1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–11

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 data set. To do so, one can perform cross validation or maximize a numerical criterion to make the final choice. In the literature, penalized criteria such as AIC and BIC 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). Thus, it is necessary to define a criterion suited to a given objective. Meanwhile, a non-asymptotic theory of penalization has been developed during the last two decades using oracle inequalities, and offers a simple way to assess the quality of a given model selection criterion.

© B.-E. Chérief-Abdellatif.

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 and show that it still holds in the general case of independent and identically distributed (i.i.d.) data. 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, and we briefly explain how it can lead to consistency rates for probabilistic principal component analysis.

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 $\big(\mathbb{X},\mathcal{X}\big)$. We denote $X_1^n=(X_1,...,X_n)$. We consider a countable collection $\{\mathcal{M}_K/K\in\mathbb{N}^*\}$ 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.

We use a Bayesian approach, and we define a prior π over the full parameter space $\cup_{K\in\mathbb{N}^*}\Theta_K$ (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 $\mathcal{M}_K$:

$$\pi=\sum_{K\in\mathbb{N}^*}\pi_K\Pi_K.$$

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

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

We also remind that the α-Renyi divergence between P and R 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 $\pi_{n,\alpha}^K(.|X_1^n)$ on parameter $\theta_K\in\Theta_K$ given model $\mathcal{M}_K$ using prior Π_K and likelihood L_n:

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

This definition is a slight variant of the regular Bayesian posterior (for which $\alpha=1$) that is more robust to model misspecification, see Grünwald and Van Ommen (2017).

The Variational Bayes approximation $\tilde{\pi}_{n,\alpha}^K(.|X_1^n)$ of the tempered posterior associated with model K is then defined as the distribution into some set $\mathcal{F}_K$ that maximizes the Evidence Lower Bound:

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

where the function inside the argmax operator is the ELBO $\mathcal{L}_K(\rho_K)$. In the following, we will just call ELBO $\mathcal{L}(K)$ the closest approximation to the log-evidence, i.e. the value of the ELBO evaluated at its maximum:

$$\mathcal{L}(K) = \alpha \int \ell_n(\theta_K) \tilde{\pi}_{n,\alpha}^K(d\theta_K | X_1^n) - \mathcal{K}(\tilde{\pi}_{n,\alpha}^K(.|X_1^n), \Pi_K).$$

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

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

which is a slight variant of the classical definition. Nevertheless, note that taking a uniform distribution over the prior weights π_K's on a finite number of models leads to maximizing the ELBO as recommended in Blei et al. (2017). We will discuss later the case $\pi_K = 2^{-K}$ for a countable collection of models. Note that the penalty term is not just an artefact in order to ease the theoretical proof, but it is a complexity term that reflect our prior beliefs over the models.

We will provide in the next section a theoretical justification to such a selection criterion and derive an oracle inequality for $\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n)$ in the spirit of the one in Chérief-Abdellatif and Alquier (2018).

3. Main result

Theorem 1 *For any $\alpha \in (0,1)$,*

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

Moreover, as soon as there exists $K_0 \in \mathbb{N}^*$ and $\theta^0 \in \Theta_{K_0}$ such that $P^0 = P_{\theta^0}$, for which there exists r_n such that there is a distribution $\rho_{K_0,n} \in \mathcal{F}_{K_0}$ such that:

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

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

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

A few words on the assumptions of Theorem 1. First, the existence of K_0 and $\theta^0 \in \Theta_{K_0}$ such that $P^0 = P_{\theta_0}$ is equivalent to the existence of a true model $\mathcal{M}_{K_0}$ to which the true distribution P^0 belongs, while Condition 3.1 is standard to study concentration of the posterior in the Bayesian community, see Alquier and Ridgway (2017); Chérief-Abdellatif and Alquier (2018).

