Consistency of variational Bayes inference for estimation and model selection in mixtures

Badr-Eddine Chérief-Abdellatif[*] and Pierre Alquier[*,†]

CREST, ENSAE
5 avenue Henry le Châtelier
91120 Palaiseau – FRANCE
e-mail: badr.eddine.cherief.abdellatif@ensae.fr; alquier@ensae.fr

Abstract: Mixture models are widely used in Bayesian statistics and machine learning, in particular in computational biology, natural language processing and many other fields. Variational inference, a technique for approximating intractable posteriors thanks to optimization algorithms, is extremely popular in practice when dealing with complex models such as mixtures. The contribution of this paper is two-fold. First, we study the concentration of variational approximations of posteriors, which is still an open problem for general mixtures, and we derive consistency and rates of convergence. We also tackle the problem of model selection for the number of components: we study the approach already used in practice, which consists in maximizing a numerical criterion (the Evidence Lower Bound). We prove that this strategy indeed leads to strong oracle inequalities. We illustrate our theoretical results by applications to Gaussian and multinomial mixtures.

MSC 2010 subject classifications: Primary 62F12; secondary 65C60, 62F15, 62F35.
Keywords and phrases: Mixture models, frequentist evaluation of Bayesian methods, variational approximations, model selection.

Received May 2018.

Contents

1. Introduction

This paper studies the statistical properties of variational inference as a tool to tackle two problems of interest: estimation and model selection in mixture models. Mixtures are often used for modelling population heterogeneity, leading to practical clustering methods [12, 32]. Moreover they have enough flexibility to approximate accurately almost every density [6, 29]. Mixtures are used in many various areas such as computer vision [5], genetics [36], economics [19], transport data analysis [14] and others. We refer the reader to [18] for an account of the recent advances on mixtures. The most famous procedure for mixture density estimation in the frequentist literature is probably Expectation-Maximization [20], a maximum-likelihood algorithm that yields increasingly higher likelihood. At the same time, the Bayesian paradigm has raised great interest among researchers and practitioners, especially through the Variational Bayes (VB) framework which aims at maximizing a quantity referred to as Evidence Lower Bound on the marginal likelihood (ELBO). Variational Bayes inference is a useful tool for approximating intractable posteriors. It is known to work well in practice for mixture models: one of the most recent survey on VB [11] chooses mixtures as an example of choice to illustrate the power of the method. Moreover [11] states: "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". The main contribution of this paper is to prove that VB is consistent for estimation in mixture models, and that the ELBO maximization strategy used in practice is consistent for model selection. Thus we solve the question raised by [11].

Variational Bayes is a method for computing intractable posteriors in Bayesian statistics and machine learning. Markov Chain Monte Carlo (MCMC) algorithms remain the most widely used methods in computational Bayesian statistics. Nevertheless, they are often too slow for practical uses when the dataset is very large. A more and more popular alternative consists in finding a deterministic approximation of the target distribution called Variational Bayes approximation. The idea is to minimize the Kullback-Leibler divergence of a tractable distribution ρ with respect to the posterior, which is also equivalent to maximizing the ELBO. This optimization procedure is much faster and efficient than MCMC sampling with numerous applications in different fields: matrix completion for collaborative filtering [2], computer vision [41], computational biology [13] and natural language processing [28], to name a few prominent examples.

However, variational inference is mainly used for its practical efficiency and only little attention has been put in the literature towards theoretical properties of the VB approximation until very recently. In [4] the properties of variational approximations of Gibbs distributions used in machine learning are derived. The results are essentially valid for bounded loss functions, which makes them difficult to use beyond the problem of supervised classification. Based on some technical advances from [8], [3] removed the boundedness assumption in [4], allowing to study more general statistical models. In [48], the authors extended the range of models covered by [4]. This allowed them to study mixture of Gaussian distributions as an example. Many questions are still left unanswered: model selection, and the estimation of mixture of non-Gaussian distributions. For example mixture of multinomials are widely used in practice [14], as well as more intricated examples such as nonparametric mixtures [22]. Note that all the results in [8, 3, 48] are limited to so-called tempered posteriors, that is, where the likelihood is taken to some power α. Still, the use of tempered posteriors is highly recommended by many authors as a way to overcome model misspecification, see [25] and the references therein. Indeed some results in [3] are valid in a misspecified setting. Alternative approaches were developed to study VB: [44] established Bernstein-von-Mises type theorems on the variational approximation of the posterior. They provide very interesting results for parametric models but it is unclear whether these results can be extended to model selection or misspecified case. More recently, [49] succeeded in adapting the now classical results of [23] to Variational Bayes and showed that a slight modification in the three classical "prior mass and testing conditions" leads to the convergence of their variational approximations, again under the assumption that the model is true. With respect to these works, our contribution is a complete study of the consistency of VB for mixtures of general distributions. In particular, we explicit independent conditions on the prior on the weights, and on the prior on the parameters of the components. The study is done in the case $\alpha < 1$ which allows to prove results in the misspecified case.

The other point addressed in this paper is model selection. This is a natural question which can be interpreted in this context as the determination of the number of components of the mixture. This point is crucial: indeed, too many components can lead to estimates with too large variances whereas with too few components, we may obtain mixtures which are not able to fit the data properly. This is a common issue and a lot of statisticians worked on this question. In the literature, criteria such as AIC [1] and BIC [38] are popular. It is well known that in some collections of models, AIC optimizes the prediction ability while BIC recovers with high probability the true model (when there is one). These two objectives are not compatible in general [47]. Anyway, these results depend on asssumptions that are not satisfied by mixtures. It seems thus more natural to develop criteria suited to a given objective. For example, [9] proposed a procedure to select a number of components that is the most relevant for clustering. A non-asymptotic theory of penalization has been developed during the last two decades using oracle inequalities [31]. In the wake of those works, our paper studies mixture model selection based on the ELBO criterion. We prove a general oracle inequality. This result establishes the consistency of ELBO maximization when the primary objective is the estimation of the distribution of the data.

The rest of this paper is organized as follows. In Section 2 we introduce the background and the notations that will be adopted. Consistency of the Variational Bayes for estimation in a mixture model is studied in Section 3. First, we give the general results under a "prior mass" assumption, as well as a general form for the algorithm to compute the VB approximation (Subsection 3.1). We then apply these results to mixtures of multinomials (Subsection 3.2) and Gaussian mixtures (Subsection 3.3). In each case, we provide a rate of convergence of VB and discuss its numerical implementation. We extend the setting to the misspecified case in Subsection 3.4. Finally, we address the issue of selecting based on the ELBO in Section 4. We discuss possible extensions in Section 5, while Section 6 is dedicated to the proofs.

2. Background and notations

Let us introduce the notations and the framework we adopt in this paper. We observe in a measurable space $(\mathbb{X}, \mathcal{X})$ a collection of n i.i.d. random variables $X_1,...,X_n$ sampled from a probability distribution P^0. We put $(X_1, ..., X_n) = X_1^n$. The goal is to estimate the generating distribution P^0 of the X_i's by a K-components mixture model. We will study the (frequentist) properties of variational approximations of the posterior. The extension to selection of the number of components is also tackled in this paper, but we will first deal with a fixed K. We introduce a collection of distributions $\{Q_\theta / \theta \in \Theta\}$ indexed by a parameter space Θ from which we will take the different components of our mixture model. We assume that for each $\theta \in \Theta$, the probability distribution Q_θ is dominated by a reference measure μ and that the density $q_\theta = \frac{dQ_\theta}{d\mu}$ is such that the map $(x, \theta) \to q_\theta(x)$ is $\mathcal{X} \times \mathcal{T}$-measurable, $\mathcal{T}$ being some sigma-

algebra on Θ. Unless explicitly stated otherwise, all the distributions that will be considered in this paper will be characterized by their density with respect to the dominating measure μ. We can now consider the statistical mixture model of $K \geq 1$ components defined as:

$$\left\{ \sum_{j=1}^{K} p_j Q_{\theta_j} \ \middle/ \ \theta_j \in \Theta \ \text{for} \ j = 1, ..., K, \ \ p = (p_1, ..., p_K) \in \mathcal{S}_K \right\}$$

where $\mathcal{S}_K = \left\{ p = (p_1, ..., p_K) \in \mathbb{R}^K \middle/ \ p_j \geq 0 \ \text{for} \ j = 1, ..., K \ \text{and} \ \sum_{j=1}^{K} p_j = 1 \right\}$ is the $K-1$ dimensional simplex. We will write $\theta = (p_1, ..., p_K, \theta_1, ... \theta_K) \in \Theta_K$ for short, where $p \in \mathcal{S}_K$, $\theta_j \in \Theta$ for $j = 1, ..., K$ and $\Theta_K = \mathcal{S}_K \times \Theta^K$. The mixture corresponding to parameter $\theta = (p_1, ..., p_K, \theta_1, ..., \theta_K)$ will be denoted $P_\theta := \sum_{j=1}^{K} p_j Q_{\theta_j}$.

First, we consider the well-specified case, assuming that the true distribution belongs to the K-components mixture model. Thus, we define the true distribution P^0 from which data are sampled:

$$X_1, ..., X_n \sim \sum_{j=1}^{K} p_j^0 Q_{\theta_j^0} \ \text{with} \ \theta_j^0 \in \Theta \ \text{for} \ j = 1, ..., K \ \text{and} \ p^0 \in \mathcal{S}_K.$$

Hence, we want to estimate the true distribution P_{θ^0} using a Bayesian approach. Therefore, we define a prior $\pi = \pi_p \bigotimes_{j=1}^{K} \pi_j$ on θ, $\pi_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ being a probability distribution on some measurable space $(\mathcal{S}_K, \mathcal{A})$, and each $\pi_j \in \mathcal{M}_1^+(\Theta)$ a probability distribution on the measurable space $(\Theta, \mathcal{T})$. We will also consider in this paper the misspecified case where the true distribution does not belong to our statistical model i.e. is not necessarily a mixture, but the specific notations and framework will be described later.

Let us introduce some notations. The likelihood will be denoted by L_n and the log-likelihood by ℓ_n, that is, for any $\theta = (p_1, ..., p_K, \theta_1, ... \theta_K)$,

$$L_n(\theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} p_j q_{\theta_j} \ , \ \ \ell_n(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} p_j q_{\theta_j} \right).$$

The negative log-likelihood ratio r_n between two distributions P and R is given by

$$r_n(P, R) = \sum_{i=1}^{n} \log \left(\frac{R(X_i)}{P(X_i)} \right)$$

(note that $r_n(\theta, \theta')$ is used by many authors instead of $r_n(P_\theta, P_{\theta'})$ but our notation is more convenient for the extension to the misspecified case). The Kullback-Leibler (KL) divergence between two probability distributions P and R is given by

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

If some measure λ dominates both P and R distributions represented here by their densities f and g with respect to this measure, we have

$$\mathcal{K}(P, R) = \int f \log\left(\frac{f}{g}\right) d\lambda$$

and we will use $K(P, R)$ or $\mathcal{K}(f, g)$ to denote this quantity, depending on the context.

We also remind that the α-Renyi divergence between P and R,

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

When for some λ we have $f = \frac{dP}{d\lambda}$ and $g = \frac{dR}{d\lambda}$,

$$D_\alpha(P, R) = D_\alpha(f, g) = \frac{1}{\alpha - 1} \log \int f^\alpha g^{1-\alpha} d\lambda.$$

Some useful properties of Renyi divergences can be found in [43]. In particular, the Renyi divergence between two probability distributions P and R can be related to the classical total variation TV and Hellinger H distances respectively defined as $TV(P, R) = \frac{1}{2} \int |dP - dR|$ and $H(P, R)^2 = \frac{1}{2} \int (\sqrt{dP} - \sqrt{dR})^2 = 1 - e^{-\frac{1}{2}D_{1/2}(P,R)}$ through:

$$TV(P, R)^2 \le 2H(P, R)^2 \le D_{1/2}(P, R) \text{ and } D_\alpha(P, R) \xrightarrow[\alpha \to 1]{\nearrow} \mathcal{K}(P, R).$$

The tempered Bayesian posterior $\pi_{n,\alpha}(.|X_1^n)$, which is our target here, is defined for $0 < \alpha \le 1$ by

$$\pi_{n,\alpha}(d\theta|X_1^n) = \frac{e^{-\alpha r_n(P_\theta, P^0)} \pi(d\theta)}{\int e^{-\alpha r_n(P_\phi, P^0)} \pi(d\phi)} \propto L_n(\theta)^\alpha \pi(d\theta)$$

(it is also referred to as fractional posterior, for example in [8]). Note that when $\alpha = 1$, then we recover the "true" Bayesian posterior, but the case $\alpha < 1$ has many advantages: it is often more tractable from a computational perspective [34, 7], it is consistent under less stringent assumptions than required for $\alpha = 1$ [8] and it is more robust to misspecification [25].

