adopt this convention in the remainder of the section, as it also facilitates proving the identifiability of the restricted model.

In the remainder of this section, the notations $\ell(\theta)$, $\hat{\theta}_N$, ... refer to densities and estimators obtained for the restricted model. We also assume that the empirical data distribution is given by $\tilde{p}(A \mid \theta_0)$ rather than $p(A \mid \theta_0)$. With this restricted model, we have the following result:

Proposition 5.4. *Let $A \in \mathbb{R}^{n \times n}$ with rank at least p and distinct eigenvalues, and let $A_t = tA$ for $t \in \mathbb{R}$. On the restricted model with $X \in \Delta_0$, the distribution $(X \mid A = A_t)$ converges to the constant value X_A, with $X_A \in \Delta_0$ the list of eigenvectors of A corresponding to the p largest eigenvalues. In particular, $\mathbb{E}[X \mid A = A_t]$ converges to X_A.*

Proof. By definition, $\mathbb{E}[X \mid A = A_t]$ is the expectation of X w.r.t. the probability density proportional to

$$\mathbf{1}_{X \in \Delta_0} \exp\left(\langle X, F\rangle + \frac{1}{2\sigma_p^2} \left\|\mu_{tA,X}\right\|^2\right).$$

As $t \to +\infty$, the function $g_t(X) = \frac{1}{2\sigma_p^2} \left\|\mu_{tA,X}\right\|^2$ reaches its maximum to a point which converges to X_A. We have indeed:

$$g_t(X) = \left\|\mu_{tA,X}\right\|^2 = t^2 \frac{\sigma_p^2}{2\sigma_\varepsilon^4} \left\|A * X\right\|^2 + t \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \langle A * X, \mu\rangle + \frac{\sigma_p^2}{\sigma_\lambda^4} \left\|\mu^2\right\|.$$

By Proposition 2.2, $\left\|A * X\right\|^2$ is only at $X = X_A$ on Δ_0 (unicity is guaranteed as the eigenvalues of A are distinct). Let D a region of $\mathcal{V}_{np}$ with non-zero Haar measure such that $X_A \notin \overline{D}$ and let $\eta > 0$ such that if $X \in D$ then $\left\|A * X\right\|^2 \leq \left\|A * X_A\right\|^2 - 2\eta$. Let B_η be a neighborhood of X_A such that $\left\|A * X\right\| \geq \left\|A * X_A\right\| - \eta$. We have:

$$\begin{aligned}
\mathbb{P}(X \in D \mid A = A_t, \theta) &= \frac{\int_D \exp\left(\langle X, F\rangle + g_t(X)\right) [\mathrm{d}X]}{\int_{\mathcal{V}_{np}} \exp\left(\langle X, F\rangle + g_t(X)\right) [\mathrm{d}X]} \\
&\leq \frac{\int_D \exp\left(\langle X, F\rangle + g_t(X)\right) [\mathrm{d}X]}{\int_{B_\eta} \exp\left(\langle X, F\rangle + g_t(X)\right) [\mathrm{d}X]} \\
&\leq \frac{\int_D \exp\left(\langle X, F\rangle + t^2 \frac{\sigma_p^2}{2\sigma_\varepsilon^4}(\left\|A * X_A\right\|^2 - 2\eta) + t \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \left\|A * X_A\right\| \left\|\mu\right\|\right) [\mathrm{d}X]}{\int_{B_\eta} \exp\left(\langle X, F\rangle + t^2 \frac{\sigma_p^2}{2\sigma_\varepsilon^4}(\left\|A * X_A\right\|^2 - \eta) - t \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \left\|A * X_A\right\| \left\|\mu\right\|\right) [\mathrm{d}X]} \\
&\leq \exp\left(2 \left\|F\right\|_* - 2t^2 \eta \frac{\sigma_p^2}{2\sigma_\varepsilon^4} + 2t \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \left\|A * X_A\right\| \left\|\mu\right\|\right) \frac{|D|_{\mathcal{V}_{np}}}{|B_\eta|_{\mathcal{V}_{np}}} \\
&\xrightarrow{t \to +\infty} 0 .
\end{aligned}$$

Hence, by the Portmanteau theorem, the sequence of probability distributions $(X \mid A = A_t)$ converges in distribution to the constant X_A. □

Remark 5.5. Proposition 5.4 can be compared to the decreasing uncertainty on the normalized position $x/\left\|x\right\|$ of a point x going to infinity. If we used the complete model, the distribution of X would instead converge to the sum of Diracs at $(X_A)_{\pi,f}$ weighted by $p((X_A)_{\pi,f} \mid \theta)$.

With the result above, we can prove that $\dim \Theta$ linearly independent gradients $\nabla_\theta \log p(A \mid \theta_0)$ can be obtained.

Lemma 5.6. *The log-likelihood gradient $\nabla_\theta \log \tilde{p}(A \mid \theta)$ of the restricted model takes $\dim \Theta = np + p + 2$ linearly independent values.*

Proof. As explained above, the Fisher identity reminded here allows computing gradients as A grows large:

$$\nabla_\theta \log \tilde{p}(A \mid \theta) = \mathbb{E}[\nabla_\theta \log \tilde{p}(A, X, \lambda \mid \theta) \mid A].$$

In order to alleviate the notations, the expectations $\mathbb{E}$ below refer to the distribution $\tilde{p}(A, X, \lambda \mid \theta_0)$. Since we have:

$$\tilde{p}(A, X, \lambda \mid \theta) = \frac{1}{(2\pi)^{n^2/2} \sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2} \sigma_\lambda^p} \frac{\mathbf{1}_{X \in \Delta_0}}{\mathcal{C}'(F)} \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2} \left\|A - \lambda \cdot X\right\|_F^2 - \frac{1}{2\sigma_\lambda^2} \left\|\lambda - \mu\right\|^2\right],$$

we get for $(F, \mu, \sigma_\lambda, \sigma_\varepsilon)$:

1. $\nabla_F \log \tilde{p}(A \mid \theta) = -\nabla_F \log C(F) + \mathbb{E}[X \mid A]$,
2. $\nabla_\mu \log \tilde{p}(A \mid \theta) = \frac{1}{\sigma_\lambda^2}\left[\mathbb{E}[\lambda \mid A] - \mu\right]$,
3. $\nabla_{\sigma_\varepsilon^2} \log \tilde{p}(A \mid \theta) = -\frac{n^2}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4}\mathbb{E}[\|A - \lambda \cdot X\|_F^2 \mid A]$,
4. $\nabla_{\sigma_p^2} \log \tilde{p}(A \mid \theta) = -\frac{p}{2\sigma_\lambda^2} + \frac{1}{2\sigma_\lambda^4}\mathbb{E}[\|\lambda - \mu\|^2 \mid A]$.

Let $t \in \mathbb{R}$, consider the matrix $A_t = tA$ and denote $X_A \in \Delta_0$ the matrix of eigenvectors of A for the p largest eigenvalues. The expressions above simplify as $t \to +\infty$:

1. *For F*: Proposition 5.4 gives for A with p distinct non-zero leading eigenvalues:

$$\nabla_F \log \tilde{p}(A_t \mid \theta) \to -\nabla_F \log C(F) + X_A\,.$$

2. *For μ*: as seen in Section 2.3, $(\lambda \mid X, A_t) \sim \mathcal{N}(\mu_{A_t X}, \sigma_p^2)$, so that we have

$$\begin{aligned}
\frac{1}{t}\nabla_\mu \log \tilde{p}(A_t \mid \theta) &= \frac{1}{t}\frac{1}{\sigma_\lambda^2}\left[\mathbb{E}[\mathbb{E}[\lambda \mid X, A_t] \mid A] - \mu\right] \\
&= \frac{1}{t}\frac{1}{\sigma_\lambda^2}\left[\mathbb{E}[\mu_{A_t X} \mid A_t] - \mu\right] \\
&= \frac{1}{t}\frac{1}{\sigma_\lambda^2}\left[\mathbb{E}\left[\frac{\sigma_p^2}{\sigma_\varepsilon^2} tA * X + \frac{\sigma_p^2}{\sigma_\lambda^2}\mu \mid A_t\right] - \mu\right] \\
&\xrightarrow{t\to+\infty} \frac{\sigma_p^2}{\sigma_\lambda^2\sigma_\varepsilon^2} A * X_A\,.
\end{aligned}$$

3. *For σ_ε^2*: similarly, we get

$$\begin{aligned}
\frac{1}{t^2}\nabla_{\sigma_\varepsilon^2} \log \tilde{p}(A_t \mid \theta) &= -\frac{n^2}{2t^2\sigma_\varepsilon^2} + \frac{1}{2t^2\sigma_\varepsilon^4}\mathbb{E}[\mathbb{E}[\|A_t - \lambda \cdot X\|_F^2 \mid X, A_t] \mid A_t] \\
&= -\frac{n^2}{2t^2\sigma_\varepsilon^2} + \frac{1}{2t^2\sigma_\varepsilon^4}\mathbb{E}[\mathbb{E}[\|tA\|_F^2 - 2\langle \lambda, tA * X\rangle + \|\lambda\|^2 \mid X, A_t] \mid A_t] \\
&= -\frac{n^2}{2t^2\sigma_\varepsilon^2} + \frac{1}{2t^2\sigma_\varepsilon^4}\mathbb{E}[\|tA\|_F^2 - 2\langle \mu_{A_t X}, tA * X\rangle + \|\mu_{A_t X}\|^2 + p\sigma_p^2 \mid A_t] \\
&\xrightarrow{t\to+\infty} \frac{1}{2\sigma_\varepsilon^4}\left[\|A\|_F^2 - 2\frac{\sigma_p^2}{\sigma_\varepsilon^2}\|A * X_A\|^2 + \frac{\sigma_p^4}{\sigma_\varepsilon^4}\|A * X_A\|^2\right]\,.
\end{aligned}$$

4. *For σ_λ^2*:

$$\begin{aligned}
\frac{1}{t^2}\nabla_{\sigma_p^2} \log \tilde{p}(A_t \mid \theta) &= -\frac{p}{2t^2\sigma_\lambda^2} + \frac{1}{2t^2\sigma_\lambda^4}\mathbb{E}[\|\lambda - \mu\|^2 \mid A_t] \\
&= -\frac{p}{2t^2\sigma_\lambda^2} + \frac{1}{2t^2\sigma_\lambda^4}\mathbb{E}[\mathbb{E}[\|\lambda - \mu\|^2 \mid X, A_t] \mid A_t] \\
&= -\frac{p}{2t^2\sigma_\lambda^2} + \frac{1}{2t^2\sigma_\lambda^4}\mathbb{E}[\|\mu_{A_t X} - \mu\|^2 + p\sigma_p^2 \mid A_t] \\
&\xrightarrow{t\to+\infty} \frac{\sigma_p^4}{2\sigma_\lambda^4\sigma_\varepsilon^4}\|A * X_A\|^2\,.
\end{aligned}$$

In these expression, $A * X_A = (\lambda_1, \dots, \lambda_p)$ is the p leading eigenvalues of A, $\|A * X_A\|_F^2 = \lambda_1^2 + \cdots + \lambda_p^2$ and $\|A\|_F^2 = \lambda_1^2 + \cdots + \lambda_n^2$. In the remainder of the proof, we call these asymptotic rescaled values *limit gradients*. Using the formulas above, we derive the following limit gradients.

