



**HAL**  
open science

## Bayesian mixtures of multiple scale distributions

Alexis Arnaud, Florence Forbes, Russel Steele, Benjamin Lemasson,  
Emmanuel L. Barbier

► **To cite this version:**

Alexis Arnaud, Florence Forbes, Russel Steele, Benjamin Lemasson, Emmanuel L. Barbier. Bayesian mixtures of multiple scale distributions. 2018. hal-01953393v1

**HAL Id: hal-01953393**

**<https://inria.hal.science/hal-01953393v1>**

Preprint submitted on 12 Dec 2018 (v1), last revised 27 Sep 2019 (v3)

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

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

# Bayesian mixtures of multiple scale distributions

Alexis Arnaud      Florence Forbes      Russel Steele      Benjamin Lemasson  
Emmanuel Barbier

1

## Abstract

Multiple scale distributions are multivariate distributions that exhibit a variety of shapes not necessarily elliptical while remaining analytical and tractable. In this work we consider mixtures of such distributions for their ability to handle non standard typically non-gaussian clustering tasks. We propose a Bayesian formulation of the mixtures and a tractable inference procedure based on variational approximation. The interest of such a Bayesian formulation is illustrated on an important mixture model selection task, which is the issue of selecting automatically the number of components. We derive promising procedures that can be carried out in a single-run, in contrast to the more costly comparison of information criteria. Preliminary results on simulated and real data show promising performance in terms of clustering and computation time.

Keywords: Covariance matrix decomposition EM algorithm Gaussian scale mixture Bayesian analysis Bayesian model selection Variational approximation

## 1 Introduction

Multiple scale distributions refer to a recent generalization of scale mixtures of Gaussians in a multivariate setting [Forbes and Wraith, 2014]. This new family of distributions has the ability to generate a number of flexible distributional forms with closed-form densities and interesting properties. It nests in particular several symmetric multiple scale heavy tailed distributions (such as generalized multivariate Student distributions [Forbes and Wraith, 2014]) and asymmetric multiple scale generalized hyperbolic distributions [Wraith and Forbes, 2015]. The multiple scale framework has also been used by [Franczak et al., 2015] for multiple scale shifted asymmetric Laplace distributions. The multiple scale framework has the advantage of allowing different tail and skewness behaviours in each dimension of the variable space with arbitrary correlation between dimensions. This is interesting when targeting clustering applications using mixtures of such distributions (see [Forbes and Wraith, 2014, Wraith and Forbes, 2015] for illustration). In this work we consider mixtures of multiple scale distributions in a Bayesian formulation for the many advantages that the Bayesian framework offers in the mixture models context, including natural procedures to automatically select the number of mixture components.

A standard  $M$ -variate scale mixture of Gaussians is a distribution of the form:

$$p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/w) f_W(w; \boldsymbol{\theta}) dw \quad (1)$$

where  $\mathcal{N}_M(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}/w)$  denotes the  $M$ -dimensional Gaussian distribution with mean  $\boldsymbol{\mu}$ , covariance  $\boldsymbol{\Sigma}/w$  and  $f_W$  is the probability distribution of a univariate positive variable  $W$  referred to hereafter as the weight variable. A common form is obtained when  $f_W$  is a Gamma distribution  $\mathcal{G}(\nu/2, \nu/2)$  where  $\nu$  denotes the degrees of freedom (we shall denote the Gamma distribution when the variable is  $X$  by  $\mathcal{G}(x; \alpha, \gamma) =$

---

<sup>1</sup>F. Forbes, A. Arnaud are at Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP\*, LJK, 38000 Grenoble, France, \* Institute of Engineering Univ. Grenoble Alpes, email: firstname.lastname@inria.fr and R. Steele is at McGill, Montreal, Canada, email: steele@math.mcgill.ca and B. Lemasson, E. Barbier are at Grenoble Institut des Neurosciences, Inserm U1216, Univ. Grenoble Alpes, France, email: firstname.lastname@univ-grenoble-alpes.fr

$x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma x)\gamma^\alpha$  where  $\Gamma$  denotes the Gamma function). For this form, (1) is the standard density denoted by  $t_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$  of the  $M$ -dimensional Student  $t$ -distribution with parameters  $\boldsymbol{\mu}$  (real location vector),  $\boldsymbol{\Sigma}$  ( $M \times M$  real positive definite scale matrix) and  $\nu$  (positive real degrees of freedom parameter). Most of the work on multivariate scale mixture of Gaussians has focused on studying different choices for the weight distribution  $f_W$  (see *e.g.* Eltoft et al. [2006]) but the dimension of the weight variable  $W$  in most cases has been considered as univariate.

The extension proposed by Forbes and Wraith [2014] consists of introducing a multidimensional weight. To do so, the scale matrix is decomposed into  $\boldsymbol{\Sigma} = \mathbf{D}\mathbf{A}\mathbf{D}^T$ , where  $\mathbf{D}$  is the matrix of eigenvectors of  $\boldsymbol{\Sigma}$  and  $\mathbf{A}$  is a diagonal matrix with the corresponding eigenvalues of  $\boldsymbol{\Sigma}$ . This spectral decomposition is classically used in Gaussian model-based clustering [Banfield and Raftery, 1993, Celeux and Govaert, 1995]. The matrix  $\mathbf{D}$  determines the orientation of the Gaussian and  $\mathbf{A}$  its shape. Using this parameterization of  $\boldsymbol{\Sigma}$ , the scale Gaussian part in (1) is set to  $\mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Delta}_w\mathbf{A}\mathbf{D}^T)$ , where  $\boldsymbol{\Delta}_w = \text{diag}(w_1^{-1}, \dots, w_M^{-1})$  is the  $M \times M$  diagonal matrix whose diagonal components are the inverse weights  $\{w_1^{-1}, \dots, w_M^{-1}\}$ . When the weights are all one, a standard multivariate Gaussian case is recovered. The multiple scale generalization consists therefore of:

$$p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \int_0^\infty \dots \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Delta}_w\mathbf{A}\mathbf{D}^T) f_w(w_1 \dots w_M; \boldsymbol{\theta}) dw_1 \dots dw_M \quad (2)$$

where  $f_w$  is now a  $M$ -variate density function to be further specified. In the following developments, we will consider only independent weights, *i.e.*  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M\}$  with  $f_w(w_1 \dots w_M; \boldsymbol{\theta}) = f_{W_1}(w_1; \boldsymbol{\theta}_1) \dots f_{W_M}(w_M; \boldsymbol{\theta}_M)$ .

As an example, setting  $f_{W_m}(w_m; \boldsymbol{\theta}_m)$  to a Gamma distribution  $\mathcal{G}(w_m; \alpha_m, \gamma_m)$  results in a multivariate generalization of a Pearson type VII distribution (see *e.g.* Johnson et al. [1994] vol.2 chap. 28 for a definition of the Pearson type VII distribution) while setting  $f_{W_m}(w_m)$  to  $\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$  leads to a generalization of the multivariate  $t$ -distribution. In both cases, we can express easily the respective densities denoted by  $\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$  and  $\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$  with  $\boldsymbol{\nu} = \{\nu_1, \dots, \nu_M\}$ ,  $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_M\}$  and  $\boldsymbol{\gamma} = \{\gamma_1 \dots \gamma_M\}$ :

$$\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \prod_{m=1}^M \frac{\Gamma(\alpha_m + 1/2)}{\Gamma(\alpha_m)(2A_m\gamma_m\pi)^{1/2}} \left(1 + \frac{[\mathbf{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2}{2A_m\gamma_m}\right)^{-(\alpha_m + 1/2)} \quad (3)$$

Similarly,

$$\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \prod_{m=1}^M \frac{\Gamma((\nu_m + 1)/2)}{\Gamma(\nu_m/2)(A_m\nu_m\pi)^{1/2}} \left(1 + \frac{[\mathbf{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2}{A_m\nu_m}\right)^{-(\nu_m + 1)/2} \quad (4)$$

However for identifiability, model (3) needs to be further specified by fixing all  $\gamma_m$  parameters, for instance to 1. Despite this additional constraint, the decomposition of  $\boldsymbol{\Sigma}$  still induces another identifiability issue. Both (3) and (4) are invariant to a same permutation of the column of  $\mathbf{D}$ ,  $\mathbf{A}$  and elements of  $\boldsymbol{\alpha}$  or  $\boldsymbol{\nu}$ . In a frequentist setting this can be solved by imposing a decreasing order for the eigenvalues in  $\mathbf{A}$ . In a Bayesian setting one way to solve the problem is to impose on  $\mathbf{A}$  a non symmetric prior (see Section 2.2) since an appropriate prior on  $\mathbf{D}$  would be more difficult to set.

As shown in [Forbes and Wraith, 2014, Wraith and Forbes, 2015], this generalization to a multiple scale representation allows for a greater variety of shapes and in particular contours that are not necessarily elliptical. It is possible to account for very different tail behaviors across dimensions, such as a Gaussian (infinite  $\nu_m$ ) tail in one dimension and a Cauchy ( $\nu_m = 1$ ) tail in an other dimension.

In previous work [Forbes and Wraith, 2014, Wraith and Forbes, 2015], inference has been carried out based on maximum likelihood principle and using the EM algorithm. In this work, we consider a Bayesian treatment of these models for the many advantages that the Bayesian framework offers in the mixture model context. Mainly, it avoids the ill-posed nature of maximum likelihood due to the presence of singularities in the likelihood function. A mixture component may collapse by becoming centered at a single data vector sending its covariance to 0 and the model likelihood to infinity. A Bayesian treatment protects the algorithm from this problem occurring in ordinary EM. Also, Bayesian model comparison embodies the principle that states that simple models should be preferred. Typically, maximum likelihood does not provide any guidance

on the choice of the model order as more complex models can always fit the data better. For standard scale mixture of Gaussians, the usual Normal-Wishart prior can be used for the Gaussian parameters. In contrast, for multiple scale distributions, the decomposition of the covariance in (2) requires appropriate separated priors on the eigenvectors and eigenvalues of the scale matrix. Such priors do not derive easily from a standard conjugate choice. We propose a simple solution for which we can derive an inference scheme based on variational approximation and show how to apply it to select automatically an appropriate number of components in the mixtures. Following common practice that is to start from deliberately overfitting mixtures (*e.g.* Malsiner-Walli et al. [2016], Corduneanu and Bishop [2001], McGrory and Titterton [2007], Attias [1999]), we investigate the component-elimination property of the Bayesian setting to select this number from a single run. We consider three different single-run strategies that avoid the repetitive inference and comparison of all possible models. They are all based on a Bayesian formulation specified in Section 2 and on a variational approximation inference detailed in Section 3. The three strategies are described in Section 4 and compared on simulated and real data in Section 5.

