Bayesian mixtures of multiple scale distributions

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

Noname manuscript No.
(will be inserted by the editor)

Bayesian mixtures of multiple scale distributions

Alexis Arnaud · Florence Forbes · Russel Steele · Benjamin Lemasson · Emmanuel Barbier

the date of receipt and acceptance should be inserted later

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 procedures that can be carried out in a single run of the inference scheme, in contrast to the more costly comparison of information criteria. Preliminary results on simulated and real data show promising performance in terms of selection and computation time.

Keywords Gaussian scale mixture · Bayesian analysis · Bayesian model selection · EM algorithm · 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 family

F. Forbes, A. Arnaud
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
E-mail: firstname.lastname@inria.fr

R. Steele
McGill, Montreal, Canada
E-mail: steele@math.mcgill.ca

B. Lemasson, E. Barbier
Grenoble Institut des Neurosciences, Inserm U1216, Univ. Grenoble Alpes, France
E-mail: firstname.lastname@univ-grenoble-alpes.fr

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, Browne and McNicholas, 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 behaviors 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 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 for the advantages that the Bayesian framework can offer 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.

However, the Bayesian formulation is more involved as it requires the additional specification of priors on the parameters and the computation of posterior distributions which are often not available in closed-form. For standard scale mixtures of Gaussians, the usual Normal-Wishart prior can be used for the Gaussian parameters. In contrast, for multiple scale distributions, the decomposition of the scale matrix in the model definition (see (2) below) requires separated priors on the eigenvectors and eigenvalues of the matrix. Such priors do not derive easily from a standard conjugate choice. We therefore propose another solution and the corresponding inference scheme based on a variational approximation of the posterior distributions. To illustrate the proposed Bayesian implementation, we consider the task of selecting the number of components in multiple scale distribution mixtures. Although standard information criterion comparison can be applied to a range of such mixtures, the goal is to avoid repetitive inference and comparison of models. Following common practice that is to start from deliberately overfitting mixtures (*e.g.* Malsiner-Walli et al. [2016], Corduneanu and Bishop [2001], McGrory and Titterington [2007], Attias [1999]), we investigate the component-elimination property of the Bayesian setting. We propose two different strategies that make use of this component elimination property to select the number of components from a single run of the inference scheme.

The rest of the paper is organized as follows. The multiple scale distributions are briefly recalled in Section 2. Mixture of these distributions and their Bayesian formulation are specified in Section 3. A variational approximation inference is detailed in Section 4. A possible use for the number of mixture components selection and the two proposed strategies are described in Section 5, illustrated with experiments on simulated and real data in Section 6.

2 Multiple scale mixture distributions

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

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

where $\mathcal{N}_M(\ .\ ;\boldsymbol{\mu},\boldsymbol{\Sigma}/w)$ 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 ν denotes the degrees of freedom (we shall denote the Gamma distribution when the variable is X by $\mathcal{G}(x;\alpha,\gamma) = x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma x)\gamma^\alpha$ where Γ denotes the Gamma function). For this form, (1) is the 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 ν (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 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 $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$, where $\boldsymbol{D}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues. This spectral decomposition is classically used in Gaussian model-based clustering [Banfield and Raftery, 1993, Celeux and Govaert, 1995]. The matrix $\boldsymbol{D}$ determines the orientation of the Gaussian and $\boldsymbol{A}$ its shape. Using this parameterization of $\boldsymbol{\Sigma}$, the scale Gaussian part in (1) is set to $\mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)$, where $\boldsymbol{\Delta}_{\mathbf{w}} = \mathrm{diag}(w_1^{-1},\ldots,w_M^{-1})$ is the $M \times M$ diagonal matrix whose diagonal components are the inverse weights $\{w_1^{-1},\ldots,w_M^{-1}\}$. The multiple scale generalization consists therefore of:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)\ f_{\mathbf{w}}(w_1 \ldots w_M;\boldsymbol{\theta})\ \mathrm{d}w_1 \ldots \mathrm{d}w_M \tag{2}$$

where $f_{\mathbf{w}}$ is now a M-variate density function depending on some parameter $\boldsymbol{\theta}$ to be further specified. In the following developments, we will consider only independent weights, *i.e.* $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}$ with $f_{\mathbf{w}}(w_1 \ldots w_M;\boldsymbol{\theta}) = f_{W_1}(w_1;\boldsymbol{\theta}_1)\ldots f_{W_M}(w_M;\boldsymbol{\theta}_M)$. For instance, setting $f_{W_m}(w_m;\theta_m)$ 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 the 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 $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_M\}$, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$ and $\boldsymbol{\gamma} = \{\gamma_1 \ldots \gamma_M\}$:

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

Similarly,

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

However for identifiability, model (3) needs to be further specified by fixing all γ_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 columns of $\boldsymbol{D}, \boldsymbol{A}$ 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 $\boldsymbol{A}$. In a Bayesian setting one way to solve the problem is to impose on $\boldsymbol{A}$ a non symmetric prior (see Section 3.1). An appropriate prior on $\boldsymbol{D}$ would be more difficult to set.

3 Bayesian mixtures of multiple scale distributions

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

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

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

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

where $\boldsymbol{\Phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$ with $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_K\}$ denotes the mixture parameters. Additional variables can be introduced to identify the class labels: $\{Z_1, \ldots, Z_N\}$

define respectively the components of origin of $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. An equivalent modelling is therefore:

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

where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}^{-1}, \ldots, w_{iM}^{-1})$, symbol $\otimes$ means that the components of $\mathbf{W}_i$ are independent and $\mathcal{M}(1, \pi_1, \ldots, \pi_k)$ denotes the Multinomial distribution. In what follows, the weight variables will be denoted by $\mathbf{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_N\}$ and the labels by $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$.