- Taking $X \in \mathcal{V}_{np}$, we consider the limit gradient for $\mu \cdot X^i$. Up to factors t and t^2 which do not affect the linear independence, the result is:

$$\begin{pmatrix} -\nabla_F \log C(F) + X \\ \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \mu \\ \frac{1}{2\sigma_\varepsilon^4} \left[1 - 2\frac{\sigma_p^2}{\sigma_\varepsilon^2} + \frac{\sigma_p^4}{\sigma_\varepsilon^4}\right] \|\mu\|^2 \\ \frac{\sigma_p^4}{2\sigma_\lambda^4 \sigma_\varepsilon^4} \|\mu\|^2 \end{pmatrix}.$$

- Taking a vector $\lambda \in \mathbb{R}^p$ such that $\|\lambda\| = \|\mu\|$, we consider matrices of the form $\lambda \cdot I_{np}$. The resulting limit gradient at $t \to +\infty$ is:

$$\begin{pmatrix} -\nabla_F \log C(F) + I_{np} \\ \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \lambda \\ \frac{1}{2\sigma_\varepsilon^4} \left[1 - 2\frac{\sigma_p^2}{\sigma_\varepsilon^2} + \frac{\sigma_p^4}{\sigma_\varepsilon^4}\right] \|\mu\|^2 \\ \frac{\sigma_p^4}{2\sigma_\lambda^4 \sigma_\varepsilon^4} \|\mu\|^2 \end{pmatrix}.$$

- We consider the matrices $A = \mathrm{Diag}(\mu_1, \dots, \mu_p, \alpha, \dots, \alpha)$ with $0 < \alpha < \min_i |\mu_i|$. The resulting limit gradient is:

$$\begin{pmatrix} -\nabla_F \log C(F) + I_{np} \\ \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \mu \\ \frac{1}{2\sigma_\varepsilon^4} \left[1 - 2\frac{\sigma_p^2}{\sigma_\varepsilon^2} + \frac{\sigma_p^4}{\sigma_\varepsilon^4}\right] \|\mu\|^2 + \frac{1}{2\sigma_\varepsilon^4} \alpha^2 (n-p)^2 \\ \frac{\sigma_p^4}{2\sigma_\lambda^4 \sigma_\varepsilon^4} \|\mu\|^2 \end{pmatrix}.$$

- Finally, we take the matrix $\mu \cdot I_{np}$. The resulting limit gradient is:

$$\begin{pmatrix} -\nabla_F \log C(F) + I_{np} \\ \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} \mu \\ \frac{1}{2\sigma_\varepsilon^4} \left[1 - 2\frac{\sigma_p^2}{\sigma_\varepsilon^2} + \frac{\sigma_p^4}{\sigma_\varepsilon^4}\right] \|\mu\|^2 \\ \frac{\sigma_p^4}{2\sigma_\lambda^4 \sigma_\varepsilon^4} \|\mu\|^2 \end{pmatrix}.$$

Subtracting the limit gradient at $\mu \cdot I_{np}$, we can get linear combinations of gradients arbitrarily close to any vector of the forms:

$$\begin{pmatrix} X - I_{np} \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{\sigma_p^2}{\sigma_\lambda^2 \sigma_\varepsilon^2} (\lambda - \mu) \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2\sigma_\varepsilon^4} \alpha^2 (n-p)^2 \\ 0 \end{pmatrix}.$$

With $X \in \mathcal{V}_{np}$ and $\|\lambda\| = \|\mu\|$. We can now use Lemma C.8, which states that the elements of the form $X - I_{np}$ span $\mathbb{R}^{np}$. Similarly, elements of the form $\lambda - \mu$ span $\mathbb{R}^p$: taking $\lambda = -\mu$, the space contains -2μ and thus μ. Hence it also contains all elements λ with norm $\|\mu\|$, which can be rescaled to get the entire space. As a consequence, the three vector families above span the entire linear hyperplan $\mathbb{R}^{np} \times \mathbb{R}^p \times \mathbb{R} \times \{0\}$. Furthermore, the limit gradient at $\mu \cdot I_{np}$ has a non zero last coordinate and does not belong to this linear hyperplan. As a consequence, the set of all limit gradients spans the entire gradient space. Therefore, the set of all gradients, which gets arbitrarily close to limit gradients, also spans the entire gradient space. Finally, we can thus find $d = np + p + 2$ matrices $(A_i)_{i=1}^d$ such that the vectors $(\nabla_\theta \log \tilde{p}(A_i \mid \theta))_{i=1}^d$ are linearly independent. □

With the result of Lemma 5.6, we can now obtain the asymptotic normality result.

Theorem 5.7. *Assume that the empirical data distribution is given by the restricted model for some parameter $\theta_0 \in \Theta^{\mathrm{id}}$. Then the MAP estimator $\hat{\theta}_N$ over Θ^{id} for the restricted model converges almost surely to θ_0, and $\hat{\theta}_N$ is asymptotically normal:*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathscr{D}} \mathcal{N}(0, I(\theta_0)^{-1}) .$$

Proof. As verified in Lemma C.9, the restricted model is identifiable on Θ^{id}, so that the only maximizer of $\ell(\theta)$ over Θ^{id} is θ_0. The proof of the consistency Theorem 4.2 adapts without hurdle to the restricted model, proving that $\hat{\theta}_N$ converges to θ_0 almost surely.

We can now check the conditions to apply Theorem 5.1. Since $\theta \mapsto \ell(\theta)$ is smooth over Θ, it admits a second-order Taylor expansion at θ_0, and Lemma 5.3 combined with Lemma 5.6 ensures that the Hessian matrix at this point is non singular. Lemma C.10 shows that the Lipschitz condition (5.1) is satisfied by $\log \tilde{p}(A \mid \theta)$. Finally, condition (5.2) is satisfied, as the MAP estimator is such that:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \log p(A_i \mid \hat{\theta}_N) &= \frac{1}{N}\sum_{i=1}^N \log \tilde{p}(A_i \mid \hat{\theta}_N) + \frac{1}{N}\log p(\hat{\theta}_N) - \frac{1}{N}\log p(\hat{\theta}_N) \\
&\geq \sup_{\theta\in\Theta}\left(\frac{1}{N}\sum_{i=1}^N \log \tilde{p}(A_i \mid \theta) + \frac{1}{N}\log p(\theta)\right) - \frac{1}{N}\log p(\hat{\theta}_N) \\
&\geq \sup_{\theta\in\Theta}\left(\frac{1}{N}\sum_{i=1}^N \log \tilde{p}(A_i \mid \theta)\right) - \underbrace{\frac{1}{N}(\sup_{\theta\in\Theta}\log p(\theta) + \log p(\hat{\theta}_N))}_{o_{\mathbb{P}}(1/N)} .
\end{aligned}$$

Theorem 5.1 thus applies, and grants the convergence in distribution of $\sqrt{N}(\hat{\theta}_N - \theta_0)$ to the centered Gaussian with covariance $[\nabla^2_\theta \ell(\theta_0)]^{-1}\mathbb{E}[(\nabla_\theta \log p(A \mid \theta_0))(\nabla_\theta \log p(A \mid \theta_0))^\top][\nabla^2_\theta \ell(\theta_0)]^{-1} = I(\theta_0)^{-1}$. □

6. Conclusion

This papers provides theoretical guarantees for the estimation of the eigenvalue and eigenvector distributions of the adjacency matrix decomposition model introduced in [43]. The considered model is identifiable, its MAP estimator exists and converges almost surely to the points minimizing the Kullback-Leibler divergence to the empirical data distribution. By considering an alternate restricted model, we obtain the usual $1/\sqrt{N}$ convergence rate and the asymptotic normality of the MAP estimator using the theory of [54]. Our results show that asymptotic statistical analysis can be performed on manifold-valued latent variable models to obtain classical guarantees. Arguments similar to those we presented should allow obtaining results in related models where little theoretical work has been done. State-space models on Stiefel and Grassman manifolds [14], eigendecomposition models for a single network matrix [26] or mixture models [2] could lend themselves to such an analysis.

The model considered here, as most of the literature on statistics for Stiefel manifolds, is estimated with MLE or MAP. Recently, [46] proposed a Bayesian framework for von Mises-Fisher distributions which allows computing the posterior distribution of F given observations of X. An interesting question would be to analyze the behavior of this posterior distribution in a hierarchical model where X is a latent variable, in a direction similar to the works of [41] and [20].

Finally, another important question on the model we studied is the analysis of its estimation error. In practice, [43] rely on a variant of the EM algorithm to estimate the model parameters. EM-based methods are known to produce local maxima of the likelihood, which prevents from getting a rigorous theoretical analysis of the estimation error. However, even assuming that no local maximum is found, the E-step of the EM algorithm behaves in an undesirable way, as the conditional distribution of (X, λ) given A is multimodal (one mode per permutation and change of sign for the columns of X). This conditional distribution yields a very low vMF concentration far from the real one, as the samples X are spread over the manifold. A heuristic thus has to be employed in practice to ensure that X stays close to Δ_0, and get a better estimate of the MAP. This question will be part of our future work.

Acknowledgements. The authors would like to thank Éloïse Berthier, Guillaume Dalle, Maxime Godin and Pierre Marion (ordered by first and last names) for their helpful discussions, comments and suggestions.

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 (Euro-POND) and No. 826421 (TVB-Cloud). It was also funded by in part by the program "Investissements d'avenir" ANR-10-IAIHU-06 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).

Appendix A. Reminders on the Stiefel manifold

The Stiefel manifold is the space of $n \times p$ matrices X such that $X^\top X = I_p$. It has inherits a Riemannian manifold either as a submanifold of $\mathbb{R}^{n\times p}$ or as a quotient of $O_n(\mathbb{R})$ by $O_{n-p}(\mathbb{R})$. The equivalence between both corresponds to mapping X to the set of orthogonal matrices $(X, X_\perp)$, with $X_\perp$ completing X into an orthonormal basis. The induced metrics are called respectively the Euclidean metric and the canonical metric. The notions exposed here are introduced with great detail and clarity in [21].

