Component elimination strategies to fit mixtures of multiple scale distributions

Florence Forbes, Alexis Arnaud, Benjamin Lemasson, Emmanuel Barbier

Component elimination strategies to fit mixtures of multiple scale distributions

Florence Forbes[1][0000−0003−3639−0226], Alexis Arnaud[1,2], Benjamin Lemasson[2], and Emmanuel Barbier[2]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
firstname.lastname@inria.fr

[2] Grenoble Institut des Neurosciences, Inserm U1216, Univ. Grenoble Alpes, France
firstname.lastname@univ-grenoble-alpes.fr

Abstract. We address the issue of selecting automatically the number of components in mixture models with non-Gaussian components. As a more efficient alternative to the traditional comparison of several model scores in a range, we consider procedures based on a single run of the inference scheme. Starting from an overfitting mixture in a Bayesian setting, we investigate two strategies to eliminate superfluous components. We implement these strategies for mixtures of multiple scale distributions which exhibit a variety of shapes not necessarily elliptical while remaining analytical and tractable in multiple dimensions. A Bayesian formulation and a tractable inference procedure based on variational approximation are proposed. Preliminary results on simulated and real data show promising performance in terms of model selection and computational time.

Keywords: Gaussian scale mixture · Bayesian analysis · Bayesian model selection · EM algorithm · Variational approximation.

1 Inroduction

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 [10]. 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 [23, 13]. Of a slightly different nature is the so-called slope heuristic [7], 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. When the components distributions complexity increases, it may then be desirable to avoid

repetitive inference of models that will be discarded in the end. For standard Gaussian distributions however, this is not really a problem as efficient software such as Mclust [28] are available. 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 [26] 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.* [30]) 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 [18]. For instance, [13] proposes 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 [6] starts with an overestimated number of components using BIC, and then merges them hierarchically according to an entropy criterion, while [24] 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. In [25, 17] an infinite Gaussian mixture (IGMM) is presented with a computationally intensive Markov Chain Monte Carlo implementation. More recently, more flexibility in the cluster shapes has been allowed by considering infinite mixture of infinite Gaussian mixtures (I^2GMM) [32]. The flexibility is however limited to a cluster composed of sub-clusters of identical shapes and orientations, which may alter the performance of this approach. Beyond the Gaussian case, infinite Student mixture models have also been considered [31]. The Bayesian non-parametric approach is a promising technique. In this work, 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. Following common practice that is to start from deliberately overfitting mixtures (*e.g.* [21, 11, 22, 3]), we investigate component elimination strategies. Component elimination refers to a natural approach which is to exploit the vanishing component phenomenon that has been proved to occur in certain Bayesian settings [27]. This requires a Bayesian formulation of the mixture for the regularization effect due to the integration of parameters in the posterior distribution. This results in an implicit penalization for model complexity. Although this approach can be based on arbitrary mixture components, most previous investigation has been confined to Gaussian mixtures where the mixture components arise from multivariate Gaussian densities with component-specific parameters.

In this work, we address the issue of selecting automatically the number of components in a non-Gaussian case. We consider mixtures of so called multiple scale distributions for their ability to handle a variety of shapes not necessarily elliptical while remaining analytical and tractable. We propose a Bayesian formulation of these mixtures and a tractable inference procedure based on a variational approximation. We propose two different single-run strategies that make use of the component elimination property.

The rest of the paper is organized as follows. Mixture of multiple scale distributions, their Bayesian formulation and inference are specified in Section 2. The two proposed strategies are described in Section 3, illustrated with experiments on simulated data in Section 4.

2 Bayesian mixtures of multiple scale distributions

2.1 Multiple scale mixtures of Gaussians

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).