We are now in position to define the VB approximation $\tilde{\pi}_{n,\alpha}(.|X_1^n)$ of the tempered posterior with respect to some set of distributions $\mathcal{F}$: it is the projection, with respect to the Kullback-Leibler divergence, of the tempered posterior onto the mean-field variational set $\mathcal{F}$,

$$\tilde{\pi}_{n,\alpha}(.|X_1^n) = \underset{\rho \in \mathcal{F}}{\arg\min} \, \mathcal{K}\left(\rho, \pi_{n,\alpha}(.|X_1^n)\right).$$

The mean-field approximation is very popular in the Variational Bayes literature. It is based on a decomposition of the space of parameters Θ_K as a product.

Then $\mathcal{F}$ consists in compatible product distributions. Here, a natural choice [11] is $\Theta_K = \mathcal{S}_K \times \Theta \times \cdots \times \Theta$ and

$$\mathcal{F} = \left\{ \rho_p \bigotimes_{j=1}^{K} \rho_j / \rho_p \in \mathcal{M}_1^+(\mathcal{S}_K), \; \rho_j \in \mathcal{M}_1^+(\Theta) \; \forall j = 1, ..., K \right\}.$$

We will work on this particular set in the following and we will often use ρ instead of $\rho_p \bigotimes_{j=1}^{K} \rho_j$ to ease notation.

We end this section by recalling Donsker and Varadhan's variational formula. Refer for example to [16] for a proof (Lemma 1.1.3).

Lemma 2.1. *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:

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

This technical lemma is one of the main ingredients for the proof of our results, but it is also very helpful to understand variational approximations. Indeed, for $\mathbf{E} = \Theta_K$ and using the definition of $\pi_{n,\alpha}(.|X_1^n)$, we get:

$$\pi_{n,\alpha}(\cdot|X_1^n) = \operatorname*{arg\,min}_{\rho \in \mathcal{M}_+^1(\Theta_K)} \left\{ \alpha \int r_n(P_\theta, P^0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}$$

and simple calculations give

$$\tilde{\pi}_{n,\alpha}(\cdot|X_1^n) = \operatorname*{arg\,min}_{\rho \in \mathcal{F}} \left\{ \alpha \int r_n(P_\theta, P^0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}$$

$$= \operatorname*{arg\,max}_{\rho \in \mathcal{F}} \left\{ \alpha \int \ell_n(\theta) \rho(d\theta) - \mathcal{K}(\rho, \pi) \right\} \tag{2.1}$$

$$= \operatorname*{arg\,min}_{\rho \in \mathcal{F}} \left\{ -\alpha \sum_{i=1}^{n} \int \log \left(\sum_{j=1}^{K} p_j q_{\theta_j}(X_i) \right) \rho(d\theta) + \mathcal{K}(\rho_p, \pi_p) \right.$$

$$\left. + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j) \right\}. \tag{2.2}$$

The quantity maximized in (2.1) is called the ELBO in the litterature (ELBO stands for Evidence Lower Bound), and many authors actually take this as the definition of VB [11].

Remark 2.1. *In practice, the choice of α is not staightforward. Depending on the objective, some heuristic might be available: for example, [42] proposed a nice method to calibrate α in order to get confidence intervals on a parameter of interest. More generally, cross-validation can give good results. We have to acknowledge that there is no universal method to calibrate α. This could lead the reader to the idea that the proper Bayesian approach ($\alpha = 1$) is simpler to use. We insist on the fact $\alpha = 1$ can produce catastrophic results in case of misspecification [25], while we present below some results in the misspecified case with $\alpha < 1$. We believe that the calibration of α is a very important research direction.*

3. Variational Bayes estimation of a mixture

3.1. A PAC-Bayesian inequality

We start with a result for general mixtures. Later in this section we provide corollaries obtained by applying this theorem to special cases: mixture of multinomials and Gaussian mixtures.

Theorem 3.1. *For any $\alpha \in (0, 1)$,*

$$
\mathbb{E}\bigg[\int D_\alpha \bigg(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \bigg) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \bigg]
$$

$$
\leq \inf_{\rho_p \otimes_{j=1}^{K} \rho_j \in \mathcal{F}} \bigg\{ \frac{\alpha}{1-\alpha} \bigg[\int \mathcal{K}(p^0, p)\rho_p(dp) + \sum_{j=1}^{K} \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_j(d\theta_j) \bigg]
$$

$$
+ \frac{\mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j)}{n(1-\alpha)} \bigg\}.
$$

As a special case, when there exists $r_{n,K}$ such that there is are distributions $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ $(j = 1, ..., K)$ such that for $j = 1, ..., K$

$$
\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K r_{n,K}, \quad \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_{j,n}(d\theta_j) \leq r_{n,K} \tag{3.1}
$$

and

$$
\mathcal{K}(\rho_{p,n}, \pi_p) \leq K n r_{n,K}, \quad \mathcal{K}(\rho_{j,n}, \pi_j) \leq n r_{n,K}, \tag{3.2}
$$

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

$$
\mathbb{E}\bigg[\int D_\alpha \bigg(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \bigg) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \bigg] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.
$$

The proof is given in Section 6. This theorem provides the consistency of the Variational Bayes for mixture models as soon as (3.1) and (3.2) are satisfied. In [3], the authors use similar conditions ((3) and (4) in their Theorem 2.6), and

show that they are strongly linked to the assumptions on the prior used by [23, 8] to derive concentration of the posterior. Thus they cannot be removed in general. Theorem 3.1 states that finding $r_{n,K}$ fulfilling (3.1) and (3.2) independently for the weights and for each component is sufficient to obtain the rate of convergence $Kr_{n,K}$ of the VB estimator towards the true distribution.

Clearly, the theorem cannot be directly extended to the case $\alpha = 1$. As discussed above, the case $\alpha = 1$ is studied in [49]: it requires a testing condition in addition to the prior mass condition given by (3.1) and (3.2). In the case of mixtures, such a testing condition was studied in [29].

Note that there always exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that the two quantities corresponding to the weights $\int \mathcal{K}(p^0, p)\rho_{p,n}(dp)$ and $\mathcal{K}(\rho_{p,n}, \pi_p)$ are bounded as required in Theorem 3.1 for $r_{n,K} = \frac{4\log(nK)}{n}$ when the chosen prior is a Dirichlet distribution $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ under some minor restriction on $\alpha_1, ..., \alpha_K$. This result summarized below for any $K \geq 2$ helps find explicit rates of convergence for the VB approximation.

Lemma 3.2. *For $r_{n,K} = \frac{4\log(nK)}{n}$ and a prior $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K) \in \mathcal{M}_1^+(\mathcal{S}_K)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$, we can find a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that*

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq Kr_{n,K}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq Knr_{n,K}.$$

Thus, conditions (3.1) and (3.2) concerning the mixture components are always satisfied for guaranteeing consistency and obtaining convergence rates of the Variational Bayes procedure.

Remark 3.1. *When $K = 1$, Lemma 3.2 does not apply as $\frac{2}{K} > 1$. Nevertheless, as there is only one component, then any $p \in \mathcal{S}_K$ is equal to 1 and the two conditions are immediately satisfied for any prior π_p and any rate $r_{n,K}$ with $\rho_{p,n} = \pi_p$.*

The central idea of the proof of Lemma 3.2 (given in details in Section 6) is to consider the ball $\mathcal{B}$ centered at p^0 of radius $Kr_{n,K}$ defined as:

$$\mathcal{B} = \left\{ p \in \mathcal{S}_K / \ \mathcal{K}(p^0, p) \leq Kr_{n,K} \right\}.$$

Hence, when considering the restriction $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ of π_p to $\mathcal{B}$, condition (3.1) is trivially satisfied and condition (3.2) is restricted to

$$\pi_p(\mathcal{B}) \geq e^{-nKr_{n,K}}.$$

This is a very classical assumption stated in many papers to study the concentration of the posterior [23, 3, 49]. However, the computation of such a prior mass $\pi_p(\mathcal{B})$ is a major difficulty. Lemma 6.1 in [23] treated the case of L_1-balls for Dirichlet priors. Since then, only a few papers in the literature addressed

this issue. Our result extends the work in [23] to KL-balls, which is of great interest in our study. Moreover, the range of Dirichlet priors for which Lemma 3.2 is applicable is the same as the one in [23].

We conclude Subsection 3.1 by a short discussion on the implementation of the VB approximation. Indeed, VB methods are meant to be practical objects, so there would be no point in proving the consistency of a VB approximation that would not be computable in practice. Many algorithms have been studied in the literature, with good performances – see [11] and the references therein. In the case of mean-field approximation, the most popular method is to optimize iteratively with respect to all the independent components. Here this might seem difficult: it is indeed as difficult as maximizing the likelihood of a mixture. But a trick widely used in practice (see for example Section 7 in [27]) is to use the equality

$$
\text{for any } i = 1, ..., K \qquad - \log\left(\sum_{j=1}^{K} p_j q_{\theta_j}(X_i) \right)
$$
$$
= \min_{\omega^i \in \mathcal{S}_K} \left\{ - \sum_{j=1}^{K} \omega_j^i \log(p_j q_{\theta_j}(X_i)) + \sum_{j=1}^{K} \omega_j^i \log(\omega_j^i) \right\}.
$$

This equality is once again a consequence of Lemma 2.1 (take $\mathbf{E} = \{1, ..., K\}$, $\lambda = (1/K, ..., 1/K)$ and $h(j) = \log(p_j q_{\theta_j}(X_i))$). This leads to the program:

$$
\min_{\rho \in \mathcal{F}, \, w \in \mathcal{S}_K^n} \left\{ - \alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right) \right.
$$
$$
\left. + \alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \log(\omega_j^i) + \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j) \right\}.
$$

This version can be solved by coordinate descent, see Algorithm 1. Update formulas once again follow from Lemma 2.1 (for instance, line 7 can be obtained with $\mathbf{E} = \{1, ..., K\}$, $\lambda = (1/K, ..., 1/K)$ and $h(j) = \int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j)$, more details are provided in Section 6). This algorithm is, in the case $\alpha = 1$, exactly equivalent to the popular CAVI algorithm [48, 11, 28], where the ω_j^i's are interpreted as the posterior means of the latent variables Z_j^i's. A very short numerical study is provided in the Supplementary Material [17], but note that CAVI has already been extensively tested in practice [11].

3.2. Application to multinomial mixture models

We present in this section an application to the multinomial mixture model frequently used for text clustering [37], transport schedule analysis [14] and others. The parameter space is the $V - 1$ dimensional simplex $\Theta = \mathcal{S}_V$ with V a positive integer, and $q_\theta(X) = \prod_{v=1}^{V} \theta_{vj}^{\mathbb{1}(X=v)}$ for any $\theta \in \Theta$. We choose conjugate

---

Algorithm 1 Coordinate Descent Variational Bayes for mixtures

---

1: **Input**: a dataset $(X_1, ..., X_n)$, priors $\pi_p, \{\pi_j\}_{j=1}^K$ and a family $\{q_\theta / \theta \in \Theta\}$

2: **Output**: a variational approximation $\rho_p(p) \prod_{j=1}^K \rho_j(\theta_j)$

3: **Initialize** variational factors ρ_p, $\{\rho_j\}_{j=1}^K$

4: **until** convergence of the objective function **do**

5: **for** $i = 1, ..., n$ **do**

6:     **for** $j = 1, ..., K$ **do**

7:         set $w_j^i = \exp\left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right)$

8:     **end for**

9:     normalize $(w_j^i)_{1 \le j \le K}$

10: **end for**

11: set $\rho_p(dp) \propto \exp\left(\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(p_j) \right) \pi_p(dp)$

12: **for** $j = 1, ..., K$ **do**

13:     set $\rho_j(d\theta_j) \propto \exp\left(\alpha \sum_{i=1}^n \omega_j^i \log(q_{\theta_j}(X_i)) \right) \pi_j(d\theta_j)$

14: **end for**

Dirichlet priors as in [37] $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{D}_V(\beta_1, ..., \beta_V)$ with $\frac{2}{K} \le \alpha_j \le 1$ for $j = 1, ..., K$ and $\frac{2}{V} \le \beta_\ell \le 1$ for $\ell = 1, ..., V$.