Tangent space. Let $X \in \mathcal{V}_{np}$. The relation satisfied by matrices H in the tangent space $T_X\mathcal{V}_{np}$ is obtained by differentiating the relation $X^\top X = I_p$: this yields $H^\top X + X^\top H = 0$. This definition of the tangent space can be made more explicit by writing H under the form $H = (X, X_\perp)\binom{A}{B} = XA + X_\perp B$, with $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{(n-p)\times p}$. Such a decomposition is always possible, as $(X, X_\perp)$ is an orthogonal matrix. Using this expression in the equation of the tangent space yields $A^\top = -A$. As a consequence, $T_X\mathcal{V}_{np}$ can be defined as the set of $XA + X_\perp B$, with A a skew-symmetric matrix.

Function gradients. Given a function $f : \mathcal{V}_{np} \to \mathbb{R}$, the manifold gradient of f at X is the matrix-valued function $\nabla_{\mathcal{V}} f$. It is defined by the property that, if X_t is a smooth curve on $\mathcal{V}_{np}$ with $X_0 = X$ and $\dot{X}_0 = H \in T_X\mathcal{V}_{np}$, then $\frac{\mathrm{d}f(X_t)}{\mathrm{d}t}(0) = \langle \nabla_{\mathcal{V}} f(X), H\rangle_X$. Here, $\langle\cdot,\cdot\rangle_X$ denotes the inner product on $T_X\mathcal{V}_{np}$ of the Riemannian manifold structure of $\mathcal{V}_{np}$. Note that the definition of the gradient depends on the metric choice, which is worth mentioning as this choice varies from one paper to another.

An important case is the situation where f can be extended to the whole matrix space. This allows computing the Euclidean gradient of f. Then, depending on the metric choice, explicit formulas are available for the manifold gradient. With respect to the canonical metric, we have [21]:

$$\nabla_{\mathcal{V}} f(X) = \nabla f(X) - X\nabla f(X)^\top X\,.$$

Cayley transform. In Riemannian geometry, the standard way of mapping elements of $T_X\mathcal{V}_{np}$ to the base manifold $\mathcal{V}_{np}$ is the Riemannian exponential map, defined with geodesic equations. Although explicit formulas

are available for the exponential map on $\mathcal{V}_{np}$ (see again [21]), they rely on matrix exponential and little is known on the properties of the inverse mapping.

In contrast, the Cayley transform C_X behaves better in that regard. It also sends elements from $T_X\mathcal{V}_{np}$ to $\mathcal{V}_{np}$ and behaves similarly to the exponential map close to X, in the sense that

$$C_X(H) = X + 2H + o(\|H\|_X)\,.$$

Denoting $K = HX^\top - XH^\top$, the Cayley transform at X is defined by:

$$C_X(H) = (I_n + K)(I_n - K)^{-1}X \in \mathcal{V}_{np}\,.$$

C_X was studied in more detailed for $X = I_{np}$ in [29]. In practice, the Cayley transform is used in optimization to perform gradient descent [24], as it allows projecting the descent direction $\nabla_{\mathcal{V}} f(X)$ onto the manifold and requires only simple linear algebra computations. We prefer it to the exponential map because it has a simple expression, is invertible, and covers the entire manifold apart from a set with measure zero.

Appendix B. Proof of the consistency of the MAP estimator

We define

$$\mathbb{E}^* = \sup_{\theta\in\Theta^\infty} \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta)]\,.$$

The proof relies on the Alexandrov compactification $\overline{\Theta^\infty}$ of Θ^∞, which adds an infinity point for the coordinates σ_ε (for the cases $\sigma_\varepsilon \in \{0, +\infty\}$), σ_λ (for the case $\sigma_\lambda = +\infty$) and μ (for all the cases where $\|\mu\| = +\infty$).

Part A. We prove that, for all $\theta_\infty \in \overline{\Theta^\infty}$ such that $\delta(\xi(\theta_\infty), \Xi_*^\infty) \geq \varepsilon$, there exists an open neighborhood $\mathcal{U} \subset \overline{\Theta^\infty}$ of θ_∞ such that

$$\mathbb{E}_{P(\mathrm{d}A)}\left[\sup_{\theta\in\mathcal{U}\cap\Theta^\infty} \log p(A \mid \theta)\right] < \mathbb{E}^*\,. \tag{B.1}$$

Let $\mathcal{U}_h$ be a decreasing sequence of open sets such that $\cap_{h\geq 0}\mathcal{U}_h = \{\theta_\infty\}$, and let

$$f_h(A) = \sup_{\theta\in\mathcal{U}_h\cap\Theta^\infty} \log p(A \mid \theta)\,.$$

Two cases arise:

1. If $\theta_\infty \in \Theta^\infty$. Since $\theta \mapsto \log p(A \mid \theta)$ is continuous, we have:

$$f_h(A) \xrightarrow{h\to+\infty} \log p(A \mid \theta_\infty)\,.$$

And the sequence $f_h(A)$ is decreasing for every A. Furthermore, Lemma C.6 ensures that the sequence is bounded from above (with the upper bound obtained by taking the whole space for $\mathcal{U}$). Hence the monotone convergence theorem applies, and we get:

$$\lim_{h\to+\infty} \mathbb{E}_{P(\mathrm{d}A)}[f_h(A)] = \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta_\infty)] < \mathbb{E}^*$$

since $\theta_\infty \notin \Theta_*^\infty$. Therefore, it is sufficient to take h large enough to have equation (B.1) satisfied.

2. If $\theta_\infty \notin \Theta^\infty$, *i.e.* the variance parameters $(\sigma_\lambda, \sigma_\varepsilon)$ take extreme values, we prove by contradiction that $\lim_{h\to\infty} f_h(A) = -\infty$ a.s. Let us assume that there exists a measurable set $E \in \mathscr{B}(\mathbb{R}^{n\times n})$ such that $\mathbb{P}(A \in E) > 0$ and, for all $A \in E$, $\inf_h f_h(A) > -\infty$. Since $f_h(A)$ is decreasing for every A in E, the infimum is reached at infinity.
For each h, let $(\theta_{h,m}) \in (\mathcal{U}_h \cap \Theta^\infty)^{\mathbb{N}}$ be a sequence such that:

$$\lim_{m\to+\infty} \log p(A \mid \theta_{h,m}) = \sup_{\theta\in\mathcal{U}_h\cap\Theta^\infty} \log p(A \mid \theta) = f_h(A) \geq \inf_h f_h(A)\,.$$

By taking for each h a value of $\theta_{h,m}$ h^{-1}-close to the function's limit, we obtain a sequence $\theta_h \in (\Theta^\eta)^{\mathbb{N}}$ such that $\theta_h \to \theta_\infty$ and

$$\liminf_{h\to+\infty} \log p(A \mid \theta_h) \geq \inf_h f_h(A) > -\infty\,.$$

Since $\theta_\infty \notin \Theta^\infty$, we have $\sigma_\lambda^\infty = +\infty$, $\sigma_\varepsilon^\infty = 0$ or $\sigma_\varepsilon^\infty = +\infty$. Hence this contradicts Lemma C.7. Therefore, $P(\mathrm{d}A)$-almost surely, $f_h(A) \to -\infty$. We can again apply Lemma C.6 and use the monotone convergence theorem, which grants

$$\lim_{h\to+\infty} \mathbb{E}_{P(\mathrm{d}A)}[f_h(A)] = -\infty < \mathbb{E}^*\,.$$

Therefore, whether θ_∞ is in Θ^∞ or not, there exists an open neighborhood $\mathcal{U}$ of θ_∞ such that

$$\mathbb{E}_{P(\mathrm{d}A)}\left[\sup_{\theta\in\mathcal{U}\cap\Theta^\infty} \log p(A \mid \theta)\right] < \mathbb{E}^*\,.$$

Part B. Define K_ε as:

$$K_\varepsilon = \{\theta \in \overline{\Theta^\infty} \mid \delta(\xi(\theta), \Xi_*^\infty) \geq \varepsilon\}\,.$$

By definition of the Alexandrov compactification and by the continuity of δ, K_ε is a compact set, hence we can find a finite open cover $(\mathcal{U}_{h\leq H})$ of it, where each $\mathcal{U}_h$ satisfies equation (B.1). Let $N \in \mathbb{N}$. For all $\theta \in K_\varepsilon$:

$$\sup_{\theta\in K_\varepsilon\cap\Theta^\infty} \sum_{i=1}^N \log p(A_i \mid \theta) \leq \sup_{1\leq h\leq H} \sum_{i=1}^N \sup_{\theta\in\mathcal{U}_h\cap\Theta^\infty} \log p(A_i \mid \theta)\,.$$

Since the observations A_i are independent (**H2**), by the law of large numbers and by the definition of $\mathcal{U}_h$:

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{i=1}^N \sup_{\theta\in\mathcal{U}_h\cap\Theta^\infty} \log p(A_i \mid \theta) < \mathbb{E}^*\,.$$

Hence

$$\limsup_{N\to+\infty} \sup_{\theta\in K_\varepsilon\cap\Theta^\infty} \frac{1}{N} \sum_{i=1}^N \log p(A_i \mid \theta) < \mathbb{E}^*\,.$$

Part C. For each $\theta^* \in \Theta_*^\infty$, the law of large numbers gives $\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^N \log p(A_i \mid \theta^*) = \mathbb{E}^*$. Let θ^k be a sequence of parameters with finite values such that $\theta^k \to \theta^*$. Then we have, for all k:

$$p(A^N \mid \hat{\theta}_N) = \frac{p(\hat{\theta}_N \mid A^N)p(A^N)}{p(\hat{\theta}_N)} \geq \frac{p(\theta^k \mid A^N)p(A^N)}{p(\hat{\theta}_N)} = \frac{p(A^N \mid \theta^k)p(\theta^k)}{p(\hat{\theta}_N)}$$
$$\sum_{i=1}^{N} \log p(A_i \mid \hat{\theta}_N) \geq \sum_{i=1}^{N} \log p(A_i \mid \theta^k) + (\log p(\theta^k) - \log p(\hat{\theta}_N)) .$$

And, since $\log p(\theta)$ is upper bounded by M, this leads to:

$$\frac{1}{N}(\log p(\theta^k) - \log p(\hat{\theta}_N)) \geq \frac{1}{N} \log \frac{p(\theta^k)}{M} .$$

Hence $\liminf_{N\to+\infty} \frac{1}{N}(\log p(\theta^k) - \log p(\hat{\theta}_N)) \geq 0$ and, almost surely, for all k:

$$\liminf_{N\to+\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(A_i \mid \hat{\theta}_N) \geq \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta^k)] .$$

And, from the continuity granted by Lemma C.7, $\lim_{k\to+\infty} \mathbb{E}_{P(\mathrm{d}A)}[\log p(A \mid \theta^k)] = \mathbb{E}^*$, so that almost surely:

$$\liminf_{N\to+\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(A_i \mid \hat{\theta}_N) \geq \mathbb{E}^* . \tag{B.2}$$

Part D. Finally, if $\hat{\theta}_N \in K_\varepsilon$ for all $N \in \mathbb{N}$, then:

$$\sum_{i=1}^{N} \log p(A_i \mid \hat{\theta}_N) \leq \sup_{\theta \in K_\varepsilon \cap \Theta^\infty} \sum_{i=1}^{N} \log p(A_i \mid \theta) .$$

Which implies almost surely:

$$\limsup_{N\to+\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(A_i \mid \hat{\theta}_N) \leq \limsup_{N\to+\infty} \sup_{\theta \in K_\varepsilon \cap \Theta^\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(A_i \mid \theta) < \mathbb{E}^* . \tag{B.3}$$

Which directly contradicts the point of part C. Furthermore, if $\hat{\theta}_N \in K_\varepsilon$ is only true up to a subsequence, the argument remains valid, as all the limits in this proof as $N \to +\infty$ can be taken with respect to any extracted subsequence chosen a *priori*. Therefore, and since we proved in Theorem 4.1 that $\hat{\theta}_N$ is finite and $\{\theta \in \Theta^\infty \mid \delta(\xi(\theta), \Xi_*^\infty) \geq \varepsilon\} \subset K_\varepsilon$, $\delta(\hat{\xi}_N, \Xi_*^\infty) \geq \varepsilon$ as $N \to +\infty$ almost surely, for all $\varepsilon > 0$. As a consequence, $\delta(\hat{\xi}_N, \Xi_*^\infty) \to 0$ almost surely.