## 2 Bayesian mixture of multiple scale distributions

In this section, we outline the Bayesian model for a mixture of multiple scale Pearson VII distributions. In a Bayesian setting, it is more convenient to use the precision matrix  $\mathbf{T}$  decomposed into  $\mathbf{T} = \mathbf{D}\mathbf{A}\mathbf{D}^T$ , which is the inverse of the covariance matrix  $\Sigma = \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T$ , in the parameterization. Note that in (3) and in previous work,  $\mathbf{A}$  is then replaced by  $\mathbf{A}^{-1}$ . Moreover, we consider for identifiability that all  $\gamma_m$  are set to 1. In a  $K$ -component mixture, the distributions we consider are therefore of the form, for each  $k = 1 \dots K$ ,

$$\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}_k, \mathbf{D}_k, \mathbf{A}_k, \boldsymbol{\alpha}_k) = \prod_{m=1}^M \frac{\Gamma(\alpha_{km} + 1/2)A_m}{\Gamma(\alpha_{km})(2\pi)^{1/2}} \left( 1 + \frac{A_m[\mathbf{D}_k^T(\mathbf{y} - \boldsymbol{\mu}_k)]_m^2}{2} \right)^{-(\alpha_{km} + 1/2)} \quad (5)$$

### 2.1 A hierarchical Bayesian model

Let us consider an *i.i.d* sample  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$  from a  $K$ -component mixture of multiple scale distributions as defined in (5). With the usual notation for the mixing proportions  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$  and  $\boldsymbol{\psi}_k = \{\boldsymbol{\mu}_k, \mathbf{T}_k, \boldsymbol{\alpha}_k\}$  for  $k = 1 \dots K$ , we consider,

$$p(\mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^K \pi_k \mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}_k, \mathbf{T}_k, \boldsymbol{\alpha}_k)$$

where  $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$  with  $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_K\}$  denotes the mixture parameters. An additional variable  $Z$  can be introduced to identify the class labels:  $\{Z_1, \dots, Z_N\}$  define respectively the components of origin of  $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ . An equivalent modelling is therefore:

$$\begin{aligned} \forall i \in \{1 \dots N\}, \quad \mathbf{Y}_i | \mathbf{W}_i = \mathbf{w}_i, Z_i = k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{D}_k \boldsymbol{\Delta}_{\mathbf{w}_i} \mathbf{A}_k^{-1} \mathbf{D}_k^T), \\ \mathbf{W}_i | Z_i = k &\sim \mathcal{G}(\alpha_{k1}, 1) \otimes \dots \otimes \mathcal{G}(\alpha_{kM}, 1), \\ \text{and } Z_i &\sim \mathcal{M}(1, \pi_1, \dots, \pi_k), \end{aligned}$$

where  $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}^{-1}, \dots, w_{iM}^{-1})$ , symbol  $\otimes$  means that the components of  $\mathbf{W}_i$  are independent and  $\mathcal{M}(1, \pi_1, \dots, \pi_k)$  denotes the Multinomial distribution.

### 2.2 Priors on component-specific parameters

To complete the Bayesian formulation, we assign priors on the parameters. However, it is common (see *e.g.* Archambeau and Verleysen [2007]) not to impose priors on the parameters  $\boldsymbol{\alpha}_k$  since no convenient conjugate



prior exist for these parameters. Then the scale matrix decomposition imposes that we set priors on  $\boldsymbol{\mu}_k$  and  $\mathbf{D}_k, \mathbf{A}_k$ . For the means  $\boldsymbol{\mu}_k$ , the standard Gaussian prior can be used:

$$\boldsymbol{\mu}_k \mid \mathbf{A}_k, \mathbf{D}_k \sim \mathcal{N}(\mathbf{m}_k, \mathbf{D}_k \boldsymbol{\Lambda}_k^{-1} \mathbf{A}_k^{-1} \mathbf{D}_k^T), \quad (6)$$

where  $\mathbf{m}_k$  (vector) and  $\boldsymbol{\Lambda}_k$  (diagonal matrix) are hyperparameters. For  $\mathbf{A}_k$  and  $\mathbf{D}_k$  a natural solution would be to use the distributions induced by the standard Wishart prior on  $\mathbf{T}_k$  but this appears not to be tractable in inference scheme even in a variational framework. The difficulty lies in considering an appropriate and tractable prior for  $\mathbf{D}_k$ . There exists a number of priors on the Stiefel manifold among which a good candidate could be the Bingham prior and extensions investigated by Hoff [2009]. However, it is not straightforward to derive from it a tractable E- $\Phi^1$  step (see below) that could provide a variational posterior distribution. Nevertheless, this kind of priors could be added in the M- $\mathbf{D}$ -step. The simpler solution adopted in the present work consists of considering  $\mathbf{D}_k$  as an unknown fixed parameter and imposing a prior only on  $\mathbf{A}_k$ , which is a diagonal matrix containing the positive eigenvalues of  $\mathbf{T}_k$ . It is natural to choose:

$$\mathbf{A}_k \sim \otimes_{m=1}^M \mathcal{G}(\lambda_{km}, \delta_{km}), \quad (7)$$

where  $\boldsymbol{\lambda}_k = \{\lambda_{km}, m = 1 \dots M\}$  and  $\boldsymbol{\delta}_k = \{\delta_{km}, m = 1 \dots M\}$  are hyperparameters. It follows the joint prior

$$p(\boldsymbol{\mu}, \mathbf{A}; \mathbf{D}) = \prod_{k=1}^K p(\boldsymbol{\mu}_k \mid \mathbf{A}_k; \mathbf{D}_k) p(\mathbf{A}_k) \quad (8)$$

where the first term is given by (6) and the second term by (7).

### 2.3 Priors on mixing weights

We examine two cases. First, following Corduneanu and Bishop [2001], no prior is imposed on  $\boldsymbol{\pi}$  (Section 3.1). Then a standard Dirichlet prior  $\mathcal{D}(\tau_1, \dots, \tau_K)$  is used in a second case (Section 3.2).

For the complete model, the whole set of parameters is denoted by  $\Phi$ . In our first setting,  $\Phi = \{\Phi^1, \Phi^2\}$  is decomposed into a set  $\Phi^1 = \{\Phi_1^1, \dots, \Phi_K^1\}$  with  $\Phi_k^1 = \{\boldsymbol{\mu}_k, \mathbf{A}_k\}$  of parameters for which we have priors and a set  $\Phi^2 = \{\boldsymbol{\pi}, \mathbf{D}, \boldsymbol{\alpha}\}$  of unknown parameters considered as fixed. In addition, hyperparameters are denoted by  $\Phi^3 = \{\Phi_1^3, \dots, \Phi_K^3\}$  with  $\Phi_k^3 = \{\mathbf{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$ . When a Dirichlet prior is used for  $\boldsymbol{\pi}$ , the parameters definitions change to  $\Phi_k^1 = \{\boldsymbol{\mu}_k, \mathbf{A}_k, \pi_k\}$ ,  $\Phi^2 = \{\mathbf{D}, \boldsymbol{\alpha}\}$  and  $\Phi_k^3 = \{\tau_k, \mathbf{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$ .

## 3 Inference by variational Expectation-Maximization

The main task in Bayesian inference is to compute the posterior probability of the latent variables  $\mathbf{X} = \{\mathbf{W}, \mathbf{Z}\}$  and the parameter  $\Phi$  for which only the  $\Phi^1$  part is considered as random. We are therefore interested in computing the posterior  $p(\mathbf{X}, \Phi^1 \mid \mathbf{y}, \Phi^2)$ . This posterior is intractable and approximated here using a variational approximation  $q(\mathbf{X}, \Phi^1)$  with a factorized form  $q(\mathbf{X}, \Phi^1) = q_X(\mathbf{X}) q_{\Phi^1}(\Phi^1)$ . We propose to use an Expectation-Maximization (EM) framework to compute  $q$ . At iteration ( $r$ ), the current fixed parameters values are denoted by  $\Phi^{2(r-1)}$ . The EM alternating procedure proceeds as follows, with  $\mathcal{D}$  the set of probability distributions that factorize as  $q(\mathbf{X}, \Phi^1) = q_X(\mathbf{X}) q_{\Phi^1}(\Phi^1)$ ,

$$\begin{aligned} \text{E-step:} & \quad q^{(r)}(\mathbf{X}, \Phi^1) = \arg \max_{q \in \mathcal{D}} \mathcal{F}(q, \Phi^{2(r-1)}) \\ \text{M-step:} & \quad \Phi^{2(r)} = \arg \max_{\Phi^2} \mathcal{F}(q^{(r)}, \Phi^2), \end{aligned}$$

where  $\mathcal{F}$  is the usual free energy

$$\mathcal{F}(q, \Phi^2) = E_q[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2)] - E_q[\log q(\mathbf{X}, \Phi^1)]. \quad (9)$$

The full expression of the free energy is not necessary to maximize it and to derive the variational EM algorithm. However, computing the free energy is useful first because it allows a stopping criterion and a safety check for the variational implementation as the free energy should increase at each iteration. Then it can be used as specified in section 4.2 as a replacement of the likelihood to provide a model selection procedure. The detailed expression is then given in Appendix A and B, respectively for the case without and with prior on the weights.