3.1 Priors on component-specific parameters

To complete the Bayesian formulation, we assign priors on parameters in $\boldsymbol{\psi}$. 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 $\boldsymbol{D}_k, \boldsymbol{A}_k$. For the means $\boldsymbol{\mu}_k$, the standard Gaussian prior can be used:

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

where $\boldsymbol{m}_k$ (vector) and $\boldsymbol{\Lambda}_k$ (diagonal matrix) are hyperparameters and we shall use the notation $\boldsymbol{m} = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_1, \ldots \boldsymbol{\Lambda}_K\}$. For $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ a natural solution would be to use the distributions induced by the standard Wishart prior on $\boldsymbol{T}_k$ but this appears not to be tractable in inference scheme based on a variational framework. The difficulty lies in considering an appropriate and tractable prior for $\boldsymbol{D}_k$. 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-$\boldsymbol{\Phi}^1$ step (see Section 4) that could provide a variational posterior distribution. Nevertheless, this kind of priors could be added in the M-$\boldsymbol{D}$-step. The simpler solution adopted in the present work consists of considering $\boldsymbol{D}_k$ as an unknown fixed parameter and imposing a prior only on $\boldsymbol{A}_k$, which is a diagonal matrix containing the positive eigenvalues of $\boldsymbol{T}_k$. It is natural to choose:

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

where $\boldsymbol{\lambda}_k = \{\lambda_{km}, m = 1 \ldots M\}$ and $\boldsymbol{\delta}_k = \{\delta_{km}, m = 1 \ldots M\}$ are hyperparameters with $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_K\}$ and $\boldsymbol{\delta} = \{\boldsymbol{\delta}_1, \ldots \boldsymbol{\delta}_K\}$ as additional notation. It follows the joint prior on $\boldsymbol{\mu}_{1:K} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, $\boldsymbol{A}_{1:K} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K\}$ given $\boldsymbol{D}_{1:K} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K\}$

$$p(\boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{A}_k; \boldsymbol{D}_k) \, p(\boldsymbol{A}_k) \tag{8}$$

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

3.2 Priors on mixing weights

As it will become clearer in Section 5 devoted to the selection of the number of components, we consider two cases for the mixing weights. First, following Corduneanu and Bishop [2001], no prior is imposed on $\boldsymbol{\pi}$ (Section 4.1). Then a standard Dirichlet prior $\mathcal{D}(\tau_1, \ldots, \tau_K)$ is used in a second case with $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K\}$ the Dirichlet parameters (Section 4.2). Note that the no prior case is actually equivalent to choose a uniform prior $\mathcal{D}(1, \ldots, 1)$ and to estimate $\boldsymbol{\pi}$ as the maximum a posteriori (MAP), while in the Bayesian setting, a full posterior is computed for $\boldsymbol{\pi}$ and point estimates rather derived using the posterior mean.

For the complete model, the whole set of parameters is denoted by $\boldsymbol{\Phi}$. In the first setting, $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}^1, \boldsymbol{\Phi}^2\}$ is decomposed into a set $\boldsymbol{\Phi}^1 = \{\boldsymbol{\Phi}^1_1, \ldots \boldsymbol{\Phi}^1_K\}$ with $\boldsymbol{\Phi}^1_k = \{\boldsymbol{\mu}_k, \boldsymbol{A}_k\}$ of parameters for which we have priors and a set $\boldsymbol{\Phi}^2 = \{\boldsymbol{\Phi}^2_1, \ldots \boldsymbol{\Phi}^2_K\}$ with $\boldsymbol{\Phi}^2_k = \{\pi_k, \boldsymbol{D}_k, \boldsymbol{\alpha}_k\}$ of unknown parameters considered as fixed. In addition, hyperparameters are denoted by $\boldsymbol{\Phi}^3 = \{\boldsymbol{\Phi}^3_1, \ldots \boldsymbol{\Phi}^3_K\}$ with $\boldsymbol{\Phi}^3_k = \{\boldsymbol{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$. When a Dirichlet prior is used for $\boldsymbol{\pi}$, the parameters definitions change to $\boldsymbol{\Phi}^1_k = \{\boldsymbol{\mu}_k, \boldsymbol{A}_k, \pi_k\}$, $\boldsymbol{\Phi}^2_k = \{\boldsymbol{D}_k, \boldsymbol{\alpha}_k\}$ and $\boldsymbol{\Phi}^3_k = \{\tau_k, \boldsymbol{m}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\lambda}_k, \boldsymbol{\delta}_k\}$.

4 Inference using variational Expectation-Maximization

The main task in Bayesian inference is to compute the posterior probability of the latent variables $\boldsymbol{X} = \{\boldsymbol{W}, \boldsymbol{Z}\}$ and the parameter $\boldsymbol{\Phi}$ for which only the $\boldsymbol{\Phi}^1$ part is considered as random. We are therefore interested in computing the posterior $p(\boldsymbol{X}, \boldsymbol{\Phi}^1 \mid \boldsymbol{y}, \boldsymbol{\Phi}^2)$. This posterior is intractable and approximated here using a variational approximation $q(\boldsymbol{X}, \boldsymbol{\Phi}^1)$ with a factorized form $q(\boldsymbol{X}, \boldsymbol{\Phi}^1) = q_X(\boldsymbol{X})\ q_{\Phi^1}(\boldsymbol{\Phi}^1)$ in the set $\mathcal{D}$ of product probability distributions. The so-called variational EM procedure (VEM) proceeds as follows. At iteration (r), the current parameters values are denoted by $\boldsymbol{\Phi}^{2(r-1)}$ and VEM alternates between two steps,

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

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

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

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. It provides a stopping criterion and a sanity check for implementation as the free energy should increase at each iteration. Then it can be used as specified in section 5.2 as a replacement of the likelihood to provide a model

selection procedure. The detailed expression is given in Appendices C and D, 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 $\boldsymbol{X}$:

$$\textbf{E-}\boldsymbol{\Phi}^1\textbf{-step:} \qquad q_{\Phi^1}^{(r)}(\boldsymbol{\Phi}^1) \propto \exp(E_{q_X^{(r-1)}}[\log p(\boldsymbol{\Phi}^1|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\Phi}^{2(r-1)})]) \tag{10}$$

$$\textbf{E-X-step:} \qquad q_X^{(r)}(\boldsymbol{X}) \propto \exp(E_{q_{\phi^1}^{(r)}}[\log p(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r-1)})]) \; . \tag{11}$$

Then the M-step reduces to:

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

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

4.1 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 $p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3)$ writes as

$$p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K})\; p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}_{1:K})\; p(\boldsymbol{Z}; \boldsymbol{\pi})\; p(\boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K}, \boldsymbol{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

Applying the previous general formulas (10) and (11), the 2 sub-steps E-$\boldsymbol{\Phi}^1$ and E-$\boldsymbol{X}$ steps provide respectively the variational posterior of $\boldsymbol{\Phi}^1$, which has the same structure as the prior distribution (8),

$$q_{\Phi^1}^{(r)}(\boldsymbol{\Phi}^1) = \prod_{k=1}^{K} q_{\mu_k|A_k}^{(r)}(\boldsymbol{\mu}_k \mid \boldsymbol{A}_k)\; q_{A_k}^{(r)}(\boldsymbol{A}_k) \; ,$$

and the variational posterior of $\boldsymbol{X}$ in the form

$$q_X^{(r)}(\boldsymbol{X}) = \prod_{i=1}^{N} q_{X_i}^{(r)}(\boldsymbol{W}_i, \boldsymbol{Z}_i) = \prod_{i=1}^{N} q_{W_i|Z_i}^{(r)}(\boldsymbol{W}_i|\boldsymbol{Z}_i) q_{Z_i}^{(r)}(\boldsymbol{Z}_i) \; .$$

The detailed expressions are given in Appendix A.