The following corollary of Theorem 3.1 states that convergence of the VB approximation for the multinomial mixture model is achieved at rate $\frac{KV \log(nV)}{n}$ as soon as $V^V \ge K$, which is the case in many text mining models such as Latent Dirichlet Allocation [10] for which the size of the vocabulary is very large:

Corollary 3.3. *For any $\alpha \in (0, 1)$,*

$$
\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0} \right) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \right]
$$
$$
\le \frac{1+\alpha}{1-\alpha}\left[\frac{8KV \log(nV)}{n} \bigvee \frac{8K \log(nK)}{n} \right].
$$

The proof is in Section 6. We also specialize Algorithm 1 to the present setting (see Algorithm 2). Here ψ denotes the Digamma function, $\psi(x) = \frac{d}{dx}\log[\Gamma(x)]$ where Γ stands for the Gamma function $\Gamma(x) = \int_0^\infty \exp(-t)t^{x-1}dt$.

3.3. Application to Gaussian mixture models

Let us now address the case of the Gaussian mixture model. This is one of the most popular mixture models for many applications including model based clustering [12, 32] and VB approximations have been studied in depth for this model [33]. First, we will give rates of convergence of the VB approximation of the tempered posterior when the variance is known, and then when the variance is unknown.

First, we consider mixtures of V^2-variance Gaussians. The parameter space is $\Theta = \mathbb{R}$, and each component $j = 1, ..., K$ is parameterized by its mean $\theta_j = \mu_j$.

Algorithm 2 Coordinate Descent Variational Bayes for multinomial mixtures

---

1: **Initialize** variational parameters $(\phi_1, ..., \phi_K) \in \mathbb{R}_+^K$, $(\gamma_{1j}, ..., \gamma_{Vj}) \in \mathbb{R}_+^V$ and corresponding variational distributions $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$, $\rho_j = \mathcal{D}_V(\gamma_{1j}, ..., \gamma_{Vj})$ for $j = 1, ..., K$
2: **until** convergence of the objective function **do**
3: **for** $i = 1, ..., n$ **do**
4:    **for** $j = 1, ..., K$ **do**
5:       set $w_j^i = \exp\left(\psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell) + \psi(\gamma_{X_i,j}) - \psi(\sum_{v=1}^{V} \gamma_{vj}) \right)$
6:    **end for**
7:    normalize $(w_j^i)_{1 \le j \le K}$
8: **end for**
9: set $\phi_j = \alpha_j + \alpha \sum_{i=1}^{n} w_j^i \quad$ for $\quad j = 1, ..., K$
10: set $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$
11: **for** $j = 1, ..., K$ **do**
12:    set $\gamma_{vj} = \beta_v + \alpha \sum_{i=1}^{n} w_j^i \mathbb{1}(X_i = v) \quad$ for $\quad v = 1, ..., V$
13:    set $\rho_j = \mathcal{D}_V(\gamma_{1j}, ..., \gamma_{Vj})$
14: **end for**

We select priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{N}(0, \mathcal{V}^2)$ with $\frac{2}{K} \le \alpha_j \le 1$ for $j = 1, ..., K$ and $\mathcal{V}^2 > 0$. The following result gives a rate of convergence $Kr_{n,K}$ of the VB approximation:

Corollary 3.4. *Let us define* $r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \right]$. *Then, for any* $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \right) \tilde{\pi}_{n,\alpha}(d\theta|X_1^n) \right] \le \frac{1+\alpha}{1-\alpha} 2Kr_{n,K}.$$

One can see that for n large enough, the convergence rate is $\frac{K\log(nK)}{n}$, which comes from the estimation of the weights of the mixture.

We can also provide a similar result when the variance of each component is unknown. The convergence rate remains the same, and is entirely characterized by the weights consistency rate. The parameter space is now $\Theta = \mathbb{R} \times (0, +\infty)$, and each component $j = 1, ..., K$ is parameterized by its pair mean/variance $\theta_j = (\mu_j, \sigma_j^2)$. We consider again a Dirichlet prior $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ with $\frac{2}{K} \le \alpha_j \le 1$ for $j = 1, ..., K$ on $p \in \mathcal{S}_K$, and we will provide our results for two different priors π_j frequently used in the literature: a Normal-Inverse-Gamma prior [40] and a factorized prior [45]. We define the Normal-Inverse-Gamma distribution as follows:

Definition 3.1. *The Normal-Inverse-Gamma* $\mathcal{NIG}(\mu, \theta^2, a, b)$ *is the distribution which density w with respect to Lebesgue measure is defined by* $w(x,y) = g(x|\mu, \frac{y}{\theta^2})h(y|a,b)$, *where* $g(.|\mu, \sigma^2)$ *is the density function of a Gaussian distribution of mean μ and variance σ^2, and* $h(.|a,b)$ *is the density distribution of an Inverse-Gamma of parameters a and b.*

Corollary 3.5. *Let us fix* $\alpha \in (0,1)$.

- *For a Normal-Inverse-Gamma prior* $\pi_j = \mathcal{NIG}(0, \mathcal{V}^{-2}, 1, \gamma^2)$ *for each* $j = 1, ..., K$. *With*

$$r_{n,K} \;=\; \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n}\Bigg[2\log(n\sqrt{\mathcal{V}}) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \Bigg],$$

$$\mathbb{E}\Bigg[\int D_\alpha \Bigg(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \Bigg) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \Bigg] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

- *For the factorized prior* $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \bigotimes \mathcal{IG}(1, \gamma^2)$ *for each* $j = 1, ..., K$. *With*

$$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n}\Bigg[2\log(n\sqrt{\mathcal{V}}) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} + \frac{1}{2}\log\left(\frac{(\sigma_j^0)^2}{\gamma^4} \right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \Bigg],$$

$$\mathbb{E}\Bigg[\int D_\alpha \Bigg(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \Bigg) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \Bigg] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

One can see that even when the variance has to be estimated, the convergence rate still achieves $\frac{K\log(nK)}{n}$ for n large enough, whatever the form of the prior - factorized or not.

We give in Algorithm 3 a version of Algorithm 1 for unit-variance Gaussian mixtures with priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{N}(0, \mathcal{V}^2)$ where $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ and $\mathcal{V}^2 > 0$.

Algorithm 3 Coordinate Descent Variational Bayes for unit-variance Gaussian mixtures

---

1: **Initialize** variational parameters $(\phi_1, ..., \phi_K) \in (\mathbb{R}_+^*)^K$, $(n_j, s_j^2) \in \mathbb{R} \times \mathbb{R}_+^*$ and corresponding variational distributions $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$, $\rho_j = \mathcal{N}(n_j, s_j^2)$ for $j = 1, ..., K$
2: **until** convergence of the objective function **do**
3: **for** $i = 1, ..., n$ **do**
4:     **for** $j = 1, ..., K$ **do**
5:         set $w_j^i = \exp\left(\psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell) - \frac{1}{2}\{ s_j^2 + (n_j - X_i)^2 \} \right)$
6:     **end for**
7:     normalize $(w_j^i)_{1 \leq j \leq K}$
8: **end for**
9: set $\phi_j = \alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i$ for $j = 1, ..., K$
10: set $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$
11: **for** $j = 1, ..., K$ **do**
12:     set $n_j = \frac{\alpha \sum_{i=1}^{n} \omega_j^i X_i}{1/\mathcal{V}^2 + \alpha \sum_{i=1}^{n} \omega_j^i}$ and $s_j^2 = \frac{1}{1/\mathcal{V}^2 + \alpha \sum_{i=1}^{n} \omega_j^i}$
13:     set $\rho_{\mu,j} = \mathcal{N}(n_j, s_j^2)$
14: **end for**

3.4. Extension to the misspecified case

From now we do not assume any longer that the true distribution P^0 belongs to the K-mixtures model. We still consider a prior $\pi = \pi_p \bigotimes_{j=1}^{K} \pi_j$ on $\theta \in \Theta_K$ for which $\pi_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\pi_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$.

For some value $r_{n,K}$, we introduce the set $\Theta_K(r_{n,K})$ of parameters $\theta^* \in \Theta_K$ such that:

- there exists a set $\mathcal{A}_{n,K} \subset \mathcal{S}_K$ satisfying:

  - for each $p \in \mathcal{A}_{n,K}$, for each $j = 1, ..., K$, $\log\left(\frac{p_j^*}{p_j}\right) \leq K r_{n,K}$,

  - $\pi_p(\mathcal{A}_{n,K}) \geq e^{-nK r_{n,K}}$.

- there are distributions $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ $(j = 1, ..., K)$ such that for $j = 1, ..., K$:

$$\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right] \rho_{j,n}(d\theta_j) \leq r_{n,K} \quad , \qquad \mathcal{K}(\rho_{j,n}, \pi_j) \leq n r_{n,K}. \quad (3.3)$$

Let us discuss this definition. To begin with, the first item of the definition of $\Theta_K(r_{n,K})$ can seem quite restrictive. It is even a much more stronger assumption than (3.1) and (3.2). Nevertheless, the way to find the required measures $\rho_{p,n}$ in Lemma 3.2 in the well-specified case implies constructing in the proof such sets $\mathcal{A}_{n,K}$ for the true weight parameter p^0. As a consequence, it might seem reasonable to replace conditions (3.1) and (3.2) by the first part of the definition of $\Theta_K(r_{n,K})$. On the other hand, the condition given by (3.3) looks like those of Theorem 2.7 in [3]. Once again, the difference is that inequalities must be satisfied here for each component. A condition on both the true distribution P^0 and the parameter θ^* considered is required through the expectation term. Besides, condition (3.3) is equivalent to (3.1) and (3.2) when the model is well-specified.

Theorem 3.6. *For any $\alpha \in (0, 1)$,*

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j Q_{\theta_j}, P^0\right) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n)\right]$$

$$\leq \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \Theta_K(r_{n,K})} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

Remark 3.2. *If there is no $r_{n,K}$ such that $\Theta_K(r_{n,K})$ is not empty, then the right-hand side is equal to infinity (by convention) for any value of $r_{n,K}$ and the inequality is useless. Nevertheless, this is not the case in models used in practice. We show an example below.*

It is worth mentioning that even if this is not exactly an oracle inequality as the risk function in the left-hand side (α-Renyi divergence) is lower than the right-hand side one (Kullback-Leibler divergence), but the theorem still

remains of great interest. Indeed, when the minimizer of $\mathcal{K}(P^0, P_\theta)$ with respect to $\theta \in \Theta_K(r_{n,K})$ exists and is such that the corresponding Kullback-Leibler divergence is small, then our oracle inequality is informative as it gives a small bound on the expected risk of the Variational Bayes.

To illustrate the relevance of Theorem 3.6, we provide the following result that is applicable for a wide range of generating distributions when considering the family of unit-variance Gaussian mixtures with priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K) \in \mathcal{M}_1^+(\mathcal{S}_K)$ with $\frac{2}{K} \le \alpha_j \le 1$ for $j = 1, ..., K$ ($K \ge 2$) and $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \in \mathcal{M}_1^+(\mathbb{R})$ for $j = 1, ..., K$ with $\mathcal{V}^2 > 0$:

Corollary 3.7. *Assume that the true distribution P^0 is such that $\mathbb{E}|X| < +\infty$. Let $L > 0$.*
For $r_{n,K} = \frac{4 \log(nK)}{n} \bigvee_{j=1}^K \frac{1}{n} \left[\frac{1}{2} \log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log(\mathcal{V}) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2} \right]$, we get
$\mathcal{S}_K \times [-L, L]^K \subset \Theta_K(r_{n,K})$ and for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j Q_{\theta_j}, P^0 \right) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \right]$$
$$\le \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \mathcal{S}_K \times [-L,L]^K} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

Remark 3.3. *If the true distribution is a mixture of unit-variance Gaussians with components means between $-L$ and L, then $\mathbb{E}|X| < +\infty$ and the first term of the right-hand side of the inequality is equal to zero, which gives directly for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j Q_{\theta_j}, P^0 \right) \tilde{\pi}_{n,\alpha}(d\theta | X_1^n) \right] \le \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

4. Variational Bayes model selection

In this section, we extend the problem to a larger family of distributions. We want to model the generating distribution P^0 using mixtures with an unknown number of components in a possibly misspecified setting. Thus, we consider a countable collection $\{\mathcal{M}_K / K \in \mathbb{N}^*\}$ of statistical mixture models

$$\mathcal{M}_K = \left\{ P_{\theta_K} = \sum_{j=1}^K p_{j,K} Q_{\theta_{j,K}} \ / \ \theta_K \in \Theta_K \right\}$$

with $\Theta_K = \mathcal{S}_K \times \Theta^K$, $\mathcal{S}_K = \{p_K = (p_{1,K}, ..., p_{K,K}) \in [0,1]^K / \sum_{j=1}^K p_{j,K} = 1\}$ and the general notation $\theta_K = (p_K, \theta_{1,K}, ..., \theta_{K,K})$. We would like to emphasize that the notations are slightly different as the size of each component parameter depends on the model complexity K. The entire parameter space Ω is the union of all parameter spaces Θ_K associated with each model index K: $\Omega = \cup_{K=1}^\infty \Theta_K$,

and we can think of a whole statistical model $\mathcal{M} = \cup_{K=1}^{\infty} \mathcal{M}_K$ as the union of all collections $\mathcal{M}_K$. First, we can notice that different models $\mathcal{M}_K$ never overlap as parameters in each one do not have the same length. Nonetheless, parameters in complex models (models $\mathcal{M}_K$ with large K) can be sparse and therefore contain the "same information" as parameters in less complex ones, i.e. can lead to the same distribution P_θ.

