Proceedings of Machine Learning Research 96:1–21, 2018. 1st Symposium on Advances in Approximate Bayesian Inference

Consistency of ELBO maximization for model selection

Badr-Eddine Chérief-Abdellatif BADR.EDDINE.CHERIEF.ABDELLATIF@ENSAE.FR
CREST, ENSAE, Université Paris Saclay

Abstract

The Evidence Lower Bound (ELBO) is a quantity that plays a key role in variational inference. It can also be used as a criterion in model selection. However, though extremely popular in practice in the variational Bayes community, there has never been a general theoretic justification for selecting based on the ELBO. In this paper, we show that the ELBO maximization strategy has strong theoretical guarantees, and is robust to model misspecification while most works rely on the assumption that one model is correctly specified. We illustrate our theoretical results by an application to the selection of the number of principal components in probabilistic PCA.

Keywords: Variational inference, Evidence lower bound, Model selection.

1. Introduction

Approximate Bayesian inference is at the core of modern Bayesian statistics and machine learning. While exact Bayesian inference is often intractable, variational inference has proved to provide an efficient solution when dealing with large datasets and complex probabilistic models. Variational Bayes (VB) aims at maximizing a numerical quantity referred to as Evidence Lower Bound on the marginal likelihood (ELBO), and thus makes use of optimization techniques to converge faster than Monte Carlo sampling approach. Blei et al. (2017) provides a comprehensive survey on variational inference. Although VB is mainly used for its practical efficiency, little attention has been put towards its theoretical properties during the last years. While Alquier et al. (2016) studied the properties of variational approximations of Gibbs distributions used in machine learning for bounded loss functions, Alquier and Ridgway (2017); Zhang and Gao (2017); Wang and Blei (2018); Bhattacharya et al. (2018); Chérief-Abdellatif and Alquier (2018) extended the results to more general statistical models.

At the same time, model selection remains a major problem of interest in statistics that naturally arises in the course of scientific inquiry. The statistician aims at selecting a model among several candidates given an observed dataset. To do so, one can perform cross validation as in Vehtari et al. (2014) or maximize a numerical criterion to make the final choice, see the review of Rao and Wu (2001). In the literature, penalized criteria such as AIC and BIC respectively introduced by Akaike (1974) and Schwarz (1978) are popular. While AIC aims at optimizing the prediction performance, BIC is more suitable for recovering with high probability the true model (when such a model exists), see Yang (2005). Thus, it is necessary to define a criterion suited to a given objective. Meanwhile, a non-asymptotic theory of penalization using oracle inequalities has been developed during

the last two decades, and offers a simple way to assess the quality of a given model selection criterion. We refer the interested reader to Massart (2007) for more details.

In this paper, we are interested in finding an estimate of the distribution of the data, and we need to choose from among competing models. Blei et al. (2017) states that "the [evidence lower] bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory". Since then, authors of Chérief-Abdellatif and Alquier (2018) have provided an analysis of model selection based on the ELBO in the case of mixture models. We extend their result to the general case of independent and identically distributed (i.i.d.) data, and we provide an oracle inequality on the ELBO criterion that justifies the consistency of ELBO maximization when the objective is the estimation of the distribution of the data. In particular, as soon as there exists a true model, we show that the ELBO criterion is adaptive and that the selected estimator achieves the convergence rate of the variational approximation associated with the true model.

The rest of this paper is organized as follows. Section 2 introduces the setting and the key concepts needed to understand our results. In Section 3, we prove that the ELBO criterion provides a variational approximation that is consistent with the sample size as soon as there exists a true model. We also extend the result to misspecified models. We finally illustrate the main theorem of this paper by an application to the selection of the number of principal components in probabilistic Principal Component Analysis (PCA) in Section 4. All the proofs are deferred to the appendix.

2. Framework

Let us introduce the notations and the framework we adopt in this paper. We consider a collection of i.i.d. random variables X_1,...,X_n distributed according to some probability distribution P^0 in a measurable space $\left(\mathbb{X}, \mathcal{X}\right)$. We denote $X_1^n = (X_1, ..., X_n)$. We consider a countable collection $\{\mathcal{M}_K / K \geq 1\}$ of statistical mixture models $\mathcal{M}_K = \{P_{\theta_K} \,/\, \theta_K \in \Theta_K\}$ where Θ_K is the parameter set associated with index K. We make no assumptions on Θ_K's nor on P_{θ_K}. Parameter spaces may overlap or have inclusion relationships. Let $\mathcal{M}_1^+(\Theta_K)$ be the set of all probability distributions over Θ_K.

We use a Bayesian approach, and we define a prior π over the full parameter space $\cup_{K\geq 1}\Theta_K$ (equipped with some suited sigma-algebra). First, we specify a prior weight π_K assigned to model $\mathcal{M}_K$, and then a conditional prior $\Pi_K(.)$ on $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$:

$$\pi = \sum_{K\geq 1} \pi_K \Pi_K.$$

The Kullback-Leibler divergence between two probability distributions P and R is

$$\mathrm{KL}(P, R) = \begin{cases} \int \log\left(\frac{dP}{dR}\right) dP \;\text{ if } R \text{ dominates } P, \\ +\infty \;\text{ otherwise.} \end{cases}$$

For any $\alpha \neq 1$, authors of Van Erven and Harremos (2014) detail the properties of the α-Rényi divergence between two probability distributions P and R which is equal to:

$$D_\alpha(P, R) = \begin{cases} \frac{1}{\alpha-1} \log \int \left(\frac{dP}{dR}\right)^{\alpha-1} dP \;\text{ if } R \text{ dominates } P, \\ +\infty \;\text{ otherwise.} \end{cases}$$

We define the tempered posterior distribution $\pi^K_{n,\alpha}(.|X_1^n)$ on parameter $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$ using prior Π_K and likelihood L_n for any $\alpha \in (0,1)$:

$$\pi^K_{n,\alpha}(d\theta_K|X_1^n) \propto L_n(\theta_K)^\alpha \Pi_K(d\theta_K).$$

This definition is a slight variant of the regular Bayesian posterior (for which $\alpha = 1$), and is also referred to as Bayesian fractional posterior in Bhattacharya et al. (2016). This posterior is easier to sample from, more robust to model misspecification and requires less stringent conditions to obtain consistency, see respectively Behrens et al. (2012), Grünwald and Van Ommen (2017) and Bhattacharya et al. (2016).

The Variational Bayes approximation $\tilde{\pi}^K_{n,\alpha}(.|X_1^n)$ of the tempered posterior associated with model $\mathcal{M}_K$ is then defined as the projection, with respect to the Kullback-Leibler divergence, of the tempered posterior onto some set $\mathcal{F}_K$:

$$\tilde{\pi}^K_{n,\alpha}(.|X_1^n) = \arg\min_{\rho_K \in \mathcal{F}_K} \mathrm{KL}(\rho_K, \pi^K_{n,\alpha}(.|X_1^n)).$$

The choice of the variational set $\mathcal{F}_K$ is crucial: the variational approximation must be close enough to the target distribution (as an approximation of the tempered posterior) but not too close (in order to be tractable). A classical variational set $\mathcal{F}_K$ is the parametric family which leads to a tractable parametric approximation, e.g. a Gaussian distribution. Another popular set $\mathcal{F}_K$ in the VB community is the mean-field approximation that is based on a partition of the space of parameters, and which consists in a factorization of the variational approximation over the partition.

Alternatively, the variational approximation is often defined as the distribution into $\mathcal{F}_K$ that maximizes the Evidence Lower Bound:

$$\tilde{\pi}^K_{n,\alpha}(.|X_1^n) = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K)\rho_K(d\theta_K) - \mathrm{KL}\big(\rho_K, \Pi_K\big) \right\}$$

where the function inside the argmax operator is the ELBO (as a function of K and ρ_K) and ℓ_n is the log-likelihood. In the following, we will just call ELBO(K) the closest approximation to the log-evidence, i.e. the value of the ELBO evaluated at its maximum:

$$\mathrm{ELBO}(K) = \alpha \int \ell_n(\theta_K)\tilde{\pi}^K_{n,\alpha}(d\theta_K|X_1^n) - \mathrm{KL}(\tilde{\pi}^K_{n,\alpha}(.|X_1^n), \Pi_K).$$

In the variational Bayes community, researchers and practitioners use the ELBO in order to select the model from which they will consider the final variational approximation $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$, as stated in Blei et al. (2017). We propose to consider a penalized version of the ELBO criterion

$$\hat{K} = \arg\max_{K \geq 1} \left\{ \mathrm{ELBO}(K) - \log\left(\frac{1}{\pi_K}\right) \right\}$$

which is a slight variant of the classical definition, although choosing a uniform prior over a finite number of models leads to maximizing the ELBO. Note that the penalty term is not just an artefact in order to ease the theoretical proof, but it is a complexity term that reflects our prior beliefs over the different models.