The extension proposed by [14] consists of introducing a multidimensional weight. To do so, the scale matrix is decomposed into eigenvectors and eigenvalues. This spectral decomposition is classically used in Gaussian model-based clustering [5, 9]. In a Bayesian setting, it is equivalent but more convenient to use matrix $\boldsymbol{T}$ the inverse of the scale matrix. We therefore consider the decomposition $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$ where $\boldsymbol{D}$ is the matrix of eigenvectors of $\boldsymbol{T}$ (equivalently of $\boldsymbol{\Sigma}$) and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues. The matrix $\boldsymbol{D}$ determines the orientation of the Gaussian and $\boldsymbol{A}$ its shape. Using this parameterization of $\boldsymbol{T}$, the scale Gaussian part in (1) is set to $\mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}^{-1}\boldsymbol{D}^T)$, 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:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \!\!\ldots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}^{-1}\boldsymbol{D}^T)\; f_{\mathbf{w}}(w_1 \ldots w_M;\boldsymbol{\theta})\; \mathrm{d}w_1 \ldots \mathrm{d}w_M \tag{2}$$

where $f_{\mathbf{w}}$ is now a M-variate density depending on some parameter $\boldsymbol{\theta}$ to be further specified. In what follows, 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.* [20] vol.2 chap. 28 for a definition of the Pearson type VII distribution). For identifiability, this model 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 due to invariance to a same permutation of the columns of $\boldsymbol{D}, \boldsymbol{A}$ and elements of $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$. 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 2.2). An appropriate prior on $\boldsymbol{D}$ would be more difficult to set. The distributions we consider are therefore of the form,

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

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 (3). 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 $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$.

2.2 Priors on parameters

In a Bayesian formulation, we assign priors on parameters in $\boldsymbol{\Phi}$. However, it is common (see *e.g.* [1]) 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:

$$\boldsymbol{\mu}_k \mid \boldsymbol{A}_k, \boldsymbol{D}_k \sim \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{D}_k \boldsymbol{\Lambda}_k^{-1} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^T) \ , \tag{4}$$

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 [19]. However, it is not straightforward to derive from it a tractable E-$\boldsymbol{\Phi}^1$ step (see Section 2.3) 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:

$$\boldsymbol{A}_k \sim \otimes_{m=1}^{M} \mathcal{G}(\lambda_{km}, \delta_{km}) \ , \tag{5}$$

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 $\boldsymbol{D}_{1:K} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K\}$

$$p(\boldsymbol{\mu}_{1:K}, \boldsymbol{A}_{1:K}; \boldsymbol{D}_{1:K}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{A}_k; \boldsymbol{D}_k)\ p(\boldsymbol{A}_k) \tag{6}$$

where the first term in the product is given by (4) and the second term by (5).

Then a standard Dirichlet prior $\mathcal{D}(\tau_1, \ldots, \tau_K)$ is used for the mixing weights $\boldsymbol{\pi}$ with $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K\}$ the Dirichlet hyperparameters.

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

2.3 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

$$\mathcal{F}(q, \boldsymbol{\Phi}^2) = E_q[\log p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^2)] - E_q[\log q(\boldsymbol{X}, \boldsymbol{\Phi}^1)]. \tag{7}$$

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 3.1 as a replacement of the likelihood to provide a model selection procedure. Its detailed expression is given in a companion paper [2].

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

$$\textbf{E-}\boldsymbol{\Phi}^1\textbf{-step:} \quad q_{\Phi^1}^{(r)}(\boldsymbol{\Phi}^1) \propto \exp(E_{q_X^{(r-1)}}[\log p(\boldsymbol{\Phi}^1 | \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\Phi}^{2(r-1)})]) \tag{8}$$

$$\textbf{E-X-step:} \quad q_X^{(r)}(\boldsymbol{X}) \propto \exp(E_{q_{\phi^1}^{(r)}}[\log p(\boldsymbol{X} | \boldsymbol{y}, \boldsymbol{\Phi}^1; \boldsymbol{\Phi}^{2(r-1)})])\ . \tag{9}$$

Then the M-step reduces to:

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

The resulting variational EM algorithm is specified in [2] in two cases depending on the prior used for the mixing weights. For component elimination, the central quantity is $q_\pi^{(r)}(\boldsymbol{\pi})$ the approximate variational posterior of $\boldsymbol{\pi}$ that itself involves $q_Z^{(r)}(\boldsymbol{Z}) = \prod_i q_{Z_i}^{(r)}(Z_i)$ the variational posterior of the labels.