The prior specification is a crucial point. As mentioned above, each parameter depends on the number of components. Then, we specify a prior weight π_K assigned to the model $\mathcal{M}_K$ and a conditional prior $\Pi_K(.)$ on $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$. More precisely, we define our conditional prior on $\theta_K = (p_K, \theta_{1,K}, ..., \theta_{K,K})$ as follows: given K, the weight parameter $p_K = (p_{1,K}, ..., p_{K,K})$ is supposed to follow a distribution $\pi_{p,K}$ on $\mathcal{M}_1^+(\mathcal{S}_K)$; finally, given K, we set independent priors $\pi_{j,K}$ for the component parameters $\theta_{j,K}$ where each $\pi_{j,K}$ is a probability distribution on $\mathcal{M}_1^+(\Theta)$. In a nutshell:

$$\pi = \sum_{K=1}^{+\infty} \pi_K \Pi_K$$

with

$$\Pi_K(\theta_K) = \pi_{p,K}(p_K) \prod_{j=1}^{K} \pi_{j,K}(\theta_{j,K}).$$

We have to adapt the notations for the VB approximations. The tempered posteriors $\pi_{n,\alpha}^K(.|X_1^n)$ on parameter $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$, is defined again as

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

The Variational Bayes $\tilde{\pi}_{n,\alpha}^K(.|X_1^n)$ is the projection of the tempered posterior onto some set $\mathcal{F}_K$ following the mean-field assumption: the variational factor corresponding to the weight parameter $p_K = (p_{1,K}, ..., p_{K,K})$ is any distribution ρ_p on $\mathcal{M}_1^+(\mathcal{S}_K)$; besides, we consider independent variational distributions $\rho_j(\theta_{j,K})$ for the component parameters $\theta_{j,K}$ where each ρ_j is a probability distribution on $\mathcal{M}_1^+(\Theta)$. Then, $\mathcal{F}_K = \{\rho_p \bigotimes_{j=1}^{K} \rho_j / \rho_p \in \mathcal{M}_1^+(\mathcal{S}_K), \ \rho_j \in \mathcal{M}_1^+(\Theta) \ \forall j = 1, ..., K\}$, and

$$\tilde{\pi}_{n,\alpha}^K(.|X_1^n) = \underset{\rho_K \in \mathcal{F}_K}{\arg\min} \mathcal{K}\left(\rho_K, \pi_{n,\alpha}^K(.|X_1^n)\right).$$

We recall that an alternative way to define the variational estimate is to use the Evidence Lower Bound via the optimization program (2.1):

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

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

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

The objective is to propose a data-driven estimate $\hat{K}$ of the number of components from which we will pick up our final VB estimate $\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n)$ and derive an oracle inequality in the spirit of [31]. It is stated in [11] that $\arg\max_{K \geq 1} \mathcal{L}(K)$ is widely used in practice, without any theoretical justification. We propose

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

which is a penalized version of the ELBO. Note that taking (π_K) as uniform on a finite set $\{1, 2, \ldots, K_{\max}\}$ leads to the procedure described in [11]. We discuss below the choice $\pi_K = 2^{-K}$.

We can now state the following result which provides an oracle-type inequality for $\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n)$:

Theorem 4.1. *For any $\alpha \in (0,1)$,*

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

$$\leq \inf_{K \geq 1} \left\{ \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \Theta_K(r_{n,K})} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)} \right\}.$$

This oracle inequality shows that our variational distribution adaptively satisfies the best possible balance between bias (misspecification error) and variance (estimation error). If we assume that there is actually a K_0 and $\theta^* \in \Theta_{K_0}$ such that $P^0 = P_{\theta^*}$ then the theorem will imply

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta|X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} 2K_0 r_{n,K_0} + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

Note that this does not mean that $\hat{K} = K_0$, but this means that the convergence rate of P_θ to P^0 $\tilde{\pi}_{n,\alpha}^{\hat{K}}(.|X_1^n)$ is as good as if we actually knew P_0. The objective of estimating K_0 is a completely different task [47]. Estimating K_0 would also require identifiability conditions that are not necessary for our results.

The variance term is composed of two parts. The first one, $K r_{n,K}$ up to a multiplicative constant, corresponds to the rate obtained when approximating the true distribution with mixtures of model $\mathcal{M}_K$. The second part of the overall rate can be interpreted as a complexity term over the different models reflecting our prior belief. For instance, if we want to penalize more complex models, we can take $\pi_K = 2^{-K}$ and the corresponding term will be of order K/n. In practice, as soon as $\frac{1}{n} \lesssim r_{n,K}$, then this penalty term is negligible when compared to the approximating rate $K r_{n,K}$: this means that this choice can be considered safe, as it does not interfere with the estimation rate.

5. Conclusion

Using variational inference, we studied consistency of variational approximations for estimation and model selection in mixtures. When considering tempered

posteriors, we showed that Variational Bayes is consistent and we gave statistical guarantees to model selection based on the ELBO. For further investigation, it would be interesting to explore the case of Bayesian posteriors when $\alpha = 1$. The recent work of Zhang and Gao [49] gives the tools for tackling such an issue, and allows one to consider risk functions different from α-Renyi divergence. But the conditions would be more stringent, and misspecification would be more problematic in this case.

Another point of interest is the study of the non-convex optimization program (2.2). Indeed, the proposed coordinate optimization can lead to a local extremum, and this implies that one needs to pay attention to initialization. The same problem also occurs in the Expectation-Maximization (EM) algorithm. In practice, users often run EM or CAVI several times with different initial distributions. Many practical ideas were proposed to target the global extrema more efficiently with EM [35] and could be extended to CAVI. But the question of convergence remains open in theory.

Finally, note that our results are remarkable as there are almost no conditions on the mixtures considered. In this paper we have focused on estimating the true probability distribution P^0, even in the well-specified case. We have no results on the estimation of the parameters. In the case of mixtures, these results are extremely difficult to obtain even for Gaussian mixtures [46]. They require restrictions on the parameters set and lead to different rates of convergence. The consistency of VB for the estimation of the parameters remains open.

6. Proofs

6.1. Some useful lemmas

We provide in this section two useful lemmas required in many proofs below.

6.1.1. An upper bound on the Kullback-Leibler divergence between two mixtures

The lemma below was first stated by [39] for mixtures of Gaussians, [21] checked that the proof remains valid for general mixtures. It is a tool widely used in signal processing [27]. We provide the proof for the sake of completeness.

Lemma 6.1. Let $p, p^0 \in \mathcal{S}_K$ and $\theta_j, \theta_j^0 \in \Theta$ for $j = 1, ..., K$. Then,

$$\mathcal{K}\left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K} p_j q_{\theta_j}\right) \leq \mathcal{K}(p^0, p) + \sum_{j=1}^{K} p_j^0 \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})$$

Proof. For any nonnegative numbers $\alpha_1, ..., \alpha_K$ and positive $\beta_1, ..., \beta_K$, we have:

$$\left(\sum_{j=1}^{K} \alpha_j\right) \log\left(\frac{\sum_{j=1}^{K} \alpha_j}{\sum_{j=1}^{K} \beta_j}\right) = \left(\sum_{j=1}^{K} \beta_j\right)\left(\frac{\sum_{j=1}^{K} \alpha_j}{\sum_{j=1}^{K} \beta_j}\right) \log\left(\frac{\sum_{j=1}^{K} \alpha_j}{\sum_{j=1}^{K} \beta_j}\right)$$

$$= \left(\sum_{j=1}^{K} \beta_j \right) \left(\sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} \frac{\alpha_j}{\beta_j} \right) \log \left(\sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} \frac{\alpha_j}{\beta_j} \right)$$

$$= \left(\sum_{j=1}^{K} \beta_j \right) f \left(\sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} \frac{\alpha_j}{\beta_j} \right)$$

where f is the convex function $x \longmapsto x \log(x)$. As $\sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} = 1$, then using Jensen's inequality:

$$\left(\sum_{j=1}^{K} \alpha_j \right) \log \left(\frac{\sum_{j=1}^{K} \alpha_j}{\sum_{j=1}^{K} \beta_j} \right) = \left(\sum_{j=1}^{K} \beta_j \right) f \left(\sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} \frac{\alpha_j}{\beta_j} \right)$$

$$\leq \left(\sum_{j=1}^{K} \beta_j \right) \sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} f \left(\frac{\alpha_j}{\beta_j} \right)$$

$$= \left(\sum_{j=1}^{K} \beta_j \right) \sum_{j=1}^{K} \frac{\beta_j}{\sum_{l=1}^{K} \beta_l} \frac{\alpha_j}{\beta_j} \log \left(\frac{\alpha_j}{\beta_j} \right)$$

$$= \sum_{j=1}^{K} \alpha_j \log \left(\frac{\alpha_j}{\beta_j} \right).$$

The inequality remains valid when some or all β_j's are zero. Indeed, assume that $\beta_j = 0$. If $\alpha_j \neq 0$, then the jth term of the sum in the right-hand side is $\alpha_j \log(\alpha_j/\beta_j) = +\infty$, and the result is obvious. Otherwise, $\alpha_j = 0$, hence the jth term of each sum in the inequality is zero as $\alpha_j \log(\alpha_j/\beta_j) = 0$, and the inequality can be obtained considering only the other numbers.

Thus, for $p, p^0 \in \mathcal{S}_K$ and $\theta_j, \theta_j^0 \in \Theta$ for $j = 1, ..., K$:

$$\mathcal{K} \left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K} p_j q_{\theta_j} \right) = \int \left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \right) \log \left(\frac{\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}}{\sum_{j=1}^{K} p_j q_{\theta_j}} \right)$$

$$\leq \int \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \log \left(\frac{p_j^0 q_{\theta_j^0}}{p_j q_{\theta_j}} \right)$$

$$= \int \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \log \left(\frac{p_j^0}{p_j} \right) + \int \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \log \left(\frac{q_{\theta_j^0}}{q_{\theta_j}} \right)$$

$$= \sum_{j=1}^{K} p_j^0 \log \left(\frac{p_j^0}{p_j} \right) \left(\int q_{\theta_j^0} \right) + \sum_{j=1}^{K} p_j^0 \int q_{\theta_j^0} \log \left(\frac{q_{\theta_j^0}}{q_{\theta_j}} \right)$$

$$= \mathcal{K}(p^0, p) + \sum_{j=1}^{K} p_j^0 \mathcal{K}(q_{\theta_j^0}, q_{\theta_j}),$$

which ends the proof. $\square$

6.1.2. KL-divergence between Gaussian distributions and between Normal-Inverse-Gamma distributions

We give in this section the Kullback-Leibler divergence between 1-dimensional Gaussian distributions and between Normal-Inverse-Gamma distributions.