The E-step above divides into two steps. At iteration  $(r)$ , denoting in addition by  $q_X^{(r-1)}$  the current variational distribution for  $\mathbf{X}$ :

$$\begin{aligned} \mathbf{E}\text{-}\Phi^1\text{-step:} & \quad q_{\Phi^1}^{(r)}(\Phi^1) \propto \exp(E_{q_X^{(r-1)}}[\log p(\Phi^1|\mathbf{y}, \mathbf{X}; \Phi^{2(r-1)})]) \\ \mathbf{E}\text{-}\mathbf{X}\text{-step:} & \quad q_X^{(r)}(\mathbf{X}) \propto \exp(E_{q_{\Phi^1}^{(r)}}[\log p(\mathbf{X}|\mathbf{y}, \Phi^1; \Phi^{2(r-1)})]) . \end{aligned}$$

Then the M-step reduces to:

$$\mathbf{M}\text{-step:} \quad \Phi^{2(r)} = \arg \max_{\Phi^2} E_{q_X^{(r)} q_{\Phi^1}^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2)] .$$

The resulting variational EM algorithm is further specified below in two cases depending on the prior used for the mixing weights.

### 3.1 No prior on mixing coefficients

This corresponds to a setting adopted by Corduneanu and Bishop [2001] where the mixing coefficients are estimated using type-II maximum likelihood. In this case, the complete likelihood  $p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2, \Phi^3)$  writes as

$$p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2, \Phi^3) = p(\mathbf{y}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\mu}, \mathbf{A}; \mathbf{D})p(\mathbf{W}|\mathbf{Z}; \boldsymbol{\alpha})p(\mathbf{Z}; \boldsymbol{\pi})p(\boldsymbol{\mu}, \mathbf{A}; \mathbf{D}, \mathbf{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

Applying the previous general formulas, the E and M steps derive as follows.

**E- $\Phi^1$  step.** In this setting, the variational posterior has the same structure as the prior distribution (8). More specifically,

$$q_{\Phi^1}^{(r)}(\Phi^1) = \prod_{k=1}^K q_{\Phi^1}^{(r)}(\Phi_k^1) \quad \text{with} \quad q_{\Phi^1}^{(r)}(\Phi_k^1) = q_{\Phi^1}^{(r)}(\boldsymbol{\mu}_k, \mathbf{A}_k) = q_{\boldsymbol{\mu}_k|\mathbf{A}_k}^{(r)}(\boldsymbol{\mu}_k | \mathbf{A}_k) q_{\mathbf{A}_k}^{(r)}(\mathbf{A}_k) \quad \text{where}$$

$$q_{\boldsymbol{\mu}_k|\mathbf{A}_k}^{(r)}(\boldsymbol{\mu}_k | \mathbf{A}_k) = \mathcal{N}(\boldsymbol{\mu}_k; \tilde{\mathbf{m}}_k^{(r)}, \tilde{\boldsymbol{\Sigma}}_k^{(r)}) \quad (10)$$

$$q_{\mathbf{A}_k}^{(r)}(\mathbf{A}_k) = \prod_{m=1}^M \mathcal{G}(A_{km}; \tilde{\lambda}_{km}^{(r)}, \tilde{\delta}_{km}^{(r)}) . \quad (11)$$

Variational posteriors are defined via variational parameters  $\tilde{\mathbf{m}}_k^{(r)}$ ,  $\tilde{\boldsymbol{\Sigma}}_k^{(r)}$  and  $\tilde{\lambda}_{km}^{(r)}$ ,  $\tilde{\delta}_{km}^{(r)}$ . These parameters involve  $q_X^{(r-1)}$  via  $q_{Z_i}^{(r-1)}(k) = q_X^{(r-1)}(Z_i = k)$  and

$$\tilde{\boldsymbol{\Delta}}_{ki}^{(r-1)} = \text{diag}(\tilde{w}_{ki1}^{(r-1)}, \dots, \tilde{w}_{kiM}^{(r-1)}) \quad (12)$$

where  $\tilde{w}_{kim}^{(r-1)} = E_{q_X^{(r-1)}}[W_{im}|Z_i = k] = \tilde{\alpha}_{km}^{(r-1)}/\tilde{\gamma}_{kim}^{(r-1)}$ . The specific expressions of  $\tilde{\alpha}_{km}^{(r-1)}$  and  $\tilde{\gamma}_{kim}^{(r-1)}$  are given in eq. (20) and (21) of the E- $\mathbf{X}$  step below but using values from iteration  $(r-1)$ . The covariance

matrix  $\tilde{\Sigma}_k^{(r)}$  depends on  $\mathbf{A}_k$  as in the prior (6) while after simplification  $\tilde{\mathbf{m}}_k^{(r)}$  does not:

$$\tilde{\Sigma}_k^{(r)} = \mathbf{D}_k^{(r-1)} \tilde{\mathbf{N}}_k^{(r)-1} \mathbf{A}_k^{-1} \mathbf{D}_k^{(r-1)T} \quad (13)$$

$$\begin{aligned} \tilde{\mathbf{m}}_k^{(r)} &= \tilde{\Sigma}_k \mathbf{D}_k^{(r-1)} \mathbf{A}_k \left( \mathbf{\Lambda}_k \mathbf{D}_k^{(r-1)T} \mathbf{m}_k + \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) \tilde{\mathbf{\Delta}}_{ki}^{(r-1)} \mathbf{D}_k^{(r-1)T} \mathbf{y}_i \right) \\ &= \mathbf{D}_k^{(r-1)} \tilde{\mathbf{N}}_k^{(r)-1} \left( \mathbf{\Lambda}_k \mathbf{D}_k^{(r-1)T} \mathbf{m}_k + \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) \tilde{\mathbf{\Delta}}_{ki}^{(r-1)} \mathbf{D}_k^{(r-1)T} \mathbf{y}_i \right) \end{aligned} \quad (14)$$

$$\text{with} \quad \tilde{\mathbf{N}}_k^{(r)} = \mathbf{\Lambda}_k + \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) \tilde{\mathbf{\Delta}}_{ki}^{(r-1)}. \quad (15)$$

Then (11) is defined by the parameters:

$$\tilde{\lambda}_{km}^{(r)} = \lambda_{km} + 1/2 \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) \quad (16)$$

$$\tilde{\delta}_{km}^{(r)} = \delta_{km} + 1/2 [\mathbf{M}_k]_{m,m} \quad (17)$$

where  $[\mathbf{M}_k]_{m,m}$  denotes the  $m^{\text{th}}$  diagonal element on the following matrix  $\mathbf{M}_k$ :

$$\begin{aligned} \mathbf{M}_k &= \mathbf{D}_k^{(r-1)T} \left( \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) \mathbf{y}_i (\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(r)})^T \mathbf{D}_k^{(r-1)} \tilde{\mathbf{\Delta}}_{ki}^{(r-1)} \right) \\ &\quad + \mathbf{D}_k^{(r-1)T} \mathbf{m}_k (\mathbf{m}_k - \tilde{\mathbf{m}}_k^{(r)})^T \mathbf{D}_k^{(r-1)} \mathbf{\Lambda}_k. \end{aligned}$$

**E-X step.**  $q_X^{(r)}(\mathbf{X}) = \prod_{i=1}^N q_{X_i}^{(r)}(\mathbf{W}_i, \mathbf{Z}_i)$  with

$$q_{\mathbf{W}_i | \mathbf{Z}_i}^{(r)}(\mathbf{W}_i | \mathbf{Z}_i = k) = \prod_{m=1}^M q_{\mathbf{W}_{im} | \mathbf{Z}_i}^{(r)}(\mathbf{W}_{im} | \mathbf{Z}_i = k) = \prod_{m=1}^M \mathcal{G}(\mathbf{W}_{im}; \tilde{\alpha}_{km}^{(r)}, \tilde{\gamma}_{kim}^{(r)}) \quad (18)$$

$$q_{\mathbf{Z}_i}^{(r)}(\mathbf{Z}_i = k) = q_{Z_i}^{(r)}(k) \propto \pi_k^{(r-1)} \exp(\tilde{\rho}_k^{(r)}/2) \prod_{m=1}^M \frac{\Gamma(\tilde{\alpha}_{km}^{(r)})}{\Gamma(\alpha_{km}^{(r-1)}) \tilde{\gamma}_{kim}^{(r)}}. \quad (19)$$

The right-hand side term above is easy to normalized. The variational parameters  $\tilde{\alpha}_{km}^{(r)}, \tilde{\gamma}_{kim}^{(r)}$  are given by:

$$\tilde{\alpha}_{km}^{(r)} = \alpha_{km}^{(r-1)} + \frac{1}{2} \quad (20)$$

$$\tilde{\gamma}_{kim}^{(r)} = 1 + \frac{1}{2} \left( \tilde{\mathbf{A}}_{km}^{(r)} [\mathbf{D}_k^{(r-1)T} (\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(r)})]_m^2 + [\tilde{\mathbf{N}}_k^{(r)}]_{m,m}^{-1} \right) \quad (21)$$

where  $\tilde{\mathbf{N}}_k^{(r)}$  is given in (15),  $\tilde{\rho}_k^{(r)}$  and  $\tilde{\mathbf{A}}_{km}^{(r)}$  are easily computed from (11) ( $\Upsilon$  is the Digamma function):

$$\tilde{\rho}_k^{(r)} = E_{q_{\mathbf{A}_k}^{(r)}}[\log |\mathbf{A}_k|] = \sum_{m=1}^M \Upsilon(\tilde{\lambda}_{km}^{(r)}) - \log \tilde{\delta}_{km}^{(r)} \quad (22)$$

$$\tilde{\mathbf{A}}_k^{(r)} = E_{q_{\mathbf{A}_k}^{(r)}}[\mathbf{A}_k] \quad \text{that is for } m = 1 \dots M, \quad \tilde{\mathbf{A}}_{km}^{(r)} = \tilde{\lambda}_{km}^{(r)} / \tilde{\delta}_{km}^{(r)}. \quad (23)$$

**M-step.** The M-step divides into 3 sub-steps, where  $\boldsymbol{\pi}$ ,  $\mathbf{D}$  and  $\boldsymbol{\alpha}$  are updated separately. The sum  $\sum_{i=1}^N q_{Z_i}^{(r)}(k)$  is denoted by  $n_k^{(r)}$ .

