Robust Bayesian clustering

Cédric Archambeau Michel Verleysen [⋆]

Machine Learning Group, Université catholique de Louvain,
B-1348 Louvain-la-Neuve, Belgium.

Abstract

A new variational Bayesian learning algorithm for Student-t mixture models is introduced. This algorithm leads to (i) robust density estimation, (ii) robust clustering and (iii) robust automatic model selection. Gaussian mixture models are learning machines which are based on a divide-and-conquer approach. They are commonly used for density estimation and clustering tasks, but are sensitive to outliers. The Student-t distribution has heavier tails than the Gaussian distribution and is therefore less sensitive to any departure of the empirical distribution from Gaussianity. As a consequence, the Student-t distribution is suitable for constructing robust mixture models. In this work, we formalize the Bayesian Student-t mixture model as a latent variable model in a different way than Svensén and Bishop (2004). The main difference resides in the fact that it is not necessary to assume a factorized approximation of the posterior distribution on the latent indicator variables and the latent scale variables in order to obtain a tractable solution. Not neglecting the correlations between these unobserved random variables leads to a Bayesian model having an increased robustness. Furthermore, it is expected that the lower bound on the log-evidence is tighter. Based on this bound, the model complexity, i.e. the number of components in the mixture, can be inferred with a higher confidence.

Key words: Bayesian learning, graphical models, approximate inference, variational inference, mixture models, density estimation, clustering, model selection, student-t distribution, robustness to outliers

1 Introduction

Probability density estimation is a fundamental tool for extracting the information embedded in raw data. For instance, efficient and robust density estimators are the foundation for Bayesian (i.e., optimal) classification and

[⋆] M.V. is a Research Director of the Belgian National Fund for Scientific Research.

statistical pattern recognition. Finite Gaussian mixture models (GMM) are commonly used in this context (see for example McLachlan and Peel, 2000). They provide an appealing alternative to nonparametric density estimators (Parzen, 1962) as they do not assume the overall shape of the underlying density either. However, unlike nonparametric techniques, they are based on a divide-and-conquer approach, meaning that subpopulations of the observed data are modeled by parametric distributions, the resulting density being often far from any standard parametric form. Thus, unlike the nonparametric methods, the complexity of the model is fixed in advance, avoiding a prohibitive increase of the number of parameters with the size of the data set.

The GMM has been successfully applied to a wide range of applications. Its success is partly explained by the fact that maximum likelihood (ML) estimates of the parameters can be found by means of the popular expectation-maximization (EM) algorithm, which was formalized by Dempster, Laird, and Rubin (1977). The problem with ML is that it favors models of ever increasing complexity. This is due to the undesirable property of ML of being ill-posed since the likelihood function is unbounded (see for example Archambeau, Lee, and Verleysen, 2003; Yamazaki and Watanabe, 2003). In order to determine the optimal model complexity, resampling techniques such as cross-validation or the bootstrap (Efron and Tibshirani, 1993), are therefore required. Yet, these techniques are computationally intensive. An alternative is provided by the Bayesian framework. In this approach, the parameters are treated as unknown random variables and the predictions are averaged over the ensemble of models they define. Let us denote the set of observed data by $X = \{\mathbf{x}_n\}_{n=1}^N$. The quantity of interest is the evidence of the data given the model structure $\mathcal{H}_M$ of complexity M:

$$p(X|\mathcal{H}_M) = \int_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta}, \mathcal{H}_M) p(\boldsymbol{\theta}|\mathcal{H}_M) d\boldsymbol{\theta} \ , \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter vector and $p(X|\boldsymbol{\theta}, \mathcal{H}_M)$ is the data likelihood. In the case of mixture models, M is the number of components in the mixture. Unfortunately, taking the distribution of the parameters into account leads usually to intractable integrals. Therefore, approximations are required. Sampling techniques such as Markov Chain Monte-Carlo are for example used for this purpose (see Richardson and Green, 1997, for its application to the GMM with unknown M). However, these techniques are rather slow and it is generally difficult to verify if they have converged properly. More recently, Attias (1999) addressed this problem from a variational Bayesian perspective. By assuming that the joint posterior on the latent variables[1] and the parameters factorizes, the integrals become tractable. As a result, a lower bound on the

[1] Although latent variables cannot be observed, they may either interact through the model parameters in the data generation process, or are just mathematical artifacts that are introduced into the model in order to simplify it in some way.

log-evidence can be computed. This bound can be maximized (thus made as tight as possible) by means of an EM-like iterative procedure, called variational Bayes, which is guaranteed to increase monotonically at each iteration.

Nonetheless, a major limitation of the GMM is its lack of robustness to outliers. Providing robustness to outlying data is essential in many practical problems, since the estimates of the means and the precisions (i.e., the inverse covariance matrices) can be severely affected by atypical observations. In addition, in the case of the GMM, the presence of outliers or any other departure of the empirical distribution from Gaussianity can lead to selecting a false model complexity. More specifically, additional components are used (and needed) to capture the tails of the distribution. Robustness can be introduced by embedding the Gaussian distribution in a wider class of elliptically symmetric distributions, called the Student-t distributions. They provide a heavy-tailed alternative to the Gaussian family. The Student-t distribution is defined as follows:

$$\mathcal{S}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma\left(\frac{d+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{d}{2}}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{d+\nu}{2}} \quad , \qquad (2)$$

where d is the dimension of the feature space, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are respectively the component mean and the component precision and $\Gamma(\cdot)$ denotes the gamma function. Parameter $\nu > 0$ is the degree of freedom, which can be viewed as a robustness tuning parameter. The smaller ν is, the heavier the tails are. When ν tends to infinity, the Gaussian distribution is recovered. A finite Student-t mixture model (SMM) is then defined as a weighted sum of multivariate Student-t distributions:

$$p(\mathbf{x}|\boldsymbol{\theta}_{\mathcal{S}}) = \sum_{m=1}^{M}\pi_m\mathcal{S}(\mathbf{x}|\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m,\nu_m) \; , \qquad (3)$$

where $\boldsymbol{\theta}_{\mathcal{S}} \equiv (\pi_1,\ldots,\pi_M,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_M,\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_M,\nu_1,\ldots,\nu_M)$. The mixing proportions $\{\pi_m\}_{m=1}^{M}$ are non-negative and must sum to one.

In the context of mixture modeling, a crucial role is played by the responsibilities. Each such quantity is simultaneously associated to a data point and a mixture component. It corresponds to the posterior probability that a particular data point was generated by a particular mixture component. In other words, the responsibilities are soft labels for the data. During the training phase, these labels are used in order to estimate the model parameters. Therefore, it is essential to estimate them reliably, especially when considering robust approaches. In this paper, we introduce an alternative robust Bayesian paradigm to finite mixture models (in the exponential family), which focuses on this specific problem. In general, one way to achieve robustness is to avoid making unnecessary approximations. In the Bayesian framework, it means that dependencies between random variables should not be neglected as it leads to underestimating the uncertainty. In previous attempts to construct robust Bayesian mixture models, independency of all the latent variables was assumed. However, we show that this hypothesis is not necessary; removing it results in more consistent estimates for the responsibilities.