Lemma 6.2. *We denote u and v the density functions of the respective Gaussian distributions $\mathcal{N}(\mu_u, \sigma_u^2)$ and $\mathcal{N}(\mu_v, \sigma_v^2)$. Similarly, we denote p and q the two densities of $\mathcal{NIG}(\mu_1, \theta_1^2, a_1, b_1)$ and $\mathcal{NIG}(\mu_2, \theta_2^2, a_2, b_2)$. Then:*

$$\mathcal{K}(u,v) = \frac{1}{2}\log\left(\frac{\sigma_v^2}{\sigma_u^2}\right) + \frac{\sigma_u^2}{2\sigma_v^2} + \frac{(\mu_v - \mu_u)^2}{2\sigma_v^2} - \frac{1}{2}$$

and

$$\mathcal{K}(p,q) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2}$$
$$+ (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2\log\left(\frac{b_1}{b_2}\right) + a_1\frac{b_2 - b_1}{b_1}.$$

Proof. The first equality is extremely classical so we don't provide the proof. For the second one,

$$\mathcal{K}(p,q) = \int_{\mathbb{R}_+^*}\int_{\mathbb{R}} p(x,y)\log\left(\frac{p(x,y)}{q(x,y)}\right)dxdy$$

$$= \int_{\mathbb{R}_+^*}\int_{\mathbb{R}} p(x|Y=y)p_Y(y)\log\left(\frac{p(x|Y=y)}{q(x|Y=y)}\frac{p_Y(y)}{q_Y(y)}\right)dxdy$$

$$= \int_{\mathbb{R}_+^*} p_Y(y)\left(\int_{\mathbb{R}} p(x|Y=y)\log\left(\frac{p(x|Y=y)}{q(x|Y=y)}\right)dx\right)dy$$

$$+ \int_{\mathbb{R}_+^*} p_Y(y)\log\left(\frac{p_Y(y)}{q_Y(y)}\right)dy$$

$$= \mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\left[\mathcal{K}(p(.|Y), q(.|Y))\right] + \mathcal{K}(p_Y, q_Y).$$

Using the KL-divergence between Gaussians:

$$\mathcal{K}(p(.|Y), q(.|Y)) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2Y} - \frac{1}{2}$$

hence

$$\mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\left[\mathcal{K}(p(.|Y), q(.|Y))\right] = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2}$$
$$+ \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2}\mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\left[\frac{1}{Y}\right] - \frac{1}{2}$$

i.e.

$$\mathbb{E}_{Y \sim \mathcal{IG}(a_1, b_1)}\left[\mathcal{K}(p(.|Y), q(.|Y))\right] = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2},$$

and using the KL-divergence between Inverse-Gamma distributions

$$\mathcal{K}(p_Y, q_Y) = (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2 \log\left(\frac{b_1}{b_2}\right) + a_1\frac{b_2 - b_1}{b_1}$$

where Γ and ψ are respectively the Gamma and Digamma functions, we have:

$$\mathcal{K}(p, q) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2}$$
$$+ (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2 \log\left(\frac{b_1}{b_2}\right) + a_1\frac{b_2 - b_1}{b_1}. \qquad \square$$

6.2. Proof of Theorem 3.1

This result relies on an application of Theorem 2.6 in [3] to mixture models. The proof of Theorem 2.6 in [3] itself relies mostly on a deviation inequality from [8] and on PAC-Bayesian theory [15, 16].

Proof. Fix $0 < \alpha < 1$. Theorem 2.6 from [3] gives:

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0}\right)\tilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right]$$
$$\leq \inf_{\rho \in \mathcal{F}}\left\{\frac{\alpha}{1 - \alpha}\int \mathcal{K}\left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K} p_j q_{\theta_j}\right)\rho(d\theta) + \frac{\mathcal{K}(\rho, \pi)}{n(1 - \alpha)}\right\}.$$

Thanks to Lemma 6.1

$$\mathcal{K}\left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K} p_j q_{\theta_j}\right) \leq \mathcal{K}(p^0, p) + \sum_{j=1}^{K} p_j^0 \mathcal{K}(q_{\theta_j^0}, q_{\theta_j}).$$

Then

$$\mathcal{K}(\rho, \pi) = \mathcal{K}\left(\rho_p \bigotimes_{j=1}^{K} \rho_j, \pi_p \bigotimes_{j=1}^{K} \pi_j\right) = \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j)$$

the last inequality being obtained thanks to Theorem 28 in [43]. Gathering all the pieces together leads to

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0}\right)\tilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right]$$

$$\leq \inf_{\rho \in \mathcal{F}} \left\{ \frac{\alpha}{1-\alpha} \left[\int \mathcal{K}(p^0, p) \rho_p(dp) + \sum_{j=1}^{K} \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j}) \rho_j(d\theta_j) \right] \right.$$

$$\left. + \frac{\mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j)}{n(1-\alpha)} \right\}$$

that is the result stated in Theorem 3.1. □

6.3. Proof of Lemma 3.2

Proof. Let us define $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ by the following formula $\rho_{p,n}(dp) \propto \mathbf{1}(p \in \mathcal{B}) \pi_p(dp)$ with

$$\mathcal{B} = \left\{ p \in \mathcal{S}_K / \mathcal{K}(p^0, p) \leq K r'_{n,K} \right\}$$

and

$$r'_{n,K} = \max \left(\frac{1}{K(n-1)}, \frac{\log(n(K-1)\Gamma(A)^{\frac{K}{K-1}}/M_p^0)}{n} \right)$$

where $A = \frac{2}{K}$ and $M_p^0 = \max\{p_j^0 / j = 1, ..., K\}$. We adopt the notation $S = \sum_{j=1}^{K} \alpha_j$ in the following. Recall that by assumption $K \geq 2$ and hence $A = \frac{2}{K} \leq 1$.

First, $\int \mathcal{K}(p^0, p) \rho_{p,n}(dp) \leq K r'_{n,K}$.

Then, let us show that $\mathcal{K}(\rho_{p,n}, \pi_p) \leq K n r'_{n,K}$. For that, let us define

$$\mathcal{A} = \left\{ p \in \mathbb{R}^K / p_j^0 e^{-K r'_{n,K}} \leq p_j \leq p_j^0 e^{-K r'_{n,K}} + \frac{p_K^0}{n(K-1)} \right.,$$

$$\left. \text{for } j = 1, ..., K-1, \ p_K = 1 - \sum_{j=1}^{K-1} p_j \right\}$$

where K is such that $p_K^0 = \max\{p_j^0 / j = 1, ..., K\}$ (this assumption can always be fulfilled by reordering and relabelling the vector components). Then, $p_K^0 \geq \frac{1}{K}$ (otherwise, the sum of the components of p^0 would be strictly lower than 1 and the vector would not be included in $\mathcal{S}_K$). We will show that $\mathcal{A} \subset \mathcal{B}$ and that $\pi_p(\mathcal{A}) \geq e^{-K n r'_{n,K}}$. Then, we will conclude thanks to the following formula: $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{B}))$.

First, let us show that $\mathcal{A} \subset \mathcal{B}$.

Let $p \in \mathcal{A}$. As $p_K = 1 - \sum_{j=1}^{K-1} p_j$, we just need to check that $\mathcal{K}(p^0, p) \leq K r'_{n,K}$ and that $p_j \geq 0$ for each $j = 1, ..., K$.

The first part can be proven using the definition of $\mathcal{A}$. According to the $K-1$ left-hand side inequalities in the definition of $\mathcal{A}$,

$$\mathcal{K}(p^0, p) = \sum_{j=1}^{K-1} p_j^0 \log\left(\frac{p_j^0}{p_j}\right) + p_K^0 \log\left(\frac{p_K^0}{p_K}\right)$$

$$\leq \sum_{j=1}^{K-1} p_j^0 \log(e^{Kr'_{n,K}}) + p_K^0 \log\left(\frac{p_K^0}{p_K}\right)$$

$$= \sum_{j=1}^{K-1} p_j^0 Kr'_{n,K} + p_K^0 \log\left(\frac{p_K^0}{p_K}\right)$$

$$= (1 - p_K^0)Kr'_{n,K} + p_K^0 \log\left(\frac{p_K^0}{p_K}\right).$$

All we need to show now is that $\log\left(\frac{p_K^0}{p_K}\right) \leq Kr'_{n,K}$. This comes from the following inequalities:

$$\log\left(\frac{p_K^0}{p_K}\right) = \log\left(\frac{p_K^0}{1 - \sum_{j=1}^{K-1} p_j}\right) \leq \log\left(\frac{p_K^0}{1 - \sum_{j=1}^{K-1} p_j^0 e^{-Kr'_{n,K}} - \frac{p_K^0}{n}}\right)$$

$$= \log\left(\frac{p_K^0}{1 - (1 - p_K^0)e^{-Kr'_{n,K}} - \frac{p_K^0}{n}}\right)$$

$$\leq \frac{p_K^0}{1 - (1 - p_K^0)e^{-Kr'_{n,K}} - \frac{p_K^0}{n}} - 1$$

$$= \frac{p_K^0 - 1 + (1 - p_K^0)e^{-Kr'_{n,K}} + \frac{p_K^0}{n}}{1 - (1 - p_K^0)e^{-Kr'_{n,K}} - \frac{p_K^0}{n}}$$

i.e. $\log\left(\frac{p_K^0}{p_K}\right) \leq \dfrac{p_K^0 - 1 + (1 - p_K^0)e^{-Kr'_{n,K}} + \frac{p_K^0}{n}}{1 - (1 - p_K^0)e^{-Kr'_{n,K}} - \frac{p_K^0}{n}}$

$$= \frac{\frac{p_K^0}{n} - (1 - p_K^0)(1 - e^{-Kr'_{n,K}})}{p_K^0(1 - \frac{1}{n}) + (1 - p_K^0)(1 - e^{-Kr'_{n,K}})}$$

$$= \frac{\frac{1}{n} - (\frac{1}{p_K^0} - 1)(1 - e^{-Kr'_{n,K}})}{(1 - \frac{1}{n}) + (\frac{1}{p_K^0} - 1)(1 - e^{-Kr'_{n,K}})}$$

$$\leq \frac{\frac{1}{n}}{1 - \frac{1}{n}} = \frac{1}{n - 1}$$

$$\leq Kr'_{n,K}.$$

Hence $\mathcal{K}(p^0, p) \leq (1 - p_K^0)Kr'_{n,K} + p_K^0 \log\left(\frac{p_K^0}{p_K}\right) \leq (1 - p_K^0)Kr'_{n,K} + p_K^0 Kr'_{n,K} = Kr'_{n,K}$.

On the other hand, for $j = 1, ..., K - 1$, $p_j \geq p_j^0 e^{-Kr'_{n,K}} \geq 0$ and:

$$p_K = 1 - \sum_{j=1}^{K-1} p_j \geq 1 - \sum_{j=1}^{K-1}\left(p_j^0 e^{-Kr'_{n,K}} + \frac{p_K^0}{n(K-1)}\right)$$

$$= 1 - \left((1 - p_K^0)e^{-Kr'_{n,K}} + \frac{p_K^0}{n}\right)$$

$$\geq 1 - (1 - p_K^0)e^{-Kr'_{n,K}} - p_K^0$$
$$= (1 - p_K^0)(1 - e^{-Kr'_{n,K}})$$
$$\geq 0.$$

Then, $p \in \mathcal{B}$, and finally $\mathcal{A} \subset \mathcal{B}$.

Now, let us show that $\pi_p(\mathcal{A}) \geq e^{-Knr'_{n,K}}$.

Let us denote f the density of the $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ Dirichlet distribution:

$$f(p) = \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K} p_j^{\alpha_j - 1} \, \mathbf{1}(p \in \mathcal{S}_K).$$

Thus, we can lower bound $\pi_p(\mathcal{A})$:

$$\pi_p(\mathcal{A}) = \int_{\mathcal{A}} f(p_1, ..., p_K) dp = \int_{\mathcal{A}} \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K} p_j^{\alpha_j - 1} \, \mathbf{1}(p \in \mathcal{S}_K) dp$$

$$\geq \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \frac{p_K^0}{n(K-1)}} p_j^{\alpha_j - 1} \, dp_j$$

as for $p \in \mathcal{A}$, $0 \leq p_j^0 e^{-Kr'_{n,K}} \leq p_j \leq p_j^0 e^{-Kr'_{n,K}} + \frac{p_K^0}{n(K-1)} \leq 1$ for each $j = 1, ..., K - 1$ (as $\mathcal{A} \subset \mathcal{B}$), and then $p_j^{\alpha_j - 1} \geq 1$.

Then, by definition of $r'_{n,K}$, $\frac{p_K^0}{n(K-1)} \geq \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}$, and using inequalities $\Gamma(A) \geq \Gamma(\alpha_j)$ as $A \leq \alpha_j \leq 1$ and $\Gamma(S) \geq 1$ as $S \geq 2$,

$$\pi_p(\mathcal{A}) \geq \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}} p_j^{\alpha_j - 1} \, dp_j$$

$$\geq \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}} dp_j$$

$$= \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \prod\limits_{j=1}^{K-1} \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}$$

$$= \frac{\Gamma(S)}{\prod\limits_{j=1}^{K} \Gamma(\alpha_j)} \Gamma(A)^K e^{-n(K-1)r'_{n,K}}$$

$$\geq e^{-nKr'_{n,K}}.$$

Hence, as $\mathcal{A} \subset \mathcal{B}$, $\pi_p(\mathcal{B}) \geq \pi_p(\mathcal{A}) \geq e^{-nKr'_{n,K}}$, and finally, $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{B})) \leq Knr'_{n,K}$.