We will provide in the next section a theoretical justification to such a selection criterion and show that the selected variational estimator $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$ is consistent under mild conditions as soon as there exists a true model. We will adopt the definition of *consistency* used in Alquier and Ridgway (2017) and Chérief-Abdellatif and Alquier (2018) that is, the Bayesian estimator is said to be consistent if, in expectation (with respect to the random variables distributed according to P^0), the average Rényi loss between a distribution in the selected model and the true distribution (over the Bayesian estimator) goes to zero as $n \to +\infty$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \xrightarrow[n\to+\infty]{} 0.$$

This definition is closely related to the notion of *concentration* which is defined in Ghosal et al. (2000) as the asymptotic concentration of the Bayesian estimator around the true distribution, and which is usually used to assess frequentist guarantees for Bayesian estimators. It is sometimes also referred to as *contraction* (or even *consistency*). See Appendix A for more details on the connection between the notions of *consistency* and *concentration*.

3. Consistency of the ELBO criterion

In this section, unless explicitly stated otherwise, we assume that there exists a true model $\mathcal{M}_{K_0}$ that contains the true distribution P^0, i.e. that there exists K_0 and $\theta^0 \in \Theta_{K_0}$ such that $P^0 = P_{\theta^0}$.

A key assumption introduced in Ghosal et al. (2000) in order to obtain the concentration of the regular posterior distribution $\pi^{K_0}_{n,1}(.|X_1^n)$ associated with the true model $\mathcal{M}_{K_0}$ is a *prior mass condition* which states that the prior Π_{K_0} must give enough mass to some neighborhood (in the Kullback-Leibler sense) of the true parameter. Bhattacharya et al. (2016) showed that this condition was sufficient when considering tempered posteriors $\pi^{K_0}_{n,\alpha}(.|X_1^n)$. Alquier and Ridgway (2017) extended this assumption in order to obtain the concentration and the consistency of variational approximations of the tempered posteriors $\tilde{\pi}^{K_0}_{n,\alpha}(.|X_1^n)$. In addition to the previous prior mass condition, this extension requires the variational set $\mathcal{F}_{K_0}$ to contain probability distributions concentrated around the true parameter. Note that when $\mathcal{F}_{K_0} = \mathcal{M}_1^+(\Theta_{K_0})$, this goes back to the standard prior mass condition. This extended prior mass condition is standard in the variational Bayes community, see Alquier and Ridgway (2017); Chérief-Abdellatif and Alquier (2018), and can be formulated as follows:

Assumption : *We assume that there exists r_n for which there is a distribution $\rho_{K_0,n} \in \mathcal{F}_{K_0}$ such that*:

$$\int \mathrm{KL}(P^0, P_{\theta_{K_0}})\rho_{K_0,n}(d\theta_{K_0}) \leq r_n \text{ and } \mathrm{KL}(\rho_{K_0,n}, \Pi_{K_0}) \leq nr_n. \tag{3.1}$$

Remark 1 *Define the KL-ball $\mathcal{B}$ centered at θ_0 of radius r_n:*

$$\mathcal{B} = \{\theta \in \Theta_{K_0} / \mathrm{KL}(P_{\theta_0}, P_\theta) \leq r_n\},$$

and consider the restriction $\rho_{K_0,n}$ of Π_{K_0} to $\mathcal{B}$. Then it is clear that when $\rho_{K_0,n} \in \mathcal{F}_{K_0}$, Assumption 3.1 becomes equivalent to the former prior mass condition of Ghosal et al. (2000), i.e. $\Pi_{K_0}(\mathcal{B}) \geq e^{-nr_n}$. The computation of the prior mass $\Pi_{K_0}(\mathcal{B})$ is a major difficulty.

It has been raised as a question of interest in Ghosal et al. (2000), and is addressed for categorical distributions and Dirichlet priors in Ghosal et al. (2000) (but for an L_1-ball) and in Chérief-Abdellatif and Alquier (2018) (for a KL-ball). Unfortunately, $\rho_{K_0,n}$ does not belong to $\mathcal{F}_{K_0}$ in general and the computation of the prior mass is no longer sufficient. Nevertheless, the strategy of computing the prior mass of KL-balls remains of interest when dealing with mixture models and mean-field approximation sets, see Chérief-Abdellatif and Alquier (2018) where the authors showed that studying the prior mass condition of Ghosal et al. (2000) independently on the weights and on each component becomes sufficient.

Remark 2 *When $\mathcal{F}_{K_0}$ is parametric, it is often possible to overcome the difficulty presented above in order to find a rate r_n as in Assumption 3.1. Indeed, the point is to express the distribution $\rho_{K_0,n}$ using the general parametric form of the variational family, and to find relevant values of the parameters that will lead to fast rates of convergence r_n. This is the strategy we follow in Section 4 for probabilistic PCA. See Alquier and Ridgway (2017); Chérief-Abdellatif and Alquier (2018) for other examples of such computations.*

Alquier and Ridgway (2017) showed that the variational approximation $\tilde{\pi}^{K_0}_{n,\alpha}(.|X_1^n)$ associated with a true model is consistent under Assumption 3.1 and that the convergence rate is equal to r_n. Nevertheless, in model selection, we do not necessarily know which model is true and the challenge is to be able to find one such that the corresponding approximation is consistent at a comparable convergence rate. We show that the variational approximation $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$ associated with the selected model is also consistent at rate r_n as soon as Assumption 3.1 is satisfied:

Theorem 3 *Assume that Assumption 3.1 is satisfied. Then for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}r_n + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

The inequality in Theorem 3 shows the adaptivity of our procedure. Indeed, whatever the value of $\hat{K}$ (which can be different from K_0), we obtain the consistency of the selected variational approximation at the same rate of convergence than the estimator associated with the true model (as soon as the additional term in the upper bound is lower than r_n, which is the case for prior weights used in practice). We recall that we look for a good estimation of the true distribution P^0 and not for an estimation of the true model index K_0 which is a different task that would require identifiability assumptions that are stronger than those in our theorem. The overall rate is composed of the convergence rate associated with the true model $\mathcal{M}_{K_0}$, and of a complexity term that reflects the prior belief over the (unknown) true model. For example, if we range a countable number of models according to our prior belief, and we take $\pi_K = 2^{-K}$, then the corresponding term will be of order K_0/n. More generally, when $\frac{1}{n} \lesssim r_n$, we obtain the consistency at the rate associated with the true model.

As a short example, Chérief-Abdellatif and Alquier (2018) investigated the case of mixture models. For instance, authors obtained a convergence rate equal to $K_0 \log(nK_0)/n$ for Gaussian mixtures when there exists a true K_0-components mixture model. We study another example in Section 4.