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.

3 Single-run number of component selection

In this work, we consider approaches that start from an overfitting mixture with more components than expected in the data. In this case, as described by [16], 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.* [27]).

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.* [27]. Although the framework of [27] 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 [11] who use priors only for the component mean and covariance parameters. See [2] for details on the investigation of this alternative case with no prior on the mixing weights.

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 [21]. 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, [27] 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 [21], much smaller value of τ_0 are needed (*e.g.* 10^{-5}). It was even observed by [29] 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 [13]. This latter work does not start from a Bayesian formulation but is based on a Minimum Message Length (MML) principle. [13] 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 [29] corresponds to a very strong prior sparsity.

In a Bayesian setting with symmetric sparse Dirichlet priors $\mathcal{D}(\tau_0, \ldots, \tau_0)$, the theoretical study of [27] therefore 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),

$$E[\pi_k|\boldsymbol{y}] \approx E_{q_{\boldsymbol{\pi}}}[\pi_k] = \frac{\tilde{\tau}_k}{\sum_{l=1}^{K} \tilde{\tau}_l} = \frac{\tau_k + n_k}{\sum_{k=1}^{K} \tau_k + N} \tag{10}$$

where $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$ and the expression for $q_{Z_i}(k)$ is detailed in [2]. If we are in the no weight prior case, then the expectation simplifies to

$$\pi_k \approx \frac{n_k}{N} \tag{11}$$

with the corresponding expression of $q_{Z_i}(k)$ also given in [2].

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 the sum $n_k = \sum_{i=1}^{N} q_{Z_i}(k)$. As an illustration, the choice of a negative τ_0 by [13] 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 [13] is stronger than that used by [22] which annihilates a component when the sum n_k reduces to 1 or the one of [11] 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 [22] use a Bayesian framework with variational inference and their rule corresponds to thresholding the variational posterior weights (10) 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, [21] 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 [12] using the so-called posterior similarity matrix ([15]). 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 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. Also it avoids the choice of a threshold for separating between large or small weights. The tested procedures are more specifically described in the next section.

3.1 Tested procedures

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

Thresholding based algorithm: 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 [27]. This corresponds to the approach adopted in [21] and [22]. The difference between the later two being how they check for vanishing coefficients. Our variational inference leads more naturally to the solution of [22] which is to check the weight posterior means, that is whether at each iteration (r),

$$n_k^{(r)} < (K\tau_0 + N)\rho_t - \tau_0 \tag{12}$$

where ρ_t is the chosen threshold on the posterior means. When ρ_t is set such that (12) 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*.

Free Energy based algorithm: We also consider a criterion based on the free energy (7) to detect components to eliminate. This choice is based on the observation that when 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 [4] and [8] that the free energy penalizes model complexity and that it converges to the well known Bayesian Information Criterion (BIC) and Minimum Description Length (MDL) criteria, when the sample size increases, illustrating the interest of this measure for model selection.

The free energy expression used is given in [2]. The heuristic denoted by *SparseDirichlet+FEtest* 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 (*SparseDirichlet*).

4 Experiments

In addition to the 3 methods *SparseDirichlet+πtest*, *SparseDirichlet+FEtest* and *SparseDirichlet*, referred to below as $\mathcal{MP}$ single-run procedures, we consider standard Gaussian mixtures using the Mclust package [28] 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 (4). The δ_{km} are then set to 1 and λ_{km} to values $5 \times 10^{-4} = \lambda_1 < \lambda_2 < \ldots < \lambda_M = 10^{-3}$. 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.

4.1 Simulated data

We consider several models (more details can be found in [2]), 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 1 and the simulated samples illustrated in Figure 1. 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 with a time divided by 3 compared to the non Bayesian $\mathcal{MP}$ mixture procedure, although the later benefits from a more optimized implementation. For the first strategy, the 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 (SparseDirichlet line in Table 1). 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 2.

Table 1. 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.