**M- $\boldsymbol{\pi}$ -step.** This step leads to the standard formula for mixtures. For  $k = 1 \dots K$ ,  $\boldsymbol{\pi}$  is updated as:

$$\pi_k^{(r)} = \sum_{i=1}^N q_{Z_i}^{(r)}(k)/N = n_k^{(r)}/N .$$

**M- $\boldsymbol{\alpha}$ -step.** This step is less standard but equivalent to the update found in non Bayesian mixture of multiple scale distributions. The details can be found in the Supplementary material of [Forbes and Wraith, 2014]. In practice  $\boldsymbol{\alpha}$  is updated as follows. The estimates do not exist in closed form, but are given as a solution of the equations below, for each  $k = 1 \dots K$  and  $m = 1 \dots M$ :

$$\Upsilon(\alpha_{km}) = \Upsilon(\tilde{\alpha}_{km}^{(r)}) - \frac{1}{n_k^{(r)}} \sum_{i=1}^N q_{Z_i}^{(r)}(k) \log(\tilde{\gamma}_{kim}^{(r)})$$

The resolution of these equations in  $\alpha_{km}$  provides the  $\alpha_{km}^{(r)}$ .

**M- $\mathbf{D}$ -step.** Each  $\mathbf{D}_k$  can be updated separately as follows. Intermediate quantities are introduced to simplify the notation. For  $i = 1 \dots N + 1$ :

$$\begin{aligned} \forall i = 1 \dots N, \quad \mathbf{V}_{ki}^{(r)} &= q_{Z_i}^{(r)}(k)(\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(r)})(\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(r)})^T \\ \mathbf{V}_{k(N+1)}^{(r)} &= (\mathbf{m}_k - \tilde{\mathbf{m}}_k^{(r)})(\mathbf{m}_k - \tilde{\mathbf{m}}_k^{(r)})^T \\ \tilde{\boldsymbol{\Delta}}_{k(N+1)}^{(r)} &= \boldsymbol{\Lambda}_k \end{aligned}$$

As already defined in (12) and (15),

$$\forall i = 1 \dots N, \quad \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} = \text{diag}(\tilde{w}_{ki1}^{(r)}, \dots, \tilde{w}_{kiM}^{(r)})$$

and  $\tilde{\mathbf{N}}_k^{(r+1)} = \boldsymbol{\Lambda}_k + \sum_{i=1}^N q_{Z_i}^{(r)}(k) \tilde{\boldsymbol{\Delta}}_{ki}^{(r)}$ .

$$\begin{aligned} \mathbf{D}_k^{(r)} &= \arg \min_{\mathbf{D}_k \in \mathcal{O}} \sum_{i=1}^{N+1} \text{trace}(\mathbf{D}_k \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} \tilde{\mathbf{A}}_k^{(r)} \mathbf{D}_k^T \mathbf{V}_{ki}^{(r)}) \\ &\quad + E_{q_{\mathbf{A}}}[\text{trace}(\mathbf{D}_k \tilde{\mathbf{N}}_k^{(r+1)} \mathbf{A}_k \mathbf{D}_k^T \mathbf{D}_k^{(r-1)} (\tilde{\mathbf{N}}_k^{(r)})^{-1} \mathbf{A}_k^{-1} \mathbf{D}_k^{(r-1)T})] . \end{aligned}$$

The exact computation of the expectation above is feasible but would result in an expression where the elements of  $\mathbf{D}_k$  would be separated. As an alternative, we consider  $J$  *i.i.d* simulations of  $\mathbf{A}_k$  according to distribution  $q_{\mathbf{A}_k}^{(r)}$  which is a product of Gamma distributions given in (11). Denoting by  $\mathbf{A}_{kj}$  for  $j = 1 \dots J$  these simulations,  $\mathbf{D}_k^{(r)}$  can be approximated by,

$$\begin{aligned} \mathbf{D}_k^{(r)} &\approx \arg \min_{\mathbf{D}_k \in \mathcal{O}} \sum_{i=1}^{N+1} \text{trace}(\mathbf{D}_k \tilde{\boldsymbol{\Delta}}_{ki}^{(r)} \tilde{\mathbf{A}}_k^{(r)} \mathbf{D}_k^T \mathbf{V}_{ki}^{(r)}) \\ &\quad + \frac{1}{J} \sum_{j=1}^J \text{trace}(\mathbf{D}_k \tilde{\mathbf{N}}_k^{(r+1)} \mathbf{A}_{kj} \mathbf{D}_k^T \mathbf{D}_k^{(r-1)} (\tilde{\mathbf{N}}_k^{(r)})^{-1} \mathbf{A}_{kj}^{-1} \mathbf{D}_k^{(r-1)T}) . \end{aligned}$$

The Monte-Carlo approximation of the expectation has the advantage to allow for the optimization of  $\mathbf{D}_k$  on the Stiefel manifold  $\mathcal{O}$ . In [Celeux and Govaert, 1995, Forbes and Wraith, 2014, Wraith and Forbes, 2015], an algorithm by Flury and Gautschi [1986] was used but we consider here a more recent procedure using an accelerated line search method proposed by Browne and McNicholas [2014].

### 3.2 Dirichlet prior on mixing coefficients

In this section  $\boldsymbol{\pi}$  is now considered as a random variable. More specifically the prior over  $\boldsymbol{\Phi}^1$  writes

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{A}; \boldsymbol{\tau}, \mathbf{D}) = p(\boldsymbol{\pi}; \boldsymbol{\tau}) \prod_{k=1}^K p(\boldsymbol{\mu}_k | \mathbf{A}_k; \mathbf{D}_k) p(\mathbf{A}_k) \quad (24)$$

where  $p(\boldsymbol{\pi}; \boldsymbol{\tau}) = \mathcal{D}(\boldsymbol{\pi}; \tau_1, \dots, \tau_K) = \frac{\Gamma(\sum_{k=1}^K \tau_k)}{\prod_{k=1}^K \Gamma(\tau_k)} \prod_{k=1}^K \pi_k^{\tau_k - 1}$  is a Dirichlet distribution.

With this modification, only the  $p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}; \boldsymbol{\tau})$  term changes in the complete likelihood which becomes,

$$p(\mathbf{y}, \mathbf{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3) = p(\mathbf{y} | \mathbf{Z}, \mathbf{W}, \boldsymbol{\mu}, \mathbf{A}; \mathbf{D}) p(\mathbf{W} | \mathbf{Z}; \boldsymbol{\alpha}) p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}; \boldsymbol{\tau}) p(\boldsymbol{\mu}, \mathbf{A}; \mathbf{D}, \mathbf{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

Applying the previous general formulas, the E and M steps derive as follows.

**E- $\boldsymbol{\Phi}^1$  step.** With the same form for the prior and variational posterior, it comes

$$q_{\boldsymbol{\Phi}^1}^{(r)}(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{A}) = q^{(r)}(\boldsymbol{\pi}) \prod_{k=1}^K q_{\boldsymbol{\mu}_k, \mathbf{A}_k}^{(r)}(\boldsymbol{\mu}_k, \mathbf{A}_k)$$

where  $q_{\boldsymbol{\mu}_k, \mathbf{A}_k}^{(r)}(\boldsymbol{\mu}_k, \mathbf{A}_k)$  has the same expression as given by (10) and (11). The new term is

$$q_{\boldsymbol{\pi}}^{(r)}(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \tilde{\tau}_1^{(r)}, \dots, \tilde{\tau}_K^{(r)})$$

with  $\tilde{\tau}_k^{(r)} = \tau_k + \sum_{i=1}^N q_{Z_i}^{(r-1)}(k) = \tau_k + n_k^{(r-1)}$ .

**E- $\mathbf{X}$  step.** This step is only partly impacted by the addition of a prior on  $\boldsymbol{\pi}$ . It comes as in the previous section,  $q_X^{(r)}(\mathbf{X}) = \prod_{i=1}^N q_{X_i}^{(r)}(\mathbf{W}_i, \mathbf{Z}_i)$  with the term below unchanged and given by (18), (20) and (21),

$$q_{W_i | Z_i}^{(r)}(\mathbf{W}_i | Z_i = k) = \prod_{m=1}^M q_{W_{im} | Z_i}^{(r)}(\mathbf{W}_{im} | Z_i = k) = \prod_{m=1}^M \mathcal{G}(\mathbf{W}_{im}; \tilde{\alpha}_{km}^{(r)}, \tilde{\gamma}_{kim}^{(r)}).$$

In contrast, the posterior on  $\mathbf{Z}$  is changed into

$$q_{Z_i}^{(r)}(Z_i = k) = q_{Z_i}^{(r)}(k) \propto \tilde{\pi}_k^{(r)} \exp(\tilde{\rho}_k^{(r)}/2) \prod_{m=1}^M \frac{\Gamma(\tilde{\alpha}_{km}^{(r)})}{\Gamma(\alpha_{km}^{(r-1)}) \tilde{\gamma}_{kim}^{(r) \tilde{\alpha}_{km}^{(r)}}}. \quad (25)$$

where the modification reduces to changing  $\pi_k^{(r-1)}$  into  $\tilde{\pi}_k^{(r)}$  that can be derived from the previous E- $\boldsymbol{\Phi}^1$  step as

$$\log \tilde{\pi}_k^{(r)} = E_{q_{\boldsymbol{\pi}}^{(r)}}[\log \pi_k] = \Upsilon(\tilde{\tau}_k^{(r)}) - \Upsilon\left(\sum_{l=1}^K \tilde{\tau}_l^{(r)}\right) = \Upsilon(\tau_k + n_k^{(r)}) - \Upsilon\left(\sum_{k=1}^K \tau_k + N\right).$$

The last term being constant with respect to  $(r)$  and  $k$ , it follows that  $\tilde{\pi}_k^{(r)}$  is proportional to  $\exp(\Upsilon(\tau_k + n_k^{(r)}))$  and it is enough to use this later expression in (25).

**M-step.** The M-step formulation remains the same as before without the M- $\boldsymbol{\pi}$  step.

In what follows, we illustrate the use of the Bayesian formulation and its variational EM implementation on the issue of selecting the number of components in the mixture.