We can also extend this result to misspecified models. In the model selection literature, only little attention has been put to misspecification when the true distribution does not belong to any of the models, see Lv and Liu (2013). Now, we do not assume any longer that there exists a true model, and we show that our ELBO criterion is robust to model misspecification:

Theorem 4 *For each index K, let us define the set $\Theta_K(r_{K,n})$ of parameters $\theta_K^* \in \Theta_K$, for which there is a distribution $\rho_{K,n} \in \mathcal{F}_K$ such that:*

$$\int \mathbb{E}\left[\log \frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right] \rho_{K,n}(d\theta_K) \leq r_{K,n} \text{ and } \mathrm{KL}(\rho_{K,n}, \Pi_K) \leq n r_{K,n}. \tag{3.2}$$

Then for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta | X_1^n)\right] \leq \inf_{K \geq 1} \left\{ \frac{\alpha}{1-\alpha} \inf_{\theta_K^* \in \Theta_K(r_{K,n})} \mathrm{KL}(P^0, P_{\theta_K^*}) + \frac{1+\alpha}{1-\alpha} r_{K,n} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)} \right\}.$$

Note that when there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\theta^0}$ with $\theta^0 \in \Theta_{K_0}$, then under Assumption 3.1, we get $\theta^0 \in \Theta_{K_0}(r_{K_0,n})$, and we recover Theorem 3. Furthermore, the oracle inequality in Theorem 4 shows that the selected variational approximation adaptively achieves the best upper bound among the different models $\mathcal{M}_K$, where each upper bound is a trade-off between two terms: a bias due to the error of approximating the true distribution by a distribution in model $\mathcal{M}_K$, and a variance term $r_{K,n}$ (as soon as the penalty term is lower than $r_{K,n}$) that is defined in Condition 3.2.

4. Application to probabilistic PCA

We consider here the probabilistic Principal Component Analysis (PCA) problem as an application of our work. From now on, matrices will be denoted in bold capital letters. We assume the model

$$X_i = \boldsymbol{W} Z_i + \sigma^2 \boldsymbol{I}_d$$

with i.i.d. Gaussian random variables $Z_i \sim \mathcal{N}(0, \boldsymbol{I}_K)$, where $\boldsymbol{I}_d$ and $\boldsymbol{I}_K$ are respectively the d- and K-dimensional identity matrices ($K < d$), $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ is the K-rank matrix that contains the principal axes and σ^2 is a noisy term that is known. We suppose here that data are centred. Hence, the distribution of each X_i is

$$P_{\boldsymbol{W}} := \mathcal{N}(0, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d).$$

We are not interested here in estimating the principal axes $\boldsymbol{W}$ and selecting the number of components K, but in estimating the true distribution of the X_i's.

Each model corresponds to a rank K. We place an equal prior weight over each integer $K = 1, ..., d$. Hence the optimization problem is equivalent to maximizing the ELBO as in Blei et al. (2017). Given rank K, we place a prior over the K-rank matrix $\boldsymbol{W}$ to infer a distribution over principal axes. We choose independent Gaussian priors $\mathcal{N}(0, s^2 \boldsymbol{I}_d)$

on the columns $W_1, ..., W_K$ of $\boldsymbol{W}$. We also consider Gaussian independent variational approximations $\mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j)$ for the columns of $\boldsymbol{W}$. Then, as soon as there exists a true model, i.e. there exists K_0 and $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ such that the true distribution of each X_i is $P_{\boldsymbol{W}_0} = \mathcal{N}(0, \boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)$, under the assumption that the coefficients of $\boldsymbol{W}_0$ are bounded, then Theorem 5 provides an explicit rate of convergence of our variational estimator even when K_0 is unknown:

Theorem 5 *For any $\alpha \in (0,1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ and such that the coefficients of $\boldsymbol{W}_0$ are bounded, then:*

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{W}}, P_{\boldsymbol{W}_0}) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\boldsymbol{W}|X_1^n) \right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n} \right).$$

The proof as well as the computation of the ELBO are detailed in the appendix. Note that this corollary can directly lead to a result in Frobenius distance between covariance matrices $\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d$ and $\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d$ instead of the Rényi divergence between the corresponding distributions even when $\boldsymbol{W}$ and $\boldsymbol{W}_0$ are not equal-sized matrices. We denote $\|.\|_F$ the Frobenius norm and $\|.\|_2$ the spectral norm of a matrix, which are respectively defined as the square root of the sum of the absolute squares of the elements of a matrix and as its largest singular value.

The following corollary assesses the consistency of the selected variational approximation to the true covariance matrix in Frobenius norm. The idea, borrowed from Alquier and Ridgway (2017), is to project matrices onto some set of bounded matrices under the assumption that the spectral norm of the true matrix $\boldsymbol{W}_0$ is also bounded:

Corollary 6 *For any $\alpha \in (0,1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is upper bounded by a positive constant $B > 0$, then:*

$$\mathbb{E}\left[\int \left\| clip_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0 \boldsymbol{W}_0^T \right\|_F^2 \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\boldsymbol{W}|X_1^n) \right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n} \right)$$

where $clip_B(\boldsymbol{A})$ is the matrix which (i,j)-entry is equal to $\begin{cases} \boldsymbol{A}_{i,j} & \text{if } |\boldsymbol{A}_{i,j}| \leq B^2 \\ B^2 & \text{if } \boldsymbol{A}_{i,j} \geq B^2 \\ -B^2 & \text{otherwise.} \end{cases}$

The requirement in our corollary is that the spectral norm of the true matrix $\boldsymbol{W}_0$ is bounded by some positive constant B, which implies the boundedness of the coefficients of the matrix as required in Theorem 5. In particular, the coefficients of the matrix $\boldsymbol{W}_0 \boldsymbol{W}_0^T$ are bounded by B^2:

$$\begin{aligned} |(\boldsymbol{W}_0 \boldsymbol{W}_0^T)_{i,j}| = \left| \sum_{k=1}^{K_0} (\boldsymbol{W}_0)_{i,k} (\boldsymbol{W}_0)_{j,k} \right| &\leq \left(\sum_{k=1}^{K_0} (\boldsymbol{W}_0)_{i,k}^2 \right)^{1/2} \left(\sum_{k=1}^{K_0} (\boldsymbol{W}_0)_{j,k}^2 \right)^{1/2} \\ &= \frac{\|\boldsymbol{W}_0 e_i\|_2}{\|e_i\|_2} \frac{\|\boldsymbol{W}_0 e_j\|_2}{\|e_j\|_2} \leq \|\boldsymbol{W}_0\|_2^2 \leq B^2 \end{aligned}$$

using Cauchy-Schwarz inequality and the property $\|\boldsymbol{W}_0\|_2 = \max_{x\neq 0} \frac{\|\boldsymbol{W}_0 x\|_2}{\|x\|_2}$ where e_ℓ is the vector of $\mathbb{R}^d$ which components are all equal to 0 except for the ℓ-th one that is set to 1. Hence it seems sensible to project (with respect to the Frobenius distance) any estimator $\boldsymbol{W}\boldsymbol{W}^T$ onto the set of all matrices whose entries lie in the interval $[-B^2, B^2]$, which is exactly what the $clip_B$ application does. Note that the spectral norm of Matrix $\boldsymbol{W}_0$ is equal to the largest eigenvalue of $\boldsymbol{W}_0\boldsymbol{W}_0^T$, so our assumption comes back to upper bounding the eigenvalues of the covariance matrix $\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d$, which is a classical assumption when estimating covariance matrices, see for instance Cai et al. (2015).

It is also possible to obtain a consistent pointwise covariance matrix estimator with the same convergence rate:

Corollary 7 *For any $\alpha \in (0,1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is bounded by B. Let us define a pointwise estimator of the covariance matrix:*

$$\hat{\boldsymbol{\Sigma}} = \int clip_B(\boldsymbol{W}\boldsymbol{W}^T)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\boldsymbol{W}|X_1^n) + \sigma^2\boldsymbol{I}_d.$$

Then,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\Sigma}} - (\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right\|_F^2\right] = \mathcal{O}\left(\frac{dK_0\log(dn)}{n}\right).$$

Discussion

In this paper we proved the consistency of ELBO maximization in model selection. By penalizing the variational lower bound using our prior beliefs over the different models, we showed that under mild conditions, the variational approximation associated with the selected model is consistent at the same convergence rate than the approximation associated with the true model. Moreover, the oracle inequality in Theorem 4 proved that the selected approximation is robust to misspecification. An application to the selection of the number of principal components in probabilistic PCA was provided as a short example.

We discuss in Appendix A the connection between the notions of *consistency* and *concentration*. This justifies the use of the α parameter in the definition of the evidence lower bound, as the regular posterior distribution is not robust to model misspecification. Indeed, authors of Grünwald and Van Ommen (2017) explain that there are pathologic cases where the regular posterior does not concentrate to the true distribution.