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
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

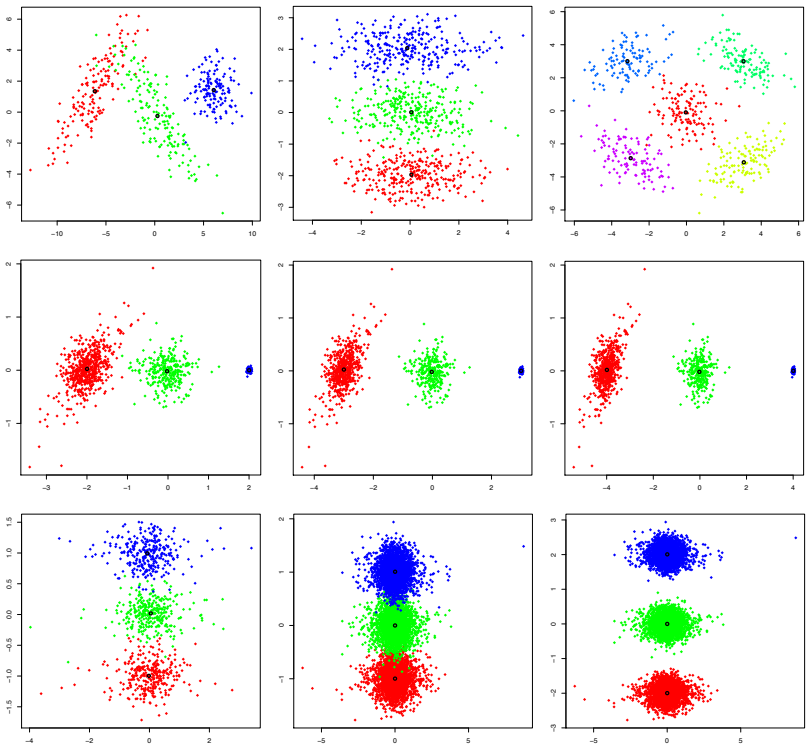

Fig. 1. Examples of simulated samples. First line: 3 Gaussian mixtures with 3 and 5 components. Second line: $\mathcal{MP}$ mixtures with different dof and increasing separation from left to right. Third line: $\mathcal{MP}$ mixtures with increasing separation, from left to right, and increasing number of points, $N = 900$ for the first plot, $N = 9000$ for the last two.

5 Discussion and conclusion

We investigated, 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 [27] and Type II maximum likelihood as proposed by [11] (not reported here but investigated in [2]). 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

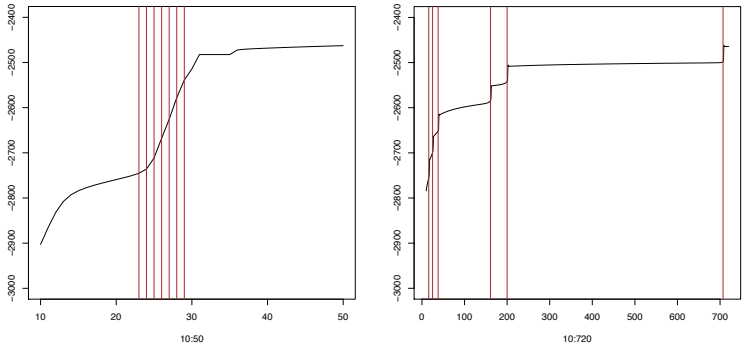

Fig. 2. Illustration of the two component elimination strategies: Free energy gain strategy, iterations 10 to 50 (left) and too small component proportion test, iterations 10 to 720 (right). Eliminations are marked with red lines. Most of them occur at earlier iterations when using the free energy test.

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.

As non Gaussian components, we investigated in particular the case of multiple scale distributions [14], which have been shown to perform well in the modelling of non-elliptical clusters with potential outliers and tails of various heaviness. We proposed a Bayesian formulation of mixtures of such multiple scale distributions and derived an inference procedure based on a variational EM algorithm to implement the single-run procedures.

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 (results reported in [2]).

To confirm these observations, more tests 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 [27], 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.

6 Supplementary Material

All details on the variational EM and free energy computations, plus additional illustrations can be found in a companion paper [2].