## 4 Identifying the number of mixture components

When the number of components is unknown, apart from the Reversible Jump Markov Chain Monte Carlo method of Richardson and Green [1997] which allows jumps between different numbers of components, two types of approaches can be distinguished depending on whether the strategy is to increase or decrease the number of components. The first ones can be referred to as greedy algorithms (*e.g.* Verbeek et al. [2003]) where the mixture is built component-wise, starting with the optimal one-component mixture and increasing the number of components until a stopping criterion is met. In this work, we will rather consider approaches that start from an overfitting model with more components than expected in the data. In this case, as described by Frühwirth-Schnatter [2006], identifiability will be violated in two possible ways. Identifiability issues can arise either because some of the components weights have to be zero (then component-specific parameters cannot be identified) or because some of the components have to be equal (then their weights cannot be identified). In practice, these two possibilities are not equivalent as checking for vanishing components is easier and is likely to lead to more stable behavior than testing for redundant components (see *e.g.* Rousseau and Mengersen [2011]).

Both increasing and decreasing methods can be considered in a Bayesian and maximum likelihood setting. However, in a Bayesian framework, in contrast to maximum likelihood, considering a posterior distribution on the mixture parameters requires integrating out the parameters and this acts as a penalization for more complex models. The posterior is essentially putting mass on the sparsest way to approximate the true density, see *e.g.* Rousseau and Mengersen [2011]. Although the framework of Rousseau and Mengersen [2011] is fully Bayesian with priors on all mixture parameters, it seems that this penalization effect is also effective when only some of the parameters are integrated out. This is observed by Corduneanu and Bishop [2001] who use priors only for the component mean and covariance parameters. This justifies in our setting the investigation of a case with no prior on the mixing weights (Section 3.1).

The idea of using overfitting finite mixtures with too many components  $K$  has been used in many papers. In a deliberately overfitting mixture model, a sparse prior on the mixture weights will empty superfluous components during estimation [Malsiner-Walli et al., 2016]. To obtain sparse solutions with regard to the number of mixture components, an appropriate prior on the weights  $\boldsymbol{\pi}$  has to be selected. Guidelines have been given in previous work when the prior for the weights is a symmetric Dirichlet distribution  $\mathcal{D}(\tau_1, \dots, \tau_K)$  with all  $\tau_k$ 's equal to a value  $\tau_0$ . To empty superfluous components automatically the value of  $\tau_0$  has to be chosen appropriately. In particular, Rousseau and Mengersen [2011] proposed conditions on  $\tau_0$  to control the asymptotic behavior of the posterior distribution of an overfitting mixture with respect to the two previously mentioned regimes. One regime in which a high likelihood is set to components with nearly identical parameters and one regime in which some of the mixture weights go to zero. More specifically, if  $\tau_0 < d/2$  where  $d$  is the dimension of the component specific parameters, when  $N$  tends to infinity, the posterior expectation of the weight of superfluous components converges to zero. In practice,  $N$  is finite and as observed by Malsiner-Walli et al. [2016], much smaller value of  $\tau_0$  are needed (*e.g.*  $10^{-5}$ ). It was even observed by Tu [2016] that negative values of  $\tau_0$  were useful to induce even more sparsity when the number of observations is too large with respect to the prior impact. Dirichlet priors with negative parameters, although not formally defined, are also mentioned by Figueiredo and Jain [2002]. This latter work does not start from a Bayesian formulation but is based on a Minimum Message Length (MML) principle. Figueiredo and Jain [2002] provide an M-step that performs component annihilation, thus an explicit rule for moving from the current number of components to a smaller one. A parallel is made with a Dirichlet prior with  $\tau_0 = -d/2$  which according to Tu [2016] corresponds to a very strong prior sparsity.

### 4.1 Single-run number of component selection

In a Bayesian setting with symmetric sparse Dirichlet priors  $\mathcal{D}(\tau_0, \dots, \tau_0)$ , the theoretical study of Rousseau and Mengersen [2011] therefore justifies to consider the posterior expectations of the weights  $E[\pi_k | \mathbf{y}]$  and to prune out the too small ones. In practice this raises at least two additional questions: which expression to use for the estimated posterior means and how to set a threshold under which the estimated means are considered too small. The posterior means estimation is generally guided by the chosen inference scheme.

For instance in our variational framework with a Dirichlet prior on the weights, the estimated posterior mean  $E[\pi_k|\mathbf{y}]$  takes the following form (the  $(r)$  notation is removed to signify the convergence of the algorithm),

$$\begin{aligned} E[\pi_k|\mathbf{y}] &\approx E_{q_\pi}[\pi_k] = \frac{\tilde{\tau}_k}{\sum_{l=1}^K \tilde{\tau}_l} \\ &= \frac{\tau_k + n_k}{\sum_{k=1}^K \tau_k + N} \end{aligned} \quad (26)$$

where  $n_k = \sum_{i=1}^N q_{Z_i}(k)$  and  $q_{Z_i}(k)$  is given by (25). If we are in the no weight prior case, then the expectation simplifies to

$$\pi_k \approx \frac{n_k}{N} \quad (27)$$

with  $q_{Z_i}(k)$  given by (19).

Nevertheless, whatever the inference scheme or prior setting, we are left with the issue of detecting when a component can be set as empty. There is usually a close relationship between the component weight  $\pi_k$  and the number of observations assigned to component  $k$ . This later number is itself often replaced by the sum  $n_k = \sum_{i=1}^N q_{Z_i}(k)$ . As an illustration, the choice of a negative  $\tau_0$  by Figueiredo and Jain [2002] corresponds to a rule that sets a component weight to zero when  $n_k = \sum_{i=1}^N q_{Z_i}(k)$  is smaller than  $d/2$ . This prevents the algorithm from approaching the boundary of the parameter space. When one of the components becomes too weak, meaning that it is not supported by the data, it is simply annihilated. One of the drawbacks of standard EM for mixtures is thus avoided. The rule of Figueiredo and Jain [2002] is stronger than that used by McGrory and Titterton [2007] which annihilates a component when the sum  $n_k$  reduces to 1 or the one of Corduneanu and Bishop [2001] which corresponds to the sum  $n_k$  lower than a very small fraction of the sample size, *i.e.*  $\sum_{i=1}^N q_{Z_i}(k)/N < 10^{-5}$  where  $N$  varies from 400 to 900 in their experiments. Note that McGrory and Titterton [2007] use a Bayesian framework with variational inference and their rule corresponds to thresholding the variational posterior weights (26) to  $1/N$  because they set all  $\tau_k$  to 0 in their experiments.

In this work, we will also adopt a thresholding approach but note that alternatives have been developed that would worth testing to avoid the issue of setting a threshold for separating large and small weights. In their MCMC sampling, Malsiner-Walli et al. [2016] propose to consider the number of non-empty components at each iteration and to estimate the number of components as the most frequent number of non-empty components. This is not directly applicable in our variational treatment as it would require to generate hard assignments to components at each iteration instead of dealing with their probabilities. In contrast, we could adopt techniques from the Bayesian non parametrics literature which seek for optimal partitions, such as the criterion of Dahl [2006] using the so-called posterior similarity matrix (Fritsch and Ickstadt [2009]). This matrix could be approximated easily in our case by computing the variational estimate of the probability that two observations are in the same component.

## 4.2 Tested procedures

We compare three types of single-run methods to estimate the number of components in a mixture of multiple scale distributions.

### 4.2.1 Fully Bayesian algorithm with sparse Dirichlet prior: "SparseDirichlet"

A first method is directly derived from a fully Bayesian setting with a sparse symmetric Dirichlet prior likely to induce vanishing coefficients as supported by the theoretical results of Rousseau and Mengersen [2011]. This corresponds to the approach adopted in Malsiner-Walli et al. [2016] and McGrory and Titterton [2007]. The difference between the later two being how they check for vanishing coefficients. Our variational

inference leads more naturally to the solution of McGrory and Titterington [2007] which is to check the weight posterior means, that is whether at each iteration ( $r$ ),

$$n_k^{(r)} < (K\tau_0 + N)\rho_t - \tau_0 \quad (28)$$

where  $\rho_t$  is the chosen threshold on the posterior means. When  $\rho_t$  is set such that (28) leads to  $n_k^{(r)} < 1$ , this method is referred to, in the next Section, as *SparseDirichlet+ $\pi$ test*. For comparison, the algorithm run with no intervention is called *SparseDirichlet*.

#### 4.2.2 Type II maximum likelihood on mixing weights: "TypeIIML"

A second method corresponds to the method proposed by Corduneanu and Bishop [2001]: no prior on the weights and a criterion on the estimated weights to detected vanishing coefficients. It corresponds to applying (28) with  $\tau_0 = 0$ . This method is referred below as *TypeIIML* when the algorithm is run until convergence and *TypeIIML+ $\pi$ test* when (28) is used at each iteration with  $\rho_t = 1/N$ .

#### 4.2.3 Free Energy based algorithm: "FEtest"

At last, we consider a criterion based on the free energy (9) to detect components to eliminate. In the no weight prior case, this is to handle potentially redundant components rather than vanishing ones and is also the opportunity to test a greedy potentially accelerated heuristic. This choice is based on the observation that when no prior is used for the weights, we cannot control the hyperparameters (*e.g.*  $\tau_k$ ) to guide the algorithm in the vanishing components regime. Thus the algorithm may as well go to the redundant component regime. The goal is then to test whether this alternative method is likely to handle this behavior. The proposal is to start from a clustering solution with too many components and to try to remove them using a criterion based on the gain in free energy. In this setting, the components that are removed are not necessarily vanishing components but also redundant ones. The free energy expression used is given in Appendix A. With no prior on the weights, the algorithm is referred to as *TypeIIML+FEtest*. The same idea can be applied in the fully Bayesian setting, referred to here as *SparseDirichlet+FEtest*. The heuristic can be described as follows (see the next Section for implementation details).