A point of interest when dealing with model selection is the question of recovering the true model (when it exists). This issue falls beyond the scope of this paper which treats the question of estimating the true distribution, and can be the object of future works. The true model recovery would require stronger assumptions, but the implementation in Section 5 in Bishop (1999) suggests that those may hold for probabilistic PCA.

Also, it would be interesting to study cross-validation instead of ELBO maximization. However, the tools used in this work such as the theory of penalized criteria and oracle inequalities were particularly suited to the ELBO, and thus a different theory should be used in order to obtain the consistency of validation log-likelihood in the VB framework. This question is left for future research.

Acknowledgments

I would like to warmly thank Pierre Alquier, Lionel Riou-Durand and the anonymous referees for their inspiring comments and suggestions on this work.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239):1–41, 2016.

G. Behrens, N. Friel, and M. Hurn. Tuning tempered transitions. *Statistics and computing*, 22(1):65–78, 2012.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125, to appear in the Annals of Statistics*, 2016.

A. Bhattacharya, D. Pati, and Y. Yang. On statistical optimality of variational Bayes. *Proceedings of Machine Learning Research*, 84 - AISTAT, 2018.

C. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514. IEE, January 1999.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, Apr 2015. ISSN 1432-2064. doi: 10.1007/s00440-014-0562-z.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

B. Chérief-Abdellatif and P. Alquier. Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035, 2018. ISSN 1935-7524. doi: 10.1214/18-EJS1475.

S. Forth, P. Hovland, E. Phipps, J. Utke, and A. Walther. *Recent Advances in Algorithmic Differentiation*. Springer Publishing Company, Incorporated, 2014. ISBN 3642439918, 9783642439919.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

P. D. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

J. Lv and J. S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167, 2013. doi: 10.1111/rssb.12023.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, Edited by Jean Picard.

K.I. Moridomi, K. Hatano, and E. Takimoto. Online linear optimization with the log-determinant regularizer. *IEICE Transactions on Information and Systems*, E101D(6): 1511–1520, 6 2018. ISSN 0916-8532. doi: 10.1587/transinf.2017EDP7317.

C. R. Rao and Y. Wu. *On model selection*, volume Volume 38 of *Lecture Notes–Monograph Series*, pages 1–57. Institute of Mathematical Statistics, Beachwood, OH, 2001. doi: 10.1214/lnms/1215540960.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

A. Vehtari, V. Tolvanen, T. Mononen, and O. Winther. Bayesian leave-one-out cross validation approximations for gaussian latent variable models. *Journal of Machine Learning Research*, 17, 12 2014.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, pages 1–85, 2018.

Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519v1*, 2017.

Appendix A. Connection between consistency and concentration.

In this appendix, we highlight the connection between the notions of *consistency* used in Alquier and Ridgway (2017) and Chérief-Abdellatif and Alquier (2018) and *concentration*. We consider a true model $\mathcal{M}_{K_0}$ to which the true distribution $P^0 = P_{\theta^0}$ belongs, $\theta^0 \in \Theta_{K_0}$. We recall that the Bayesian estimator $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$ is said to be consistent if, in expectation (with respect to the random variables distributed according to P^0), the average Rényi loss between a distribution in the selected model and the true distribution (over the Bayesian estimator) goes to zero as $n \to +\infty$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \xrightarrow[n\to+\infty]{} 0.$$

Similarly, we give the definition of *concentration* at rate s_n of the selected variational approximation to P^0 as stated in Ghosal et al. (2000), that is, in probability (with respect to the random variables distributed according to P^0), the approximation concentrates asymptotically around the true distribution as $n \to +\infty$, i.e. in probability:

$$\tilde{\pi}^{\hat{K}}_{n,\alpha}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right) \xrightarrow[n\to+\infty]{} 0$$

for any constant $M > 0$. The reference metric here is the α-Rényi divergence.

We show in this appendix that the consistency of the selected variational approximation to P^0 at rate r_n implies the concentration of the selected variational approximation to P^0 at any rate s_n such that $r_n = o(s_n)$ and $s_n \to 0$ as $n \to +\infty$, as for instance $s_n = r_n \log(\log(n))$ when the consistency rate r_n is slower than a log-logarithmic one.

To do so, we assume that the selected variational approximation is consistent to P^0 at rate r_n, i.e.:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \leq r_n.$$

Then, using Markov's inequality for any s_n such that $r_n = o(s_n)$ and $s_n \to 0$ and any constant $M > 0$:

$$\mathbb{E}\left[\tilde{\pi}^{\hat{K}}_{n,\alpha}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right)\right] \leq \frac{\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\right]}{Ms_n} \leq \frac{r_n}{Ms_n} \xrightarrow[n\to+\infty]{} 0.$$

Hence, we obtain the convergence in mean of $\tilde{\pi}^{\hat{K}}_{n,\alpha}\big(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\big)$ to 0, which implies the convergence in probability of $\tilde{\pi}^{\hat{K}}_{n,\alpha}\big(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\big)$ to 0, i.e. the concentration of $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$ to P^0 at rate s_n.

Appendix B. Proof of Theorem 3.

First, we need Donsker and Varadhan's famous variational formula. Refer for example to Catoni (2007) for a proof (Lemma 1.1.3).

Lemma 8 *For any probability λ on some measurable space $(\boldsymbol{E}, \mathcal{E})$ and any measurable function $h : \boldsymbol{E} \to \mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\lambda < \infty$,*

$$\log \int \mathrm{e}^h \mathrm{d}\lambda = \sup_{\rho \in \mathcal{M}_1^+(\boldsymbol{E})} \left\{ \int h \mathrm{d}\rho - KL(\rho, \lambda) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, if h is upper-bounded on the support of λ, then the supremum on the right-hand side is reached by the distribution of the form:

$$\lambda_h(d\beta) = \frac{\mathrm{e}^{h(\beta)}}{\int \mathrm{e}^h \mathrm{d}\lambda} \lambda(\mathrm{d}\beta).$$

We come back to the proof of Theorem 3. We adapt the proof of Theorem 4.1 in Chérief-Abdellatif and Alquier (2018).

Proof For any $\alpha \in (0,1)$ and $\theta \in \Omega := \cup_{K \geq 1} \Theta_K$, using the definition of Rényi divergence and $D_\alpha(P^{\otimes n}, R^{\otimes n}) = n D_\alpha(P, R)$ as data are i.i.d.:

$$\mathbb{E}\left[\exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \right] = 1$$

where $r_n(P_\theta, P^0) = \sum_{i=1}^n \log(P^0(X_i)/P_\theta(X_i))$ is the negative log-likelihood ratio. Then we integrate and use Fubini's theorem,

$$\mathbb{E}\left[\int \exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \pi(d\theta) \right] = 1.$$

Using Lemma 8,

$$\mathbb{E}\left[\exp\left(\sup_{\rho \in \mathcal{M}_1^+(\Omega)} \left\{ \int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \rho(d\theta) - \mathrm{KL}(\rho, \pi) \right\} \right) \right] = 1.$$

Then, using Jensen's inequality,

$$\mathbb{E}\left[\sup_{\rho \in \mathcal{M}_1^+(\Omega)} \left\{ \int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \rho(d\theta) - \mathrm{KL}(\rho, \pi) \right\} \right] \leq 0.$$

Now, we consider $\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n)$ as a distribution on $\mathcal{M}_1^+(\Omega)$ with all its mass on $\Theta_{\hat{K}}$,

$$\mathbb{E}\left[\int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n) - \mathrm{KL}(\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n), \pi) \right] \leq 0.$$

We use $\mathrm{KL}(\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n), \pi) = \mathrm{KL}(\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n), \Pi_{\hat{K}}) + \log(\frac{1}{\pi_{\hat{K}}})$, and we rearrange terms:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n) \right]$$
$$\leq \mathbb{E}\left[\frac{\alpha}{1-\alpha} \int \frac{r_n(P_\theta, P^0)}{n} \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n) + \frac{\mathrm{KL}(\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n), \Pi_{\hat{K}})}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_{\hat{K}}})}{n(1-\alpha)} \right].$$

By definition of $\hat{K}$,

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\Big] \leq \mathbb{E}\Big[\inf_{K\geq 1}\Big\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\tilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n)+\frac{\mathrm{KL}(\tilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Big\}\Big]$$

which gives

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\Big\{\mathbb{E}\Big[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\tilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n)+\frac{\mathrm{KL}(\tilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Big]\Big\}$$

and hence, by definition of $\tilde{\pi}_{n,\alpha}^{K}(.|X_1^n)$,

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\Big\{\mathbb{E}\Big[\inf_{\rho\in\mathcal{F}_K}\Big\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\rho(d\theta)+\frac{\mathrm{KL}(\rho,\Pi_K)}{n(1-\alpha)}\Big\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Big]\Big\}.$$

which leads to,

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\inf_{\rho\in\mathcal{F}_K}\Big\{\mathbb{E}\Big[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\rho(d\theta)+\frac{\mathrm{KL}(\rho,\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Big]\Big\}.$$

Finally,

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\Big\{\inf_{\rho_K\in\mathcal{F}_K}\Big\{\frac{\alpha}{1-\alpha}\int\mathrm{KL}(P^0,P_{\theta_K})\rho_K(d\theta_K)+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\Big\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Big\}.$$

The theorem is a direct corollary of this inequality as soon as Assumption 3.1 is satisfied. ■

Appendix C. Proof of Theorem 4.