The oracle inequality in Theorem 1 shows that the selected variational approximation is adaptive. This means that when there exists a true model $\mathcal{M}_{K_0}$, i.e. that there exists K_0 and $\theta^0 \in \Theta_{K_0}$ such that $P^0 = P_{\theta_0}$, even if $\hat{K}$ is not equal to K_0, then the convergence rate of P_θ to P^0 under $\tilde{\pi}^{\hat{K}}_{n,\alpha}(.|X_1^n)$ is as good as if we knew P^0 (as soon as the penalty is lower than r_n, which is the case for prior weights used in practice). Note that the estimation of such K_0 is a different task that would require identifiability assumptions that are stronger than those in our theorem, and that our result is robust to misspecification. We can also see that the convergence rate is composed of the rate obtained when approximating the true distribution with distributions in model $\mathcal{M}_K$, and of a complexity term over the model index set that reflects the prior belief over the different models. 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/n. More generally, when $\frac{1}{n} \lesssim r_n$, we obtain the consistency at the estimation 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 consistency rate equal to $\frac{K_0 \log(nK_0)}{n}$ for univariate Gaussian mixtures when there exists a true K_0-components mixture model. We study another example in the next section.

4. Application to probabilistic PCA

We consider here the probabilistic Principal Component Analysis (PCA) problem as an application of our work. We assume the model $X_i = WZ_i + \sigma^2 I_d$ with i.i.d. random variables $Z_i \sim \mathcal{N}(0, I_K)$, where I_d and I_K are respectively the d- and K-dimensional identity matrices $(K < d)$, $W \in \mathbb{R}^{d\times K}$ 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 $P_W := \mathcal{N}(0, WW^T + \sigma^2 I_d)$. We are interested in estimating the principal axes W and selecting the number of components K. The optimal number of components will be chosen using the ELBO criterion and the principal axes will be estimated using the corresponding variational approximation.

Each model will correspond 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 W to infer a distribution over principal axes. We choose independent Gaussian priors $\mathcal{N}(0, s^2 I_d)$ on the columns $W_1, ..., W_K$ of W. We also consider Gaussian independent variational approximations $\mathcal{N}(\mu_j, \Sigma_j)$ for the columns of W. Then, as soon as there exists a true model, i.e. there exists K_0 and $W_0 \in \mathbb{R}^{d\times K_0}$ such that the true distribution of each X_i is $P_{W_0} = \mathcal{N}(0, W_0 W_0^T + \sigma^2 I_d)$, under the assumption that the coefficients of W_0 are bounded, then Theorem 1 provides an explicit rate of convergence of our variational estimator even when K_0 is unknown:

$$\mathbb{E}\left[\int D_\alpha(P_W, P_{W_0}) \tilde{\pi}^{\hat{K}}_{n,\alpha}(dW|X_1^n)\right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n}\right).$$

The proof as well as the computation of the ELBO are detailed in the appendix.

Acknowledgments

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

References

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.

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

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.

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.

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.

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

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

Appendix A. Proof of Theorem 1.

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

Lemma 2 *For any probability λ on some measurable space $(\boldsymbol{E},\mathcal{E})$ and any measurable function $h:\boldsymbol{E}\to\mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\lambda<\infty$,*

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

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 1. We adapt the proof of Theorem 4.1 in Chérief-Abdellatif and Alquier (2018).

Proof For any $\alpha \in (0,1)$ and $\theta \in \Omega := \cup_{K \in \mathbb{N}^*} \Theta_K$, using the definition of Renyi divergence and $D_\alpha(P^{\otimes n}, R^{\otimes n}) = nD_\alpha(P, R)$ 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 2,

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

Then, using Jensen's inequality,

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

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 $\Theta_{\hat{K}}$,

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

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

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

By definition of $\hat{K}$,

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

which gives

$$\mathbb{E}\bigg[\int D_\alpha(P_\theta,P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n)\bigg]$$
$$\leq \inf_{K\in\mathbb{N}^*}\left\{\mathbb{E}\bigg[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\tilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n)+\frac{\mathcal{K}(\tilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\bigg]\right\}$$

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

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

which leads to,

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

Finally,

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

The special case in the theorem is a direct corollary of the more general form.

■

Appendix B. Proof of the consistency rate of VB for probabilistic PCA.