1. Iteration  $r = 0$ : Initialization of the  $K^{(0)}$  clusters and probabilities using for instance repetitions of k-means or trimmed k-means.
2. Iteration  $r \geq 1$ :
  - (a) E and M steps updating from parameters at iteration  $r - 1$
  - (b) Updating of the resulting Free Energy value
  - (c) In parallele, for each cluster  $k \in \{1 \dots K^{(r-1)}\}$ 
    - i. Re-normalization of the cluster probabilities when cluster  $k$  is removed from current estimates at iteration  $r - 1$ : the sum over the remaining  $K^{(r-1)} - 1$  clusters must be equal to 1
    - ii. Updating of the corresponding E and M steps and computation of the associate Free Energy value
  - (d) Selection of the mixture with the highest Free Energy among the  $K^{(r-1)}$ -component mixture (step (b)) or one of the  $(K^{(r-1)} - 1)$ -component mixtures (step (c)).
  - (e) Updating of  $K^{(r)}$  accordingly, to  $K^{(r-1)}$  or  $K^{(r-1)} - 1$ .
3. When no more cluster deletion occur (*eg.* during 5 steps), we switch to the EM algorithm (*TypeIIML* or *SparseDirichlet*).



## 5 Experiments

In addition to the 6 methods mentioned above and referred to below as  $\mathcal{MP}$  single-run procedures, we consider standard Gaussian mixtures using the Mclust package [Scrucca et al., 2016] including a version with priors on the means and covariance matrices. The Bayesian Information Criterion (BIC) is then used to select the number of components from  $K = 1$  to 10. The respective methods are denoted below by *GM+BIC* and *Bayesian GM+BIC*. Regarding mixtures of  $\mathcal{MP}$  distributions, we also consider their non Bayesian version, using BIC to select  $K$ , denoted below by *MMP+BIC*.

In practice, values need to be chosen for hyperparameters. These include the  $\mathbf{m}_k$  that are set to 0, the  $\mathbf{\Lambda}_k$  that are set to  $\epsilon \mathbf{I}_M$  with  $\epsilon$  small (set to  $10^{-4}$ ) so has to generate a large variance in (6). The  $\delta_{km}$  are then set to 1 and  $\lambda_{km}$  to values  $5 \times 10^{-4} = \lambda_1 < \lambda_2 < \dots < \lambda_M = 10^{-3}$ . When necessary, the  $\tau_k$ 's are set to  $10^{-3}$  to favor sparse mixtures.

Initialization is also an important step in EM algorithms. All single-run methods are initialized with the same initial cluster assignments using  $K = 10$  obtained with 10 repetitions of trimmed k-means for 10 iterations each and excluding 10% of outliers. For Gaussian mixtures, the initialization procedure is that embedded in Mclust. For each run of the procedures, we allow 100 re-starts and choose the best one after 1000 iterations. For  $\mathcal{MP}$  models, initial values of the  $\alpha_{km}$ 's are set to 1.

Another important point for single-run procedures, is how to finally enumerate remaining components. For simplicity, we report components that are expressed by the maximum a posteriori (MAP) rule, which means components for which there is at least one data point assigned to them with the highest probability.

### 5.1 Simulated data

We first start with some simulated data from a mixture of  $\mathcal{MP}$  distributions in dimension 2 with 3 components respectively centered at  $[0, -2]$ ,  $[0, 0]$  and  $[0, 2]$  with the same scale matrix  $[2, 0; 0, 0.2]$  and parameters  $\alpha_{k1} = 1$  and  $\alpha_{k2} = 100$  for  $k = 1, 2, 3$ . This example is a  $\mathcal{MP}$  version of a Gaussian mixture used by Corduneanu and Bishop [2001] and McGrory and Titterton [2007] (see Figure 1 (a)). The sample size is  $N = 900$  and 10 samples are simulated and used to test the different procedures. Table 1 summarizes the final observed or selected number of components for each procedure. For Gaussian mixtures  $K = 6$  and 7 are the most selected values by BIC. The presence of data points in the tails induces the addition of components to capture all points that cannot be well explained by the 3 main visual components (Figures 1 (c) and (d)). In the MMP case, tails are rather well captured but BIC is hesitating between the 4 and 3 component solutions. For the 6  $\mathcal{MP}$  single-run procedures, the final number of components is almost always 3 and the clusterings are all very similar to the one shown in Figure 1 (e) (*TypeIIML+ $\pi$ test* case). Mean computational times over the 10 repetitions are reported in seconds in Table 1 (last column). The procedure using the  $\mathcal{MP}$  mixture and BIC is the longest due to repetitive runs for each value of  $K$  between 1 and 10. Computational gain is observed as expected when using one of the 4 procedures with component elimination. In particular, combining a sparse Dirichlet prior and free energy-based elimination seems to provide the largest gain with a running time (525s) divided by 5 compared to the *MMP+BIC* procedure (2767s).

A second example consists of 3 similar  $\mathcal{MP}$  distributions but with closer means namely  $[0, -1]$ ,  $[0, 0]$  and  $[0, 1]$  and  $\alpha_{k1} = 2$ ,  $\alpha_{k2} = 100$  for  $k = 1, 2, 3$  (Figure 2 (a)). In terms of clustering and computational times, conclusions are similar as illustrated in Figure 2 and Table 2. All  $\mathcal{MP}$  methods find  $K = 3$  most of the time including the procedure using BIC.

### 5.2 Standard dataset

#### 5.2.1 Old Faithful Geyser data

This data set contains 272 observations on 2 variables which are the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park. The scatter plot in Figure 3 (a) shows two moderately separated groups. For model selection this example has been studied in particular by Stephens [2000] with Gaussian mixtures. It was found that when more than 2 clusters are

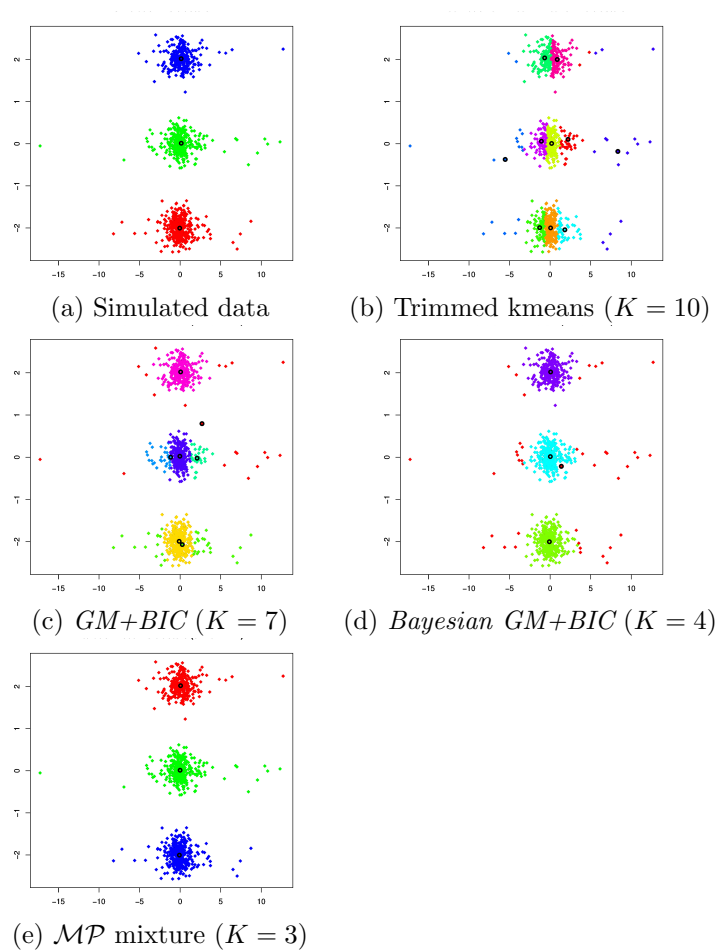


Figure 1: (a): Mixture of 3  $MP$  distributions with  $N = 900$ , (b): 10 component initialization using trimmed k-means, (c):  $GM+BIC$  clustering ( $K = 7$ ), (d):  $Bayesian GM+BIC$  clustering ( $K = 4$ ), (e)  $MP$  mixture clustering ( $K = 3$ ).

Procedure (10 runs)	Selected number of components										Average time (in seconds)
	1	2	3	4	5	6	7	8	9	10	
GM+BIC	.	.	.	2	.	3	3	.	1	1	604
BayesianGM+BIC	.	.	.	8	1	.	.	.	.	1	252
MMP+BIC	.	.	5	5	.	.	.	.	.	.	2767
TypeIIML	.	.	9	1	.	.	.	.	.	.	1894
TypeIIML+ $\pi$ test	.	.	10	.	.	.	.	.	.	.	718
TypeIIML+FEtest	.	.	10	.	.	.	.	.	.	.	748
SparseDirichlet	.	.	9	1	.	.	.	.	.	.	1891
SparseDirichlet+ $\pi$ test	.	.	10	.	.	.	.	.	.	.	678
SparseDirichlet+FEtest	.	.	10	.	.	.	.	.	.	.	<b>525</b>

Table 1: Final observed or selected number of components for each procedure on 10 samples, and mean computational times in seconds.

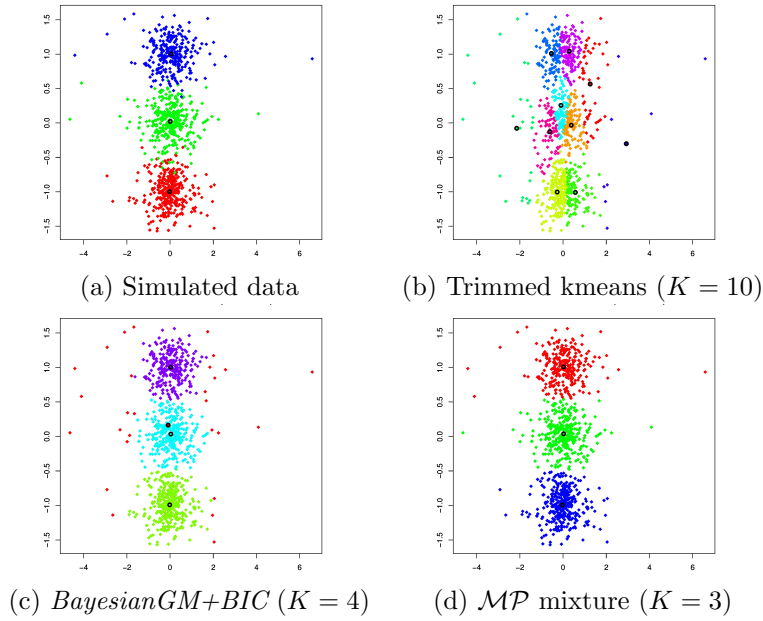


Figure 2: a): Mixture of 3 closer  $\mathcal{MP}$  distributions with  $N = 900$ , (b): 10 component initialization using trimmed k-means, (c): *Bayesian GM+BIC* clustering ( $K = 4$ ), (d):  $\mathcal{MP}$  mixture clustering ( $K = 3$ ).