Proof

Fix $\alpha\in(0,1)$ and let us prove Theorem 4. Let us recall that $\Theta_K(r_{K,n})$ is defined as the set of parameters $\theta_K^*\in\Theta_K$, for which there is a distribution $\rho_{K,n}\in\mathcal{F}_K$ such that:

$$\int\mathbb{E}\Big[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\Big]\rho_{K,n}(d\theta_K)\leq r_{K,n}\ \text{ and }\ \mathrm{KL}(\rho_{K,n},\Pi_K)\leq nr_{K,n}.$$

We begin from:

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\left\{\inf_{\rho_K\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int \mathrm{KL}(P^0,P_{\theta_K})\rho_K(d\theta_K)+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

Then, we write for any K, any $\theta_K\in\Theta_K$, $\theta_K^*\in\Theta_K$:

$$\mathrm{KL}(P^0,P_{\theta_K})=\mathrm{KL}(P^0,P_{\theta_K^*})+\mathbb{E}\left[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right]$$

which gives:

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\Big] \leq \inf_{K\geq 1}\left\{\inf_{\theta_K^*\in\Theta_K}\left\{\frac{\alpha}{1-\alpha}\mathrm{KL}(P^0,P_{\theta_K^*})+\inf_{\rho_K\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int\mathbb{E}\left[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right]\rho_K(d\theta_K)+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\right\}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

Hence, using the definition of $\Theta_K(r_{K,n})$ and upper bounding the right-hand-side of the previous inequality by an inf over $\Theta_K(r_{K,n})$, we conclure:

$$\mathbb{E}\Big[\int D_\alpha(P_\theta,P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta|X_1^n)\Big]\leq\inf_{K\geq1}\left\{\frac{\alpha}{1-\alpha}\inf_{\theta^*\in\Theta_K(r_{K,n})}\mathrm{KL}(P^0,P_{\theta_K^*})+\frac{1+\alpha}{1-\alpha}r_{K,n}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

■

Appendix D. Proof of Theorem 5.

Proof

We still consider the framework of probabilistic PCA in Section 4. We assume that there exists a true rank K_0 and a matrix $\boldsymbol{W}_0\in\mathbb{R}^{d\times K_0}$ with bounded coefficients such that the true distribution of each X_i is $\mathcal{N}(0,\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)$, and we place a prior $\Pi_{K_0}=\mathcal{N}(0,s^2\boldsymbol{I}_d)^{\otimes K_0}$ and a variational approximation $\rho_{K_0}=\rho^{\otimes K_0}$ on W given $K=K_0$ where we denote $\rho=\mathcal{N}(0,\frac{1}{dn^2}\boldsymbol{I}_d)$. We recall that $\pi_K=\frac{1}{d}$ for any $K=1,...,d$.

To obtain the rate of convergence $r_n=dK_0\log(nd)/n$ for probabilistic PCA, we just need to show that the quantities in Assumption 3.1 are upper bounded by r_n (up to a constant) as we have $\log(1/\pi_{K_0})/n$ much smaller than r_n:

$$\int\mathrm{KL}\Big(\mathcal{N}(0,\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d),\mathcal{N}(0,\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)\Big)\rho_{K_0}(d\theta_K)\quad,\quad\frac{\mathrm{KL}(\rho_{K_0},\Pi_{K_0})}{n}.$$

We have two terms. The first one, i.e. the Kullback-Leibler term, provides a rate of convergence of $dK_0 \log(dn)/n$ as:

$$\begin{aligned}
\mathrm{KL}(\rho_{K_0}, \Pi_{K_0}) &= \sum_{j=1}^{K_0} \mathrm{KL}\Big(\mathcal{N}(0, \frac{1}{dn^2}\boldsymbol{I}_d), \mathcal{N}(0, s^2\boldsymbol{I}_d)\Big) \\
&= \frac{K_0}{2}\left(\frac{1}{n^2 s^2} - d + d\log(s^2) + d\log(dn^2)\right) \\
&\leq \frac{K_0}{2n^2 s^2} - \frac{dK_0}{2} + \frac{dK_0 \log(s^2)}{2} + dK_0 \log(dn).
\end{aligned}$$

The integral is much more complicated to deal with. We will show that it leads to a rate faster than $dK_0 \log(dn)/n$. If we denote $\mathbb{E}$ the expectation with respect to ρ_{K_0}, then the integral will be equal to:

$$\frac{1}{2}\mathbb{E}\left[\mathrm{Tr}\left((\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)^{-1}(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right)\right] - \frac{d}{2} + \frac{1}{2}\mathbb{E}\left[\log\left(\frac{\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)}{\det(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)}\right)\right].$$

The expectation of the log-ratio is easy to upper bound. We denote $\lambda_1, ..., \lambda_d$ the positive eigenvalues of the positive definite matrix $\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d$. Then for each $j = 1, ..., d$, $\lambda_j \geq \sigma^2$ and using Jensen's inequality and the log-concavity of the determinant:

$$\begin{aligned}
\mathbb{E}\left[\log\left(\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)\right)\right] &\leq \log\left(\det\left(\mathbb{E}[\boldsymbol{W}\boldsymbol{W}^T] + \sigma^2\boldsymbol{I}_d\right)\right) \\
&= \log\left(\det\left(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d + \frac{1}{dn^2}\boldsymbol{I}_d\right)\right) \\
&= \sum_{j=1}^{d}\log\left(\lambda_j + \frac{1}{dn^2}\right) \\
&= \sum_{j=1}^{d}\log(\lambda_j) + \sum_{j=1}^{d}\log\left(1 + \frac{1}{\lambda_j dn^2}\right) \\
&= \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right)\right] + \sum_{j=1}^{d}\log\left(1 + \frac{1}{\lambda_j dn^2}\right) \\
&\leq \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right)\right] + \sum_{j=1}^{d}\frac{1}{\lambda_j dn^2} \\
&\leq \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right)\right] + \frac{1}{n^2\sigma^2}
\end{aligned}$$

and then the expectation of the log-ratio provides a rate of convergence of $1/n^2$:

$$\mathbb{E}\left[\log\left(\frac{\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)}{\det(\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)}\right)\right] \leq \frac{1}{n^2\sigma^2}.$$

The remainder can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\Big[\mathrm{Tr}\Big((\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}(\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)\Big)\Big]-d& \\
&=\mathbb{E}\Big[\mathrm{Tr}\Big((\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}(\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T)\Big)\Big]\\
&\leq\mathbb{E}\Big[\|(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}\|_F\times\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]\\
&\leq\sqrt{d}\mathbb{E}\Big[\|(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}\|_2\times\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]\\
&=\sqrt{d}\mathbb{E}\Big[\sigma_{\max}\big((\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)^{-1}\big)\times\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]\\
&=\sqrt{d}\mathbb{E}\Big[\sigma_{\min}(\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)^{-1}\times\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]
\end{aligned}$$

i.e.

$$\begin{aligned}
\mathbb{E}\Big[\mathrm{Tr}\Big((\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}(\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)\Big)\Big]-d&\leq\sqrt{d}\mathbb{E}\Big[(\sigma^2)^{-1}\times\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]\\
&=\frac{\sqrt{d}}{\sigma^2}\mathbb{E}\Big[\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]
\end{aligned}$$

where $\|.\|_F$ is the Frobenius norm on matrices, $\|.\|_2$ the spectral norm, and $\sigma_{\min}(\boldsymbol{A})$, $\sigma_{\max}(\boldsymbol{A})$ the lowest and largest singular values of a matrix $\boldsymbol{A}$. We use the fact that for a symmetric semi-definite positive matrix: $\sigma_{\max}(\boldsymbol{A}^{-1})=\big(\sigma_{\min}(\boldsymbol{A})\big)^{-1}$ and $\sigma_{\min}(\boldsymbol{A}+\sigma^2\boldsymbol{I}_d)\geq\sigma^2$, as well as the inequality $\|\boldsymbol{A}\|_F\leq\sqrt{d}\|\boldsymbol{A}\|_2$ for any $d\times d$ matrix $\boldsymbol{A}$.