We just proved the lemma but with the rate $r'_{n,K}$ instead of the value $r_{n,K}$ used in the lemma. We can conlude by noticing that the result is valid for every r such that $r'_{n,K} \leq r$, and that in particular $r'_{n,K} \leq r_{n,K}$. This last result comes from the inequality:

$$\Gamma(A) \leq \frac{\Gamma(1 + \frac{A}{2})}{\left(\frac{A}{2}\right)^{1 - \frac{A}{2}}}$$

which is a direct application of the left-hand side of inequality (3.2) part 3 in [30] with $x = \frac{A}{2} > 0$ and $\lambda = \frac{A}{2} \in (0, 1)$. As, $1 + \frac{A}{2} \in [1, 2]$, then $\Gamma(1 + \frac{A}{2}) \leq 1$, and $\frac{1}{\left(\frac{A}{2}\right)^{1 - \frac{A}{2}}} = K^{1 - \frac{A}{2}} \leq K$. Thus:

$$\Gamma(A) \leq K$$

and as $K \geq 2$ and $p_K^0 \geq \frac{1}{K}$, it follows that

$$\log((K-1)\Gamma(A)^{\frac{K}{K-1}}/p_K^0) \leq \log(K(K)^{\frac{K}{K-1}}K) \leq \log(K(K)^2 K) \leq \log(K^4)$$

i.e. $r'_{n,K} \leq \max(\frac{1}{K(n-1)}, \frac{\log(nK^4)}{n}) \leq \max(\frac{1}{K(n-1)}, \frac{4\log(nK)}{n})$. Besides, $\frac{n}{n-1} = 1 + \frac{1}{n-1} \leq 2$ implies $\frac{1}{K(n-1)} \leq \frac{1}{2(n-1)} \leq \frac{1}{n} \leq \frac{4\log(2)}{n} \leq \frac{4\log(nK)}{n}$, and finally $r'_{n,K} \leq \frac{4\log(nK)}{n} = r_{n,K}$. $\qquad \square$

6.4. Proof of Corollary 3.3

Proof. According to Lemma 3.2, there exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K\frac{4\log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq Kn\frac{4\log(nK)}{n}.$$

Similarly, the same result states that there exists distributions $\rho_{j,n} \in \mathcal{M}_1^+(\mathcal{S}_V)$ for $j = 1, ..., K$ such that

$$\int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_{j,n}(d\theta_j) \leq \frac{4V\log(nV)}{n}$$

and

$$\mathcal{K}(\rho_{j,n}, \pi_j) \leq n\frac{4V\log(nV)}{n}.$$

We conclude using Theorem 3.1:

$$\mathbb{E}\left[\int D_\alpha(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0})\tilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right]$$

$$\leq \frac{1+\alpha}{1-\alpha}\left[\frac{8KV\log(nV)}{n} \bigvee \frac{8K\log(nK)}{n}\right]. \qquad \square$$

6.5. Proof of Corollary 3.4

Proof. For $R_{j,n} = \frac{1}{n} \vee \frac{1}{n} \left[\frac{1}{2} \log \left(\frac{n}{2} \right) + \frac{V^2}{n\mathcal{V}^2} + \log \left(\frac{\mathcal{V}}{V} \right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \right]$ (for $j = 1, ..., K$), there exists distributions $\rho_{j,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ for $j = 1, ..., K$ such that

$$\int \mathcal{K}(q_{\mu_j^0}, q_{\mu_j}) \rho_{j,n}(d\mu_j) \leq R_{j,n}$$

and

$$\mathcal{K}(\rho_{j,n}, \pi_j) \leq nR_{j,n}.$$

Indeed, let us define $\rho_{j,n}$ as a Gaussian distribution of mean μ_j^0 and variance $\frac{2V^2}{n}$. According to Lemma 6.2:

$$\mathcal{K}(q_{\mu_j^0}, q_{\mu_j}) = \frac{(\mu_j - \mu_j^0)^2}{2V^2}.$$

Then,

$$\int \mathcal{K}(q_{\mu_j^0}, q_{\mu_j}) \rho_{j,n}(d\mu_j) = \frac{1}{2V^2} \mathbb{E}_{\mu_j \sim \rho_{j,n}}[(\mu_j - \mu_j^0)^2] = \frac{1}{2V^2} \times \frac{2V^2}{n} = \frac{1}{n} \leq R_{j,n}.$$

We can apply Lemma 6.2 again to conclude:

$$\begin{aligned}
\mathcal{K}(\rho_{j,n}, \pi_j) &= \frac{1}{2} \log \left(\frac{n\mathcal{V}^2}{2V^2} \right) + \frac{V^2}{n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \\
&= \frac{1}{2} \log \left(\frac{n}{2} \right) + \frac{V^2}{n\mathcal{V}^2} + \log \left(\frac{\mathcal{V}}{V} \right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \\
&= n \times \frac{1}{n} \left[\frac{1}{2} \log \left(\frac{n}{2} \right) + \frac{V^2}{n\mathcal{V}^2} + \log \left(\frac{\mathcal{V}}{V} \right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \right] \\
&\leq nR_{j,n}.
\end{aligned}$$

In addition, Lemma 3.2 tells us that there exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p) \rho_{p,n}(dp) \leq K \frac{4 \log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq nK \frac{4 \log(nK)}{n}.$$

For $r_{n,K} = \frac{4 \log(nK)}{n} \vee_{j=1}^K R_{j,n} = \frac{4 \log(nK)}{n} \vee \frac{1}{n} \vee_{j=1}^K \frac{1}{n} \left[\frac{1}{2} \log \left(\frac{n}{2} \right) + \frac{V^2}{n\mathcal{V}^2} + \log \left(\frac{\mathcal{V}}{V} \right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \right]$ i.e. $r_{n,K} = \frac{4 \log(nK)}{n} \vee_{j=1}^K \frac{1}{n} \left[\frac{1}{2} \log \left(\frac{n}{2} \right) + \frac{V^2}{n\mathcal{V}^2} + \log \left(\frac{\mathcal{V}}{V} \right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \right]$, we finally obtain the required inequality using Theorem 3.1. $\qquad \square$

6.6. Proof of Corollary 3.5

6.6.1. Normal-Inverse-Gamma prior

Proof. First, let us focus on the first result, when the chosen prior is the Normal-Inverse-Gamma $\pi_j = \mathcal{NIG}(0, \mathcal{V}^{-2}, 1, \gamma^2)$ for each $j = 1, ..., K$. In order to obtain the required rate

$$r_{n,K} = \frac{4 \log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n} \left[2 \log(n\sqrt{\mathcal{V}}) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2 \mathcal{V}^2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) \right.$$
$$\left. + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2} \log(2\pi) \right],$$

we proceed as previously and find a variational density on both the mean and the variance such that the two different terms $\int \mathcal{K}(q_{(\mu_j^0, (\sigma_j^0)^2)}, q_{(\mu_j, \sigma_j^2)}) \rho_{j,n}(d\mu_j, d\sigma_j^2)$ and $\mathcal{K}(\rho_{j,n}, \pi_j)$ are upper bounded for $j = 1, ..., K$.

Let us define $\rho_{j,n}$ as a Normal-Inverse-Gamma distribution $\mathcal{NIG}(\mu_j^0, \lambda_n, a_n, b_n)$ where λ_n, a_n and b_n are hyperparameters that we will make precise later. Using Lemma 6.2:

$$\mathcal{K}(q_{(\mu_j^0, (\sigma_j^0)^2)}, q_{(\mu_j, \sigma_j^2)}) = \frac{1}{2} \log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2} \right) + \frac{(\sigma_j^0)^2}{2\sigma_j^2} + \frac{(\mu_j - \mu_j^0)^2}{2\sigma_j^2} - \frac{1}{2}.$$

Then,

$$\int \mathcal{K}(q_{(\mu_j^0, (\sigma_j^0)^2)}, q_{(\mu_j, \sigma_j^2)}) \rho_{j,n}(d\mu_j) = \frac{1}{2} \mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2} \right) \right]$$
$$+ \mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{(\sigma_j^0)^2}{2\sigma_j^2} \right] + \mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{(\mu_j - \mu_j^0)^2}{2\sigma_j^2} \right] - \frac{1}{2}.$$

As

$$\mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{(\sigma_j^0)^2}{2\sigma_j^2} \right] = \frac{(\sigma_j^0)^2}{2} \mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{1}{\sigma_j^2} \right] = \frac{(\sigma_j^0)^2}{2} \frac{a_n}{b_n},$$

$$\frac{1}{2} \mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2} \right) \right] = \frac{1}{2} \left(\log(b_n) - \psi(a_n) \right) - \frac{1}{2} \log((\sigma_j^0)^2)$$

and

$$\mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{(\mu_j - \mu_j^0)^2}{2\sigma_j^2} \right] = \mathbb{E}_{\sigma_j^2 \sim \mathcal{IG}(a_n, b_n)} \left[\frac{1}{2\sigma_j^2} \cdot \mathbb{E}_{\mu_j \sim \mathcal{N}(\mu_j^0, \frac{\sigma_j^2}{\lambda_n})} [(\mu_j - \mu_j^0)^2] \right]$$

i.e.

$$\mathbb{E}_{(\mu_j, \sigma_j^2) \sim \rho_{j,n}} \left[\frac{(\mu_j - \mu_j^0)^2}{2\sigma_j^2} \right] = \mathbb{E}_{\sigma_j^2 \sim \mathcal{IG}(a_n, b_n)} \left[\frac{1}{2\sigma_j^2} \cdot \frac{\sigma_j^2}{\lambda_n} \right] = \frac{1}{2\lambda_n},$$

we get:

$$
\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = -\frac{1}{2} + \frac{(\sigma_j^0)^2}{2}\frac{a_n}{b_n} + \frac{1}{2\lambda_n}
$$
$$
+ \frac{1}{2}\big(\log(b_n) - \psi(a_n)\big) - \frac{1}{2}\log((\sigma_j^0)^2).
$$

Now, we compute the term $\mathcal{K}(\rho_{j,n}, \pi_j)$ using the fomula giving the Kullback-Leibler divergence between two Gaussian-Inverse-Gamma distributions. Using Lemma 6.2:

$$
\mathcal{K}(\rho_{j,n}, \pi_j) = \frac{1}{2}\log\left(\frac{\lambda_n}{\mathcal{V}^{-2}}\right) + \frac{\mathcal{V}^{-2}}{2\lambda_n} + \frac{\mathcal{V}^{-2}(\mu_j^0)^2}{2}\frac{a_n}{b_n} - \frac{1}{2}
$$
$$
+ (a_n - 1)\psi(a_n) + \log\left(\frac{1}{\Gamma(a_n)}\right) + \log\left(\frac{b_n}{\gamma^2}\right) + a_n\frac{\gamma^2 - b_n}{b_n}.
$$

Then, for $\lambda_n = n$, $a_n = n$ and $b_n = n(\sigma_j^0)^2$,

$$
\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2n} + \frac{1}{2}\big(\log(n) - \psi(n)\big)
$$
$$
\leq \frac{1}{2n} + \frac{1}{4n} + \frac{1}{24n^2} \leq R_{j,n}
$$

and

$$
\mathcal{K}(\rho_{j,n}, \pi_j)
$$
$$
= \frac{1}{2}\log(n\mathcal{V}^2) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - n(\sigma_j^0)^2}{(\sigma_j^0)^2}
$$
$$
+ (n-1)\psi(n) + \log(n) - \log\Gamma(n)
$$
$$
\leq \frac{1}{2}\log(n\mathcal{V}^2) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - n
$$
$$
+ n\psi(n) + \log(n) - \log(n-1)!
$$
$$
\leq \frac{1}{2}\log(n\mathcal{V}^2) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - n
$$
$$
+ \left(n\log(n) - \frac{n}{2n} - \frac{n}{12n^2} + \frac{n}{120n^4}\right) + \log(n)
$$
$$
+ \left(-\frac{1}{2}\log(2\pi) + n - 1 - n\log(n-1) + \frac{1}{2}\log(n-1)\right)
$$
$$
\leq \frac{1}{2}\log(n\mathcal{V}^2) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{3}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2}
$$
$$
+ n\log\left(\frac{n}{n-1}\right) - \frac{1}{2} + \log(n) - \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(n-1)
$$

$$\leq \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{3}{2} + \left(\log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \right)$$

$$+ 2 - \frac{1}{2} + \left(\frac{1}{2}\log(n\mathcal{V}^2) + \frac{3}{2}\log(n) \right)$$

$$= \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} + \left(\log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \right)$$

$$+ 2\log(n\sqrt{\mathcal{V}})$$

$$= n \times \frac{1}{n} \left[2\log(n\sqrt{\mathcal{V}}) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) \right.$$

$$\left. + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \right] \leq nR_{j,n}$$

with $R_{j,n} = \frac{1}{n} \bigvee \frac{1}{n} \left[2\log(n\sqrt{\mathcal{V}}) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \right]$
where we used Theorem 5 in [24] and inequality (1.15) in [26]:

$$\forall t > 0, \ \ \log(t) - \frac{1}{2t} - \frac{1}{12t^2} < \psi(t) < \log(t) - \frac{1}{2t} - \frac{1}{12t^2} + \frac{1}{120t^4},$$

$$\forall n \geq 2, \ \ \ n! > \sqrt{2\pi}e^{-n}n^{n+1/2}.$$

Recall again that by Lemma 3.2, there exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K\frac{4\log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq nK\frac{4\log(nK)}{n}.$$

We can finally conclude using again Theorem 3.1 with $r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} R_{j,n}$ i.e.

$$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n} \left[2\log(n\sqrt{\mathcal{V}}) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2} \right) \right.$$

$$\left. + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi) \right]. \qquad \square$$

6.6.2. Factorized prior

Proof. Let us focus now on the case of independant priors $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \bigotimes \mathcal{IG}(1, \gamma^2)$ for $j = 1, ..., K$. The proof is almost the same as previously.