Procedure (10 runs)	Selected number of components										Average time (in seconds)
	1	2	3	4	5	6	7	8	9	10	
GM+BIC	.	.	.	10	.	.	.	.	.	.	860
BayesianGM+BIC	.	.	.	9	1	.	.	.	.	.	405
MMP+BIC	.	.	10	.	.	.	.	.	.	.	2584
TypeIIML	.	.	10	.	.	.	.	.	.	.	1917
TypeIIML+ $\pi$ test	.	.	10	.	.	.	.	.	.	.	719
TypeIIML+FEtest	.	.	10	.	.	.	.	.	.	.	771
SparseDirichlet	.	.	10	.	.	.	.	.	.	.	1907
SparseDirichlet+ $\pi$ test	.	.	10	.	.	.	.	.	.	.	<b>688</b>
SparseDirichlet+FEtest	.	.	10	.	.	.	.	.	.	.	751

Table 2: Final observed or selected number of components for each procedure on 10 samples, and mean computational times in seconds.

fit the extra components are there to model the deviation from normality in the two obvious groups rather than to model interpretable extra clusters. This is consistent with what we observe with Gaussian models (*Mclust+BIC* and *Bayesian Mclust+BIC*), finding 3 components (Figure 3 (c)) while our  $\mathcal{MP}$  model with BIC and the free energy based elimination procedure show consistently 2 selected clusters (Figure 3 (d)). All other methods select 3 components but the clustering differs from that of the Gaussian models (Figures 3 (c) and (e)). All procedures were initialized with a 10 component assignment similar to that shown in Figure 3 (b). In terms of complexity, it appears that the fastest procedures are the free energy based ones (*FEtest*) (176s) with a time divided by 5 compared to the *MMP+BIC* one (1012s).

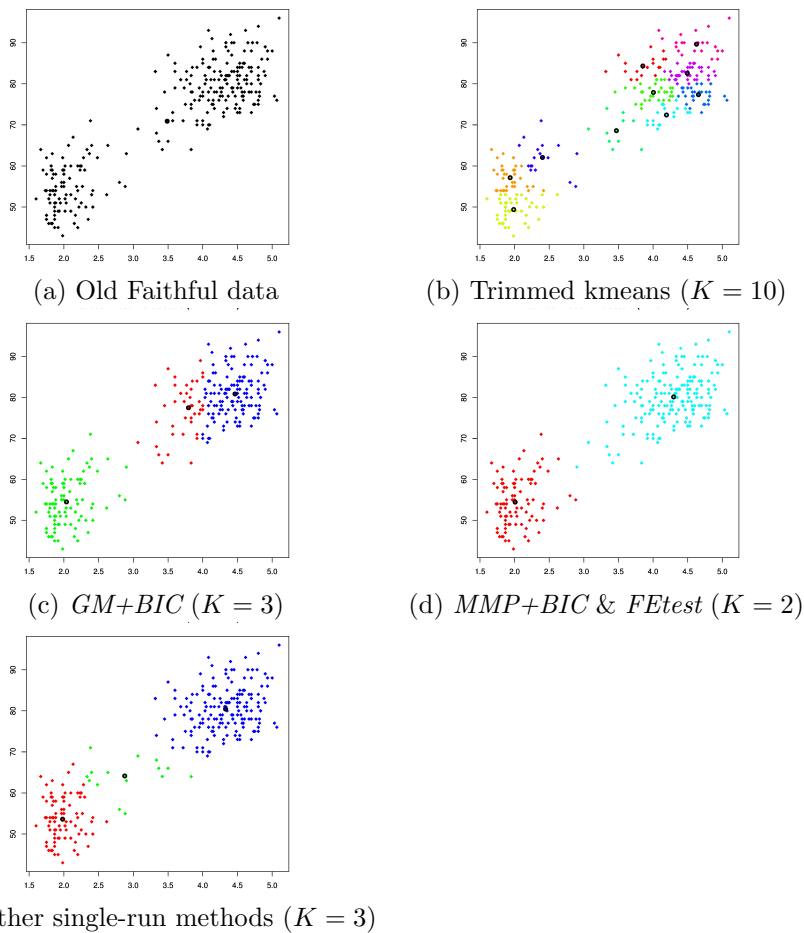


Figure 3: (a): Old Faithful data, (b): 10 component initialization using trimmed k-means, (c): Gaussian mixture clustering as selected by BIC ( $K = 3$ ), (d):  $\mathcal{MP}$  clustering obtained with BIC and free energy based elimination ( $K = 2$ ); (e):  $\mathcal{MP}$  clustering for the other single-run procedures ( $K = 3$ ).

## 6 Discussion

We investigated, in the context of mixtures of non-Gaussian distributions, different single-run procedures to select automatically the number of components. The most time efficient single run strategies boil down to replace a discrete model choice (*e.g.* selecting  $K$  using BIC) by a continuous control parameter, here the threshold  $\rho_t$  over the posterior weights or the gain in free energy. The advantage of single run procedures is to avoid time consuming comparison of scores for each model. However, there are different ways to

implement this idea: full Bayesian settings which have the advantage to be supported by some theoretical justification [Rousseau and Mengersen, 2011] and Type II maximum likelihood as proposed by Corduneanu and Bishop [2001]. In this work, we proposed in addition another heuristic based on the gain in free energy. On preliminary experiments, we observed that both in terms of selection and computation time, Type II maximum likelihood on the weights was competitive with the use of a Dirichlet prior. In addition, both approaches were more efficient than BIC comparisons, in particular for  $\mathcal{MP}$  mixtures which are more costly to estimate than Gaussian mixtures. Free energy based methods appeared to be slightly faster but not very significantly so compared to posterior weight thresholding methods. To confirm these observations, more tests in particular on larger and real data sets would be required to better compare and understand the various characteristics of each procedures.

## A Expression of the free energy

The expression of the free energy at each iteration is needed to apply the procedure mentioned in section 4.2. It is given here in the absence of weight prior (Section 3.1). The free energy expression differs only slightly when a Dirichlet prior is added (see Appendix B). The free energy can be decomposed into two terms. At each iteration ( $r$ ),

$$\mathcal{F}(q^{(r)}, \Phi^{2(r)}) = E_{q^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^{2(r)})] - E_{q^{(r)}}[\log q^{(r)}(\mathbf{X}, \Phi^1)],$$

where the second term is made of entropies and the first term has been already computed in the M-step.

### A.1 Entropy terms

In this section, we provide the expression of  $-E_{q^{(r)}}[\log q^{(r)}(\mathbf{X}, \Phi^1)]$  when it is equal to

$$\begin{aligned} -E_{q^{(r)}}[\log q^{(r)}(\mathbf{X}, \Phi^1)] &= H[q_{\mathbf{X}}^{(r)}] + H[q_{\Phi^1}^{(r)}] \\ &= \sum_{i=1}^N H[q_{X_i}^{(r)}] + \sum_{k=1}^K H[q_{\Phi_k^1}^{(r)}]. \end{aligned}$$

In the expression above,  $H[q_{\Phi_k^1}^{(r)}]$  is the entropy of the Normal-Wishart distribution defined in (10) and (11),

$$\begin{aligned} H[q_{\Phi_k^1}^{(r)}] &= \frac{1}{2} \left( M \log 2\pi e - \log |\tilde{\mathbf{N}}_k^{(r)}| - \sum_{m=1}^M \Upsilon(\tilde{\lambda}_{km}^{(r)}) - \log \tilde{\delta}_{km}^{(r)} \right) \\ &\quad + \sum_{m=1}^M \left( \tilde{\lambda}_{km}^{(r)} - \log \tilde{\delta}_{km}^{(r)} + \log \Gamma(\tilde{\lambda}_{km}^{(r)}) + (1 - \tilde{\lambda}_{km}^{(r)}) \Upsilon(\tilde{\lambda}_{km}^{(r)}) \right), \end{aligned}$$

where  $\tilde{\mathbf{N}}_k^{(r)}$  is given by equation (15),  $\tilde{\lambda}_{km}^{(r)}$  and  $\tilde{\delta}_{km}^{(r)}$  by (16) and (17).

Then each term  $H[q_{X_i}^{(r)}]$  is the sum of a product-of-Gamma entropy and a multinomial entropy,

$$\begin{aligned} H[q_{X_i}^{(r)}] &= \sum_{k=1}^K q_{Z_i}^{(r)}(k) \sum_{m=1}^M \left( \tilde{\alpha}_{km}^{(r)} + \log \Gamma(\tilde{\alpha}_{km}^{(r)}) + (1 - \tilde{\alpha}_{km}^{(r)}) \Upsilon(\tilde{\alpha}_{km}^{(r)}) - \log \tilde{\gamma}_{kim}^{(r)} \right) \\ &\quad - \sum_{k=1}^K q_{Z_i}^{(r)}(k) \log q_{Z_i}^{(r)}(k) \end{aligned}$$

where  $q_{Z_i}^{(r)}(k)$  is given by (19),  $\tilde{\alpha}_{km}^{(r)}$  by (20) and  $\tilde{\gamma}_{kim}^{(r)}$  by (21).