This article is organized as follows. In Section 2, the Bayesian Student-t mixture model is formalized as a latent variable model, which enables us to use the variational Bayesian framework to learn the model parameters as in the conjugate-exponential family (Ghahramani and Beal, 2001). In Section 3, the variational update rules are derived, as well as the lower bound on the log-evidence. Finally in Section 4, the approach is validated experimentally. It is shown empirically that the proposed variational inference scheme for the SMM leads to a model having a higher robustness to outliers than previous approaches. The robustness has a positive impact on the automatic model selection (based on the variational lower bound), as well as the quality of the parameter estimates. These properties are crucial when tackling real-life problems, which might be very noisy.

2 The latent variable model

The SMM can be viewed as a latent variable model in the sense that the component label associated to each data point is unobserved. Let us denote the set of label indicator vectors by $Z = \{\mathbf{z}_n\}_{n=1}^{N}$, with $z_{nm} \in \{0, 1\}$ and such that $\sum_{m=1}^{M} z_{nm} = 1$, $\forall n$. In contrast to the GMM, the observed data X augmented by the indicator vectors Z is still incomplete, meaning that there are still random variables that are not observed. This can be understood by noting that (2) can be re-written as follows:

$$\mathcal{S}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_{0}^{+\infty} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, u\boldsymbol{\Lambda})\mathcal{G}(u|\tfrac{\nu}{2}, \tfrac{\nu}{2})du \ , \tag{4}$$

where $u > 0$. The Gaussian and the Gamma distribution are respectively given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right\} \ , \tag{5}$$

$$\mathcal{G}(u|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}u^{\alpha-1} \exp(-\beta u) \ . \tag{6}$$

Equation (4) is easily verified by noting that the Gamma distribution is conjugate to the Gaussian distribution. Under this alternative representation, the Student-t distribution is thus an infinite mixture of Gaussian distributions with the same mean, but different precisions. The scaling factor u of the precisions is following a Gamma distribution with parameters depending only on ν. In contrast to the Gaussian distribution, there is no closed form solution for estimating the parameters of a single Student-t distribution based on the maximum likelihood principle. However, as discussed by Liu and Rubin (1995), the EM algorithm can be used to find an approximate ML solution by viewing u as an implicit latent variable on which a Gamma prior is imposed. This result was extended to mixtures of Student-t distributions by Peel and McLachlan (2000).

Based on (4), one can see that for each data point $\mathbf{x}_n$ and for each component m, the scale variable u_{nm} given z_{nm} is unobserved. For a fixed number of components M, the latent variable model of the SMM can therefore be specified as follows:

$$p(\mathbf{z}_n|\boldsymbol{\theta}_{\mathcal{S}}, \mathcal{H}_M) = \prod_{m=1}^{M} {\pi_m}^{z_{nm}} \ , \tag{7}$$

$$p(\mathbf{u}_n|\mathbf{z}_n, \boldsymbol{\theta}_{\mathcal{S}}, \mathcal{H}_M) = \prod_{m=1}^{M} \mathcal{G}(u_{nm}|\tfrac{\nu_m}{2}, \tfrac{\nu_m}{2})^{z_{nm}} \ , \tag{8}$$

$$p(\mathbf{x}_n|\mathbf{u}_n, \mathbf{z}_n, \boldsymbol{\theta}_{\mathcal{S}}, \mathcal{H}_M) = \prod_{m=1}^{M} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, u_{nm}\boldsymbol{\Lambda}_m)^{z_{nm}} \ , \tag{9}$$

where the set of scale vectors is denoted by $U = \{\mathbf{u}_n\}_{n=1}^{N}$. Marginalizing over the latent variables results in (3). At this point, the Bayesian formulation of the SMM is complete when imposing a particular prior on the parameters. As it will become clear in the next section, it is convenient to choose the prior as being conjugate to the likelihood terms (7–9). Therefore, the prior on the mixture proportions is chosen to be jointly Dirichlet $\mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\kappa}_0)$ and the joint prior on the mean and the precision of each component is chosen to be Gaussian-Wishart $\mathcal{NW}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m|\boldsymbol{\theta}_{\mathcal{NW}_0})$. The former is conjugate to the multinomial distribution $p(\mathbf{z}_n|\boldsymbol{\theta}_{\mathcal{S}}, \mathcal{H}_M)$ and the latter to each factor of $p(\mathbf{x}_n|\mathbf{u}_n, \mathbf{z}_n, \boldsymbol{\theta}_{\mathcal{S}}, \mathcal{H}_M)$. Since there is no conjugate prior for $\{\nu_m\}_{m=1}^{m}$, no prior is imposed on them. Moreover, the hyperparameters are usually chosen such that broad priors are obtained. The resulting joint prior on the model parameters is given by

$$p(\boldsymbol{\theta}_{\mathcal{S}}|\mathcal{H}_M) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\kappa}_0) \prod_{m=1}^{M} \mathcal{NW}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m|\boldsymbol{\theta}_{\mathcal{NW}_0}) \ . \tag{10}$$

The Gaussian-Wishart distribution is the product of a Gaussian and a Wishart distribution: $\mathcal{NW}(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m|\boldsymbol{\theta}_{\mathcal{NW}_0}) = \mathcal{N}(\boldsymbol{\mu}_m|\mathbf{m}_0, \eta_0\boldsymbol{\Lambda}_m)\mathcal{W}(\boldsymbol{\Lambda}_m|\gamma_0, \mathbf{S}_0)$. The Dirichlet and the Wishart distribution are respectively defined as follows:

$$\mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\kappa}) = c_{\mathcal{D}}(\boldsymbol{\kappa}) \prod_{m=1}^{M} {\pi_m}^{\kappa_m - 1} \ , \tag{11}$$

$$\mathcal{W}(\boldsymbol{\Lambda}|\gamma, \mathbf{S}) = c_{\mathcal{W}}(\gamma, \mathbf{S})\, |\boldsymbol{\Lambda}|^{\frac{\gamma - d - 1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\{\mathbf{S}\boldsymbol{\Lambda}\}\right) \ , \tag{12}$$

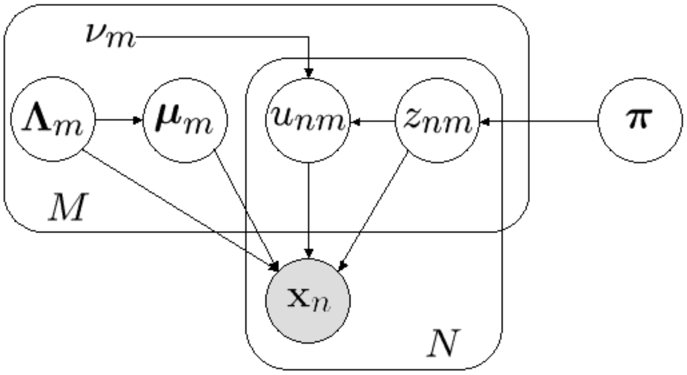


Fig. 1. Graphical representation of the Bayesian Student-t mixture model. The shaded node is observed. The arrows represent conditional dependencies between the random variables. The plates indicate independent copies. Note that the scale variables and the indicator variables are contained in both plates, meaning that there is one such variable for each component and each data point. It is important to see that the scale variables depend on the discrete indicator variables. Similarly, the component means depend on the component precisions.

where $c_{\mathcal{D}}(\boldsymbol{\kappa})$ and $c_{\mathcal{W}}(\gamma, \mathbf{S})$ are normalizing constants. Figure 1 shows the directed acyclic graph of the Bayesian SMM. Each observation $\mathbf{x}_n$ is conditionally dependent on the indicator vector $\mathbf{z}_n$ and the scale vector $\mathbf{u}_n$, which are both unobserved. The scale vectors are also conditionally dependent on the indicator variables. By contrast, Svensén and Bishop (2004) assume that the scale variables are independent of the indicator variables, therefore neglecting the correlations between these random variables. Furthermore, they assume that the component means are independent from the corresponding precisions.