We define here $\rho_{j,n}$ as the product measure of Normal distribution $\mathcal{N}(\mu_j^0, \theta_n^2)$ and of an Inverse-Gamma distribution $\mathcal{IG}(a_n, b_n)$ where θ_n^2, a_n and b_n are hy-

perparameters to be described later. Then, we have again:

$$\mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)}) = \frac{1}{2}\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2\sigma_j^2} + \frac{(\mu_j - \mu_j^0)^2}{2\sigma_j^2} - \frac{1}{2}.$$

Hence,

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2}\mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right)\right]$$

$$+ \mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{(\sigma_j^0)^2}{2\sigma_j^2}\right]$$

$$+ \mathbb{E}_{\mu_j\sim\mathcal{N}(\mu_j^0,\theta_n^2)}\left[(\mu_j - \mu_j^0)^2\right]\mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{1}{2\sigma_j^2}\right] - \frac{1}{2}.$$

i.e.

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j)$$

$$= -\frac{1}{2} + \frac{a_n}{2b_n}\left((\sigma_j^0)^2 + \theta_n^2\right) + \frac{1}{2}\left(\log(b_n) - \psi(a_n)\right) - \frac{1}{2}\log((\sigma_j^0)^2).$$

Then we compute the term $\mathcal{K}(\rho_{j,n}, \pi_j)$ as the sum of the Kullback-Leibler divergence between two Gaussian distributions and between two Inverse-Gamma distributions:

$$\mathcal{K}(\rho_{j,n}, \pi_j) = \frac{1}{2}\log\left(\frac{\mathcal{V}^2}{\theta_n^2}\right) + \frac{\theta_n^2}{2\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}$$

$$+ (a_n - 1)\psi(a_n) + \log\left(\frac{1}{\Gamma(a_n)}\right) + \log\left(\frac{b_n}{\gamma^2}\right) + a_n\frac{\gamma^2 - b_n}{b_n}.$$

Then, for $\theta_n^2 = \frac{(\sigma_j^0)^2}{n}$, $a_n = n$ and $b_n = n(\sigma_j^0)^2$:

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j)$$

$$= \frac{1}{2n} + \frac{1}{2}\left(\log(n) - \psi(n)\right) \le \frac{1}{2n} + \frac{1}{4n} + \frac{1}{24n^2} \le R_{j,n}$$

and

$$\mathcal{K}(\rho_{j,n}, \pi_j)$$

$$= \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - n(\sigma_j^0)^2}{(\sigma_j^0)^2}$$

$$+ (n-1)\psi(n) + \log(n) - \log\Gamma(n)$$

$$\le \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - n$$

$$+ n\psi(n) + \log(n) - \log(n-1)!$$

$$\leq \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - n$$

$$+ \left(n\log(n) - \frac{n}{2n} - \frac{n}{12n^2} + \frac{n}{120n^4}\right) + \log(n)$$

$$+ \left(-\frac{1}{2}\log(2\pi) + n - 1 - n\log(n-1) + \frac{1}{2}\log(n-1)\right)$$

$$= \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{3}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2}{(\sigma_j^0)^2}$$

$$+ n\log(\frac{n}{n-1}) - \frac{1}{2} + \log(n) - \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(n-1)$$

$$\leq \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{3}{2} + \left(\frac{1}{2}\log\left(\frac{(\sigma_j^0)^2}{\gamma^4}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi)\right)$$

$$+ 2 - \frac{1}{2} + \left(\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{3}{2}\log(n)\right)$$

$$= \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} + \left(\frac{1}{2}\log\left(\frac{(\sigma_j^0)^2}{\gamma^4}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi)\right) + 2\log(n\sqrt{\mathcal{V}})$$

$$= n \times \frac{1}{n}\left[2\log(n\sqrt{\mathcal{V}}) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} + \frac{1}{2}\log\left(\frac{(\sigma_j^0)^2}{\gamma^4}\right) + \frac{\gamma^2}{(\sigma_j^0)^2}\right.$$

$$\left. - \frac{1}{2}\log(2\pi)\right] \leq nR_{j,n}$$

with $R_{j,n} = \frac{1}{n} \vee \frac{1}{n}\left[2\log(n\sqrt{\mathcal{V}}) + \frac{(\sigma_j^0)^2}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} + \frac{1}{2}\log\left(\frac{(\sigma_j^0)^2}{\gamma^4}\right) + \frac{\gamma^2}{(\sigma_j^0)^2} - \frac{1}{2}\log(2\pi)\right]$.

The end of the proof is the same as the one used in the Normal-Inverse-Gamma case. $\square$

6.7. Proof of Theorem 3.6

Proof. We assume that $\Theta_K(r_{n,K})$ is not empty (otherwise, this is obvious). Applying Theorem 2.7 in [3] for any $\alpha \in (0,1)$, $\theta^* \in \Theta_K(r_{n,K})$:

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

$$+ \inf_{\rho \in \mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathbb{E}\left[\log\frac{P_{\theta^*}(X)}{P_\theta(X)}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j, \pi_j)}{n(1-\alpha)}\right\}.$$

Let us take $\rho_{j,n}$ and $\mathcal{A}_{n,K}$ from the definition of $\Theta_K(r_{n,K})$, and $\rho_{p,n}(dp) \propto \mathbf{1}(p \in \mathcal{A}_{n,K})\pi_p(dp)$:

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

$$+ \frac{\alpha}{1 - \alpha} \int \mathbb{E}\left[\log \frac{P_{\theta^*}(X)}{P_\theta(X)}\right] \rho_{p,n}(dp) \prod_{j=1}^{K} \rho_{j,n}(d\theta_j)$$

$$+ \frac{\mathcal{K}(\rho_{p,n}, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_{j,n}, \pi_j)}{n(1 - \alpha)}.$$

We have $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{A}_{n,K})) \leq nKr_{n,K}$ and $\mathcal{K}(\rho_{j,n}, \pi_j) \leq nr_{n,K}$ for each j by definition of $\Theta_K(r_{n,K})$. Moreover, using the same argument contained in the proof of Lemma 6.1:

$$\log \frac{P_{\theta^*}(X)}{P_\theta(X)} = \frac{1}{P_{\theta^*}(X)} P_{\theta^*}(X) \log \frac{P_{\theta^*}(X)}{P_\theta(X)}$$

$$\leq \frac{1}{P_{\theta^*}(X)} \sum_{j=1}^{K} p_j^* q_{\theta_j^*}(X) \log \frac{p_j^* q_{\theta_j^*}(X)}{p_j q_{\theta_j}(X)}$$

$$= \sum_{j=1}^{K} \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{p_j^*}{p_j} + \sum_{j=1}^{K} \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}$$

$$\leq \sum_{j=1}^{K} \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{p_j^*}{p_j} + \sum_{j=1}^{K} \log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}$$

and thus, as the support of $\rho_{p,n}$ is on $\mathcal{A}_{n,K}$ where $\log \frac{p_j^*}{p_j} \leq Kr_{n,K}$,

$$\int \mathbb{E}\left[\log \frac{P_{\theta^*}(X)}{P_\theta(X)}\right] \rho_{p,n}(dp) \prod_{j=1}^{K} \rho_{j,n}(d\theta_j) \leq \int \mathbb{E}\left[\sum_{j=1}^{K} \frac{p_j^* q_{\theta_j^*}}{P_{\theta^*}}\right] \log \frac{p_j^*}{p_j} \rho_{p,n}(dp)$$

$$+ \sum_{j=1}^{K} \int \mathbb{E}\left[\log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right] \rho_{j,n}(d\theta_j)$$

$$\leq \int \mathbb{E}\left[\sum_{j=1}^{K} \frac{p_j^* q_{\theta_j^*}}{P_{\theta^*}}\right] Kr_{n,K} \rho_{p,n}(dp)$$

$$+ Kr_{n,K} = 2Kr_{n,K}$$

which ends the proof as it holds for any $\theta^* \in \Theta_K(r_{n,K})$. $\qquad \square$

6.8. Proof of Corollary 3.7

Proof. It is sufficient to show that $\mathcal{S}_K \times [-L, L]^K \subset \Theta_K(r_{n,K})$ for

$$r_{n,K} = \frac{4 \log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n} \left[\frac{1}{2} \log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log(\mathcal{V}) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2}\right],$$

the stated oracle inequality is a direct corollary of Theorem 3.6. For that, let us take any $\theta^* \in \mathcal{S}_K \times [-L, L]^K$ and show that it satisfies the conditions in the definition of $\Theta_K(r_{n,K})$.

The existence of a set $\mathcal{A}_{n,K}$ fulfilling the first condition has already been done in the proof of Lemma 3.2 as $\frac{4 \log(nK)}{n} \leq r_{n,K}$.

We define distributions $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ by Gaussians of mean θ_j^* and variance $\frac{2}{n}$ ($j = 1, ..., K$) and we show that for $j = 1, ..., K$:

$$\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right]\rho_{j,n}(d\theta_j) \leq r_{n,K} \ , \qquad \mathcal{K}(\rho_{j,n}, \pi_j) \leq nr_{n,K}.$$

We start from

$$\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right) = \frac{(\theta_j - \theta_j^*)^2}{2} - (X - \theta_j^*)(\theta_j - \theta_j^*)$$

and if we take the mean of this quantity with respect to P^0, we obtain:

$$\mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right] = \frac{(\theta_j - \theta_j^*)^2}{2} - (\mathbb{E}X - \theta_j^*)(\theta_j - \theta_j^*)$$

and as $\theta_j - \theta_j^*$ is a zero-mean random variable, we have:

$$\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right]\rho_{j,n}(d\theta_j)$$

$$= \frac{1}{2}\mathbb{E}_{\theta_j \sim \rho_{j,n}}[(\theta_j - \theta_j^*)^2] - (\mathbb{E}X - \theta_j^*)\mathbb{E}_{\theta_j \sim \rho_{j,n}}[\theta_j - \theta_j^*]$$

$$= \frac{1}{2} \times \frac{2}{n} \leq r_{n,K}.$$

Then, we conclude according to Lemma 6.2:

$$\mathcal{K}(\rho_{j,n}, \pi_j) = \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{2}\right) + \frac{1}{n\mathcal{V}^2} + \frac{(\theta_j^*)^2}{2\mathcal{V}^2} - \frac{1}{2}$$

$$= \frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log(\mathcal{V}) + \frac{(\theta_j^*)^2}{2\mathcal{V}^2} - \frac{1}{2}$$

$$\leq n \times \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log(\mathcal{V}) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$$

$$\leq nr_{n,K}. \qquad \qquad \square$$

6.9. Proof of Theorem 4.1

Here, we cannot directly use the results from [3]. So we prove this theorem from scratch, by following the main steps outlined in [8, 3] with some adaptation.