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 $W_0\in\mathbb{R}^{d\times K_0}$ with bounded coefficients such that the true distribution of each X_i is $\mathcal{N}(0,W_0W_0^T+\sigma^2I_d)$, and we place a prior $\Pi_{K_0}=\mathcal{N}(0,s^2I_d)^{\otimes K_0}$ and a variational approximation $\rho_{K_0}=\rho^{\otimes K_0}$ on W given $K=K_0$ where we denote $\rho=\mathcal{N}(0,\frac{1}{dn^2}I_d)$. We recall that $\pi_K=\frac{1}{d}$ for any $K=1,...,d$.

To obtain a rate of convergence of the VB for probabilistic PCA, we upper bound the right-hand side of the main inequality of Theorem 1 by

$$\frac{\alpha}{1-\alpha}\int\mathcal{K}\Big(\mathcal{N}(0,W_0W_0^T+\sigma^2I_d),\mathcal{N}(0,WW^T+\sigma^2I_d)\Big)\rho_{K_0}(d\theta_K)+\frac{\mathcal{K}(\rho_{K_0},\Pi_{K_0})}{n(1-\alpha)}+\frac{\log(d)}{n(1-\alpha)}.$$

We have three terms. The last one, corresponding to the prior beliefs over the different models, gives a rate of convergence of $\log(d)/n$. The second one, i.e. the Kullback-Leibler term, provides a rate of convergence of $dK_0 \log(dn)/n$ as:

$$\begin{aligned}\mathcal{K}(\rho_{K_0}, \Pi_{K_0}) &= \sum_{j=1}^{K_0} \mathcal{K}\Big(\mathcal{N}(0, \frac{1}{dn^2} I_d), \mathcal{N}(0, s^2 I_d)\Big) \\ &= \frac{K_0}{2}\Big(\frac{1}{n^2 s^2} - d + d\log(s^2) + d\log(dn^2)\Big) \\ &\leq \frac{K_0}{2n^2 s^2} - \frac{dK_0}{2} + \frac{dK_0 \log(s^2)}{2} + dK_0 \log(dn).\end{aligned}$$

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:

$$\frac{1}{2}\mathbb{E}\Big[\mathrm{Tr}\Big((WW^T + \sigma^2 I_d)^{-1}(W_0 W_0^T + \sigma^2 I_d)\Big)\Big] - \frac{d}{2} + \frac{1}{2}\mathbb{E}\Big[\log\Big(\frac{\det(WW^T + \sigma^2 I_d)}{\det(W_0 W_0^T + \sigma^2 I_d)}\Big)\Big].$$

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 $W_0 W_0^T + \sigma^2 I_d$. Then for each $j = 1, ..., d$, $\lambda_j \geq \sigma^2$ and using Jensen's inequality and the log-concavity of the determinant:

$$\begin{aligned}\mathbb{E}\Big[\log\Big(\det(WW^T + \sigma^2 I_d)\Big)\Big] &\leq \log\Big(\det\big(\mathbb{E}[WW^T] + \sigma^2 I_d\big)\Big) \\ &= \log\Big(\det\big(W_0 W_0^T + \sigma^2 I_d + \frac{1}{dn^2} I_d\big)\Big) \\ &= \sum_{j=1}^{d} \log\Big(\lambda_j + \frac{1}{dn^2}\Big) \\ &= \sum_{j=1}^{d} \log(\lambda_j) + \sum_{j=1}^{d} \log\Big(1 + \frac{1}{\lambda_j dn^2}\Big) \\ &= \mathbb{E}\Big[\log\Big(\det(W_0 W_0^T + \sigma^2 I_d)\Big)\Big] + \sum_{j=1}^{d} \log\Big(1 + \frac{1}{\lambda_j dn^2}\Big) \\ &\leq \mathbb{E}\Big[\log\Big(\det(W_0 W_0^T + \sigma^2 I_d)\Big)\Big] + \sum_{j=1}^{d} \frac{1}{\lambda_j dn^2} \\ &\leq \mathbb{E}\Big[\log\Big(\det(W_0 W_0^T + \sigma^2 I_d)\Big)\Big] + \frac{1}{n^2\sigma^2}\end{aligned}$$