## A.2 M-step terms

The term  $E_{q^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^{2(r)})]$  decomposes into five terms,

$$\begin{aligned} E_{q^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^{2(r)})] &= E_q[\log p(\mathbf{y}|\mathbf{X}, \Phi^1; \mathbf{D}^{(r)})] + E_{q_{Z,W}^{(r)}}[\log p(\mathbf{W}|\mathbf{Z}; \boldsymbol{\alpha}^{(r)})] \\ &\quad + E_{q_Z^{(r)}}[\log p(\mathbf{Z}; \boldsymbol{\pi}^{(r)})] + E_{q_{\mu,\mathbf{A}}^{(r)}}[\log p(\boldsymbol{\mu}|\mathbf{A}; \mathbf{D}^{(r)})] + E_{q_{\mathbf{A}}^{(r)}}[\log p(\mathbf{A})] . \end{aligned}$$

The five terms are detailed in turn below,

$$\begin{aligned} E_{q^{(r)}}[\log p(\mathbf{y}|\mathbf{X}, \Phi^1; \mathbf{D}^{(r)})] &= -1/2 \sum_{i=1}^N \sum_{k=1}^K q_{Z_i}^{(r)}(k) \left( M \log 2\pi - \tilde{\rho}_k^{(r)} - \sum_{m=1}^M (\Upsilon(\tilde{\alpha}_{km}^{(r)}) - \log \tilde{\gamma}_{kim}^{(r)}) \right. \\ &\quad \left. + (\tilde{m}_k^{(r)} - \mathbf{y}_i)^T \mathbf{D}_k^{(r)} \bar{\Delta}_{ki}^{(r)} \tilde{\mathbf{A}}_k^{(r)} \mathbf{D}_k^{(r)T} (\tilde{m}_k^{(r)} - \mathbf{y}_i) + \text{trace}(\bar{\Delta}_{ki}^{(r)} (\tilde{\mathbf{N}}_k^{(r)})^{-1}) \right) \end{aligned}$$

with  $\bar{\Delta}_{ki}^{(r)}$  given in (12),  $\tilde{m}_k^{(r)}$  in (14),  $\tilde{\mathbf{N}}_k^{(r)}$  in (15),  $\tilde{\rho}_k^{(r)}$  and  $\tilde{\mathbf{A}}_k^{(r)}$  in (22) and (23),  $\tilde{\alpha}_{km}^{(r)}$  and  $\tilde{\gamma}_{kim}^{(r)}$  in (20) and (21),  $\mathbf{D}^{(r)}$  is the solution of the M- $\mathbf{D}$ -step. .

$$\begin{aligned} E_{q_{Z,W}^{(r)}}[\log p(\mathbf{W}|\mathbf{Z}; \boldsymbol{\alpha}^{(r)})] &= \sum_{i=1}^N \sum_{k=1}^K q_{Z_i}^{(r)}(k) \sum_{m=1}^M \left( -\log \Gamma(\alpha_{km}^{(r)}) \right. \\ &\quad \left. + (\alpha_{km}^{(r)} - 1)(\Upsilon(\tilde{\alpha}_{km}^{(r)}) - \log \tilde{\gamma}_{kim}^{(r)}) - \frac{\tilde{\alpha}_{km}^{(r)}}{\tilde{\gamma}_{kim}^{(r)}} \right) \end{aligned}$$

where  $q_{Z_i}^{(r)}(k)$  is given in (19),  $\alpha_{km}^{(r)}$  are the solutions of the M- $\boldsymbol{\alpha}$  step,  $\tilde{\alpha}_{km}^{(r)}$  and  $\tilde{\gamma}_{kim}^{(r)}$  are given in (20) and (21).

$$E_{q_Z^{(r)}}[\log p(\mathbf{Z}; \boldsymbol{\pi}^{(r)})] = \left( \sum_{k=1}^K n_k^{(r)} \log n_k^{(r)} \right) - N \log N$$

with  $n_k^{(r)} = \sum_{i=1}^N q_{Z_i}^{(r)}(k)$  .

$$\begin{aligned} E_{q_{\mu,\mathbf{A}}^{(r)}}[\log p(\boldsymbol{\mu}|\mathbf{A}; \mathbf{D}^{(r)})] &= -1/2 \sum_{k=1}^K M \log 2\pi - \log |\boldsymbol{\Lambda}_k| - \tilde{\rho}_k^{(r)} \\ &\quad + (\tilde{m}_k^{(r)} - m_k)^T \mathbf{D}_k^{(r)} \boldsymbol{\Lambda}_k \tilde{\mathbf{A}}_k^{(r)} \mathbf{D}_k^{(r)T} (\tilde{m}_k^{(r)} - m_k) + \text{trace}(\boldsymbol{\Lambda}_k (\tilde{\mathbf{N}}_k^{(r)})^{-1}) \end{aligned}$$

where  $\tilde{\mathbf{N}}_k^{(r)}$ ,  $\tilde{m}_k^{(r)}$  are given in (15) and (14),  $\tilde{\rho}_k^{(r)}$  is given in (22) and  $\tilde{\mathbf{A}}_k^{(r)}$  in (23).

$$E_{q_{\mathbf{A}}^{(r)}}[\log p(\mathbf{A})] = \sum_{k=1}^K \sum_{m=1}^M \left( \lambda_{km} \log \delta_{km} - \log \Gamma(\lambda_{km}) + (\lambda_{km} - 1)(\Upsilon(\tilde{\lambda}_{km}^{(r)}) - \log \tilde{\delta}_{km}^{(r)}) - \delta_{km} \tilde{A}_{km}^{(r)} \right)$$

with  $\tilde{A}_{km}^{(r)}$  given in (23),  $\tilde{\lambda}_{km}^{(r)}$  and  $\tilde{\delta}_{km}^{(r)}$  in (16) and (17).

## B Free energy expression with a Dirichlet prior on the mixture weights

### B.1 Entropy terms

The entropy terms are the same as in the previous Section with an additional term that corresponds to the entropy of  $q_\pi^{(r)}$ :

$$H[q_\pi^{(r)}] = \log B(\tilde{\boldsymbol{\tau}}^{(r)}) - (K - \tilde{\tau}_0^{(r)})\Upsilon(\tilde{\tau}_0^{(r)}) - \sum_{k=1}^K (\tilde{\tau}_k^{(r)} - 1)\Upsilon(\tilde{\tau}_k^{(r)})$$

where  $B(\tilde{\boldsymbol{\tau}}^{(r)}) = \frac{\prod_{k=1}^K \Gamma(\tilde{\tau}_k^{(r)})}{\Gamma(\tilde{\tau}_0^{(r)})}$  and  $\tilde{\tau}_0^{(r)} = \sum_{k=1}^K \tilde{\tau}_k^{(r)}$ .

### B.2 M-step terms

Similarly to the previous Section A, the term  $E_{q^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2^{(r)})]$  decomposes now into six terms,

$$\begin{aligned} E_{q^{(r)}}[\log p(\mathbf{y}, \mathbf{X}, \Phi^1; \Phi^2^{(r)})] &= E_q[\log p(\mathbf{y}|\mathbf{X}, \Phi^1; \mathbf{D}^{(r)})] + E_{q_{\mathbf{Z}, \mathbf{W}}^{(r)}}[\log p(\mathbf{W}|\mathbf{Z}; \boldsymbol{\alpha}^{(r)})] + E_{q_{\mathbf{Z}}^{(r)} q_\pi^{(r)}}[\log p(\mathbf{Z}; \boldsymbol{\pi})] \\ &\quad + E_{q_{\boldsymbol{\mu}, \mathbf{A}}^{(r)}}[\log p(\boldsymbol{\mu}|\mathbf{A}; \mathbf{D}^{(r)})] + E_{q_{\mathbf{A}}}^{(r)}[\log p(\mathbf{A})] + E_{q_\pi^{(r)}}[\log p(\boldsymbol{\pi}; \boldsymbol{\tau})] . \end{aligned}$$

where the last term is an additional term not present in Section A and the third term has changed and is now

$$E_{q_{\mathbf{Z}}^{(r)} q_\pi^{(r)}}[\log p(\mathbf{Z}; \boldsymbol{\pi})] = \left( \sum_{k=1}^K n_k^{(r)} \Upsilon(\tilde{\tau}_k^{(r)}) \right) - N \Upsilon(\tilde{\tau}_0^{(r)})$$

The new term is,

$$E_{q_\pi^{(r)}}[\log p(\boldsymbol{\pi}; \boldsymbol{\tau})] = -\log B(\boldsymbol{\tau}) + (K - \tau_0)\Upsilon(\tilde{\tau}_0^{(r)}) + \sum_{k=1}^K (\tau_k - 1)\Upsilon(\tilde{\tau}_k^{(r)})$$

where  $\tau_0 = \sum_{k=1}^K \tau_k$ . All other terms have already been computed in Section A.

## References

- C. Archambeau and M. Verleysen. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.
- H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 21–30, 1999.
- J. Banfield and A.E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3): 803–821, 1993.
- R. Browne and P. McNicholas. Orthogonal stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, 24, 03 2014.
- G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.
- A. Corduneanu and C. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, page 2734. Morgan Kaufmann, January 2001.

- D. B Dahl. Model-based clustering for expression data via a Dirichlet process mixture model, in Bayesian Inference for Gene Expression and Proteomics. 2006.
- T. Eltoft, T. Kim, and T-W. Lee. Multivariate Scale Mixture of Gaussians Modeling. In Justinian Rosca, Deniz Erdogmus, Jose Principe, and Simon Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 799–806. Springer Berlin / Heidelberg, 2006.
- M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.
- B. N. Flury and W. Gautschi. An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.
- F. Forbes and D. Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.
- B. C. Franczak, C. Tortora, R. P. Browne, and P. D. McNicholas. Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters*, 58:69–76, 2015.
- A. Fritsch and K. Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, 06 2009.
- S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Verlag, 2006.
- P. D. Hoff. A Hierarchical Eigenmodel for Pooled Covariance Estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(5):971–992, 2009.
- N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, vol.2, 2nd edition*. John Wiley & Sons, New York, 1994.
- G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1):303–324, Jan 2016.
- C. A. McGrory and D. M. Titterton. Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. *Comput. Stat. Data Anal.*, 51(11):5352–5367, July 2007.
- S. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.
- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.
- L. Scrucca, M. Fop, T. B. Murphy, and A.E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.
- M. Stephens. Bayesian analysis of mixture models with an unknown number of componentsan alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 02 2000.
- K. Tu. Modified Dirichlet Distribution: Allowing Negative Parameters to Induce Stronger Sparsity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1986–1991, 2016.
- J. Verbeek, N. Vlassis, and B. Kröse. Efficient Greedy Learning of Gaussian Mixture Models. *Neural Computation*, 15(2):469–485, 2003.



D. Wraith and F. Forbes. Location and scale mixtures of Gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics & Data Analysis*, 90:61–73, 2015.