3 Variational Bayesian inference for the SMM

The aim in Bayesian learning is to compute (or approximate) the evidence. This quantity is obtained by integrating out the latent variables and the parameters. For a fixed model structure $\mathcal{H}_M$ of the SMM, the evidence is given by

$$p(X|\mathcal{H}_M) = \int_{\boldsymbol{\theta}_{\mathcal{S}}} \int_U \sum_Z p(X, U, Z, \boldsymbol{\theta}_{\mathcal{S}}|\mathcal{H}_M) dU d\boldsymbol{\theta}_{\mathcal{S}} \ . \tag{13}$$

This quantity is intractable. However, for any distribution $q(U, Z, \boldsymbol{\theta}_{\mathcal{S}})$, the logarithm of the evidence can be lowerbounded as follows:

$$\log p(X|\mathcal{H}_M) \geq \log p\left(X|\mathcal{H}_M\right) - \mathrm{KL}\left[q(U, Z, \boldsymbol{\theta}_{\mathcal{S}}) \| p\left(U, Z, \boldsymbol{\theta}_{\mathcal{S}}|X, \mathcal{H}_M\right)\right] \ . \tag{14}$$

The second term on the right hand side is the Kullback-Leibler divergence (KL) between the approximate posterior $q(U, Z, \boldsymbol{\theta}_{\mathcal{S}})$ and the true posterior $p\left(U, Z, \boldsymbol{\theta}_{\mathcal{S}}|X, \mathcal{H}_M\right)$. Below, we show that when assuming that $q(U, Z, \boldsymbol{\theta}_{\mathcal{S}})$ only factorizes over the latent variables and the parameters, a tractable lower bound on the log-evidence can be constructed. Performing a free-form maximization of the lower bound with respect to $q(U,Z)$ and $q(\boldsymbol{\theta}_{\mathcal{S}})$ leads to the following VBEM update rules:

$$\textbf{VBE-step} \ : q(\mathbf{u}_n,\mathbf{z}_n) \propto \exp\left(\mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log p(\mathbf{x}_n,\mathbf{u}_n,\mathbf{z}_n|\boldsymbol{\theta}_{\mathcal{S}},\mathcal{H}_M)\}\right) \ , \quad \forall n \ . \quad (15)$$

$$\textbf{VBM-step} : q(\boldsymbol{\theta}_{\mathcal{S}}) \propto p(\boldsymbol{\theta}_{\mathcal{S}}|\mathcal{H}_M) \exp\left(\mathrm{E}_{U,Z}\{\log \mathcal{L}_c(\boldsymbol{\theta}_{\mathcal{S}}|X,U,Z,\mathcal{H}_M)\}\right) \ . \quad (16)$$

In the VBE-step we have used the fact that the data are i.i.d. and in the VBM-step $\mathcal{L}_c(\boldsymbol{\theta}_{\mathcal{S}}|X,U,Z,\mathcal{H}_M)$ is the complete data likelihood. The expectations $\mathrm{E}_{U,Z}\{\cdot\}$ and $\mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\cdot\}$ are respectively taken with respect to the variational posteriors $q(U,Z)$ and $q(\boldsymbol{\theta}_{\mathcal{S}})$. From (14), it can be seen that maximizing the lower bound is equivalent to minimizing the KL divergence between the true and the variational posterior. Thus, the VBEM algorithm consists in iteratively updating the variational posteriors by making the bound as tight as possible. By construction, the bound cannot decrease. Note also that for a given model complexity, the only difference between the VBE- and VBM-steps is the number of quantities to update. For the first one, this number scales with the size of the learning set, while for the second one it is fixed.

Due to the factorized form of $p(\mathbf{x}_n,\mathbf{u}_n,\mathbf{z}_n|\boldsymbol{\theta}_{\mathcal{S}},\mathcal{H}_M)$, it is likely that $q(\mathbf{z}_n) = \prod_{m=1}^{M} q(z_{nm})^{z_{nm}}$ and similarly that $q(\mathbf{u}_n|\mathbf{z}_n) = \prod_{m=1}^{M} q(u_{nm})^{z_{nm}}$. Furthermore, since the prior on the parameters is chosen conjugate to the likelihood terms, it can be seen from the VBM-step that the corresponding variational posteriors have the same functional form:

$$q(\boldsymbol{\theta}_{\mathcal{S}}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\kappa}) \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{\mu}_m|\mathbf{m}_m,\eta_m\boldsymbol{\Lambda}_m)\mathcal{W}(\boldsymbol{\Lambda}_m|\gamma_m,\mathbf{S}_m) \ . \quad (17)$$

Given the form of the variational posteriors, the VBE-step can be computed. Taking expectations with respect to the posterior distribution of the parameters leads to the following identity:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log p(\mathbf{x}_n,\mathbf{u}_n,\mathbf{z}_n|\mathcal{H}_M)\} = \sum_{m=1}^{M} z_{nm}\Big\{ &\log\tilde{\pi}_m - \frac{d}{2}\log 2\pi + \frac{d}{2}\log u_{nm} \\ &+\frac{1}{2}\log\tilde{\Lambda}_m - \frac{u_{nm}\gamma_m}{2}(\mathbf{x}_n-\mathbf{m}_m)^{\mathrm{T}}{\mathbf{S}_m}^{-1}(\mathbf{x}_n-\mathbf{m}_m) - \frac{u_{nm}d}{2\eta_m} \\ &+\frac{\nu_m}{2}\log\frac{\nu_m}{2} - \log\Gamma\left(\frac{\nu_m}{2}\right) + \left(\frac{\nu_m}{2}-1\right)\log u_{nm} - \frac{\nu_m}{2}u_{nm}\Big\} \ . \quad (18)\end{aligned}$$

The special quantities in (18) are $\log\tilde{\pi}_m \equiv \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log\pi_m\} = \psi(\kappa_m) - \psi\left(\sum_{m'=1}^{M}\kappa_{m'}\right)$ and $\log\tilde{\Lambda}_m \equiv \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log|\boldsymbol{\Lambda}_m|\} = \sum_{i=1}^{d}\psi\left(\frac{\gamma_m+1-i}{2}\right) + d\log 2 - \log|\mathbf{S}_m|$, where $\psi(\cdot)$ denotes the digamma function.

First, the VBE-step for the indicator variables is obtained by substituting (18) in (15) and integrating out the scale variables:

$$q(z_{nm}=1) \propto \frac{\Gamma\left(\frac{d+\nu_m}{2}\right)}{\Gamma\left(\frac{\nu_m}{2}\right)(\nu_m \pi)^{\frac{d}{2}}} \tilde{\pi}_m \tilde{\Lambda}_m{}^{\frac{1}{2}} \tag{19}$$
$$\times \left[1 + \frac{\gamma_m}{\nu_m} (\mathbf{x}_n - \mathbf{m}_m)^{\mathrm{T}} \mathbf{S}_m{}^{-1} (\mathbf{x}_n - \mathbf{m}_m) + \frac{d}{\nu_m \eta_m}\right]^{-\frac{d+\nu_m}{2}} .$$

This equation resembles a weighted Student-t distribution (which is an *infinite* mixture of scaled Gaussian distributions).