The only thing left to do is to upper bound the expectation of the Frobenius norm of $\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T$ by a multiple of $\frac{\sqrt{d}K_0\log(dn)}{n}$. We use the triangle and Cauchy-Schwarz's inequalities:

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big]&\leq\mathbb{E}\Big[\|\boldsymbol{W}\boldsymbol{W}^T-\boldsymbol{W}\boldsymbol{W}_0^T\|_F\Big]+\mathbb{E}\Big[\|\boldsymbol{W}\boldsymbol{W}_0^T-\boldsymbol{W}_0\boldsymbol{W}_0^T\|_F\Big]\\
&\leq\mathbb{E}\Big[\|\boldsymbol{W}(\boldsymbol{W}-\boldsymbol{W}_0)^T\|_F\Big]+\mathbb{E}\Big[\|(\boldsymbol{W}-\boldsymbol{W}_0)\boldsymbol{W}_0^T\|_F\Big]\\
&\leq\mathbb{E}\Big[\|\boldsymbol{W}\|_F\|\boldsymbol{W}-\boldsymbol{W}_0\|_F\Big]+\mathbb{E}\Big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F\|\boldsymbol{W}_0\|_F\Big]\\
&\leq\sqrt{\mathbb{E}\big[\|\boldsymbol{W}\|_F^2\big]\mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]}+\sqrt{\mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]\mathbb{E}\big[\|\boldsymbol{W}_0\|_F^2\big]}\\
&\leq\sqrt{\mathbb{E}\big[\|\boldsymbol{W}\|_F^2\big]\mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]}+\|\boldsymbol{W}_0\|_F\sqrt{\mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]}.
\end{aligned}$$

We can upper bound $\|\boldsymbol{W}_0\|_F=\sqrt{\sum_{i=1}^d\sum_{j=1}^{K_0}(\boldsymbol{W}_0)_{i,j}^2}$ by $\sqrt{dK_0}C$ where C is an upper bound on each of the coefficients of matrix $\boldsymbol{W}_0$.

Also, we can notice that $dn^2\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2=\sum_{i=1}^d\sum_{j=1}^{K_0}\big(\sqrt{d}n(\boldsymbol{W}_{i,j}-(\boldsymbol{W}_0)_{i,j})\big)^2$ is a sum of squares of independent standard normal random variables. Thus $dn^2\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2$

follows a chi-squared distribution with dK_0 degrees of freedom and its expectation is equal to dK_0. Hence:

$$\mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]=\frac{K_0}{n^2}.$$

Similarly, as $\boldsymbol{W}_{i,j}-(\boldsymbol{W}_0)_{i,j}$ is centered, we get:

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{W}\|_F^2\big] &= \mathbb{E}\bigg[\sum_{i=1}^{d}\sum_{j=1}^{K_0}\boldsymbol{W}_{i,j}^2\bigg]\\
&= \sum_{i=1}^{d}\sum_{j=1}^{K_0}\mathbb{E}\Big[\big(\boldsymbol{W}_{i,j}-(\boldsymbol{W}_0)_{i,j}\big)^2+(\boldsymbol{W}_0)_{i,j}^2-2(\boldsymbol{W}_0)_{i,j}\big(\boldsymbol{W}_{i,j}-(\boldsymbol{W}_0)_{i,j}\big)\Big]\\
&= \mathbb{E}\big[\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2\big]+\|\boldsymbol{W}_0\|_F^2\\
&\leq \frac{K_0}{n^2}+dK_0C^2\\
&= \bigg(dC^2+\frac{1}{n^2}\bigg)K_0.
\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T\|_F\Big] &\leq \frac{\sqrt{K_0}}{n}\sqrt{K_0}\sqrt{dC^2+\frac{1}{n^2}}+\sqrt{dK_0}C\frac{\sqrt{K_0}}{n}\\
&= \frac{K_0}{n}\sqrt{dC^2+\frac{1}{n^2}}+\frac{\sqrt{d}K_0C}{n}\\
&\leq \frac{K_0}{n}\bigg(\sqrt{d}C+\frac{1}{n}\bigg)+\frac{\sqrt{d}K_0C}{n}\\
&= \frac{K_0}{n}\bigg(2\sqrt{d}C+\frac{1}{n}\bigg).
\end{aligned}$$

Hence, the order of the upper bound of the expectation of the Fobrenius norm of matrix $\boldsymbol{W}_0\boldsymbol{W}_0^T-\boldsymbol{W}\boldsymbol{W}^T$ is $\frac{\sqrt{d}K_0}{n}<\frac{\sqrt{d}K_0\log(dn)}{n}$.

Finally, the consistency rate associated with the integral term is $\frac{dK_0}{n}$, and the overall rate of convergence is $\frac{dK_0\log(dn)}{n}$. ■

Appendix E. Computation of the ELBO for probabilistic PCA.

We consider the framework of probabilistic PCA detailed in Section 4. Given rank K, we place independent Gaussian priors on the columns $W_1,...,W_K$ of $\boldsymbol{W}$ such that $\Pi_K=\mathcal{N}(0,s^2\boldsymbol{I}_d)^{\otimes K}$, and Gaussian independent variational approximations $\mathcal{N}(\mu_j,\boldsymbol{\Sigma}_j)$ for the columns of $\boldsymbol{W}$. The ELBO associated with rank K and variational approximation $\rho_K=\otimes_{j=1}^K\mathcal{N}(\mu_j,\boldsymbol{\Sigma}_j)$ is given by:

$$\mathrm{ELBO}_K(\rho_K)=\alpha\int\ell_n(\boldsymbol{W})\rho_K(d\boldsymbol{W})-\mathrm{KL}\big(\rho_K,\Pi_K\big).$$

The Kullback-Leibler term $\mathrm{KL}\big(\rho_K, \Pi_K\big)$ is equal to:

$$\frac{1}{2}\sum_{j=1}^{K}\left\{\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_j)}{s^2}+\frac{\mu_j^T\mu_j}{s^2}-\log\big(\det(\boldsymbol{\Sigma}_j)\big)\right\}-\frac{dK}{2}+\frac{dK\log(s^2)}{2}$$

while the average log-likelihood $\int \ell_n(\boldsymbol{W})\rho_K(d\boldsymbol{W})$ is:

$$-\frac{dn}{2}\log(2\pi)-\frac{n}{2}\int\log\big(\det(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)\big)\rho_K(dW)-\frac{1}{2}\sum_{i=1}^{n}\int X_i^T(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}X_i\,\rho_K(d\boldsymbol{W})$$

where both integrals can be computed thanks to Monte-Carlo sampling approximations:

$$\int\log\big(\det(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)\big)\rho_K(d\boldsymbol{W})\approx\sum_{\ell=1}^{N}\log\big(\det(\boldsymbol{W}^{(\ell)}\boldsymbol{W}^{(\ell)T}+\sigma^2\boldsymbol{I}_d)\big)$$

and

$$\int X_i^T(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}X_i\,\rho_K(d\boldsymbol{W})\approx\sum_{\ell=1}^{N}X_i^T(\boldsymbol{W}^{(\ell)}\boldsymbol{W}^{(\ell)T}+\sigma^2\boldsymbol{I}_d)^{-1}X_i$$

where $\boldsymbol{W}^{(1)},...,\boldsymbol{W}^{(N)}$ are N i.i.d. data sampled from ρ_K.