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

$$\mathbb{E}\Big[\log\Big(\frac{\det(WW^T + \sigma^2 I_d)}{\det(W_0 W_0^T + \sigma^2 I_d)}\Big)\Big] \leq \frac{1}{n^2\sigma^2}.$$

The remainder can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\mathrm{Tr}\Big((WW^T+\sigma^2 I_d)^{-1}(W_0W_0^T+\sigma^2 I_d)\Big)\Big]-d \\
&\qquad\qquad = \mathbb{E}\Big[\mathrm{Tr}\Big((WW^T+\sigma^2 I_d)^{-1}(W_0W_0^T-WW^T)\Big)\Big] \\
&\qquad\qquad \leq \mathbb{E}\Big[\|(WW^T+\sigma^2 I_d)^{-1}\|_F \times \|W_0W_0^T-WW^T\|_F\Big] \\
&\qquad\qquad \leq \sqrt{d}\,\mathbb{E}\Big[\|(WW^T+\sigma^2 I_d)^{-1}\|_2 \times \|W_0W_0^T-WW^T\|_F\Big] \\
&\qquad\qquad = \sqrt{d}\,\mathbb{E}\Big[\sigma_{\max}\big((W_0W_0^T+\sigma^2 I_d)^{-1}\big) \times \|W_0W_0^T-WW^T\|_F\Big] \\
&\qquad\qquad = \sqrt{d}\,\mathbb{E}\Big[\sigma_{\min}(W_0W_0^T+\sigma^2 I_d)^{-1} \times \|W_0W_0^T-WW^T\|_F\Big] \\
&\qquad\qquad \leq \sqrt{d}\,\mathbb{E}\Big[(\sigma^2)^{-1} \times \|W_0W_0^T-WW^T\|_F\Big] \\
&\qquad\qquad = \frac{\sqrt{d}}{\sigma^2}\mathbb{E}\Big[\|W_0W_0^T-WW^T\|_F\Big]
\end{aligned}$$

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

The only thing left to do is to upper bound the expectation of the Frobenius norm of $W_0W_0^T - WW^T$ by a multiple of $\frac{\sqrt{d}K_0\log(dn)}{n}$. We use the triangle and Cauchy-Schwarz's inequalities:

$$\begin{aligned}
\mathbb{E}\Big[\|W_0W_0^T-WW^T\|_F\Big] &\leq \mathbb{E}\Big[\|WW^T-WW_0^T\|_F\Big]+\mathbb{E}\Big[\|WW_0^T-W_0W_0^T\|_F\Big] \\
&\leq \mathbb{E}\Big[\|W(W-W_0)^T\|_F\Big]+\mathbb{E}\Big[\|(W-W_0)W_0^T\|_F\Big] \\
&\leq \mathbb{E}\Big[\|W\|_F\|W-W_0\|_F\Big]+\mathbb{E}\Big[\|W-W_0\|_F\|W_0\|_F\Big] \\
&\leq \sqrt{\mathbb{E}\big[\|W\|_F^2\big]\mathbb{E}\big[\|W-W_0\|_F^2\big]}+\sqrt{\mathbb{E}\big[\|W-W_0\|_F^2\big]\mathbb{E}\big[\|W_0\|_F^2\big]} \\
&\leq \sqrt{\mathbb{E}\big[\|W\|_F^2\big]\mathbb{E}\big[\|W-W_0\|_F^2\big]}+\|W_0\|_F\sqrt{\mathbb{E}\big[\|W-W_0\|_F^2\big]}.
\end{aligned}$$

We can upper bound $\|W_0\|_F = \sqrt{\sum_{i=1}^d\sum_{j=1}^{K_0}(W_0)_{i,j}^2}$ by $\sqrt{dK_0}C$ where C is an upper bound on each of the coefficients of matrix W_0.