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

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

Thus, integrating and using Fubini's theorem,

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

Using Lemma 2.1,

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

Note that [8, 3] also used Lemma 2.1 in their proofs, this is inspired by the PAC-Bayesian theory [15, 16]. It is interesting to note that Lemma 2.1 is at the core of VB: it is used to provide approximation algorithms, and also to prove the consistency of VB. Thanks to Jensen's inequality,

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

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

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

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

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

Thus, by definition of $\hat{K}$,

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

which leads to

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

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

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

$$\leq \inf_{K\geq 1}\left\{\mathbb{E}\left[\inf_{\rho\in\mathcal{M}_1^+(\Theta_K)}\left\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta, P^0)}{n}\rho(d\theta)+\frac{\mathcal{K}(\rho,\Pi_K)}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right]\right\}.$$

Then,

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

$$\leq \inf_{K\geq 1}\inf_{\rho\in\mathcal{M}_1^+(\Theta_K)}\left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta, P^0)}{n}\rho(d\theta)+\frac{\mathcal{K}(\rho,\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right]\right\}.$$

And finally,

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

$$\leq \inf_{K\geq 1}\inf_{\rho\in\mathcal{M}_1^+(\Theta_K)}\left\{\frac{\alpha}{1-\alpha}\int\mathcal{K}(P^0, P_\theta)\rho(d\theta)+\frac{\mathcal{K}(\rho,\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

To conclude, we just need to upper bound the function inside the infimum over all integers K's by $\frac{\alpha}{1-\alpha}\inf_{\theta^*\in\Theta_K(r_{n,K})}\mathcal{K}(P^0, P_{\theta^*})+\frac{1+\alpha}{1-\alpha}2Kr_{n,K}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}$. This is direct: if the set $\Theta_K(r_{n,K})$ is not empty (otherwise the inequality is obvious) we notice that $\mathcal{K}(P^0, P_\theta)=\mathcal{K}(P^0, P_{\theta^*})+\mathbb{E}\left[\log\frac{P_{\theta^*}(X_i)}{P_\theta(X_i)}\right]$ for any $\theta^*\in\Theta_K(r_{n,K})$ and then we follow the sketch of the proof of Theorem 3.6. $\quad\square$

6.10. Algorithms

We now provide the derivations leading to the algorithms described in the paper.

6.10.1. Algorithm 1

We apply a coordinate descent on variables $\omega^1\in\mathcal{S}_K,...,\omega^n\in\mathcal{S}_K$, $\rho_p\in\mathcal{M}_1^+(\mathcal{S}_K)$, $\rho_1\in\mathcal{M}_1^+(\Theta),...,$ and $\rho_K\in\mathcal{M}_1^+(\Theta)$ in order to solve the optimization program:

$$\min_{\rho\in\mathcal{F},\,w\in\mathcal{S}_K^n}\left\{-\alpha\sum_{i=1}^n\sum_{j=1}^K\omega_j^i\left(\int\log(p_j)\rho_p(dp)+\int\log(q_{\theta_j}(X_i))\rho_j(d\theta_j)\right)\right.$$

$$\left.+\alpha\sum_{i=1}^n\sum_{j=1}^K\omega_j^i\log(\omega_j^i)+\mathcal{K}(\rho_p,\pi_p)+\sum_{j=1}^K\mathcal{K}(\rho_j,\pi_j)\right\}.$$

We explain how to obtain Algorithm 1.

Optimization with respect to $\omega^i \in \mathcal{S}_K$:

First, we fix $\omega^\ell \in \mathcal{S}_K$ for $\ell \neq i$, $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$, and we solve the program with respect to $\omega^i \in \mathcal{S}_K$, which becomes:

$$\min_{\omega^i \in \mathcal{S}_K} \left\{ \sum_{j=1}^K \omega_j^i \left(\log(\omega_j^i) - \int \log(p_j) \rho_p(dp) - \int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) \right) \right\}.$$

Put $\mathbf{E} = \{1, ..., K\}$, $\lambda = \left(\frac{1}{K}, ..., \frac{1}{K} \right)$ and $h(j) = \int \log(p_j) \rho_p(dp) + \int \log(q_{\theta_j}(X_i)) \times \rho_j(d\theta_j)$ and use Lemma 2.1 to obtain:

$$w_j^i \propto \exp\left(\int \log(p_j) \rho_p(dp) + \int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) \right).$$

Optimization with respect to $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$:

Now, we fix $\omega^i \in \mathcal{S}_K$ for $i = 1, ..., n$, and $\rho_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$, and we solve the program with respect to $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$, which becomes:

$$\min_{\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)} \left\{ -\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \int \log(p_j) \rho_p(dp) + \mathcal{K}(\rho_p, \pi_p) \right\}.$$

Using Lemma 2.1 for $\mathbf{E} = \mathcal{S}_K$, $\lambda = \pi_p$ and $h(p) = \alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(p_j)$, we get directly the solution:

$$\rho_p(dp) \propto \exp\left(\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(p_j) \right) \pi_p(dp).$$

Optimization with respect to $\rho_j \in \mathcal{M}_1^+(\Theta)$:

Now, we fix $\omega^i \in \mathcal{S}_K$ for $i = 1, ..., n$, $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_\ell \in \mathcal{M}_1^+(\Theta)$ for $\ell \neq j$, and we solve the program with respect to $\rho_j \in \mathcal{M}_1^+(\Theta)$, which becomes:

$$\min_{\rho_j \in \mathcal{M}_1^+(\Theta)} \left\{ -\alpha \sum_{i=1}^n \omega_j^i \int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) + \mathcal{K}(\rho_j, \pi_j) \right\}.$$

Using Lemma 2.1 for $\mathbf{E} = \Theta$, $\lambda = \pi_j$ and $h(\theta_j) = \alpha \sum_{i=1}^n \omega_j^i \log(q_{\theta_j}(X_i))$, we get directly the solution:

$$\rho_j(d\theta_j) \propto \exp\left(\alpha \sum_{i=1}^n \omega_j^i \log(q_{\theta_j}(X_i)) \right) \pi_j(d\theta_j)$$

6.10.2. Application to multinomial mixture models

We simply use

$$\int \log(p_j)\rho_p(dp) = \mathbb{E}_{p \sim \rho_p}[\log(p_j)] = \psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell),$$

$$\int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) = \mathbb{E}_{\theta_j \sim \rho_j}[\log(\theta_{X_i,j})] = \psi(\gamma_{X_i,j}) - \psi\left(\sum_{v=1}^{V} \gamma_{vj}\right),$$

$$\exp\left(\alpha \sum_{i=1}^{n}\sum_{j=1}^{K} \omega_j^i \log(p_j)\right)\pi_p(p) \propto \prod_{j=1}^{K} p_j^{\alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i - 1},$$

$$\exp\left(\alpha \sum_{i=1}^{n} \omega_j^i \log(q_{\theta_j}(X_i))\right)\pi_j(\theta_j) \propto \prod_{v=1}^{V} \theta_{vj}^{\beta_v + \alpha \sum_{i=1}^{n} \omega_j^i \mathbb{1}(X_i = v) - 1}.$$

We recognize a Dirichlet distribution.

6.10.3. Application to Gaussian mixture models

For Gaussian mixtures, use

$$\int \log(p_j)\rho_p(dp) = \mathbb{E}_{p \sim \rho_p}[\log(p_j)] = \psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell),$$

$$\int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) = -\frac{1}{2}\mathbb{E}_{\theta_j \sim \rho_j}[(\theta_j - X_i)^2] + \text{cst}$$

$$= -\frac{1}{2}\{s_j^2 + (n_j - X_i)^2\} + \text{cst},$$

$$\exp\left(\alpha \sum_{i=1}^{n}\sum_{j=1}^{K} \omega_j^i \log(p_j)\right)\pi_p(p) \propto \prod_{j=1}^{K} p_j^{\alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i - 1},$$

$$\exp\left(\alpha \sum_{i=1}^{n} \omega_j^i \log(q_{\theta_j}(X_i))\right)\pi_j(\theta_j)$$

$$\propto \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{n} \omega_j^i(\theta_j - X_i)^2\right)\exp\left(-\frac{1}{2\mathcal{V}^2}\theta_j^2\right)$$

$$\propto \exp\left(-\frac{1/\mathcal{V}^2 + \alpha \sum_{i=1}^{n} \omega_j^i}{2}\left(\theta_j - \frac{\alpha \sum_{i=1}^{n} \omega_j^i X_i}{1/\mathcal{V}^2 + \alpha \sum_{i=1}^{n} \omega_j^i}\right)^2\right).$$

We recognize a Gaussian distribution.

Acknowledgments

Supplementary Material

Supplement to "Consistency of variational Bayes inference for estimation and model selection in mixtures"

(doi: [10.1214/18-EJS1475SUPP](#); .zip). The supplementary material zip contains the description of a short simulation study (supplement.pdf) and the notebook used for the simulation study (supplement.ipynb).

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974. [MR0423716](#)

[2] P. Alquier and V. Cottet. 1-bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Machine Learning*, 107(3):579–603, 2018. [MR3761297](#)

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

[4] P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239):1–41, 2016. [MR3595173](#)

[5] S. Ayer and H.S. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. *International Conference on Computer Vision*, 1995.

[6] A.G. Bacharoglou. Approximation of probability distributions by convex mixtures of Gaussian measures. *Proceedings of the American of the American Mathematical Society*, 138(7):2619–2628, 2010. [MR2607892](#)

[7] G. Behrens, N. Friel, and M. Hurn. Tuning tempered transitions. *Statistics and computing*, 22(1):65–78, 2012. [MR2865056](#)

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

[9] C. Biernacki, G. Celeux, and G. Govaert. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272, 1999.

[10] D. M. Blei, A.Y. Ng, C. Wang, and M.I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: a review for statisticians. *arXiv preprint arXiv:1601.00670*, 2017. [MR3671776](#)

[12] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71:52–78, 2014. [MR3131954](#)

[13] P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108, 2012. [MR2896713](#)

[14] L. Carel and P. Alquier. Simultaneous dimension reduction and clustering via the NMF-EM algorithm. *arXiv preprint arXiv:1709.03346*, 2017.

[15] O. Catoni. *Statistical learning theory and stochastic optimization.* Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004. MR2163920

[16] 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. MR2483528

[17] B. E. Chérief-Abdellatif and P. Alquier. *Supplement to "Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures",* DOI: 10.1214/18-EJS1475SUPP, 2018.

[18] G. Celeux, S. Frühwirth-Schnatter, and C. P. (Editors) Robert. *Handbook of mixture analysis.* CRC Press, 2018.

[19] P. Deb, W.T. Gallo, P. Ayyagari, J.M. Fletcher, and J.L. Sindelar. The effect of job loss on overweight and drinking. *Journal of Health Economics*, 2011.

[20] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. MR0501537

[21] M. N. Do. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 10(4):115–118, 4 2003.

[22] E. Gassiat, J. Rousseau, and E. Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electronic Journal of Statistics*, 12(1):703–740, 2018. MR3769193

[23] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000. MR1790007

[24] L. Gordon. A stochastic approach to the gamma function. *The American Mathematical Monthly*, 101(9):858–865, 1994. MR1300491

[25] 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. MR3724979

[26] S. Guo. Monotonicity and concavity properties of some functions involving the gamma function with applications. *JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only]*, 7, 01 2006. MR2221326

[27] J.R. Hershey and P.A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4, 2007.

[28] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013. MR3081926

[29] W. Kruijer, J. Rousseau, and A. Van Der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010. MR2735885

[30] A. Laforgia and P. Natalin. On some inequalities for the gamma function. *Advances in Dynamical Systems and Applications*, 8(2):261–267, 2013. MR3162146

[31] P. Massart. *Concentration inequalities and model selection.* Saint-Flour Summer School on Probability Theory 2003 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2007. MR2319879

[32] P. D. McNicholas. Model-based clustering. *Journal of Classification*, 33(3):331–373, 2016. MR3575621

[33] N. Nasios and A.G. Bors. Variational learning for Gaussian mixture models. In *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, volume 36, pages 849–862, 2006.

[34] R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366, 1996.

[35] A. O'Hagan, T. B. Murphy, and I. C. Gormley. Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics & Data Analysis*, 56(12):3843–3864, 2012. MR2957835

[36] W. Pan, J. Lin, and C.T. Le. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, 3:117–124, 2003.

[37] L. Rigouste, O. Cappé, and F. Yvon. Inference and evaluation of the multinomial mixture model for text clustering. In *Information Processing & Management*, volume 43, pages 1260–1280, 2007.

[38] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. MR0468014

[39] Y. Singer and M. K. Warmuth. Batch and online parameter estimation of Gaussian mixtures based on the joint entropy. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA, 1999.

[40] C. J. Stoneking. Bayesian inference of Gaussian mixture models with non-informative priors. *arXiv preprint arXiv:1405.4895*, 2014.

[41] E.B. Sudderth and M.I. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. *In Advances in Neural Information Processing Systems*, pages 1585–1592, 2009.

[42] N. Syring and R. Martin. Scaling the Gibbs posterior credible regions. *Preprint*, 2015. MR3828667

[43] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. MR3225930

[44] Y. Wang and D.M. Blei. Frequentist consistency of variational Bayes. *arXiv preprint arXiv:1705.034339v1, accepted for publication in JASA*, 2017.

[45] L. Watier, S. Richardson, and P. J. Green. Using Gaussian mixtures with unknown number of components for mixed model estimation. In *14th International Workshop on Statistical Modelling, Graz, Austria*, 1999.

[46] Y. Wu and P. Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *working paper*, 2018.

[47] 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. MR2234196

[48] Y. Yang, Pati D., and A. Bhattacharya. α-variational inference with statistical guarantees. *preprint arXiv:1710.03266v1*, 2017.

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