The corresponding VBE-step obtained by Svensén and Bishop (2004) has the form of a *single* weighted Gaussian distribution with a scaled precision, the scale being the expected value of the associated scale variable:

$$q^{(SB)}(z_{nm}=1) \propto (2\pi)^{-\frac{d}{2}} \tilde{\pi}_m \tilde{\Lambda}_m{}^{\frac{1}{2}} \mathrm{E}_U\{\log u_{nm}\}^{\frac{d}{2}} \tag{20}$$
$$\times \exp\left\{ \frac{\mathrm{E}_U\{u_{nm}\}}{2} \mathrm{E}_{\boldsymbol{\theta}_S}\{(\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Lambda}_m (\mathbf{x}_n - \boldsymbol{\mu}_m)\} \right\} .$$

It is thus assumed that most of the probability mass of the posterior distribution of each scale variable is located around its mean. This is not necessarily true for all data points as the Gamma prior might be highly skewed. In contrast, (19) results from integrating out the scale variables, which are here nuisance parameters:

$$q(z_{nm}=1) \propto (2\pi)^{-\frac{d}{2}} \tilde{\pi}_m \tilde{\Lambda}_m{}^{\frac{1}{2}} \int_0^{+\infty} u_{nm}{}^{\frac{d}{2}} \mathcal{G}(u_{nm}|\tfrac{\nu_m}{2}, \tfrac{\nu_m}{2}) \tag{21}$$
$$\times \exp\left\{ \frac{u_{nm}}{2} \mathrm{E}_{\boldsymbol{\theta}_S}\{(\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Lambda}_m (\mathbf{x}_n - \boldsymbol{\mu}_m)\} \right\} du_{nm} .$$

This means that the uncertainty on the scale variables is here properly taken into account when estimating the responsibilities.

Since the distribution $q(\mathbf{z}_n)$ must be normalized for each data point $\mathbf{x}_n$, we define the responsibilities as follows:

$$\bar{\rho}_{nm} = \frac{q(z_{nm}=1)}{\sum_{m'=1}^{M} q(z_{nm'}=1)} , \quad \forall n , \quad \forall m . \tag{22}$$

These quantities are similar in form to the responsibilities computed in the E-step in ML learning (see for example McLachlan and Peel, 2000).

Second, since the Gamma prior on the scale variables is conjugate to the exponential family, the variational posterior on the scale variables conditioned on the indicator variables has also the form of a Gamma distribution. Substituting (18) in (15) and rearranging leads to the VBE-step for the scale variables:

$$q(u_{nm}|z_{nm}=1) = \mathcal{G}(u_{nm}|\alpha_{nm}, \beta_{nm}) \ , \tag{23}$$

where

$$\alpha_{nm} = \frac{d+\nu_m}{2} \ , \tag{24}$$

$$\beta_{nm} = \frac{\gamma_m}{2}(\mathbf{x}_n - \mathbf{m}_m)^{\mathrm{T}}{\mathbf{S}_m}^{-1}(\mathbf{x}_n - \mathbf{m}_m) + \frac{d}{2\eta_m} + \frac{\nu_m}{2} \ . \tag{25}$$

The VBE-step for the scale variables consists simply in updating these hyperparameters. Again, there is a striking similarity with the corresponding E-step in ML learning (see Peel and McLachlan, 2000).

Next, let us compute the VBM-step. The expected complete data log-likelihood is given by

$$\begin{aligned}
&\mathrm{E}_{U,Z}\{\log \mathcal{L}_c(\boldsymbol{\theta}_{\mathcal{S}}|X,U,Z)\} = \\
&\quad \sum_{n=1}^{N}\sum_{m=1}^{M} \bar{\rho}_{nm}\Big\{ \log \pi_m - \frac{d}{2}\log 2\pi + \frac{d}{2}\log \tilde{u}_{nm} + \frac{1}{2}\log|\boldsymbol{\Lambda}_m| \\
&\qquad\qquad -\frac{\bar{u}_{nm}}{2}(\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}}\boldsymbol{\Lambda}_m(\mathbf{x}_n - \boldsymbol{\mu}_m) + \frac{\nu_m}{2}\log\frac{\nu_m}{2} \\
&\qquad\qquad - \log\Gamma\left(\frac{\nu_m}{2}\right) + \left(\frac{\nu_m}{2} - 1\right)\log \tilde{u}_{nm} - \frac{\nu_m}{2}\bar{u}_{nm}\Big\} \ ,
\end{aligned} \tag{26}$$

where the special quantities are $\bar{u}_{nm} \equiv \mathrm{E}_U\{u_{nm}\} = \alpha_{nm}/\beta_{nm}$ and $\log \tilde{u}_{nm} \equiv \mathrm{E}_U\{\log u_{nm}\} = \psi(\alpha_{nm}) - \log \beta_{nm}$, which are both found using the properties of the Gamma distribution. Substituting the expected complete data log-likelihood in (16) and rearranging leads to the VBM update rules for the hyperparameters:

$$\kappa_m = N\bar{\pi}_m + \kappa_0 \ , \tag{27}$$

$$\eta_m = N\bar{\omega}_m + \eta_0 \ , \tag{28}$$

$$\mathbf{m}_m = \frac{N\bar{\omega}_m\bar{\boldsymbol{\mu}}_m + \eta_0\mathbf{m}_0}{\eta_m} \ , \tag{29}$$

$$\gamma_m = N\bar{\pi}_m + \gamma_0 \tag{30}$$

$$\mathbf{S}_m = N\bar{\omega}_m\bar{\boldsymbol{\Sigma}}_m + \frac{N\bar{\omega}_m\eta_0}{\eta_m}\left(\bar{\boldsymbol{\mu}}_m - \mathbf{m}_0\right)\left(\bar{\boldsymbol{\mu}}_m - \mathbf{m}_0\right)^{\mathrm{T}} + \mathbf{S}_0 \ , \tag{31}$$

where (most of) the auxiliary variables correspond to the quantities computed in the M-step in ML learning:

$$\bar{\boldsymbol{\mu}}_m = \frac{1}{N\bar{\omega}_m} \sum_{n=1}^{N} \bar{\rho}_{nm} \bar{u}_{nm} \mathbf{x}_n \ , \tag{32}$$

$$\bar{\boldsymbol{\Sigma}}_m = \frac{1}{N\bar{\omega}_m} \sum_{n=1}^{N} \bar{\rho}_{nm} \bar{u}_{nm} \left(\mathbf{x}_n - \bar{\boldsymbol{\mu}}_m\right) \left(\mathbf{x}_n - \bar{\boldsymbol{\mu}}_m\right)^{\mathrm{T}} \ , \tag{33}$$

$$\bar{\pi}_m = \frac{1}{N} \sum_{n=1}^{N} \bar{\rho}_{nm} \ , \tag{34}$$

$$\bar{\omega}_m = \frac{1}{N} \sum_{n=1}^{N} \bar{\rho}_{nm} \bar{u}_{nm} \ . \tag{35}$$

$$\tag{36}$$

It is worth mentioning that the normalizing factor of the covariance matrices, which is here obtained automatically, is the one proposed by Kent, Tyler, and Vardi (1994) in order to accelerate the convergence of the ordinary EM algorithm.