Appendix C. Lemmas

In order to state the required lemmas, let us denote

$$A * X = (x_k^\top A x_k)_{k=1}^p, \quad \frac{1}{\sigma_p^2} = \frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\lambda^2} \quad \text{and} \quad \mu_{AX} = \sigma_p^2 \left[\frac{1}{\sigma_\varepsilon^2} A * X + \frac{1}{\sigma_\lambda^2} \mu \right] . \tag{C.1}$$

We have the following lemma.

Lemma C.1. *The model likelihood rewrites as*

$$p(A \mid \theta) = \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{\sigma_p^p}{\sigma_\lambda^p} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2\right) \mathbb{E}_X\left[\exp\left(\frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right)\right], \tag{C.2}$$

where $\mathbb{E}_X$ denotes the expectation taken with respect to X only.

Proof. From the definition of our model,

$$\begin{aligned} p(A \mid \theta) &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} p(A \mid X, \lambda, \theta) p(X \mid \theta) p(\lambda \mid \theta)\,[\mathrm{d}X]\,\mathrm{d}\lambda \\ &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2}\sigma_\lambda^p} \frac{1}{C(F)} \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\|A - \lambda\cdot X\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\lambda - \mu\|^2\right][\mathrm{d}X]\,\mathrm{d}\lambda\,. \end{aligned}$$

Furthermore:

$$\begin{aligned} \|A - \lambda\cdot X\|_F^2 &= \|A\|_F^2 - 2\sum_{k=1}^p \lambda_k \left\langle A, x_k^\top x_k\right\rangle_F + \sum_{k,l=1}^p \lambda_k\lambda_l \left\langle x_k^\top x_k, x_l^\top x_l\right\rangle_F \\ &= \|A\|_F^2 - 2\sum_{k=1}^p \lambda_k (x_k^\top A x_k) + \sum_{k,l=1}^p \lambda_k\lambda_l\delta_{kl} \\ &= \|A\|_F^2 - 2\langle\lambda, A * X\rangle + \|\lambda\|^2\,. \end{aligned}$$

So that, using $\frac{1}{\sigma_p^2} = \frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\lambda^2}$:

$$\begin{aligned} p(A \mid \theta) &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2}\sigma_\lambda^p} \frac{1}{C(F)} \\ &\qquad \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\left(\|A\|_F^2 - 2\langle\lambda, A*X\rangle + \|\lambda\|^2\right) - \frac{1}{2\sigma_\lambda^2}\left(\|\lambda\|^2 - \langle\lambda,\mu\rangle + \|\mu\|^2\right)\right][\mathrm{d}X]\,\mathrm{d}\lambda \\ &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2}\sigma_\lambda^p} \frac{1}{C(F)} \\ &\qquad \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2 + \left\langle\lambda, \frac{1}{\sigma_\varepsilon^2}A*X + \frac{1}{\sigma_\lambda^2}\mu\right\rangle - \frac{1}{2}\left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\lambda^2}\right)\|\lambda\|^2\right][\mathrm{d}X]\,\mathrm{d}\lambda \\ &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2}\sigma_\lambda^p} \frac{1}{C(F)} \\ &\qquad \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2 + \frac{1}{\sigma_p^2}\left\langle\lambda, \sigma_p^2\left[\frac{1}{\sigma_\varepsilon^2}A*X + \frac{1}{\sigma_\lambda^2}\mu\right]\right\rangle - \frac{1}{2\sigma_p^2}\|\lambda\|^2\right][\mathrm{d}X]\,\mathrm{d}\lambda\,. \end{aligned}$$

Let $\mu_{AX} = \sigma_p^2\left[\frac{1}{\sigma_\varepsilon^2}A*X + \frac{1}{\sigma_\lambda^2}\mu\right]$. We get:

$$\begin{aligned} p(A \mid \theta) &= \iint_{\mathcal{V}_{np}\times\mathbb{R}^p} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{(2\pi)^{p/2}\sigma_\lambda^p} \frac{1}{C(F)} \\ &\qquad \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2 + \frac{1}{\sigma_p^2}\langle\lambda, \mu_{AX}\rangle - \frac{1}{2\sigma_p^2}\|\lambda\|^2 \pm \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right][\mathrm{d}X]\,\mathrm{d}\lambda \end{aligned}$$

$$= \int_{\mathcal{V}_{np}} \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{C(F)} \frac{\sigma_p^p}{\sigma_\lambda^p} \exp\left[\langle F, X\rangle - \frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2 + \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right] \underbrace{\int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2}\sigma_p^p} \exp\left[-\frac{1}{2\sigma_p^2}\|\lambda - \mu_{AX}\|^2\right] \mathrm{d}\lambda}_{=1} [\mathrm{d}X] \,.$$

Thus we obtain the result:

$$\begin{aligned} p(A \mid \theta) &= \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{1}{C(F)} \frac{\sigma_p^p}{\sigma_\lambda^p} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2\right] \int_{\mathcal{V}_{np}} \exp\left[\langle F, X\rangle + \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right] [\mathrm{d}X] \\ &= \frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}} \frac{\sigma_p^p}{\sigma_\lambda^p} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2\right] \mathbb{E}_X\left[\exp\left[\frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right]\right] \,. \end{aligned}$$

□

Lemma C.2 (Bound on the log-likelihood)**.** *For all matrix A and parameters θ,*

$$\log p(A \mid \theta) \leq -\frac{n^2}{2}\log(2\pi) - (n^2 - p)\log\sigma_\varepsilon - p\log\sigma_\lambda \,.$$

Proof. Using $\frac{1}{\sigma_p^2} = \frac{1}{\sigma_\lambda^2} + \frac{1}{\sigma_\varepsilon}$, Jensen's inequality gives $\|\mu_{AX}\|^2 \leq \frac{1}{\sigma_\varepsilon^2}\|A * X\|^2 + \frac{1}{\sigma_\lambda^2}\|\mu\|^2$. Proposition 2.2 implies, for $\mu = 0$, that $\|A * X\| \leq \|A\|_F$. Hence, for all $X \in \mathcal{V}_{np}$, $\|\mu_{AX}\|^2 \leq \frac{1}{\sigma_\varepsilon^2}\|A\|_F^2 + \frac{1}{\sigma_\lambda^2}\|\mu\|^2$. This bound yields in the expression of Lemma C.1:

$$\log p(A \mid \theta) \leq -\frac{n^2}{2}\log(2\pi) - n^2\log\sigma_\varepsilon + p\log\sigma_p - p\log\sigma_\lambda \,.$$

Furthermore, from the definition of σ_p we have $\sigma_p \leq \sigma_\varepsilon$, which gives the desired bound:

$$\log p(A \mid \theta) \leq -\frac{n^2}{2}\log(2\pi) - (n^2 - p)\log\sigma_\varepsilon - p\log\sigma_\lambda \,.$$

□

Lemma C.3 (Continuity of $p(A \mid \theta)$ over Θ^∞)**.** *The likelihood $p(A \mid \theta)$ extends continuously when $s_i = +\infty$ for a subset I of r indices or when $\sigma_\lambda = 0$. In other words, $\theta \mapsto p(A \mid \theta)$ is continuous over Θ^∞. With the following notations*

- *J is the complementary of I in $\{1, \ldots, p\}$,*
- *X_I is the $n \times r$ matrix $(x_{i_1}, \ldots, x_{i_r})$,*
- *$M_I^\perp$ denotes an $n \times (n-r)$ matrix such that $M_I^\top M_I^\perp = 0$ and $M_I^\perp \in \mathcal{V}_{n,n-r}$.*
- *$q_{\mathrm{vMF}}(X, F)$ is the von Mises-Fisher density with parameter F and variable X,*
- *$F = M\mathrm{Diag}(s)$ is the parameterization of F described in Section 2,*

the extension reads:

$$p(A \mid \theta) = \begin{cases} \int_{\mathcal{V}_{n-r,p-r}} q_{\mathrm{vMF}}(Y; (M_I^\perp)^\top F_J) p(A \mid X = (M_I, M_I^\perp Y), \lambda = \mu, \theta) \, [\mathrm{d}Y] & \text{if } \sigma_\lambda = 0 \\ \iint_{\mathcal{V}_{n-r,p-r}\times\mathbb{R}^p} q_{\mathrm{vMF}}(Y; (M_I^\perp)^\top F_J) p(A \mid X = (M_I, M_I^\perp Y), \lambda = \mu, \theta) p(\lambda \mid \theta) \, [\mathrm{d}Y]\mathrm{d}\lambda & \text{otherwise.} \end{cases}$$

If all latent variables are constant, this yields the Gaussian likelihood $A \sim \mathcal{N}(\mu \cdot M, \sigma_\varepsilon^2 I_{n\times n})$.