The M-step divides into 3 sub-steps, where $\boldsymbol{\pi}, \boldsymbol{D}_{1:K}$ and $\boldsymbol{\alpha}_{1:K}$ are updated separately. Denoting by $n_k^{(r)}$ the sum $\sum_{i=1}^{N} q_{Z_i}^{(r)}(k)$, the M-$\boldsymbol{\pi}$-step leads to the standard formula for mixtures. For $k = 1 \dots K$, π_k is updated as:

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

The other sub-steps are given in Appendix A.

4.2 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

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{\tau}, \boldsymbol{D}_{1:K}) = p(\boldsymbol{\pi}; \boldsymbol{\tau}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{A}_k; \boldsymbol{D}_k) p(\boldsymbol{A}_k) \tag{12}$$

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

With this modification, only the $p(\boldsymbol{Z}; \boldsymbol{\pi})$ term changes into $p(\boldsymbol{Z}|\boldsymbol{\pi})\ p(\boldsymbol{\pi}; \boldsymbol{\tau})$ in the complete likelihood $p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2, \boldsymbol{\Phi}^3)$ which becomes,

$$p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K})\ p(\boldsymbol{W}|\boldsymbol{Z}; \boldsymbol{\alpha}_{1:K})\ p(\boldsymbol{Z}|\boldsymbol{\pi})\ p(\boldsymbol{\pi}; \boldsymbol{\tau})\ p(\boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K}, \boldsymbol{m}, \boldsymbol{\Lambda}, \boldsymbol{\lambda}, \boldsymbol{\delta}).$$

For the E-$\boldsymbol{\Phi}^1$ step, the variational posterior has the same form as the prior (12) with the variational posterior for $\boldsymbol{\pi}$ denoted by $q_{\boldsymbol{\pi}}^{(r)}(\boldsymbol{\pi})$,

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

where $q_{\mu_k, \boldsymbol{A}_k}^{(r)}(\boldsymbol{\mu}_k, \boldsymbol{A}_k)$ has the same expression as given by (16) and (17) in Appendix A. The new term is

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

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

The E-$\boldsymbol{X}$-step is given in Apprendix B. It is only partly impacted by the addition of a prior on $\boldsymbol{\pi}$, which changes only the expression of $q_{Z_i}^{(r)}$. The M-step formulation remains the same as before without the M-$\boldsymbol{\pi}$ step.

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

5 Identifying the number of mixture components

A difficult problem when fitting mixture models is to determine the number K of components to include in the mixture. A recent review on the problem with theoretical and practical aspects can be found in Celeux et al. [2018]. Traditionally, this selection is performed by comparing a set of candidate models for a range of values of K, assuming that the true value is in this range. The number of components is selected by minimizing a model selection criterion, such as the Bayesian inference criterion (BIC), minimum message length (MML), Akaike's information criteria (AIC) to cite just a few [McLachlan and Peel,

2000, Figueiredo and Jain, 2002]. Of a slightly different nature is the so-called slope heuristic [Baudry et al., 2012], which involves a robust linear fit and is not simply based on criterion comparisons. However, the disadvantage of these approaches is that a whole set of candidate models has to be obtained and problems associated with running inference algorithms (such as EM) many times may emerge. Alternatives have been investigated that select the number of components from a single run of the inference scheme. 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 to 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. More recently, there seems to be an increase interest among mixture model practitioners for model selection strategies that start instead with a large number of components and merge them [Hennig, 2010]. For instance, Figueiredo and Jain [2002] propose a practical algorithm that starts with a very large number of components, iteratively annihilates components, redistributes the observations to the other components, and terminates based on the MML criterion. The approach in Baudry et al. [2010] starts with an overestimated number of components using BIC, and then merges them hierarchically according to an entropy criterion, while Melnykov [2014] proposes a similar method that merges components based on measuring their pair-wise overlap. Another trend in handling the issue of finding the proper number of components is to consider Bayesian non-parametric mixture models. This allows the implementation of mixture models with an infinite number of components via the use of Dirichlet process mixture models (*e.g.* Rasmussen, Gorur and Rasmussen [2010], Yerebakan et al. [2014], Wei and Li [2012]). The Bayesian non-parametric approach is a promising technique. We consider a Bayesian formulation but in the simpler case of a finite number of components. We suspect all our Bayesian derivations could be easily tested in a non parametric setting with some minor adaptation left for future work.

We consider approaches that start from an overfitting mixture 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 mix-

ture 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, 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. Considering that no prior on $\boldsymbol{\pi}$ in this case is equivalent to a uniform Dirichlet prior $\mathcal{D}(1, \ldots, 1)$ if the maximum a posteriori is used, this is not surprising and what Corduneanu and Bishop [2001] observed is that the penalization is visible on the MAP. This justifies in our setting the investigation of a case with no prior on the mixing weights (Section 4.1).

However, 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 $\mathcal{D}(\tau_1, \ldots, \tau_K)$ with all τ_k's equal to a value τ_0. To empty superfluous components automatically the value of τ_0 has to be chosen appropriately. In particular, Rousseau and Mengersen [2011] proposed conditions on τ_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 τ_0 are needed (*e.g.* 10^{-5}). It was even observed by Tu [2016] that negative values of τ_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.

5.1 Single-run number of component selection

In a Bayesian setting with symmetric sparse Dirichlet priors $\mathcal{D}(\tau_0, \ldots, \tau_0)$, the theoretical study of Rousseau and Mengersen [2011] justifies to consider the

posterior expectations of the weights $E[\pi_k|\boldsymbol{y}]$ 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 $E[\pi_k|\boldsymbol{y}]$ takes the following form (the (r) notation is removed to signify the convergence of the algorithm),

$$\begin{aligned} E[\pi_k|\boldsymbol{y}] \approx E_{q_{\boldsymbol{\pi}}}[\pi_k] &= \frac{\tilde{\tau}_k}{\sum_{l=1}^{K} \tilde{\tau}_l} \\ &= \frac{\tau_k + n_k}{\sum_{k=1}^{K} \tau_k + N} \end{aligned} \tag{13}$$

where $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$ and $q_{Z_i}(k)$ is given by (30) (see Appendix A). $q_{Z_i}(k)$ is the variational a posteriori probability that observation $\boldsymbol{y}_i$ comes from component k and n_k can be interpreted as the estimated size of component k. If we are in the no weight prior case, then the expectation simplifies to

$$\pi_k \approx \frac{n_k}{N} \tag{14}$$

with $q_{Z_i}(k)$ given by (25) (see Appendix B).