Finally, since no prior is imposed on the degrees of freedom, we update them by maximizing the expected complete data log-likelihood. This leads to the same M-step as the one obtained by Peel and McLachlan (2000):

$$\log \frac{\nu_m}{2} + 1 - \psi\left(\frac{\nu_m}{2}\right) + \frac{1}{N\bar{\pi}_m} \sum_{n=1}^{N} \bar{\rho}_{nm} \left\{\log \tilde{u}_{nm} - \bar{u}_{nm}\right\} = 0 \ . \tag{37}$$

At each iteration and for each component, the fixed point can be easily found by line search. In contrast, Shoham (2002) proposed to use an approximate formula instead.

To end our discussion, we provide the expression of the variational lower bound:

$$\begin{aligned} &\mathrm{E}_{U,Z,\boldsymbol{\theta}_{\mathcal{S}}}\{\log p(X|U,Z,\boldsymbol{\theta}_{\mathcal{S}},\mathcal{H}_M)\} + \mathrm{E}_{U,Z,\boldsymbol{\theta}_{\mathcal{S}}}\{\log p(U,Z|\boldsymbol{\theta}_{\mathcal{S}},\mathcal{H}_M)\} \\ &+ \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log p(\boldsymbol{\theta}_{\mathcal{S}}|\mathcal{H}_M)\} - \mathrm{E}_{U,Z}\{\log q(U,Z)\} - \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{S}}}\{\log q(\boldsymbol{\theta}_{\mathcal{S}})\} \ . \end{aligned} \tag{38}$$

Note that the last two terms correspond to the entropies of the variational distributions. Given the functional form of the posteriors, each term of the bound can be computed (see Appendix). Since the Bayesian approach takes the uncertainty of the model parameters into account and since the lower bound is made as tight as possible during learning, it can be used as a model selection criterion.

4 Experimental results and discussion

In this section, the robustness of the variational Bayesian learning algorithm for the SMM is investigated. First, we show that this new algorithm enables us to perform robust automatic model selection based on the variational lower bound. Second, we focus on robust clustering and on the quality of the parameter estimates. Finally, some empirical evidence is given in order to explain why the proposed algorithm performs better than previous approaches.

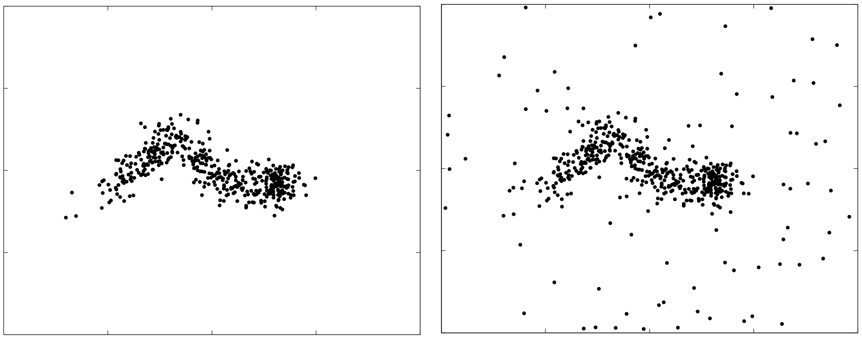
4.1 Robust automatic model selection

Before investigating the performance of the proposed approach compared to previous approaches, let us first illustrate the model selection procedure on a toy example. Consider a mixture of three multivariate Gaussian distributions with the following parameters:

$$\boldsymbol{\mu}_1 = (-6\ 1.5)^{\mathrm{T}}\ , \quad \boldsymbol{\mu}_2 = (0\ 0)^{\mathrm{T}}\ , \qquad \boldsymbol{\mu}_3 = (6\ 1.5)^{\mathrm{T}}\ ,$$
$$\boldsymbol{\Lambda}_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}^{-1}, \quad \boldsymbol{\Lambda}_2 = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}^{-1}, \quad \boldsymbol{\Lambda}_3 = \begin{bmatrix} 1.56 & 0 \\ 0 & 1.56 \end{bmatrix}^{-1} .$$

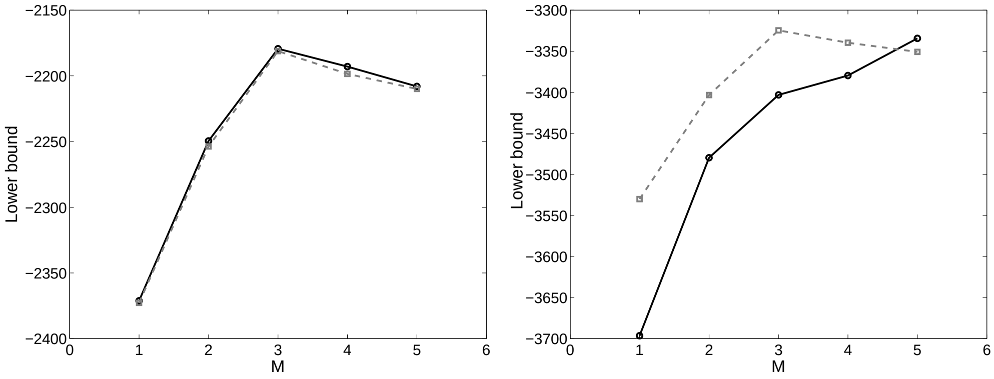
One hundred and fifty data points are drawn from each component (each component is thus equally likely). Two training set examples are shown in Figure 2. The first one contains no outliers, while the second one is the same data augmented by 25% of outliers. The outliers are drawn from a uniform distribution on the interval $[-20, 20]$, in each direction of the feature space. Figure 3 shows the variational lower bound in presence and absence of outliers, for both the Bayesian Gaussian mixture model (GMM) and Bayesian Student-t mixture model (SMM). The variational Bayesian algorithm is run 10 times. The curves in Figure 3 are thus averages. The model complexity M ranges from 1 to 5 components. When there are no outliers, the GMM and the SMM perform similarly. Both methods select the correct number of components, which is 3. In contrast, when there are atypical observations only the SMM selects the right number of components.

Next, let us consider the well-known Old Faithful Geyser data. The data are recordings of the eruption duration and the waiting time between successive eruptions. In the experiments, the data are normalized and then corrupted by a certain amount of outliers. The latter are generated uniformly on the interval $[-10, 10]$ in each direction of the feature space. Figure 4(a) shows the variational lower bound for the GMM, the type-1 SMM, which assumes that the variational posterior on the indicator variables and the scale vari-



(a) (b)