Proof. For notational convenience, we suppose that I is composed of the first r indices of $\{1,\dots,p\}$. Let $\{X_I = M_I\} \subset \mathcal{V}_{np}$ be the set of values of X such that X and M match on the columns of I. The continuity at infinity comes from the expression:

$$\begin{aligned} p(A \mid \theta) &= \mathbb{E}_{X,\lambda}[p(A \mid X, \lambda, \theta)] \\ &= \mathbb{E}_{X_I,\lambda}[\mathbb{E}_{X_J}[p(A \mid X, \lambda, \theta) \mid X_I, \lambda]]\,. \end{aligned}$$

The conditional expectation, computed below, is continuous (as the parameters in it remain finite). Furthermore, in distribution, $X_I \to M_I$ and $\lambda \to \mu$, as $s_I \to +\infty$ and $\sigma_\lambda \to 0$. Therefore in the limit the expression reduces to the conditional expectation taken at the limiting final values:

$$\mathbb{E}[p(A \mid X, \lambda, \theta) \mid X_I] = \frac{1}{\iint_{\{X_I = M_I\}\times\mathbb{R}^p} \exp(\langle F_J, X_J\rangle) p(\lambda \mid \theta)\,[\mathrm{d}X]\mathrm{d}\lambda} \iint_{\{X_I = M_I\}\times\mathbb{R}^p} \exp(\langle F_J, X_J\rangle) f(A, X, \lambda) p(\lambda \mid \theta)\,[\mathrm{d}X]\mathrm{d}\lambda\,,$$

where the measure for X here corresponds to the Hausdorff measure over $\{X_I = M_I\}$. Furthermore, we have $\{X_I = M_I\} = \{(M_I, M_I^\perp Y) \mid Y \in \mathcal{V}_{n-r,p-r}\}$ and the map $Y \mapsto (M_I, M_I^\perp Y)$ is an isometry with respect to the Haar measures (which is equal to the Hausdorff measure for Stiefel manifolds, as noted in [29]). We thus get:

$$\begin{aligned} &\iint_{\{X_I = M_I\}\times\mathbb{R}^p} \exp(\langle F_J, X_J\rangle) f(A, X, \lambda) p(\lambda \mid \theta)\,[\mathrm{d}X]\mathrm{d}\lambda \\ &\qquad = \iint_{\mathcal{V}_{n-r,p-r}\times\mathbb{R}^p} \exp(\langle (M_I^\perp)^\top F_J, Y\rangle) f(A, (M_I, M_I^\perp Y), \lambda) p(\lambda \mid \theta)\,[\mathrm{d}Y]\mathrm{d}\lambda\,, \end{aligned}$$

and similarly $\iint_{\{X_I = M_I\}\times\mathbb{R}^p} \exp(\langle F_J, X_J\rangle) p(\lambda \mid \theta)\,[\mathrm{d}X]\mathrm{d}\lambda = \mathcal{C}((M_I^\perp)^\top F_J)$. □

Lemma C.4 (Better bound on the likelihood)**.** *For $\theta \in \Theta$ and $A \in \mathbb{R}^{n\times n}$ such that $\|\mu\| > \max(2\|A\|_F, 2\sigma_\lambda\sqrt{p/2-1})$, we have the bound*

$$p(A \mid \theta) \leq \frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}} \left(\frac{2\|\mu\|^p}{\Gamma(p/2)\sigma_\lambda^p} + 1\right) \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\|\mu\|/2 - \|A\|_F)^2\right).$$

Proof. Using Proposition 2.2, which in particular grants that $\|A * X\| \leq \|A\|_F$, we have

$$\begin{aligned} p(A \mid \theta) &= \mathbb{E}_{\lambda,X}\left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|A - \lambda\cdot X\|_F^2\right)\right] \\ &= \mathbb{E}_{\lambda,X}\left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(\|A\|_F^2 - 2\langle\lambda, A * X\rangle, + \|\lambda\|^2\right)\right)\right] \\ &= \mathbb{E}_{\lambda,X}\left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(\|A\|_F^2 - \|A * X\|^2 + \|A * X - \lambda\|^2\right)\right)\right] \\ &\leq \mathbb{E}_{\lambda,X}\left[\frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|A * X - \lambda\|^2\right)\right]. \end{aligned}$$

Since $\|A * X\| \leq \|A\|_F$, we have $\|A * X - \lambda\| \geq d(\lambda, B(0, \|A\|_F)) = \max(0, \|\lambda\| - \|A\|_F)$. And since $\|\mu\| > 2\|A\|_F$, we have:

$$\begin{aligned}(2\pi\sigma_\varepsilon^2)^{n^2/2} p(A \mid \theta) &\leq \mathbb{E}_{\lambda,X}\left[\mathbf{1}_{\|\lambda\| \leq \|\mu\|/2} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\max(0, \|\lambda\| - \|A\|_F)^2\right)\right] \\ &\quad + \mathbb{E}_{\lambda,X}\left[\mathbf{1}_{\|\lambda\| > \|\mu\|/2} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\|\mu\|/2 - \|A\|_F)^2\right)\right] \\ &\leq \mathbb{P}(\|\lambda\| \leq \|\mu\|/2) + \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\|\mu\|/2 - \|A\|_F)^2\right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{P}(\|\lambda\| \leq \|\mu\|/2) &\leq \mathbb{P}\left(\|\lambda - \mu\| \in \left[\frac{1}{2}\|\mu\|, \frac{3}{2}\|\mu\|\right]\right) \\ &= \mathbb{P}\left(\frac{1}{\sigma_\lambda^2}\|\lambda - \mu\|^2 \in \left[\frac{1}{4\sigma_\lambda^2}\|\mu\|^2, \frac{9}{4\sigma_\lambda^2}\|\mu\|^2\right]\right).\end{aligned}$$

Since by definition $\lambda \sim \mathcal{N}(0, \sigma_\lambda^2 I_p)$, $\frac{1}{\sigma_\lambda^2}\|\lambda - \mu\|^2$ follows a chi-squared distribution with degree p. Its CDF is given by:

$$F(x) = \frac{\gamma(p/2, x/2)}{\Gamma(p/2)},$$

with $\gamma(p/2, x/2) = \int_{x/2}^{\infty} t^{p/2-1}e^{-t}\,\mathrm{d}t$. Therefore we have:

$$\begin{aligned}\mathbb{P}(\|\lambda\| \leq \|\mu\|/2) &\leq \mathbb{P}\left(\frac{1}{\sigma_\lambda^2}\|\lambda - \mu\|^2 \in \left[\frac{1}{4\sigma_\lambda^2}\|\mu\|^2, \frac{9}{4\sigma_\lambda^2}\|\mu\|^2\right]\right) \\ &= \frac{1}{\Gamma(p/2)}\int_{\frac{1}{4\sigma_\lambda^2}\|\mu\|^2}^{\frac{9}{4\sigma_\lambda^2}\|\mu\|^2} t^{p/2-1}e^{-t}\,\mathrm{d}t.\end{aligned}$$

Furthermore the function $t \mapsto t^{p/2-1}e^{-t}$ is decreasing for $t > p/2 - 1$ and $\|\mu\|^2 > 4\sigma_\lambda^2(p/2-1)$, so that:

$$\begin{aligned}\mathbb{P}(\|\lambda\| \leq \|\mu\|/2) &\leq \frac{1}{\Gamma(p/2)}\frac{9\|\mu\|^2 - \|\mu\|^2}{4\sigma_\lambda^2}\left(\frac{1}{4\sigma_\lambda^2}\|\mu\|^2\right)^{p/2-1}\exp\left(-\frac{1}{4\sigma_\lambda^2}\|\mu\|^2\right) \\ &\leq \frac{2\|\mu\|^p}{\Gamma(p/2)\sigma_\lambda^p}\exp\left(-\frac{1}{4\sigma_\lambda^2}\|\mu\|^2\right).\end{aligned}$$

This finally yields the claimed result:

$$\begin{aligned}p(A \mid \theta) &\leq \frac{2\|\mu\|^p}{\Gamma(p/2)\sigma_\lambda^p(2\pi\sigma_\varepsilon^2)^{n^2/2}}\exp\left(-\frac{1}{4\sigma_\lambda^2}\|\mu\|^2\right) + \frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\|\mu\|/2 - \|A\|_F)^2\right) \\ &\leq \frac{1}{(2\pi\sigma_\varepsilon^2)^{n^2/2}}\left(\frac{2\|\mu\|^p}{\Gamma(p/2)\sigma_\lambda^p} + 1\right)\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\|\mu\|/2 - \|A\|_F)^2\right).\end{aligned}$$

□

Lemma C.5 ([11], Lem. 1)**.** *Let $p < q$ be two integers. Then, for any differentiable map $f : \mathbb{R}^p \to \mathbb{R}^q$ and any compact subset K of $\mathbb{R}^p$, there exists a constant λ depending only on p and q such that*

$$\int_{\mathbb{R}^q \setminus f(K)} \log^+ \frac{1}{d(A, f(K))} \mathrm{d}A < \lambda \left(\sup_K \|Df\| + 2 \right)^q \mathrm{Diam}(K) .$$

Lemma C.6. *Assume hypotheses **H1**, **H3**. We have*

$$\mathbb{E}_{P(\mathrm{d}A)} \left[\sup_{\theta \in \Theta^\infty} (\log p(A \mid \theta))^+ \right] < +\infty .$$

Proof. For an observation A and all $\theta \in \Theta$, we have:

$$\begin{aligned} p(A \mid \theta) &= \mathbb{E}\left[p(A \mid X, \lambda) \right] \\ &= \mathbb{E}\left[\frac{1}{(\sigma_\varepsilon \sqrt{2\pi})^{n^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \|A - \lambda \cdot X\|_F^2 \right) \right] \\ &\leq \frac{1}{(\sigma_\varepsilon \sqrt{2\pi})^{n^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} d(A, R_{np})^2 \right) . \end{aligned}$$

Where R_{np} denotes the set of $n \times n$ matrices with rank less than p. This inequality remains true for $\theta \in \Theta^\infty$, as both sides extends continuously to Θ^∞. Hence for all $\theta \in \Theta^\infty$:

$$\log p(A \mid \theta) \leq -n^2 \log(\sigma_\varepsilon \sqrt{2\pi}) - \frac{1}{2\sigma_\varepsilon^2} d(A, R_{np})^2 \tag{C.3}$$

$$\leq -n^2 \log(\sigma_\varepsilon \sqrt{2\pi}) - \frac{1}{2\sigma_\varepsilon^2} d(A, R_{np})^2 . \tag{C.4}$$

This quantity is maximized for $\sigma_\varepsilon^2 = \frac{1}{n^2} d(A, R_{np})^2$. Which gives, taking the positive part, up to a finite additive constant α:

$$(\log p(A \mid \theta))^+ \leq \alpha + n^2 \log^+ \left(\frac{1}{d(A, R_{np})} \right) . \tag{C.5}$$

We now want to apply Lemma C.5 to integrate over A. To that end, we need to parameterize R_{np} with a map from a lower dimensional space. The naive mapping $\mathbb{R}^p \times \mathbb{R}^{n \times p} \to R_{np}$ mapping (λ, X) to $\lambda \cdot X$ does not work directly, as it is not "coercive", in the sense that (λ, X) can go to infinity with $\lambda \cdot X$ possibly staying bounded. This problem is overcome by restricting the X domain of the map $(\lambda, X) \mapsto \lambda \cdot X$ to a set of points close to $\mathcal{V}_{np}$.