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 π_k and the number of observations assigned to component k. This later number is itself often replaced by n_k. As an illustration, the choice of a negative τ_0 by Figueiredo and Jain [2002] corresponds to a rule that sets a component weight to zero when n_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 Titterington [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.* $n_k/N < 10^{-5}$ where N varies from 400 to 900 in their experiments. Note that McGrory and Titterington [2007] use a Bayesian framework with variational inference and their rule corresponds to thresholding the variational posterior weights (13) to $1/N$ because they set all τ_k to 0 in their experiments.

In addition to these thresholding approaches, 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. However, even for moderate numbers of components, the optimization is already very costly.

In this work, we consider two strategies for component elimination. The first one is a thresholding approach while the second one is potentially more general as it is based on increasing the overall fit of the model assessed via the variational free energy at each iteration. Both these strategies lead themselves to two variants depending on whether or not a prior is used for the mixing weights. The tested procedures are more specifically described in the next section.

5.2 Tested procedures

We compare three types of single-run methods to estimate the number of components in a mixture of multiple scale distributions. The first two types correspond to a thresholding strategy but for two different Bayesian models (Sparse Dirichlet prior or TypeII ML).

5.2.1 Bayesian algorithm with sparse Dirichlet prior: "SparseDirichlet"

A first method is directly derived from a 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 Titterington [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),

$$n_k^{(r)} < (K\tau_0 + N)\rho_t - \tau_0 \tag{15}$$

where ρ_t is the chosen threshold on the posterior means. When ρ_t is set such that (15) leads to $n_k^{(r)} < 1$, this method is referred to, in the next Section, as *SparseDirichlet+πtest*. For comparison, the algorithm run with no intervention is called *SparseDirichlet*.

5.2.2 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 (15) with $\tau_0 = 0$. This method is referred to below as *TypeIIML* when the algorithm is run until convergence and *TypeIIML+πtest* when (15) is used at each iteration with $\rho_t = 1/N$.

5.2.3 Free Energy based algorithm: "FEtest"

At last, we consider a criterion based on the free energy (9) to detect components to eliminate. This choice is based on the observation that when no prior is used for the weights, we cannot control the hyperparameters (*e.g* τ_k) to guide the algorithm in the vanishing components regime. 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 can also be redundant ones. In the proposed variational EM inference framework, the free energy arises naturally as a selection criterion. It has been stated in Attias [2000] and Beal [2003] that the free energy penalizes model complexity and that it converges to the well known Bayesian Information Criterion (BIC) and Minimum Description Length (MDL) criterion, when the sample size increases, illustrating the interest of this measure for model selection.

The free energy expressions used are given in Appendices C and D. 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*).

6 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 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*.

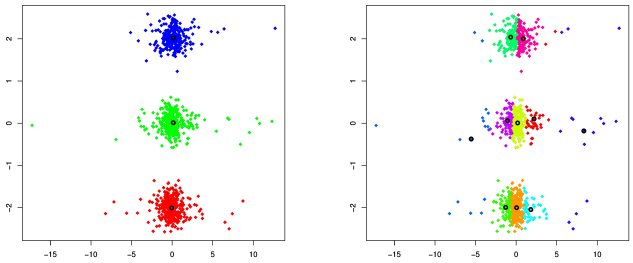
In practice, values need to be chosen for hyperparameters. These include the $\boldsymbol{m}_k$ that are set to 0, the $\boldsymbol{\Lambda}_k$ that are set to $\epsilon \boldsymbol{I}_M$ with ϵ small (set to 10^{-4}) so has to generate a large variance in (6). The δ_{km} are then set to 1 and λ_{km}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

Initialization is also an important step in EM algorithms. For one data sample, each single-run method is initialized $I = 10$ times. These $I = 10$ initializations are the same for all single-run methods. Each initialization is obtained with $K = 10$ using trimmed k-means and excluding 10% of outliers. Each trimmed kmeans output is the one obtained after running the algorithm from $R = 10$ restarts and selecting the best assignment after 10 iterations. For each run of a procedure (data sample), the $I = 10$ initializations are followed by 5000 iterations maximum of VEM before choosing the best output. For Gaussian mixtures, the initialization procedure is that embedded in Mclust. For $\mathcal{MP}$ models, initial values of the α_{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.

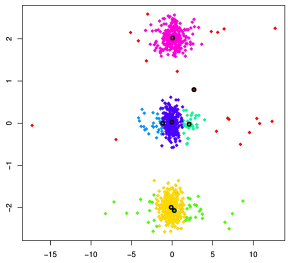
6.1 Simulated data

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 $\alpha_{k1} = 2$ 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 Titterington [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 $K = 4$ is the most selected value by BIC with sometimes values up to 7 selected. 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)). For the MMP case and the

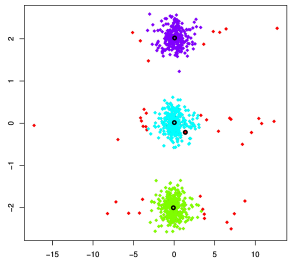


(a) Simulated data

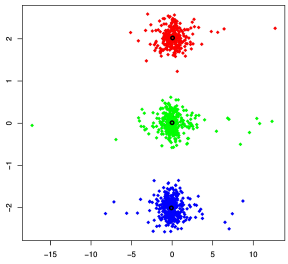
(b) Trimmed kmeans ($K = 10$)



(c) *Bayesian GM+BIC* ($K = 7$)



(d) *GM+BIC* ($K = 4$)



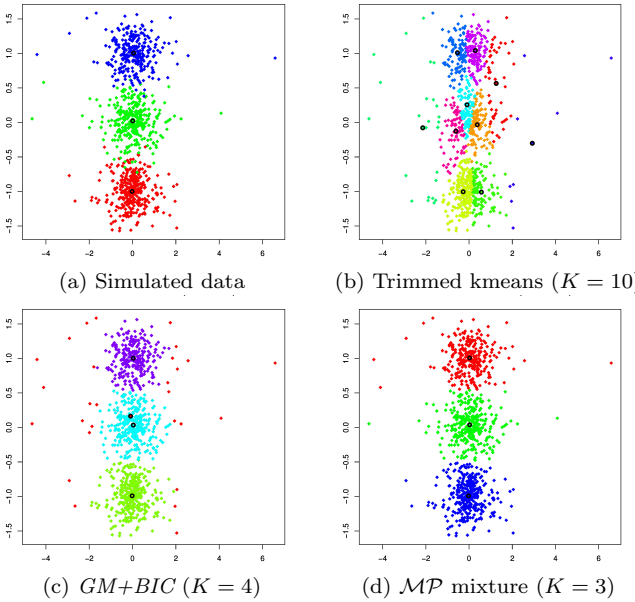
(e) $\mathcal{MP}$ mixture ($K = 3$)