The inverse matrix $(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}$ can be derived thanks to classical inversion algorithms. For instance, it is possible to do so in $\mathcal{O}(Kd^2)$ operations instead of the classical $\mathcal{O}(d^3)$ inversion procedure thanks to Sherman-Morrison formula: for any matrix $\boldsymbol{A}\in\mathbb{R}^{d\times d}$ and vectors $u,v\in\mathbb{R}^d$ such that $\boldsymbol{A}+uv^T$ is invertible,

$$(\boldsymbol{A}+uv^T)^{-1}=\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}uv^T\boldsymbol{A}^{-1}}{1+v^T\boldsymbol{A}^{-1}u}.$$

We write

$$\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d=\sigma^2\boldsymbol{I}_d+\sum_{j=1}^{K}W_jW_j^T=\left(\sigma^2\boldsymbol{I}_d+\sum_{j=1}^{K-1}W_jW_j^T\right)+W_KW_K^T$$

and iterate K times Sherman-Morrison formula. The first time, we apply it to $\boldsymbol{A}=\sigma^2\boldsymbol{I}_d+\sum_{j=1}^{K-1}W_jW_j^T$ and $u=v=\boldsymbol{W}_K$, then to $\boldsymbol{A}=\sigma^2\boldsymbol{I}_d+\sum_{j=1}^{K-2}W_jW_j^T$ and $u=v=W_{K-1}$, and so on. We finally obtain $(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}=\boldsymbol{M}_K$ where:

$$\begin{cases}\boldsymbol{M}_0=\sigma^2\boldsymbol{I}_d\\ \forall j=1,...,K,\ \ \boldsymbol{M}_j=\boldsymbol{M}_{j-1}-\frac{1}{1+W_j^TZ_j}Z_jZ_j^T\ \text{with}\ Z_j=\boldsymbol{M}_{j-1}W_j.\end{cases}$$

In order to compute the maximum value ELBO(K) of the ELBO associated with rank K, one can use a stochastic gradient descent on $(\mu_1,\boldsymbol{\Sigma}_1,...,\mu_K,\boldsymbol{\Sigma}_K)$ that will converge to a local maximum and will give the variational estimator for rank K. Then, maximizing ELBO(K) over desired values of K leads to the optimal number of principal components and to the associated optimal variational approximation.

Appendix F. Results in matrix norm for probabilistic PCA.

To prove Corollaries 6 and 7, we need the two lemmas presented behind. We introduce some notations first. We refer the interested reader to Forth et al. (2014) for more details.

Notations : Let us call $\mathcal{S}_d^+$ the set of $d \times d$ symmetric positive semi-definite matrices, and $\mathcal{X}_M = \left\{ \boldsymbol{A} \in \mathcal{S}_d^+ / \|\boldsymbol{A}\|_2 \leq M \right\}$. We define the vectorization of Matrix $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ with columns $X_1, ..., X_q$:

$$\mathrm{Vec}(\boldsymbol{A}) = (\boldsymbol{A}_1^T, ..., \boldsymbol{A}_q^T)^T \in \mathbb{R}^{p \times q}.$$

We define the Frobenius inner product of two matrices $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ and $\tilde{\boldsymbol{A}} \in \mathbb{R}^{p \times q}$, that is the sum of componentwise products:

$$\boldsymbol{A} \cdot \tilde{\boldsymbol{A}} = \mathrm{Vec}(\boldsymbol{A})^T \mathrm{Vec}(\tilde{\boldsymbol{A}}).$$

Notice that $\|\boldsymbol{A}\|_F^2 = \boldsymbol{A} \cdot \boldsymbol{A} = \mathrm{Vec}(\boldsymbol{A})^T \mathrm{Vec}(\boldsymbol{A})$.

We also introduce the Kronecker and Box products of two matrices $\boldsymbol{A} \in \mathbb{R}^{p_1 \times q_1}$ and $\tilde{\boldsymbol{A}} \in \mathbb{R}^{p_2 \times q_2}$ which are respectively the matrices $\boldsymbol{A} \otimes \tilde{\boldsymbol{A}} \in \mathbb{R}^{p_1 p_2 \times q_1 q_2}$ and $\boldsymbol{A} \boxtimes \tilde{\boldsymbol{A}} \in \mathbb{R}^{p_1 p_2 \times q_1 q_2}$ such that their coefficients are defined as:

$$(\boldsymbol{A} \otimes \tilde{\boldsymbol{A}})_{p_2(i-1)+j, q_2(k-1)+l} = \boldsymbol{A}_{i,k} \tilde{\boldsymbol{A}}_{j,l},$$
$$(\boldsymbol{A} \boxtimes \tilde{\boldsymbol{A}})_{p_2(i-1)+j, q_1(k-1)+l} = \boldsymbol{A}_{i,l} \tilde{\boldsymbol{A}}_{j,k}$$

for any integers i, j, k, l such that $1 \leq i \leq p_1$, $1 \leq j \leq q_1$, $1 \leq k \leq p_2$, $1 \leq l \leq q_2$.

We have the following properties for any matrix $\boldsymbol{P}$:

$$(\boldsymbol{A} \otimes \tilde{\boldsymbol{A}}) \mathrm{Vec}(\boldsymbol{P}) = \mathrm{Vec}(\tilde{\boldsymbol{A}} \boldsymbol{P} \boldsymbol{A}^T),$$
$$(\boldsymbol{A} \boxtimes \tilde{\boldsymbol{A}}) \mathrm{Vec}(\boldsymbol{P}) = \mathrm{Vec}(\tilde{\boldsymbol{A}} \boldsymbol{P}^T \boldsymbol{A}^T).$$

We also define the gradient $\nabla f(\boldsymbol{A}) \in \mathbb{R}^{p \times q}$ and the Hessian $\nabla^2 f(\boldsymbol{A}) \in \mathbb{R}^{pq \times pq}$ of a differentiable function $f : \mathbb{R}^{p \times q} \to \mathbb{R}$ at matrix $\boldsymbol{A}$:

$$(\nabla f(\boldsymbol{A}))_{p_2(i-1)+j, q_2(k-1)+l} = \frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}_{i,j}},$$
$$(\nabla^2 f(\boldsymbol{A}))_{p_2(j-1)+i, p_2(l-1)+k} = \frac{\partial^2 f(\boldsymbol{A})}{\partial \boldsymbol{A}_{i,j} \partial \boldsymbol{A}_{k,l}}$$

for any integers i, j, k, l such that $1 \leq i, k \leq p$, $1 \leq j, l \leq q$ where ∂f is the partial derivative of f.

We say that a differentiable function $f : \mathbb{R}^{p \times q} \to \mathbb{R}$ is s-strongly convex in $\mathcal{S} \subset \mathbb{R}^{pq \times pq}$ with respect to the norm $\|.\|$ as soon as one of the two following equivalent properties is satisfied:

$$f(\boldsymbol{A}) \geq f(\tilde{\boldsymbol{A}}) + \nabla f(\boldsymbol{A}) \cdot (\boldsymbol{A} - \tilde{\boldsymbol{A}}) + \frac{s}{2} \|\boldsymbol{A} - \tilde{\boldsymbol{A}}\|^2$$

or

$$\mathrm{Vec}(\boldsymbol{P})^T \nabla^2 f(\boldsymbol{A}) \mathrm{Vec}(\boldsymbol{P}) \geq s \|\boldsymbol{P}\|^2$$

for any matrix $\boldsymbol{A}, \tilde{\boldsymbol{A}} \in \mathcal{S}$ and any symmetric matrix $\boldsymbol{P} \in \mathbb{R}^{pq \times pq}$.

Lemma 9 *Then, function $f : \boldsymbol{A} \to -\log\left(\det(\boldsymbol{A} + M \boldsymbol{I}_d)\right)$ is $1/(M + \sigma^2)^2$ strongly convex in $\mathcal{X}_M$ with respect to the Frobenius norm.*

Proof The proof follows the same steps than the proof of Theorem 3.1 in Moridomi et al. (2018).