Let $f : \mathbb{R}^{n \times p} \times \mathbb{R}^p \to \mathbb{R}^{n \times n}$, defined by $f(U, v) = v \cdot U = U \mathrm{Diag}(v) U^\top$. Then $R_{np} = f(\mathcal{V}_{np} \times \mathbb{R}^p)$. We have furthermore $Df_{U,v}(H, w) = U \mathrm{Diag}(v) H^\top + H \mathrm{Diag}(v) U^\top + U \mathrm{Diag}(w) U^\top$, so

$$\|Df_{U,v}(H)\|_2 \leq 2 \|U\|_2 \|v\|_\infty \|H\|_2 + \|U\|_2^2 \|w\|_\infty .$$

Hence the operator norm of the differential (for the matrix operator norm) satisfies $\|Df_{u,v}\|_2 \leq C_{np} \|(U, v)\|_{\ell^1}^2$ (with C_{np} a generic product of norm equivalence constants, whose definition may implicitly vary depending on the equation).

Let $\beta \in]0,1]$. Since $\mathcal{V}_{np}$ is a compact subset of $\mathbb{R}^{n\times p}$, There exists $X_1,\dots,X_H \in \mathcal{V}_{np}$ such that the union of Frobenius balls $\cup_{h=1}^{H} B_F(X_h,\beta)$ covers $\mathcal{V}_{np}$. In particular, we have

$$f\left(\left(\cup_{h=1}^{H} B(X_h,\beta)\right)\times\mathbb{R}^p\right)\big) = R_{np}\,.$$

Let $(h,t)\in[\![1,H]\!]\times\mathbb{Z}^p$: we define B_{ht} as $B(X_h,\beta)\times B_\infty(t,1/2)$. Hypothesis **H1** gives $np+p<n^2$, hence Lemma C.5 applies to f. We get:

$$\begin{aligned}\int_{\mathbb{R}^{n\times n}\setminus f(B_{ht})} \log^+\frac{1}{d(A,f(B_{ht}))}\mathrm{d}A &< \lambda\left(\sup_{B_{ht}}\|Df\|+2\right)^{n^2}\mathrm{Diam}(B_{ht})\\ &\le \lambda\left(\sup_{(U,v)\in B_{ht}} C_{np}\,\|(U,v)\|_{\ell^1}^2+2\right)^{n^2}(\sqrt{n}+\beta)\\ &\le \lambda\left(C_{np}(\|X_h\|_{\ell^1}+\|t\|_{\ell^1}+C_{np}\beta+p)^2+2\right)^{n^2}(\sqrt{n}+1)\\ &\le (a_{np}\,\|t\|_\infty+b_{np})^{n^2}\quad(\text{as }\ \|U\|_F\le\|X_h\|_F+\beta\le\sqrt{p}+1)\,.\end{aligned}$$

With a_{np},b_{np} constants depending only on n and p. Let $D_T=\cup_{h\in[\![1,H]\!],\|t\|_\infty\le T}B_{ht}$. We have:

$$\begin{aligned}1/d(A,f(D_T)) = \sup_{(U,v)\in D_T} 1/d(A,f(U,v)) &\le \sum_{h=1}^{H}\sum_{\|t\|_\infty\le T}\sup_{(U,v)\in B_{ht}} 1/d(A,f(U,v))\\ &\le \sum_{h=1}^{H}\sum_{\|t\|_\infty\le T} 1/d(A,f(B_{ht}))\,.\end{aligned}$$

Hence, since the sets $f(B_t)$ have zero Lebesgue measure in $\mathbb{R}^{n\times n}$(as $np+p<n^2$):

$$\begin{aligned}\int_{\mathbb{R}^{n\times n}}\log^+\frac{1}{d(A,f(D_T))}P(A)\mathrm{d}A &= \sum_{h=1}^{H}\sum_{j=1}^{T}\sum_{\|t\|_\infty=j}\int_{\mathbb{R}^{n\times n}\setminus f(B_{ht})}\log^+\frac{1}{d(A,f(B_{ht}))}P(A)\mathrm{d}A\\ &\le \sum_{h=1}^{H}\sum_{j=1}^{T}\sum_{\|t\|_\infty=j}(a_{np}\,\|t\|_\infty+b_{np})^{n^2}\max_{d(A,f(B_{ht}))\le 1}P(A)\,.\end{aligned}$$

Now, if A is such that $d(A,f(B_{ht}))\le 1$, we have $\|A-f(X_h,t)\|_F\le 1+C_{np}/2$. Furthermore, since the columns of X_h are orthonormal we have $\|f(X_h,t)\|_2=\|t\|_\infty$, so that $d(A,f(B_{ht}))\le 1 \implies \|A\|_F\ge C_{np}\,\|t\|_\infty-1-C'_{np}/2\ge c(\|t\|_\infty-1)$ for some $c>0$. Hence

$$\begin{aligned}\int_{\mathbb{R}^{n\times n}}\log^+\frac{1}{d(A,f(D_T))}P(A)\mathrm{d}A &\le \sum_{j=1}^{T}\sum_{\|t\|_\infty=j}H(a_{np}j+b_{np})^{n^2}\max_{d(A,f(B_{ht}))\le 1}P(A)\\ &\le \sum_{j=1}^{T}(j+1)^{(np+p)}H(a_{np}j+b_{np})^{n^2}\sup_{\|A\|_F\ge c(j-1)}P(A)\,.\end{aligned}$$

Since P has an exponentially decaying tail beyond some compact set (Hypothesis **H3**), this sum converges to a finite value. Since the sequence $\left(\log^+\left(d(A,f(D_T))^{-1}\right)\right)_{T\in\mathbb{N}}$ is non-negative non-decreasing with limit

$\log^+\left(d(A,R_{np})^{-1}\right)$, Fatou's lemma gives:

$$\int_{\mathbb{R}^{n\times n}} \log^+\left(\frac{1}{d(A,R_{np})}\right) P(\mathrm{d}A) = \int_{\mathbb{R}^{n\times n}} \liminf_{T\to+\infty} \log^+\left(\frac{1}{d(A,f(D_T))}\right) P(\mathrm{d}A)$$
$$\leq \liminf_{T\to+\infty} \int_{\mathbb{R}^{n\times n}} \log^+\left(\frac{1}{d(A,f(D_T))}\right) P(\mathrm{d}A) < +\infty\,.$$

Thus we finally get the desired result with equation (C.5):

$$\mathbb{E}\left[(\log p(A\mid\theta))^+\right] \leq \alpha + n^2 \int_{\mathbb{R}^{n\times n}} \log^+\left(\frac{1}{d(A,R_{np})}\right) P(\mathrm{d}A) < +\infty\,.$$

□

Lemma C.7. *We have:*

1. *$P(\mathrm{d}A)$ almost-surely, for any sequence $\theta_k \in \Theta^\infty$ such that $\lim_{k\to+\infty}\theta_k \in \overline{\Theta^\infty}\setminus\Theta^\infty$,*

$$\lim_{k\to+\infty} \log p(A\mid\theta_k) = -\infty\,.$$

2. *For any sequence $\theta_k \in \Theta^\infty$ such that $\lim_{k\to+\infty}\theta_k \in \overline{\Theta^\infty}\setminus\Theta^\infty$,*

$$\lim_{k\to+\infty} \mathbb{E}_{P(\mathrm{d}A)}\left[\log p(A\mid\theta_k)\right] = -\infty\,.$$

3. *The mapping $\theta\mapsto\mathbb{E}_{P(\mathrm{d}A)}[\log p(A\mid\theta)]$ is continuous on Θ^∞ and $\Theta_*^\infty\neq\varnothing$.*

Proof. We prove the three points consecutively.

1. Let $(\theta_k)\in\Theta^\infty$ a sequence such that $\theta_\infty=\lim_{k\to+\infty}\theta_k\in\overline{\Theta^\infty}\setminus\Theta^\infty$. By definition,

$$\overline{\Theta^\infty}\setminus\Theta^\infty = \{(M,s,\mu,\sigma_\lambda,\sigma_\varepsilon)\mid s\in[0,+\infty]^p \text{ and } (\sigma_\lambda=+\infty \text{ or } \sigma_\varepsilon\in\{0,+\infty\} \text{ or } \mu=\infty)\}\,.$$

   We treat the cases separately, depending on the limits $\sigma_\lambda,\sigma_\varepsilon\in\{0,c>0,\infty\}$ and $\mu\in\mathbb{R}^p\cup\{\infty\}$.
   (a) $\sigma_\lambda\to\infty,\sigma_\varepsilon\to c$: then, by Lemma C.2, $\log p(A\mid\theta)\to-\infty$
   (b) If $\sigma_\varepsilon\to+\infty$ or $\sigma_\varepsilon\to 0$. We can use Lemma C.4: since A has density with respect to the Lebesgue measure, $\|A\|_F\neq\|\mu\|/2$ almost surely, so that $\log p(A\mid\theta)\to-\infty$ as $\sigma_\varepsilon\to+\infty$ or $\sigma_\varepsilon\to 0$.
   (c) If $\mu\to\infty$ and ($\sigma_\lambda\to c,\sigma_\varepsilon\to c$ or $\sigma_\lambda\to 0,\sigma_\varepsilon\to c$): Lemma C.4 grants that $\log p(A\mid\theta)\to-\infty$.
2. Let $(\theta_k)\in\Theta^\infty$ a sequence such that $\theta_\infty=\lim_{k\to+\infty}\theta_k\in\overline{\Theta^\infty}\setminus\Theta^\infty$. Let $f_k(A)=p(A\mid\theta_k)$. We proved above that, almost surely, $f_k(A)\to-\infty$.
   Let $m<0$. We have $\mathbf{1}_{f_k(A)\geq m}\to 0$ almost surely, hence $\mathbb{E}_{P(\mathrm{d}A)}[f_k(A)\mathbf{1}_{f_k(A)\geq m}]\to 0$ as $k\to+\infty$.

$$\mathbb{E}_{P(\mathrm{d}A)}[f_k(A)] = \mathbb{E}_{P(\mathrm{d}A)}[f_k(A)\mathbf{1}_{f_k(A)<m}] + \mathbb{E}[f_k(A)\mathbf{1}_{f_k(A)\geq m}] \underset{k\to+\infty}{\leq} m + o(1)\,.$$

   Therefore $\limsup_{k\to+\infty}\mathbb{E}_{P(\mathrm{d}A)}[f_k(A)]\leq m$ for all $m<0$, hence

$$\lim_{k\to+\infty} \mathbb{E}_{P(\mathrm{d}A)}[\log p(A\mid\theta_k)] = -\infty\,.$$

3. Let $x>0$. Lemma C.3 shows that $\log p(A\mid\theta)$ is continuous over $S_x=\{\theta\in\Theta^\infty\mid\sigma_\varepsilon\in[x,1/x],\sigma_\lambda\leq 1/x\}$, which is a compact set. It is therefore bounded, which implies that $\theta\mapsto\mathbb{E}_{P(\mathrm{d}A)}[\log p(A\mid\theta)]$ is continuous

over S_x for every x, hence continuous over Θ^∞. Furthermore, suppose that Θ^∞_* is empty. Then any maximizing sequence θ_k is such that $\lim \sigma_\lambda \to +\infty$ or $\lim \sigma_\varepsilon \in \{0, +\infty\}$, which contradicts the point proved above. Therefore $\Theta^\infty_* \neq \varnothing$.