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

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+πtest* case). Mean computational times over the 10 samples are reported in seconds in Table 1 (last column). These mean times are all including the $I = 10$ repetitions. They are only indicative because implementations differ significantly between Mclust, MMP and the other $\mathcal{MP}$ single-run procedures. Mclust is the fastest. MMP is not the slowest but this is due to a more optimized implementation than the other Bayesian models. However, despite implementation differences, 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

Procedure (10 restarts)	Selected number of components										Average time (in seconds)
	1	2	3	4	5	6	7	8	9	10	
GM+BIC	.	.	**2**	**8**	.	.	.	.	.	.	16
BayesianGM+BIC	.	.	**4**	3	1	1	1	.	.	.	34
MMP+BIC	.	.	**10**	.	.	.	.	.	.	.	3589
TypeIIML	.	.	**8**	2	.	.	.	.	.	.	5115
TypeIIML+πtest	.	.	**9**	1	.	.	.	.	.	.	1349
TypeIIML+FEtest	.	.	**10**	.	.	.	.	.	.	.	1391
SparseDirichlet	.	.	**9**	1	.	.	.	.	.	.	5051
SparseDirichlet+πtest	.	.	**10**	.	.	.	.	.	.	.	1326
SparseDirichlet+FEtest	.	.	**10**	.	.	.	.	.	.	.	1032

Table 1 Two dimensional $\mathcal{MP}$ mixture with $K = 3$ well separated components. Final observed or selected number of components for each procedure on 10 samples, and mean computational times (over the 10 samples) in seconds (for the total of the $I = 10$ repetitions). The most frequent selection is indicated by a box while the true value is in green.



(a) Simulated data (b) Trimmed kmeans ($K = 10$)

(c) *GM+BIC* ($K = 4$) (d) $\mathcal{MP}$ mixture ($K = 3$)

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

provide the largest gain with a running time (1032s) divided by more than 3 compared to the *MMP+BIC* procedure (3589s).

A second example consists of 3 similar $\mathcal{MP}$ distributions but with closer means namely $[0, -1]$, $[0, 0]$ and $[0, 1]$ (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 $K = 3$.

Procedure (10 restarts)	Selected number of components										Average time (in seconds)
	1	2	3	4	5	6	7	8	9	10	
GM+BIC	.	.	1	9	.	.	.	.	.	.	86
BayesianGM+BIC	.	.	1	9	.	.	.	.	.	.	41
MMP+BIC	.	.	10	.	.	.	.	.	.	.	2942
TypeIIML	.	.	10	.	.	.	.	.	.	.	4892
TypeIIML+πtest	.	.	10	.	.	.	.	.	.	.	1471
TypeIIML+FEtest	.	.	10	.	.	.	.	.	.	.	1336
SparseDirichlet	.	.	10	.	.	.	.	.	.	.	5004
SparseDirichlet+πtest	.	.	10	.	.	.	.	.	.	.	1311
SparseDirichlet+FEtest	.	.	10	.	.	.	.	.	.	.	1334

Table 2 Two dimensional $\mathcal{MP}$ mixture with $K = 3$ closer components. Final observed or selected number of components for each procedure on 10 samples, and mean computational times (over the 10 samples) in seconds (for the total of the $I = 10$ repetitions). The most frequent selection is indicated by a box while the true value is in green.

We consider several other models, 3 Gaussian mixtures and 10 $\mathcal{MP}$ mixtures, with 10 simulated samples each, for a total of 130 samples, K varying from 3 to 5, N from 900 to 9000, with close or more separated clusters. The results are summarized in Table 3 and the simulated samples illustrated in Figure 3. Gaussian mixture models provide the right component number in 26% to 32% of the cases, which is higher than the number of Gaussian mixtures in the test (23%). All procedures hesitate mainly between the true number and this number plus 1. We observe a good behavior of the free energy heuristic in both fully Bayesian and TypeII ML cases with a time divided by 3 compared to the non Bayesian $\mathcal{MP}$ mixture procedure, although the later benefits from a more optimized implementation. Component elimination procedures based on proportions (πtest) are less successful maybe due to slower convergence. Their dependence to the choice of a threshold value is certainly a limitation although some significant gain is observed over the cases with no component elimination (TypeIIML and SparseDirichlet lines in Table 3). Overall, eliminating components on the run is beneficial, both in terms of time and selection performance but using a penalized likelihood criterion (free energy) to do so avoid the commitment to a fix threshold and is more successful. A possible reason is that small components are more difficult to eliminate than redundant ones. Small components not only require the right threshold to be chosen but also they may appear at much latter iterations as illustrated in Figure 4.

6.2 Standard dataset

The procedures are also illustrated on a standard data set in more than 1 dimension and for which the results are easy to interpret.

Procedures (10 restarts)	Difference between selected and true number of components								Average time (in seconds)
	0	1	2	3	4	5	6	7	
GM+BIC	26.1	33.0	8.4	3.8	19.2	1.5	2.3	5.3	177
Bayesian GM+BIC	31.5	34.6	3.0	3.0	20.7	3.8	1.5	1.5	92
MMP+BIC	94.6	5.3	.	.	.	.	.	.	9506
TypeIIML	45.3	41.5	12.3	.7	.	.	.	.	10473
TypeIIML+πtest	61.5	34.6	3.0	.7	.	.	.	.	4872
TypeIIML+FEtest	98.4	.7	.	.7	.	.	.	.	3975
SparseDirichlet	54.6	39.2	5.3	.7	.	.	.	.	10355
SparseDirichlet+πtest	70.0	27.6	1.5	.7	.	.	.	.	4640
SparseDirichlet+FEtest	99.2	.	.	.7	.	.	.	.	3125

Table 3 13 models simulated 10 times each: the true number of components is varying so the columns indicate the difference between the selection and the truth. The average time (for the total of the $I = 10$ repetitions, over the 130 samples) is indicated in the last column. The most frequent selection (in %) is indicated by a box while the true value is in green.