Also, we can notice that $dn^2\|W-W_0\|_F^2 = \sum_{i=1}^d\sum_{j=1}^{K_0}\big(\sqrt{d}n(W_{i,j}-(W_0)_{i,j})\big)^2$ is a sum of squares of independent standard normal random variables. Thus $dn^2\|W-W_0\|_F^2$ follows a chi-squared distribution with dK_0 degrees of freedom and its expectation is equal to dK_0. Hence:

$$\mathbb{E}\big[\|W-W_0\|_F^2\big] = \frac{K_0}{n^2}.$$

Similarly, as $W_{i,j} - (W_0)_{i,j}$ is centered, we get:

$$\begin{aligned}\mathbb{E}\big[\|W\|_F^2\big] &= \mathbb{E}\Big[\sum_{i=1}^{d}\sum_{j=1}^{K_0} W_{i,j}^2\Big] \\ &= \sum_{i=1}^{d}\sum_{j=1}^{K_0}\mathbb{E}\Big[\big(W_{i,j}-(W_0)_{i,j}\big)^2 + (W_0)_{i,j}^2 - 2(W_0)_{i,j}\big(W_{i,j}-(W_0)_{i,j}\big)\Big] \\ &= \mathbb{E}\big[\|W-W_0\|_F^2\big] + \|W_0\|_F^2 \\ &\leq \frac{K_0}{n^2} + dK_0C^2 \\ &= \left(dC^2 + \frac{1}{n^2}\right)K_0.\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}\mathbb{E}\Big[\|W_0W_0^T - WW^T\|_F\Big] &\leq \frac{\sqrt{K_0}}{n}\sqrt{K_0}\sqrt{dC^2+\frac{1}{n^2}} + \sqrt{dK_0}C\frac{\sqrt{K_0}}{n} \\ &= \frac{K_0}{n}\sqrt{dC^2+\frac{1}{n^2}} + \frac{\sqrt{d}K_0C}{n} \\ &\leq \frac{K_0}{n}\left(\sqrt{d}C + \frac{1}{n}\right) + \frac{\sqrt{d}K_0C}{n} \\ &= \frac{K_0}{n}\left(2\sqrt{d}C + \frac{1}{n}\right).\end{aligned}$$

Hence, the order of the upper bound of the expectation of the Fobrenius norm of matrix $W_0W_0^T - WW^T$ 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 C. 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 W such that $\Pi_K = \mathcal{N}(0, s^2 I_d)^{\otimes K}$, and Gaussian independent variational approximations $\mathcal{N}(\mu_j, \Sigma_j)$ for the columns of W. The ELBO associated with rank K and variational approximation $\rho_K = \otimes_{j=1}^{K}\mathcal{N}(\mu_j, \Sigma_j)$ is given by:

$$\mathcal{L}_K(\rho_K) = \alpha\int \ell_n(W)\rho_K(dW) - \mathcal{K}\big(\rho_K, \Pi_K\big).$$

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

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

while the average log-likelihood $\int \ell_n(W)\rho_K(dW)$ is:

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

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

$$\int \log\big(\det(WW^T+\sigma^2 I_d)\big)\rho_K(dW) \approx \sum_{\ell=1}^{N}\log\big(\det(W^{(\ell)}W^{(\ell)T}+\sigma^2 I_d)\big)$$

and

$$\int X_i^T(WW^T+\sigma^2 I_d)^{-1}X_i\,\rho_K(dW) \approx \sum_{\ell=1}^{N} X_i^T(W^{(\ell)}W^{(\ell)T}+\sigma^2 I_d)^{-1}X_i$$

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

The inverse matrix $(WW^T+\sigma^2 I_d)^{-1}$ 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 $A\in\mathbb{R}^{d\times d}$ and vectors $u,v\in\mathbb{R}^d$ such that $A+uv^T$ is invertible,

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

We write

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

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

$$\begin{cases} M_0=\sigma^2 I \\ \forall j=1,...,K,\ \ M_j=M_{j-1}-\frac{Z_jZ_j^T}{1+W_j^TZ_j} \text{ with } Z_j=M_{j-1}W_j. \end{cases}$$

In order to compute the maximum value $\mathcal{L}(K)$ of the ELBO associated with rank K, one can use a stochastic gradient descent on $(\mu_1,\Sigma_1,...,\mu_K,\Sigma_K)$ that will converge to a local maximum and will give the variational estimator for rank K. Then, maximizing $\mathcal{L}(K)$ over desired values of K leads to the optimal number of principal components and to the associated optimal variational approximation.