□

Lemma C.8. *For every neighborhood V of I_{np}, $\mathrm{Span}(V \cap \mathcal{V}_{np}) = \mathbb{R}^{n\times p}$. Furthermore, the set $\{X - I_{np} \mid X \in V\}$ also spans $\mathbb{R}^{n\times p}$.*

Proof. The tangent vectors at I_{np} write $H = \binom{A}{B}$ with $A^\top = -A$. The proof relies on a second-order expansion of the Cayley retraction map at I_{np}. Following [29], we define the Cayley transform on this tangent space as a function of $K = \begin{pmatrix} A & -B^\top \\ B^\top & 0 \end{pmatrix}$:

$$C_I(H) = (I_n + K)(I_n - K)^{-1} I_{np}\,.$$

Furthermore, if a $n \times n$ matrix K is sufficiently small, we have $(I_n + K)^{-1} = I_n - K + K^2 + O(K^3)$. Taking $B = 0$, we get:

$$C_I \begin{pmatrix} \varepsilon A \\ 0 \end{pmatrix} - I_{np} = \varepsilon \begin{pmatrix} A \\ 0 \end{pmatrix} + O(\varepsilon^2)\,.$$

We can thus get linear combinations of elements of $\mathcal{V}_{np}$ arbitrarily close to elements of the form $\binom{A}{0}$ with $A^\top = -A$. Taking $A = 0$ similarly leads to:

$$C_I \begin{pmatrix} 0 \\ \varepsilon B \end{pmatrix} - I_{np} = \begin{pmatrix} -2\varepsilon^2 B^\top B \\ 2\varepsilon B \end{pmatrix} + O(\varepsilon^3)\,.$$

As with A, we obtain a linear combination $\left(C_I\binom{0}{\varepsilon B} - I_{np}\right)/\varepsilon$ arbitrarily close to matrices of the form $\binom{0}{B}$ with $B \in \mathbb{R}^{(n-p)\times p}$. Furthermore, still taking $A = 0$, we obtain:

$$C_I \begin{pmatrix} 0 \\ \varepsilon B \end{pmatrix} + C_I \begin{pmatrix} 0 \\ -\varepsilon B \end{pmatrix} - 2I_{np} = \begin{pmatrix} -4\varepsilon^2 B^\top B \\ 0 \end{pmatrix} + O(\varepsilon^3)\,.$$

We can thus get linear combinations close to elements of the form $\binom{B^\top B}{0}$. This is sufficient to get all matrices with a symmetric upper part, as any symmetric matrix can be obtained as a weighted sum of rank one matrices of the form $(x, 0, \ldots, 0)^\top \in \mathbb{R}^{(n-p)\times p}$ ($x \in \mathbb{R}^p$).

As a consequence, there are linear combinations converging to any matrix $\binom{A}{B}$, by combining symmetric and skew-symmetric components for A, and the term for B. In particular, we obtain linear combinations arbitrarily close to a basis of $\mathbb{R}^{n\times p}$, which thus also span the entire space. □

Lemma C.9. *The restricted model $\tilde{p}(A \mid \theta)$ is identifiable on Θ^{id}.*

Proof. The parameters $\sigma_\lambda, \sigma_\varepsilon$ and μ can be identified as in Theorem 3.1. It thus remains to identify $F = M\mathrm{Diag}(s)$ from the distribution of $\lambda \cdot X$. Here, the argument gets much simpler than for the full model: since X is constrained in Δ_0, the mapping $(\lambda, X) \mapsto \lambda \cdot X$ is injective over the whole support of latent variables. Therefore, the changes of variable using the formula of [53] directly give access to the density of X over Δ_0 (with the same argument as the one used to obtain $f_\lambda(X)$ for the full model).

By the hypothesis we made when introducing the restricted model, the maximum of $\langle X, F\rangle_F$ over $\mathcal{V}_{np}$ is reached in Δ_0: this point, which can thus be identified, gives the value of M, the normalized columns of F (we recall that we introduced the decomposition $F = M\mathrm{Diag}(s)$).

We use the gradient of $\tilde{p}(X \mid \theta)$ to identify the concentration parameters (s_i). Since the function is defined over $\mathcal{V}_{np}$, we only have access to the projection of its gradient onto the tangent spaces. If we denote by $G(X)$ the Euclidean gradient, the projected manifold gradient writes: $G_{\mathcal{V}}(X) = G(X) - XG(X)^\top X$ [21]. In the case of the function $p(X \mid \theta)$, the manifold gradient thus is: $G_{\mathcal{V}}(X) = \tilde{p}(X \mid \theta)(F - XF^\top X)$. As a consequence, the function $h(X) = F - XF^\top X$ is known over Δ_0. Coherently, we have $h(M) = M\mathrm{Diag}(s) - M\mathrm{Diag}(s)M^\top M = 0$. We will use the first-order variations of $h(X)$ around M allow to retrieve s.

These variations are retrieved by using the Cayley transform on tangent vectors at M (any other smooth retraction map could be used here). As reminded in Appendix A, such tangent vectors $H \in T_M\mathcal{V}_{np}$ write as $H = MA + M_\perp B$, with $A^\top = -A$. Denoting $K = HM^\top - MH^\top$, the Cayley transform at M is defined by:

$$C_M(H) = (I_n + K)(I_n - K)^{-1}M \in \mathcal{V}_{np}\,.$$

In particular, as in Lemma C.8, it satisfies $C_M(\varepsilon H) = M + \varepsilon H + O(\varepsilon^2)$. This gives:

$$\begin{aligned}
h(C_M(\varepsilon H)) &= F - C_M(H)F^\top C_M(\varepsilon H)\\
&= F - (M + \varepsilon H)F^\top(M + \varepsilon H) + O(\varepsilon^2)\\
&= \underbrace{F - MF^\top M}_{=0} - \varepsilon HF^\top M - \varepsilon MF^\top H + O(\varepsilon^2)\\
&= -\varepsilon(MA + M_\perp B)\mathrm{Diag}(s) + \varepsilon M\mathrm{Diag}(s)M^\top(MA + M_\perp B) + O(\varepsilon^2)\\
&= -\varepsilon M[\mathrm{Diag}(s)A + A\mathrm{Diag}(s)] - \varepsilon M_\perp B\mathrm{Diag}(s) + O(\varepsilon^2)\,.
\end{aligned}$$

Taking $B = 0$ and normalizing by ε, we obtain the value of $M[\mathrm{Diag}(s)A + A\mathrm{Diag}(s)]$ for every $p \times p$ skew-symmetric matrix A, which gives $\mathrm{Diag}(s)A + A\mathrm{Diag}(s)$ when multiplying by $M^\top$. For every i, j, taking for A the matrix with $A_{ij} = -A_{ji} = 1$ and zeros everywhere else gives the value of $s_i + s_j$. This gives an over-determined system of equations which allows identifying the s_i's. □

Lemma C.10. *If the empirical data distribution is given by $P(A) = \tilde{p}(A \mid \theta_0)$, then condition (5.1) for the asymptotic normality theorem of [54] is satisfied by the restricted model on a neighborhood of θ_0.*

Proof. We are looking for a function $L : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ with $\mathbb{E}[L(A)^2] < +\infty$ and such that, for θ_1 and θ_2 sufficiently close to θ_0,

$$|\log p(A \mid \theta_1) - \log p(A \mid \theta_2)| \le L(A)\,\|\theta_1 - \theta_2\|\,.$$

Transposed to the restricted model, Lemma C.1 gives the marginalized expression:

$$\log\tilde{p}(A \mid \theta) = \log\left[\frac{1}{(2\pi)^{n^2/2}\sigma_\varepsilon^{n^2}}\frac{\sigma_p^p}{\sigma_\lambda^p}\right] - \frac{1}{2\sigma_\varepsilon^2}\|A\|_F^2 - \frac{1}{2\sigma_\lambda^2}\|\mu\|^2 + \log\int_{\Delta_0}\frac{1}{\mathcal{C}'(F)}\exp\left(\langle X, F\rangle_F + \frac{1}{2\sigma_p^2}\|\mu_{AX}\|^2\right)[\mathrm{d}X]\,.$$

For two parameters θ_1 and θ_2, all terms apart from the integral over Δ_0 are bounded by $(C + \|A\|_F^2)\,\|\theta_1 - \theta_2\|$, with C a constant depending on the neighborhood around θ_0. Let $h(\theta, A, X) = \exp\left(\langle X, F\rangle_F + \frac{1}{2\sigma_p^2}\left\|\mu_{AX}^2\right\|\right)$. Denoting

$$M_{\theta,A} = \max_X \|\mu_{AX}\|^2 \le \sigma_p^2(\|A\|_F^2/\sigma_\varepsilon^2 + \|\mu\|^2/\sigma_\lambda^2)\,,$$

we have:

$$h(\theta + \mathrm{d}\theta, A, X) = \exp\left(\langle X, F + \mathrm{d}F\rangle_F + \frac{1}{2\sigma_p^2 + 2\mathrm{d}\sigma_p^2}\left\|\mu_{AX}^2\right\|\right)$$

$$= h(\theta, A, X)\left(1 + \langle X, \mathrm{d}F\rangle_F - \frac{\mathrm{d}\sigma_p^2}{2\sigma_p^4}\left\|\mu_{AX}\right\|^2 + O((1+\|A\|_F)^2\left\|\mathrm{d}\theta\right\|^2)\right)$$
$$= h(\theta, A, X)\left(1 + O((1+\|A\|_F^2)\left\|\mathrm{d}\theta\right\|)\right),$$

where the O notation contains constants depending on θ_0 and the size of its neighborhood. As a consequence:

$$\log\int_{\Delta_0} h(\theta_2, A, X)\,[\mathrm{d}X] - \log\int_{\Delta_0} h(\theta_1, A, X)\,[\mathrm{d}X] = O((1+\|A\|_F^2)\left\|\theta_2 - \theta_1\right\|).$$

Finally, the Lipschitz condition (5.1) is satisfied by $L(A) = C(1+\|A\|_F^2)$. Furthermore, by Lemma C.4, $p(A \mid \theta)$ admits second order moments, so that $\mathbb{E}[L(A)^2] < +\infty$. □

References

[1] C. Aicher, A.Z. Jacobs and A. Clauset, Learning Latent Block Structure in Weighted Networks. *J. Complex Netw.* **3** (2015) 221–248.