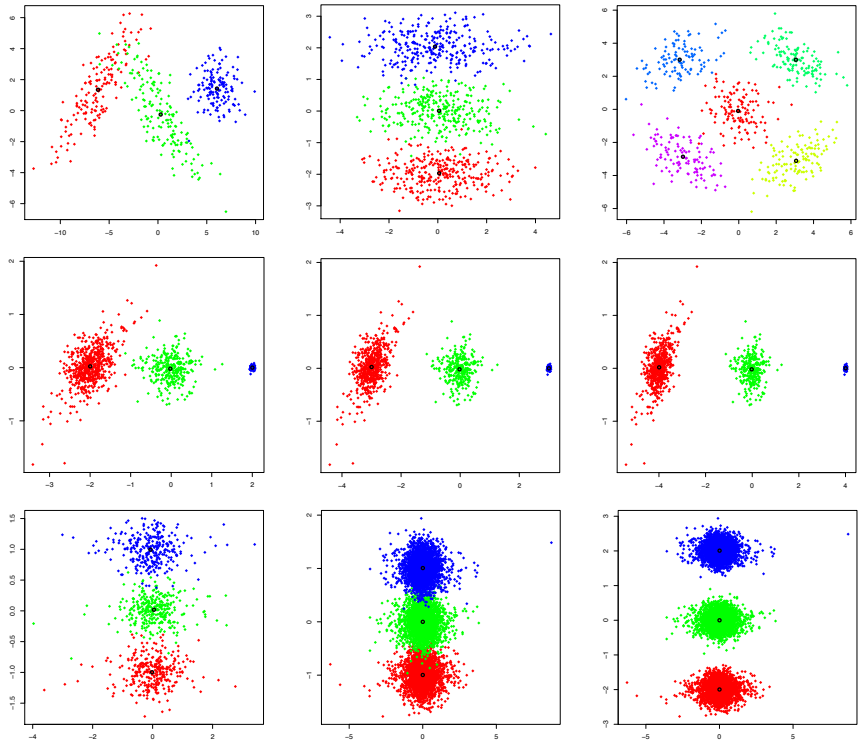


Fig. 3 Examples of simulated samples. First line: 3 Gaussian mixtures with 3 and 5 components. Second line: $\mathcal{MP}$ mixtures with different dof and increasing separation from left to right. Third line: $\mathcal{MP}$ mixtures with increasing separation, from left to right, and increasing number of points, $N = 900$ for the first plot, $N = 9000$ for the last two.

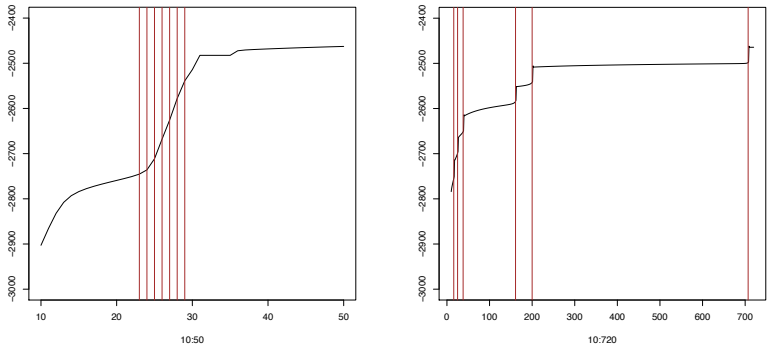


Fig. 4 Illustration of the two component elimination strategies: Free energy gain strategy, iterations 10 to 50 (left) and too small component proportion test, iterations 10 to 720 (right). Eliminations are marked with red lines. Most of them occur at earlier iterations when using the free energy test.

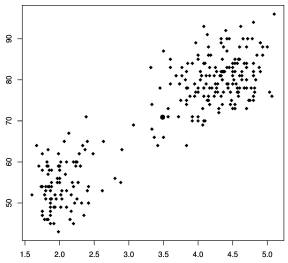
6.2.1 Old Faithful Geyser data

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 5 (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 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 (*GM+BIC* and *Bayesian GM+BIC*), finding 3 components (Figure 5 (c)) while our $\mathcal{MP}$ model with BIC (*MMP+BIC*) and the free energy based elimination procedure show consistently 2 selected clusters (Figure 5 (d)). All other methods select 3 components but the clustering differs from that of the Gaussian models (Figures 5 (c) and (e)). All procedures were initialized with a 10 component assignment similar to that shown in Figure 5 (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).

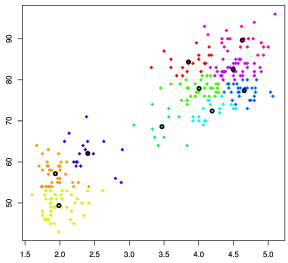
The code used for the experiments is available under the MMST item at https://team.inria.fr/mistis/software/.

7 Discussion and conclusion

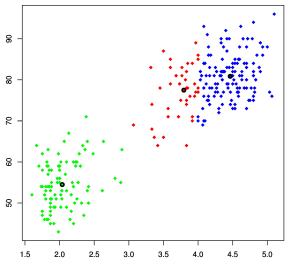
Multiple scale distributions have been shown to perform well in the modelling of non-elliptical clusters with potential outliers and tails of various heaviness. Considering a Bayesian formulation of mixtures of such multiple scale distributions, we derived an inference procedure based on a variational EM algo-



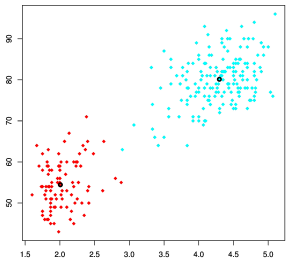
(a) Old Faithful data



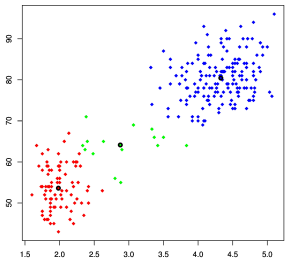
(b) Trimmed kmeans ($K = 10$)



(c) *GM+BIC* ($K = 3$)



(d) *MMP+BIC* & *FEtest* ($K = 2$)



(e) Other single-run methods ($K = 3$)

Fig. 5 (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 ($K = 3$).

rithm. Our main motivation was to investigate, in the context of mixtures of non-Gaussian distributions, different single-run procedures to select automatically the number of components. The Bayesian formulation makes this possible when starting from an overfitting mixture, where K is larger than the expected number of components. The advantage of single run procedures is to avoid time consuming comparison of scores for each mixture model from 1 to K components. 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]. For further acceleration, we investigated component elimination which consists of eliminating components on the run. They are two main ways to do so: components are eliminated as

soon as they are not supported by enough data points (their estimated weight is under some threshold) or when their removal does not penalize the overall fit. For the latest case, we proposed a heuristic based on the gain in free energy. The free energy acts as a penalized likelihood criterion and can potentially eliminate both too small components and redundant ones. Redundant components do not necessarily see their weight tend to zero and cannot be eliminated via a simple thresholding.