The Hessian of function f at any symmetric matrix in $\boldsymbol{A} \in \mathcal{X}_M$ is given by (see Forth et al. (2014)):

$$\nabla^2 f(\boldsymbol{A}) = \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big)^T \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} = (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}.$$

Then, we have for any $\boldsymbol{A} \in \mathcal{X}_M$ and any symmetric matrix $\boldsymbol{P} \in \mathbb{R}^{pq \times pq}$:

$$\begin{aligned}
\mathrm{Vec}(\boldsymbol{P})^T \nabla^2 f(\boldsymbol{A}) \mathrm{Vec}(\boldsymbol{P}) &= \mathrm{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \mathrm{Vec}(\boldsymbol{P}) \\
&= \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}\Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boldsymbol{P}^T (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \\
&= \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}\Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boldsymbol{P} (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \\
&= \mathrm{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \mathrm{Vec}(\boldsymbol{P}).
\end{aligned}$$

Note that the eigenvalues of a Kronecker product $\boldsymbol{A} \otimes \boldsymbol{P}$ are the products of an eigenvalue of $\boldsymbol{A}$ and an eigenvalue of $\boldsymbol{P}$, and the eigenvalues of $\boldsymbol{P}^{-1}$ are the inverse of the eigenvalues of $\boldsymbol{P}$. Moreover, the maximum eigenvalue of $\boldsymbol{A} + M\boldsymbol{I}_d$ is $\|\boldsymbol{A}\|_2 + \sigma^2$, so the minimum eigenvalue of $(\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}$ is equal to $(\|\boldsymbol{A}\|_2 + \sigma^2)^{-2}$. Hence, for any matrix $\boldsymbol{A} \in \mathcal{X}_M$, we get:

$$\begin{aligned}
\mathrm{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \mathrm{Vec}(\boldsymbol{P})^T &\geq (\|\boldsymbol{A}\|_2 + \sigma^2)^{-2} \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P}) \\
&\geq \frac{1}{(M + \sigma^2)^2} \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P}),
\end{aligned}$$

and we conclude using the definition of the strong convexity and $\|\boldsymbol{P}\|_F^2 = \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P})$. ■

Lemma 10 *For any $\alpha \in (0, 1)$ and any matrices $\boldsymbol{W} \in \mathbb{R}^{d \times K_1}$ and $\tilde{\boldsymbol{W}} \in \mathbb{R}^{d \times K_2}$, as soon as the spectral norms of $\boldsymbol{W}\boldsymbol{W}^T$ and $\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T$ are bounded by a constant B^2, then:*

$$D_\alpha(P_{\boldsymbol{W}}, P_{\tilde{\boldsymbol{W}}}) \geq \frac{\alpha}{16(B^2 + \sigma^2)^2} \big\| \tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T - \boldsymbol{W}\boldsymbol{W}^T \big\|_F^2.$$

Proof

We recall that function $f : \boldsymbol{A} \to -\log\big(\det(\boldsymbol{A} + M\boldsymbol{I}_d)\big)$ is $1/(M + \sigma^2)^2$ strongly convex in $\mathcal{X}_M$ with respect to the Fobrenius norm according to Lemma 9. Hence, for any matrices $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ in $\mathcal{X}_M$, we have:

$$\begin{aligned}
-\log\big(\det((1-\alpha)\boldsymbol{A} + \alpha\tilde{\boldsymbol{A}})\big) \leq -(1-\alpha)\log\big(\det(\boldsymbol{A})\big) - \alpha\log\big(\det(\tilde{\boldsymbol{A}})\big) \\
- \frac{1}{2}\alpha(1-\alpha)\frac{1}{4M^2}\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2.
\end{aligned}$$

We rearrange terms:

$$\log\left(\frac{\det\left((1-\alpha)\boldsymbol{A}+\alpha\tilde{\boldsymbol{A}}\right)}{\det(\boldsymbol{A})^{1-\alpha}\det(\tilde{\boldsymbol{A}})^{\alpha}}\right) \geq \frac{\alpha(1-\alpha)}{8M^2}\|\tilde{\boldsymbol{A}}-\boldsymbol{A}\|_F^2.$$

Now, we use the fact that:

$$D_\alpha\big(\mathcal{N}(0,\boldsymbol{A}),\mathcal{N}(0,\tilde{\boldsymbol{A}})\big) = \frac{1}{2(1-\alpha)}\log\left(\frac{\det\left((1-\alpha)\boldsymbol{A}+\alpha\tilde{\boldsymbol{A}}\right)}{\det(\boldsymbol{A})^{1-\alpha}\det(\tilde{\boldsymbol{A}})^{\alpha}}\right)$$

to get for any matrices $\boldsymbol{W}\in\mathbb{R}^{d\times K_1}$ and $\tilde{\boldsymbol{W}}\in\mathbb{R}^{d\times K_2}$ such that $\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T+\sigma^2\boldsymbol{I}_d\|_2\leq M$ and $\|\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d\|_F\leq M$:

$$D_\alpha\big(P_{\boldsymbol{W}},P_{\tilde{\boldsymbol{W}}}\big)\geq\frac{\alpha}{16M^2}\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T+\sigma^2\boldsymbol{I}_d-\boldsymbol{W}\boldsymbol{W}^T-\sigma^2\boldsymbol{I}_d\|_F^2.$$

Moreover, for any matrix $\boldsymbol{W}\in\mathbb{R}^{d\times K}$ such that the spectral norm of $\boldsymbol{W}\boldsymbol{W}^T$ is bounded by B^2, we have $\|\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d\|_2\leq B^2+\sigma^2$. We conclude using the previous inequality for $M=B^2+\sigma^2$.

■

Now, let us go back to the proof of Corollary 6.

Proof

We assume that there exists a true model $\mathcal{M}_{K_0}$ such that $P^0=P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0\in\mathbb{R}^{d\times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is bounded by B (hence the coefficients of $\boldsymbol{W}_0$ are also bounded). As clip_B is a projection onto a closed convex set with respect to the Frobenius norm, we have for any matrix $\boldsymbol{W}\in\mathbb{R}^{d\times\hat{K}}$:

$$\big\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T)-\mathrm{clip}_B(\boldsymbol{W}_0\boldsymbol{W}_0^T)\big\|_F\leq\big\|\boldsymbol{W}\boldsymbol{W}^T-\boldsymbol{W}_0\boldsymbol{W}_0^T\big\|_F$$

and as the coefficients of $\boldsymbol{W}_0\boldsymbol{W}_0^T$ are bounded by B^2:

$$\big\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T)-\boldsymbol{W}_0\boldsymbol{W}_0^T\big\|_F=\big\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T)-\mathrm{clip}_B(\boldsymbol{W}_0\boldsymbol{W}_0^T)\big\|_F.$$

According to Lemma 10, we get for any matrix $\boldsymbol{W}\in\mathbb{R}^{d\times\hat{K}}$:

$$\big\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T)-\boldsymbol{W}_0\boldsymbol{W}_0^T\big\|_F^2\leq\frac{16(B^2+\sigma^2)^2}{\alpha}D_\alpha\big(P_{\boldsymbol{W}_1},P_{\boldsymbol{W}_2}\big).$$

Thus:

$$\mathbb{E}\bigg[\int\big\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T)-\boldsymbol{W}_0\boldsymbol{W}_0^T\big\|_F^2\tilde{\pi}^{\hat{K}}_{n,\alpha}(dW|X_1^n)\bigg]$$
$$\leq\frac{16(B^2+\sigma^2)^2}{\alpha}\mathbb{E}\bigg[\int D_\alpha(P_W,P_{\boldsymbol{W}_0})\tilde{\pi}^{\hat{K}}_{n,\alpha}(dW|X_1^n)\bigg]$$

and we use Theorem 5:

$$\mathbb{E}\bigg[\int D_\alpha(P_{\boldsymbol{W}},P_{\boldsymbol{W}_0})\tilde{\pi}^{\hat{K}}_{n,\alpha}(dW|X_1^n)\bigg]=\mathcal{O}\bigg(\frac{dK_0\log(dn)}{n}\bigg).$$

which ends the proof.

■

We can obtain Corollary 7 using a simple convexity argument.