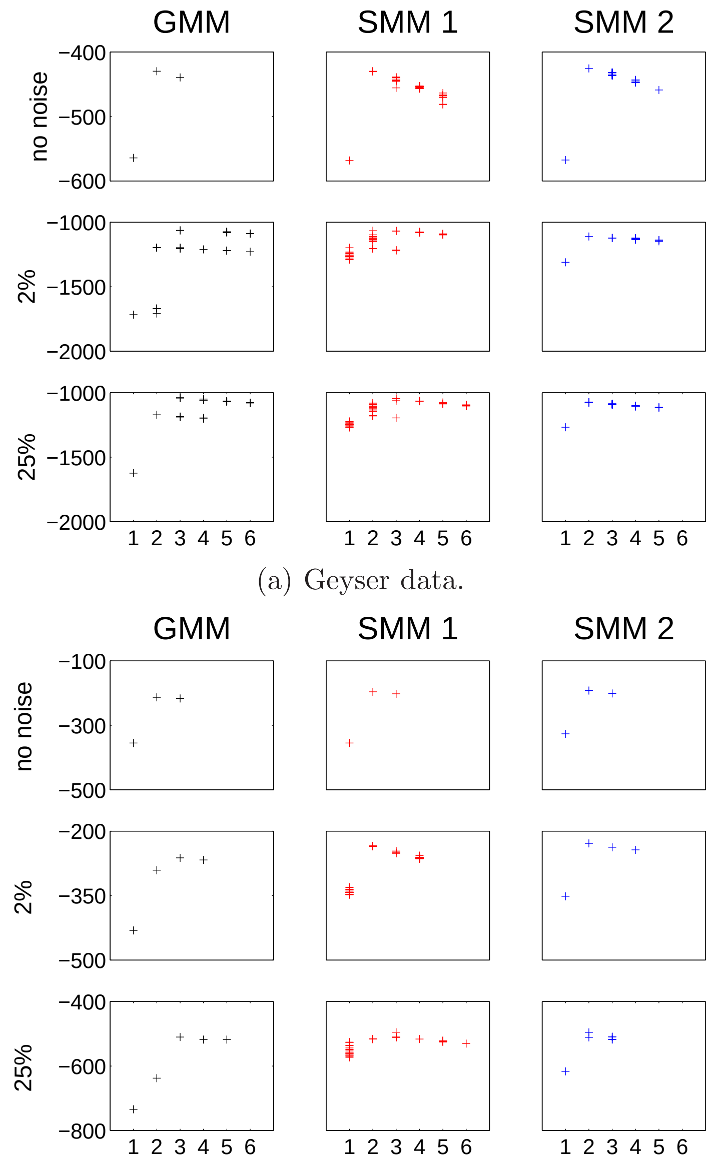
Fig. 2. Training sets. The data shown in (a) are generated from a mixture of three Gaussian distributions with different mean and precision. In (b) the same data are corrupted by 25% of atypical observations (uniform random noise).



(a) (b)

Fig. 3. Variational lower bound on the log-evidence versus the number of components M. The solid and the dashed lines correspond respectively to the lower bounds obtained for the GMM and the SMM. The curves show the average on 10 trials, (a) in absence and (b) in presence of outliers. The model complexity is selected according to the maximum of the lower bound.

ables factorizes (Svensén and Bishop, 2004), and the type-2 SMM, which does not make this assumption. The number of components is varied from 1 to 6. For each model complexity 20 runs are considered. Note that in some cases components are automatically pruned out when they do not have sufficient support. In absence of outliers, the bound of the three methods is maximal for two components. In presence of 2% of outliers the bound of the type-1 SMM has almost the same value for two and three components. This was also observed by Svensén and Bishop (2004). For the type-2 SMM, the bound is still maximal for two components. The GMM however favors 3 components. When the amount of noise further increases (25%), only the type-2 SMM selects 2 components. As a matter of fact, the value of the bound seems almost not affected by a further increase of the noise. Thus, not neglecting the correlation between the indicator variables and the scale variables clearly increases the



(a) Geyser data.

(b) Enzyme data.

Fig. 4. Variational lower bound for (a) the Old Faithful Geyser data and (b) the Enzyme data versus the number of components. An increasing number of outliers is successively added to the training set. Results are reported for the Bayesian GMM, as well as the type-1 and type-2 Bayesian SMM. Twenty runs are considered for all model complexities. The value of the bound obtained for each run is indicated by a cross. It is important to realize that the number of crosses for each model complexity might differ from method to method. The reason is that mixture components are pruned out during the learning process when there is too little evidence for them. Therefore, we did not use a standard representation such as box and whiskers plots to show the variability of the results but preferred this more intuitive representation.

robustness. A similar behavior was observed on the Enzyme data (Richardson and Green, 1997). The results are presented in Figure 4.

In practice, the type-2 SMM is less affected by the outliers as expected. However, a tighter lower bound was not always observed, but the bound appeared to be more stable from run to run than for the type-1 SMM. This suggests that the type-2 SMM is less sensitive to local maxima in the objective function.

4.2 Robust clustering

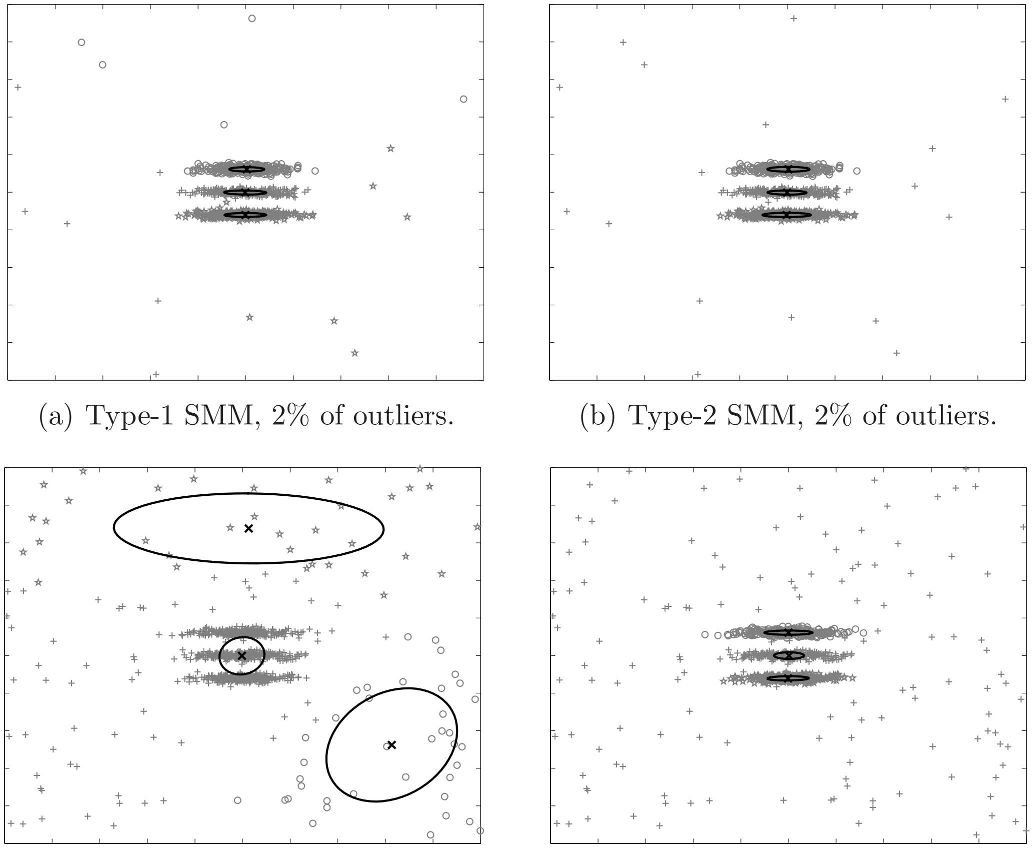
In order to assess the robustness of the type-2 SMM, we first consider the 3-component bivariate mixture of Gaussian distributions from Ueda and Nakano (1998). The mixture proportions are all equal to 1/3, the mean vectors are $(0\ -2)^{\mathrm{T}}$, $(0\ 0)^{\mathrm{T}}$ and $(0\ 2)^{\mathrm{T}}$, and the covariance matrix of each component is equal to $\mathrm{diag}\{2, 0.2\}$. The label assignment of the data points are presented in Figure 5. Two situations are considered. In presence of a small proportion of outliers (2%), the two Bayesian SMMs perform similarly. However, note that the type-2 SMM assigns the same label to all the outliers, while the type-1 SMM partitions the feature space in three parts. In presence of lots of outliers (15%) only the type-2 SMM provides a satisfactory solution. Still all outliers are assigned the same label, i.e. the label of the middle component. By contrast, the type-1 SMM cannot make a distinction between the data clumps and the outliers.