On preliminary experiments, we observed that eliminating components on the run is beneficial, both in terms of time and selection performance. Free energy based methods appeared to perform better than posterior weight thresholding methods: using a penalized likelihood criterion (free energy) avoids the commitment to a fix threshold and is not limited to the removal of small components. However, a fully Bayesian setting is probably not necessary as both in terms of selection and computation time, Type II maximum likelihood on the weights was competitive with the use of a Dirichlet prior with a slight advantage to the latter.

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 procedure. Theoretical justification for thresholding approaches, as provided by Rousseau and Mengersen [2011], applies for Gaussian mixtures but may not hold in our case of non-elliptical distributions. A more specific study would be required and could provide additional guidelines as how to set the threshold in practice. Also time comparison in our study is only valid for the Bayesian procedures for which the implementation is similar while the other methods using BIC have been better optimized, but this does not change the overall conclusion as regards computational efficiency.

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.

H. Attias. A variational Bayesian framework for graphical models. In *Proc. Advances in Neural Information Processing Systems 12*, pages 209–215, Denver, Colorado, United States, 2000. MIT Press.

J. Banfield and A. E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803–821, 1993.

J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

R. P. Browne and P. D. McNicholas. Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, 24, 03 2014.

R. P. Browne and P. D. McNicholas. A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2):176–198, 2015.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

G. Celeux, S. Frühwirth-Schnatter, and C. Robert. *Model Selection for Mixture Models-Perspectives and Strategies*. Handbook of Mixture Analysis, CRC press, 12 2018.

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.

D. Gorur and C.E. Rasmussen. Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

C. Hennig. Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1):3–34, 2010.

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. Titterington. Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. *Comput. Stat. Data Anal.*, 51(11):5352–5367, July 2007.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

V. Melnykov. Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics*, 2014.

C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*.

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 components: an 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.

X. Wei and C. Li. The infinite student t-mixture for robust modeling. *Signal Processing*, 92(1):224–234, 2012.

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.

H. Z Yerebakan, B. Rajwa, and M. Dundar. The infinite mixture of infinite Gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 28–36, 2014.

A No prior on mixing coefficients: E and M steps

A.1 E-step.

E-$\boldsymbol{\Phi}^1$ step. In this setting, the variational posterior has the same structure as the prior distribution (8). More specifically,
$q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1) = \prod_{k=1}^{K} q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1_k)$ with $q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\Phi}^1_k) = q^{(r)}_{\boldsymbol{\Phi}^1}(\boldsymbol{\mu}_k, \boldsymbol{A}_k) = q^{(r)}_{\boldsymbol{\mu}_k|A_k}(\boldsymbol{\mu}_k \mid \boldsymbol{A}_k)\, q^{(r)}_{A_k}(\boldsymbol{A}_k)$ where

$$q^{(r)}_{\boldsymbol{\mu}_k|A_k}(\boldsymbol{\mu}_k \mid \boldsymbol{A}_k) = \mathcal{N}(\boldsymbol{\mu}_k; \tilde{\boldsymbol{m}}^{(r)}_k, \tilde{\boldsymbol{\Sigma}}^{(r)}_k) \tag{16}$$

$$q^{(r)}_{A_k}(\boldsymbol{A}_k) = \prod_{m=1}^{M} \mathcal{G}(A_{km}; \tilde{\lambda}^{(r)}_{km}, \tilde{\delta}^{(r)}_{km})\ . \tag{17}$$

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

$$\tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} = diag(\tilde{w}^{(r-1)}_{ki1}, \ldots, \tilde{w}^{(r-1)}_{kiM}) \tag{18}$$

where $\tilde{w}^{(r-1)}_{kim} = E_{q^{(r-1)}_X}[W_{im}|Z_i = k] = \tilde{\alpha}^{(r-1)}_{km}/\tilde{\gamma}^{(r-1)}_{kim}$. The specific expressions of $\tilde{\alpha}^{(r-1)}_{km}$ and $\tilde{\gamma}^{(r-1)}_{kim}$ are given in eq. (26) and (27) of the E-$\boldsymbol{X}$ step below but using values from iteration $(r-1)$. The covariance matrix $\tilde{\boldsymbol{\Sigma}}^{(r)}_k$ depends on $\boldsymbol{A}_k$ as in the prior (6) while after simplification $\tilde{\boldsymbol{m}}^{(r)}_k$ does not:

$$\tilde{\boldsymbol{\Sigma}}^{(r)}_k = \boldsymbol{D}^{(r-1)}_k \tilde{\boldsymbol{N}}^{(r)-1} \boldsymbol{A}^{-1}_k \boldsymbol{D}^{(r-1)T}_k \tag{19}$$

$$\begin{aligned}\tilde{\boldsymbol{m}}^{(r)}_k &= \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{D}^{(r-1)}_k \boldsymbol{A}_k \left(\boldsymbol{\Lambda}_k \boldsymbol{D}^{(r-1)T}_k \boldsymbol{m}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \boldsymbol{D}^{(r-1)T}_k \boldsymbol{y}_i\right) \\ &= \boldsymbol{D}^{(r-1)}_k \tilde{\boldsymbol{N}}^{(r)-1} \left(\boldsymbol{\Lambda}_k \boldsymbol{D}^{(r-1)T}_k \boldsymbol{m}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki} \boldsymbol{D}^{(r-1)T}_k \boldsymbol{y}_i\right)\end{aligned} \tag{20}$$

$$\text{with} \qquad \tilde{\boldsymbol{N}}^{(r)}_k = \boldsymbol{\Lambda}_k + \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tilde{\boldsymbol{\Delta}}^{(r-1)}_{ki}\ . \tag{21}$$

Then (17) is defined by the parameters:

$$\tilde{\lambda}^{(r)}_{km} = \lambda_{km} + 1/2 \sum_{i=1}^{N} q^{(r-1)}_{Z_i}(k) \tag{22}$$

$$\tilde{\delta}^{(r)}_{km} = \delta_{km} + 1/2\ [\boldsymbol{M}_k]_{m,m} \tag{23}$$

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

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

E-$\boldsymbol{X}$ step. $q^{(r)}_X(\boldsymbol{X}) = \prod_{i=1}^{N} q^{(r)}_{X_i}(\boldsymbol{W}_i, \boldsymbol{Z}_i)$ with

$$q^{(r)}_{W_i|Z_i}(\boldsymbol{W}_i|Z_i = k) = \prod_{m=1}^{M} q^{(r)}_{W_{im}|Z_i}(\boldsymbol{W}_{im}|\boldsymbol{Z}_i = k) = \prod_{m=1}^{M} \mathcal{G}(\boldsymbol{W}_{im}; \tilde{\alpha}^{(r)}_{km}, \tilde{\gamma}^{(r)}_{kim}) \tag{24}$$