References

1. Archambeau, C., Verleysen, M.: Robust Bayesian clustering. Neural Networks **20**(1), 129–138 (2007)
2. Arnaud, A., Forbes, F., Steele, R., Lemasson, B., Barbier, E.L.: Bayesian mixtures of multiple scale distributions (Jul 2019), https://hal.inria.fr/hal-01953393, working paper or preprint
3. Attias, H.: 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. pp. 21–30 (1999)
4. Attias, H.: A variational Bayesian framework for graphical models. In: Proc. Advances in Neural Information Processing Systems 12. pp. 209–215. MIT Press, Denver, Colorado, United States (2000)
5. Banfield, J., Raftery, A.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics **49**(3), 803–821 (1993)
6. Baudry, J.P., Raftery, E.A., Celeux, G., Lo, K., Gottardo, R.: Combining mixture components for clustering. Journal of Computational and Graphical Statistics **19**(2) (2010)
7. Baudry, J.P., Maugis, C., Michel, B.: Slope heuristics: overview and implementation. Statistics and Computing **22**(2), 455–470 (2012)
8. Beal, M.J.: Variational algorithms for approximate Bayesian inference. Ph.D. thesis, University of London (2003)
9. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognition **28**(5), 781–793 (1995)
10. Celeux, G., Fruhwirth-Schnatter, S., Robert, C.: Model Selection for Mixture Models-Perspectives and Strategies. Handbook of Mixture Analysis, CRC press (12 2018)
11. Corduneanu, A., Bishop, C.: Variational Bayesian Model Selection for Mixture Distributions. In: Proceedings Eighth International Conference on Artificial Intelligence and Statistics. p. 2734. Morgan Kaufmann (January 2001)
12. Dahl, D.B.: Model-based clustering for expression data via a Dirichlet process mixture model, in Bayesian Inference for Gene Expression and Proteomics (2006)
13. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(3), 381–396 (2002)
14. Forbes, F., Wraith, D.: 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)

15. Fritsch, A., Ickstadt, K.: Improved criteria for clustering based on the posterior similarity matrix. Bayesian Analysis **4**(2), 367–391 (06 2009)
16. Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer Verlag (2006)
17. Gorur, D., Rasmussen, C.: Dirichlet process Gaussian mixture models: Choice of the base distribution. Journal of Computer Science and Technology **25**(4), 653–664 (2010)
18. Hennig, C.: Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification **4**(1), 3–34 (2010)
19. Hoff, P.D.: A Hierarchical Eigenmodel for Pooled Covariance Estimation. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **71**(5), 971–992 (2009)
20. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, vol.2, 2nd edition. John Wiley & Sons, New York (1994)
21. Malsiner-Walli, G., , Frühwirth-Schnatter, S., Grün, B.: Model-based clustering based on sparse finite Gaussian mixtures. Statistics and Computing **26**(1), 303–324 (Jan 2016)
22. McGrory, C.A., Titterington, D.M.: Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. Comput. Stat. Data Anal. **51**(11), 5352–5367 (Jul 2007)
23. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley (2000)
24. Melnykov, V.: Merging mixture components for clustering through pairwise overlap. Journal of Computational and Graphical Statistics (2014)
25. Rasmussen, C.E.: The infinite Gaussian mixture model. In: NIPS. vol. 12, pp. 554–560 (1999)
26. Richardson, S., Green, P.J.: 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)
27. Rousseau, J., Mengersen, K.: 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)
28. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 205–233 (2016)
29. Tu, K.: 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. pp. 1986–1991 (2016)
30. Verbeek, J., Vlassis, N., Kröse, B.: Efficient Greedy Learning of Gaussian Mixture Models. Neural Computation **15**(2), 469–485 (2003)
31. Wei, X., Li, C.: The infinite student t-mixture for robust modeling. Signal Processing **92**(1), 224–234 (2012)
32. Yerebakan, H.Z., Rajwa, B., Dundar, M.: The infinite mixture of infinite Gaussian mixtures. In: Advances in Neural Information Processing Systems. pp. 28–36 (2014)