Next, we consider again the Old Faithful Geyser data. The goal is to illustrate that the parameter estimates of the type-2 SMM are less sensitive to outliers. Figure 6 shows the location of the means of the two components in presence and in absence of outliers. The ellipses correspond to a single standard deviation. It can easily be observed that the estimates of the means and the precisions are less affected by the outliers in the case of the type-2 SMM.

4.3 Effect of the factorization of the latent variable posteriors

As already mentioned, the type-1 SMM assumes that the variational posterior on the latent indicator variables and the latent scale variables factorize, as well as the priors on the component means and precisions. However, we have demonstrated in Section 3 that these factorizations are not necessary. In particular, taking into account the correlations between the indicator and the scale variables leads to a model with (i) an increased robustness to atypical observations and (ii) to a tighter lower bound on the log-evidence.

From (32–35) it can be observed that the responsibilities play an essential role in the estimation of the component parameters. As a consequence, accurate estimates are mandatory when considering robust mixture models. As shown in Section 3, when we assume that the scale variables are conditionally depen-



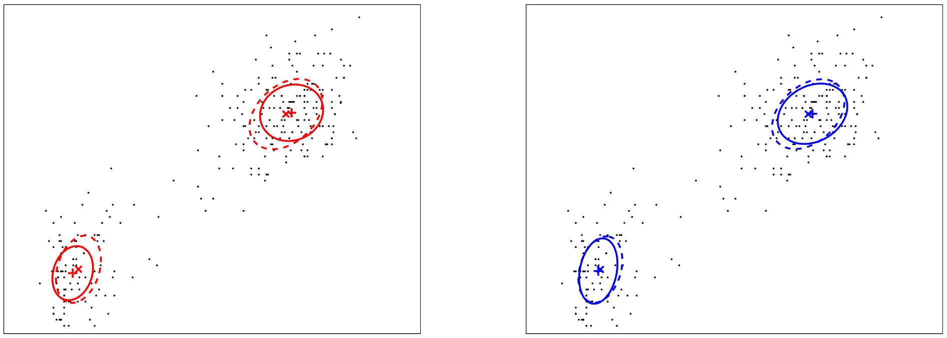
(a) Type-1 SMM, 2% of outliers.

(b) Type-2 SMM, 2% of outliers.

(c) Type-1 SMM, 15% of outliers.

(d) Type-2 SMM, 15% of outliers.

Fig. 5. Reconstructed data labels by the Bayesian type-1 and type-2 SMMs. (a) and (b) are the models obtained when 2% of outliers is added to the training set, while (c) and (d) are the ones obtained in presence of 15% of outliers.



(a) Type-1 SMM.

(b) Type-2 SMM.

Fig. 6. Old Faithful Geyser data. The markers '×' and '+' indicate respectively the means in absence and presence of outliers (25%). The dashed curves and the solid curves correspond respectively to single standard deviation in absence and presence of outliers. The models are constructed with 2 components.

dent on the indicator variables (type-2 SMM), we end up with responsibilities of the following form:

$$\bar{\rho}_{nm} \propto \tilde{\pi} \times \text{Student}-t \text{ distribution} . \tag{39}$$

The Student-t is an infinite mixture of scaled Gaussian distribution and the prior on the scales is assumed to be Gamma distributed. Clearly, the uncertainty on the scale variable is explicitly taken into account when estimating the responsibilities by (39), as the scale variables (which can be here viewed as nuisance parameters) are integrated out.

By contrast, Svensén and Bishop (2004) neglect the dependencies between the scale and the indicator variables (type-1 SMM) and therefore find responsibilities of the following form:

$$\bar{\rho}_{nm}^{(SB)} \propto \tilde{\pi} \times \text{scaled Gaussian distribution} , \tag{40}$$

where the scale is equal to $\bar{u}_{nm}$. Thus, in this approach we find that each data point is assumed to be generated from a *single* Gaussian distribution, its covariance matrix being scaled. In other words, it is assumed that the posterior distribution of the corresponding scale variable is highly peaked around its mean and therefore that the mean is a good estimate for the scale. Of course, this is not true for all data points.

In Figure 7, the typical variational posterior of a single data point $\mathbf{x}_n$ is shown. It can be observed that the type-1 SMM assigns the probability mass almost exclusively to one component (here to component $m = 2$) and that the posterior for that component is more peaked than the posterior of the type-2 SMM. This suggests that the empirical variance is (even more) underestimated when assuming that the scale variables are independent from the indicator variables. Since the uncertainty is underestimated, the robustness of the model is reduced. This was also observed experimentally.

Obtaining a tighter and more reliable variatonal lower bound is also important. When using the lower bound as a model selection criterion, it is implicitly assumed that the gap between the log-evidence and the bound is identical after convergence for models of different complexity. In general, this is not true. Usually, variational Bayesian inference tends to overpenalize complex models, as the factorized approximations lead to a posterior that is more compact (i.e., less complex) than the true posterior. This can be understood by seeing that maximizing the lower bound is done by minimizing the KL divergence between the variational posterior and the true posterior. However, the KL divergence is taken with respect to the support of the variational distribution and not with respect to the support of the true posterior. Therefore, the optimal variational posterior underestimates the correlations between the latent variables and the parameters, and in turn leads to an approximation of the joint posterior that

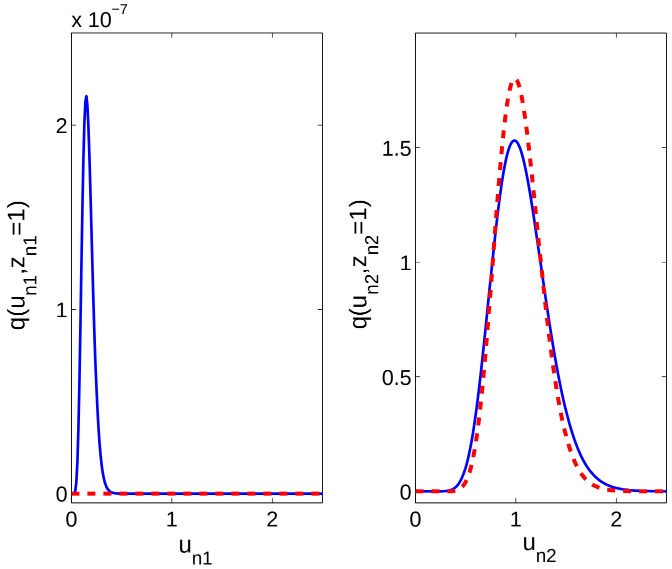


Fig. 7. The typical joint variational posterior $q(\mathbf{u}_n|\mathbf{z}_n)$ of the indicator and the scale variable for a single data point $\mathbf{x}_n$. The mixture has two components. The data is the Old Faithful Geyser data. The solid curve does not neglect the correlation between both latent variables (type-2 SMM), while the dashed curve does (type-1 SMM).

is more peaked. Now, the type-1 SMM makes the additional assumption that the distribution on latent variable factorizes as well. As a result, the type-1 SMM makes additional approximations compared to the type-2 SMM, such that the approximate posterior is even more compact. In practice, this leads, for example depending on the initialization, to a less reliable estimate of the lower bound (see Figure 4).