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

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

$$\tilde{\alpha}_{km}^{(r)} = \alpha_{km}^{(r-1)} + \frac{1}{2} \tag{26}$$

$$\tilde{\gamma}_{kim}^{(r)} = 1 + \frac{1}{2}\left(\tilde{A}_{km}^{(r)} [\boldsymbol{D}_k^{(r-1)T}(\boldsymbol{y}_i - \tilde{\boldsymbol{m}}_k^{(r)})]_m^2 + [\tilde{\boldsymbol{N}}_k^{(r)}]_{m,m}^{-1}\right) \tag{27}$$

where $\tilde{\boldsymbol{N}}_k^{(r)}$ is given in (21), $\tilde{\rho}_k^{(r)}$ and $\tilde{\boldsymbol{A}}_{km}^{(r)}$ are easily computed from (17) (Υ is the Digamma function):

$$\tilde{\rho}_k^{(r)} = E_{q_{A_k}^{(r)}}[\log |\boldsymbol{A}_k|] = \sum_{m=1}^{M} \Upsilon(\tilde{\lambda}_{km}^{(r)}) - \log \tilde{\delta}_{km}^{(r)} \tag{28}$$

$$\tilde{\boldsymbol{A}}_k^{(r)} = E_{q_{A_k}^{(r)}}[\boldsymbol{A}_k] \quad \text{that is for } m = 1 \dots M, \quad \tilde{\boldsymbol{A}}_{km}^{(r)} = \tilde{\lambda}_{km}^{(r)}/\tilde{\delta}_{km}^{(r)} . \tag{29}$$

A.2 M-step.

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 the $\boldsymbol{\alpha}_k$'s are 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 $m = 1 \dots M$:

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

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

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

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

As already defined in (18) and (21),

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

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

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

The exact computation of the expectation above is feasible but would result in an expression where the elements of $\boldsymbol{D}_k$ would be separated. As an alternative, we consider J *i.i.d* simulations of $\boldsymbol{A}_k$ according to distribution $q_{A_k}^{(r)}$ which is a product of Gamma distributions given in (17). Denoting by $\boldsymbol{A}_{kj}$ for $j = 1\dots J$ these simulations, $\boldsymbol{D}_k^{(r)}$ can be approximated by,

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

The Monte-Carlo approximation of the expectation has the advantage to allow for the optimization of $\boldsymbol{D}_k$ 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].

B Dirichlet prior on mixing coefficients: E-step

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

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

where $q_{\mu_k,\boldsymbol{A}_k}^{(r)}(\boldsymbol{\mu}_k, \boldsymbol{A}_k)$ has the same expression as given by (16) and (17). 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-$\boldsymbol{X}$ step. This step is only partly impacted by the addition of a prior on $\boldsymbol{\pi}$. It comes as in the previous section, $q_X^{(r)}(\boldsymbol{X}) = \prod_{i=1}^{N} q_{X_i}^{(r)}(\boldsymbol{W}_i, \boldsymbol{Z}_i)$ with the term below unchanged and given by (24), (26) and (27),

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

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

$$q_{Z_i}^{(r)}(Z_i = k) = q_{Z_i}^{(r)}(k) \propto \tilde{\pi}_k^{(r)} \exp(\tilde{\rho}_k^{(r)}/2) \prod_{m=1}^{M} \frac{\Gamma(\tilde{\alpha}_{km}^{(r)})}{\Gamma(\alpha_{km}^{(r-1)})\ \tilde{\gamma}_{kim}^{(r)\tilde{\alpha}_{km}^{(r)}}} . \tag{30}$$

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

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

The last term being constant with respect to (r) and k, it follows that $\tilde{\pi}_k^{(r)}$ is proportional to $\exp(\Upsilon(\tau_k + n_k^{(r-1)}))$ and it is enough to use this later expression in (30).

C No prior on mixing coefficients: free energy expression

The expression of the free energy at each iteration is needed to apply the procedure mentioned in section 5.2. It is given here in the absence of weight prior (Section 4.1). The free energy expression differs only slightly when a Dirichlet prior is added (see Appendix D). The free energy can be decomposed into two terms. At each iteration (r),

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

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

C.1 Entropy terms

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

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

In the expression above, $H[q_{\boldsymbol{\Phi}_k^1}^{(r)}]$ is the entropy of the Normal-Wishart distribution defined in (16) and (17),

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

where $\tilde{\boldsymbol{N}}_k^{(r)}$ is given by equation (21), $\tilde{\lambda}_{km}^{(r)}$ and $\tilde{\delta}_{km}^{(r)}$ by (22) and (23).

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

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

where $q_{Z_i}^{(r)}(k)$ is given by (25), $\tilde{\alpha}_{km}^{(r)}$ by (26) and $\tilde{\gamma}_{kim}^{(r)}$ by (27).

C.2 M-step terms

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

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

The five terms are detailed in turn below,

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

with $\bar{\boldsymbol{\Delta}}^{(r)}_{ki}$ given in (18), $\tilde{m}^{(r)}_k$ in (20), $\tilde{\boldsymbol{N}}^{(r)}_k$ in (21), $\tilde{\rho}^{(r)}_k$ and $\tilde{\boldsymbol{A}}^{(r)}_k$ in (28) and (29), $\tilde{\alpha}^{(r)}_{km}$ and $\tilde{\gamma}^{(r)}_{kim}$ in (26) and (27), $\boldsymbol{D}^{(r)}_k$ is the solution of the M-$\boldsymbol{D}$-step. .

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

where $q^{(r)}_{Z_i}(k)$ is given in (25), $\alpha^{(r)}_{km}$ are the solutions of the M-$\boldsymbol{\alpha}$ step, $\tilde{\alpha}^{(r)}_{km}$ and $\tilde{\gamma}^{(r)}_{kim}$ are given in (26) and (27).

$$E_{q^{(r)}_Z}[\log p(\boldsymbol{Z}; \boldsymbol{\pi}^{(r)})] = (\sum_{k=1}^{K} n^{(r)}_k \log n^{(r)}_k) - N \log N$$

with $n^{(r)}_k = \sum_{i=1}^{N} q^{(r)}_{Z_i}(k)$.

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

where $\tilde{\boldsymbol{N}}^{(r)}_k$, $\tilde{m}^{(r)}_k$ are given in (21) and (20), $\tilde{\rho}^{(r)}_k$ is given in (28) and $\tilde{\boldsymbol{A}}^{(r)}_k$ in (29).

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

with $\tilde{A}^{(r)}_{km}$ given in (29), $\tilde{\lambda}^{(r)}_{km}$ and $\tilde{\delta}^{(r)}_{km}$ in (22) and (23).

D Dirichlet prior on mixing coefficients: free energy expression

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

D.2 M-step terms

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

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

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

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

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 C.