[2] M. Ali and J. Gao, Classification of matrix-variate fisher–bingham distribution via maximum likelihood estimation using manifold valued data. *Neurocomputing* **295** (2018) 72–85.

[3] S. Allassonniere, Y. Amit and A. Trouvé, Toward a coherent statistical framework for dense deformable template estimation. *J. Royal Stat. Soc. B* **69** (2007) 3–29.

[4] E.S. Allman, C. Matias and J.A. Rhodes, Identifiability of parameters in latent structure models with many observed variables. *Ann. Stat.* **37** (2009) 3099–3132.

[5] T.W. Anderson and Y. Amemiya, The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Stat.* **16** (1988) 759–771.

[6] O.E. Barndorff-Nielsen, Identifiability of mixtures of exponential families. *J. Math. Anal. Appl.* **12** (1965) 115–121.

[7] O.E. Barndorff-Nielsen, Information and exponential families. In: Statistical theory, Wiley series in probability and mathematical statistics. Wiley, Chichester, New York (1978).

[8] P.J. Bickel, Y. Ritov and T. Rydén, Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Stat.* **26** (1998) 1614–1635.

[9] S. Bonhomme and J.-M. Robin, Consistent noisy independent component analysis. *J. Econ.* **149** (2009) 12–25.

[10] J. Chen, G. Han, H. Cai, J. Ma, M. Kim, P. Laurienti and G. Wu, Estimating common harmonic waves of brain networks on Stiefel manifold, in A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu and L. Joskowicz (editors), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, Springer International Publishing, Cham (2020) 367–376.

[11] J. Chevallier, V. Debavelaere and S. Allassonnière, A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. *SIAM J. Imag. Sci.* **14** (2021) 349–388.

[12] Y. Chikuse, Concentrated matrix Langevin distributions. *J. Multivar. Anal.* **85** (2003) 375–394.

[13] Y. Chikuse, Statistics on Special Manifolds, Lecture Notes in Statistics, Springer-Verlag, New York (2003).

[14] Y. Chikuse, State space models on special manifolds. *J. Multivar. Anal.* **97** (2006) 1284–1294.

[15] R. Douc, Non Singularity of the Asymptotic Fisher Information Matrix in Hidden Markov Models. arXiv:math/0511631 (2005).

[16] R. Douc, E. Moulines, J. Olsson and R. van Handel, Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Stat.* **39** (2011) 474–513.

[17] R. Douc, F. Roueff and T. Sim, Necessary and sufficient conditions for the identifiability of observation-driven models. *J. Time Ser. Anal.* **42** (2021) 140–160.

[18] N.S. D'Souza, M.B. Nebel, N. Wymbs, S. Mostofsky and A. Venkataraman, A generative-discriminative basis learning framework to predict clinical severity from resting state functional MRI data, in A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger (editors), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham (2018), vol. 11072, 163–171.

[19] N.S. D'Souza, M.B. Nebel, N. Wymbs, S. Mostofsky and A. Venkataraman, Integrating neural networks and dictionary learning for multidimensional clinical characterizations from functional connectomics data, in D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan (editors), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham (2019), vol. 11766, 709–717.

[20] L.L. Duan, G. Michailidis and M. Ding, Spiked Laplacian Graphs: Bayesian Community Detection in Heterogeneous Networks. arXiv:1910.02471 [stat] (2020).

[21] A. Edelman, T.A. Arias and S.T. Smith, The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** (1998) 303–353.

[22] K. Fan, On a theorem of weyl concerning eigenvalues of linear transformations I. *Proc. Natl. Acad. Sci.* **35** (1949) 652–655.

[23] P.J. Forrester, Log-gases and random matrices (LMS-34). Vol. 34 of *London Mathematical Society Monographs*. Princeton University Press (2010).
[24] C. Fraikin, K. Hüper and P.V. Dooren, Optimization over the Stiefel Manifold, in vol. 7 of PAMM: Proceedings in Applied Mathematics and Mechanics. Wiley Online Library (2007) 1062205–1062206.
[25] Y. Gu and G. Xu, Identifiability of Hierarchical Latent Attribute Models. arXiv:1906.07869 [cs, stat] (2021).
[26] P.D. Hoff, Simulation of the matrix Bingham—von Mises—Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Stat.* **18** (2009) 438–456.
[27] H. Holzmann, A. Munk and B. Stratmann, Identifiability of finite mixtures - with applications to circular distributions. *Sankhyā* **66** (2004) 440–449.
[28] S. Janson, Graphons, Cut Norm and Distance, Couplings and Rearrangements, Vol. 4 of New York Journal of Mathematics. *NYJM Monographs, State University of New York, University at Albany, Albany, NY* **4** (2013) 76.
[29] M. Jauch, P.D. Hoff and D.B. Dunson, Random Orthogonal Matrices and the Cayley Transform. *Bernoulli* **26** (2020) 1560–1586.
[30] P.E. Jupp and K.V. Mardia, Maximum Likelihood Estimators for the Matrix Von Mises-Fisher and Bingham Distributions. *Ann. Stat.* **7** (1979) 599–606.
[31] J.T. Kent, Identifiability of Finite Mixtures for Directional Data, *Ann. Stat.* **11** (1983).
[32] C.G. Khatri and K.V. Mardia, The von Mises–Fisher Matrix Distribution in Orientation Statistics. *J. R. Stat Soc. Ser. B (Methodological)* **39** (1977) 95–106.
[33] A. Khetan and M. Mj, Cheeger Inequalities for Graph Limits, arXiv:1807.02225 [math] (2018).
[34] T.N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in ICLR 2017 (2017).
[35] A. Kume, S.P. Preston and A.T.A. Wood, Saddlepoint Approximations for the Normalizing Constant of Fisher–Bingham Distributions on Products of Spheres and Stiefel Manifolds. *Biometrika* **100** (2013) 971–984.
[36] P. Latouche and S. Robin, Variational Bayes Model Averaging for Graphon Functions and Motif Frequencies Inference in W-graph Models, *Stat. Comput.* **26** (2016) 1173–1185.
[37] M. Lavielle and L. Aarons, What Do We Mean by Identifiability in Mixed Effects Models?. *J. Pharmacokinet. Pharmacodyn.* **43** (2016) 111–122.
[38] E.L. Lehmann and G. Casella, Theory of Point Estimation, Springer Texts in Statistics, 2nd edn., Springer, New York (2003).
[39] X. Li, N.C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola and J.S. Duncan, Graph Neural Network for Interpreting Task-fMRI Biomarkers, in D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan (editors), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science, Springer International Publishing, Cham (2019) 485–493.
[40] X. Liang, L. Wang, L.-H. Zhang and R.-C. Li, On Generalizing Trace Minimization. arXiv:2104.00257 [cs, math] (2021).
[41] L. Lin, V. Rao and D. Dunson, Bayesian Nonparametric Inference on the Stiefel Manifold. *Stat. Sin.* **27** (2017) 535–553.
[42] L. Lovász, Large Networks and Graph Limits. *Colloquium Publications*, vol. 60, American Mathematical Society, Providence, Rhode Island (2012).
[43] C. Mantoux, B. Couvy-Duchesne, F. Cacciamani, S. Epelbaum, S. Durrleman and S. Allassonnière, Understanding the Variability in Graph Data Sets through Statistical Modeling on the Stiefel Manifold, *Entropy* **23** (2021) 490.
[44] S.S. Mukherjee and S. Chakrabarti, Graphon Estimation from Partially Observed Network Data. arXiv:1906.00494 [cs, stat] (2019).
[45] S.C. Olhede and P.J. Wolfe, Network Histograms and Universality of Blockmodel Approximation. *Proc. Natl. Acad. Sci.* **111** (2014) 14722–14727.
[46] S. Pal, S. Sengupta, R. Mitra and A. Banerjee, Conjugate Priors and Posterior Inference for the Matrix Langevin Distribution on the Stiefel Manifold. *Bayesian Anal.* **15** (2020) 871–908.
[47] T.P. Peixoto, Bayesian Stochastic Blockmodeling, in P. Doreian, V. Batagelj and A. Ferligoj (editors), Advances in Network Clustering and Blockmodeling, Wiley Series in Computational and Quantitative Social Science, 289–332, Wiley (2020) .
[48] Z. Ren, T. Sun, C.-H. Zhang and H.H. Zhou, Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models. *Ann. Stat.* **43** (2015).
[49] A.A. Shabalin and A.B. Nobel, Reconstruction of a Low-Rank Matrix in the Presence of Gaussian Noise. *J. Multivar. Anal.* **118** (2013) 67–76.
[50] B. Sischka and G. Kauermann, EM-based Smooth Graphon Estimation Using MCMC and Spline-Based Approaches. *Soc. Netw.* **68** (2022) 279–295.
[51] E. Tabrizi, E.B. Samani and M. Ganjali, A Note on the Identifiability of Latent Variable Models for Mixed Longitudinal Data. *Stat. Probab. Lett.* **167** (2020) 108882.
[52] H. Teicher, Identifiability of Finite Mixtures. *Ann. Math. Stat.* **34** (1963) 1265–1269.
[53] T. Traynor, Change of Variables for Hausdorff Measure (from the Beginning). *Università degli Studi di Trieste. Dipartimento di Scienze Matematiche* **26 suppl.** (1994) 327–347.
[54] A.W. van der Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, 1st edn., Cambridge Univ. Press, Cambridge (1998).

[55] J. Xu, Rates of Convergence of Spectral Methods for Graphon Estimation, in International Conference on Machine Learning (2018) 5433–5442.
[56] S.J. Yakowitz and J.D. Spragins, On the Identifiability of Finite Mixtures. *Ann. Math. Stat.* **39** (1968) 209–214.

Subscribe to Open (S2O)
A fair and sustainable open access model



This journal is currently published in open access under a Subscribe-to-Open model (S2O). S2O is a transformative model that aims to move subscription journals to open access. Open access is the free, immediate, online availability of research articles combined with the rights to use these articles fully in the digital environment. We are thankful to our subscribers and sponsors for making it possible to publish this journal in open access, free of charge for authors.

Please help to maintain this journal in open access!

Check that your library subscribes to the journal, or make a personal donation to the S2O programme, by contacting subscribers@edpsciences.org

More information, including a list of sponsors and a financial transparency report, available at: https://www.edpsciences.org/en/maths-s2o-programme