5 Conclusion

In this article, we derive new variational update rules for Bayesian mixtures of Student-t distributions. It was demonstrated that it is not required to assume a factorized variational posterior on the indicator and the scale variables. Taking the correlation between these latent variables into account leads to a variational posterior that is less compact than the one obtained in previous approaches; therefore it underestimates less the uncertainty in the latent variables. Although the resulting lower bound is not always tighter, the correct model complexity is selected in a more consistent way, as it is less sensitive to local maxima of the objective function. Finally, it was shown experimentally that the resulting model is less sensitive to outliers, which leads to very robust mixture modeling in practice.

Appendix

The following expressions are obtained for each term of the variational lower bound (38):

$$
\begin{aligned}
&\sum_{Z}\iint q_U(U|Z)q_Z(Z)q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})\log p(X|U,Z,\boldsymbol{\theta}_\mathcal{S},\mathcal{H}_M)dU d\boldsymbol{\theta}_\mathcal{S}\\
&\qquad = \textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\bar{\rho}_{nm}\Big\{-\frac{d}{2}\log 2\pi+\frac{d}{2}\log\tilde{u}_{nm}+\frac{1}{2}\log\tilde{\Lambda}_m\\
&\qquad\quad -\frac{\bar{u}_{nm}\gamma_m}{2}(\mathbf{x}_n-\mathbf{m}_m)^{\mathrm{T}}{\mathbf{S}_m}^{-1}(\mathbf{x}_n-\mathbf{m}_m)-\frac{\bar{u}_{nm}d}{2\eta_m}\Big\}\ ,
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
&\sum_{Z}\iint q_U(U|Z)q_Z(Z)q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})\log p(U|Z,\boldsymbol{\theta}_\mathcal{S},\mathcal{H}_M)dU d\boldsymbol{\theta}_\mathcal{S}\\
&\qquad = \textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\bar{\rho}_{nm}\Big\{\frac{\nu_m}{2}\log\frac{\nu_m}{2}-\log\Gamma\left(\frac{\nu_m}{2}\right)\\
&\qquad\quad +\left(\frac{\nu_m}{2}-1\right)\log\tilde{u}_{nm}-\frac{\nu_m}{2}\bar{u}_{nm}\Big\}\ ,
\end{aligned}
\tag{42}
$$

$$
\begin{aligned}
&\sum_{Z}\int q_Z(Z)q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})\log p(Z|\boldsymbol{\theta}_\mathcal{S},\mathcal{H}_M)\boldsymbol{\theta}_\mathcal{S}\\
&\qquad = \textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\bar{\rho}_{nm}\log\tilde{\pi}_m\ ,
\end{aligned}
\tag{43}
$$

$$
\begin{aligned}
&\int q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})\log p(\boldsymbol{\theta}_\mathcal{S}|\mathcal{H}_M)d\boldsymbol{\theta}_\mathcal{S}\\
&\qquad = \log c_\mathcal{D}(\boldsymbol{\kappa}_0)+\textstyle\sum_{m=1}^{M}(\kappa_0-1)\log\tilde{\pi}_m+\sum_{m=1}^{M}\Big\{-\frac{d}{2}\log 2\pi\\
&\qquad\quad +\frac{d}{2}\log\eta_0-\frac{\gamma_m\eta_0}{2}(\mathbf{m}_m-\mathbf{m}_0)^{\mathrm{T}}{\mathbf{S}_m}^{-1}(\mathbf{m}_m-\mathbf{m}_0)-\frac{\eta_0 d}{2\eta_m}\\
&\qquad\quad +\log c_{\mathcal{NW}}(\gamma_0,\mathbf{S}_0)+\frac{\gamma_0-d}{2}\log\tilde{\Lambda}_m-\frac{\gamma_m}{2}\mathrm{tr}\{\mathbf{S}_0{\mathbf{S}_m}^{-1}\}\Big\}\ ,
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&\sum_{Z}\int q_U(U|Z)q_Z(Z)\log q_U(U|Z)dU\\
&\qquad = \textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\bar{\rho}_{nm}\Big\{-\log\Gamma\left(\alpha_{nm}\right)+\left(\alpha_{nm}-1\right)\psi\left(\alpha_{nm}\right)\\
&\qquad\quad +\log\beta_{nm}-\alpha_{nm}\Big\}\ ,
\end{aligned}
\tag{45}
$$

$$
\begin{aligned}
&\sum_{Z}q_Z(Z)\log q_Z(Z)\\
&\qquad = \textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\bar{\rho}_{nm}\log\bar{\rho}_{nm}\ ,
\end{aligned}
\tag{46}
$$

$$
\begin{aligned}
&\int q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})\log q_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{\theta}_\mathcal{S})d\boldsymbol{\theta}_\mathcal{S}\\
&\qquad = \log c_\mathcal{D}(\boldsymbol{\kappa})+\textstyle\sum_{m=1}^{M}(\kappa_m-1)\log\tilde{\pi}_m+\sum_{m=1}^{M}\Big\{-\frac{d}{2}\log 2\pi\\
&\qquad\quad +\frac{d}{2}\log\eta_m-\frac{d}{2}+\log c_{\mathcal{NW}}(\gamma_m,\mathbf{S}_m)+\frac{\gamma_m-d}{2}\log\tilde{\Lambda}_m-\frac{\gamma_m d}{2}\Big\}\ .
\end{aligned}
\tag{47}
$$

References

Archambeau, C., Lee, J. A., Verleysen, M., 2003. On the convergence problems of the EM algorithm for finite Gaussian mixtures. In: Eleventh European Symposium on Artificial Neural Networks. D-side, pp. 99–106.

Attias, H., 1999. A variational Bayesian framework for graphical models. In: Solla, S. A., Leen, T. K., Müller, K.-R. (Eds.), Advances in Neural Information Processing Systems 12. MIT Press, pp. 209–215.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statisitcal Society B 39, 1–38.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London.

Ghahramani, Z., Beal, M. J., 2001. Propagation algorithms for variational bayesian learning. In: Leen, T. K., Dietterich, T. G., Tresp, V. (Eds.), Advances in Neural Information Processing Systems 13. MIT Press, pp. 507–513.

Kent, J. T., Tyler, D. E., Vardi, Y., 1994. A curious likelihood identity for the multivariate t-distribution. Communications in Statistics – Simulation and Computation 23 (2), 441–453.

Liu, C., Rubin, D. B., 1995. ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica 5, 19–39.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. John Willey and Sons, New York.

Parzen, E., 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Peel, D., McLachlan, G. J., 2000. Robust mixture modelling using the t distribution. Statistics and Computing 10, 339–348.

Richardson, S., Green, P., 1997. On Bayesian analysis of mixtures with unknown number of components. Journal of the Royal Statistical Society B 59, 731–792.

Shoham, S., 2002. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. Pattern Recognition 35 (5), 1127–1142.

Svensén, M., Bishop, C. M., 2004. Robust Bayesian mixture modelling. Neurocomputing 64, 235–252.

Ueda, N., Nakano, R., 1998. Deterministic annealing EM algorithm. Neural Networks 11, 271–282.

Yamazaki, K., Watanabe, S., 2003. Singularities in mixture models and upper bounds of stochastic complexity. Neural Networks 16 (7), 1029–1038.