Robust Bayesian Mixture Modelling

Markus Svensén and Christopher M. Bishop,

Microsoft Research, 7 J J Thomson Avenue, Cambridge, CB3 0FB, UK

Abstract

Bayesian approaches to density estimation and clustering using mixture distributions allow the automatic determination of the number of components in the mixture. Previous treatments have focussed on mixtures having Gaussian components, but these are well known to be sensitive to outliers, which can lead to excessive sensitivity to small numbers of data points and consequent over-estimates of the number of components. In this paper we develop a Bayesian approach to mixture modelling based on Student-t distributions, which are heavier tailed than Gaussians and hence more robust. By expressing the Student-t distribution as a marginalisation over additional latent variables we are able to derive a tractable variational inference algorithm for this model, which includes Gaussian mixtures as a special case. Results on a variety of real data sets demonstrate the improved robustness of our approach.

This paper is published in Neurocomputing, volume 64, pages 235–252. Due to different formatting, there is no exact page correspondance between this version and the published version, but the content should be identical. The published version is available at `www.sciencedirect.com`.

1 Introduction

Mixture models are ubiquitous in virtually every facet of statistical analysis, machine learning and data mining. For data sets comprising continuous variables, the most common approach involves mixture distributions having Gaussian components fitted by maximum likelihood, for which the EM algorithm has a closed-form M-step.

Email addresses: `markussv@microsoft.com` (Markus Svensén), `cmbishop@microsoft.com` (Christopher M. Bishop).

URLs: `http://research.microsoft.com/~markussv` (Markus Svensén), `http://research.microsoft.com/~cmbishop` (Christopher M. Bishop).

A central issue in mixture modelling is the choice of the number of components in the mixture. Maximum likelihood is unable to address this issue since it favours ever more complex models, leading to over-fitting. Another difficulty concerns the infinities which plague the likelihood function, associated with the collapsing of Gaussian components onto individual data points. These problems can be resolved elegantly by adopting a Bayesian framework in which we marginalize over the model parameters with respect to appropriate priors. The resulting model likelihood can then be maximized with respect to the number of components in the mixture if the goal is model selection, or combined with a prior over the number of components if the goal is model averaging. While exact Bayesian inference for Gaussian mixtures is intractable, it has been addressed through Markov chain Monte Carlo (MCMC). Diebolt and Robert (8) described the case with a fixed number of components, although this framework could incorporate model selection by simply comparing the (MCMC-approximated) marginal likelihood (also known as the *evidence* (12)) of models with different number of components. Richardson and Green (15) proposed an alternative MCMC-based model fitting procedure which involves also the number of components, and also other MCMC-schemes have been proposed, e.g. (14). The main drawbacks of MCMC-techniques are that they are generally computationally demanding, and that it can be difficult to diagnose convergence. More recently, variational methods have emerged as a deterministic alternative for doing Bayesian inference, with much better scaling properties in terms of computational cost (10; 16). They have been applied to Gaussian mixture models (1; 4), thereby avoiding the singularity problems of maximum likelihood.

A major limitation of Gaussian mixture models, however, is their lack of robustness to outliers. This is easily understood by recalling that maximization of the likelihood function under an assumed Gaussian distribution is equivalent to finding the least-squares solution, whose lack of robustness is well known. In the Bayesian model determination context, the presence of outliers or other departure of the empirical distribution from Gaussianity can lead to errors in the determination of the number of clusters in the data. In particular, the Gaussian mixture model tends to over-estimate the number of clusters since it uses additional components to capture the tails of the distributions.

In an earlier paper (3) we proposed a Bayesian treatment of mixture models based on components having a Student distribution (13) which has heavier tails compared to the exponentially decaying tails of a Gaussian. In this paper, we extend the earlier description of this model to give a full account of the variational inference framework employed. The next section will describe the Bayesian mixture of Student distributions, and in Section 3, we describe how to infer a tractable variational posterior distribution over the parameters in the Student mixture model, with full details given in the appendix. Section 4 presents results on applying this model to real data sets, focussing on its robustness in presence of outliers as compared to a Bayesian Gaussian mixture model.

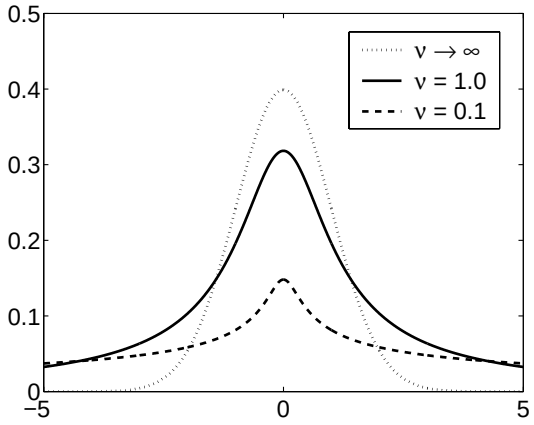

Fig. 1. The plot shows the univariate Student distribution $\mathcal{S}(x|\mu, \Lambda, \nu)$ with μ and Λ fixed for various values of ν, in which $\nu \to \infty$ corresponds to a Gaussian.

2 Bayesian Student Mixture Models

Our approach to robust Bayesian mixture modelling is based on component distributions given by a multivariate Student distribution, also known as a t-distribution,

$$\mathcal{S}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\nu/2 + d/2)|\boldsymbol{\Lambda}|^{1/2}}{\Gamma(\nu/2)(\nu\pi)^{d/2}} \left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+d)/2} \tag{1}$$

where

$$\Gamma(y) = \int_0^\infty z^{y-1} e^{-z}\, dz \tag{2}$$

is the Gamma function and

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) \tag{3}$$

is the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$. $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ denote the mean and precision (inverse covariance) matrix, respectively, d denotes the dimensionality of $\mathbf{x}$, and ν is a parameter known as the number of 'degrees of freedom'. The Student distribution represents a generalization of the Gaussian, and in the limit $\nu \to \infty$ it reduces to a Gaussian with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$. For finite values of ν this distribution has heavier tails than the corresponding Gaussian having the same $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, as shown in Figure 1.

In contrast to the Gaussian, there is no closed form solution for maximizing likelihood under a Student distribution. However, there is a useful representation of the Student distribution as an infinite mixture of scaled Gaussians. In particular we can write the Student distribution in the form

$$\mathcal{S}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, u\boldsymbol{\Lambda}) \mathcal{G}(u|\nu/2, \nu/2)\, du \tag{4}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ denotes the Gaussian distribution,

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

and Δ^2 is defined by (3). The Gamma distribution is given by

$$\mathcal{G}(u|a, b) = \frac{1}{\Gamma(a)} a^b u^a e^{-bu}.$$

Using (2), it is straightforward to evaluate the left hand side of (4) to obtain (1). We can think of (4) as introducing an implicit latent variable u for each observation of $\mathbf{x}$, and this can be exploited to find maximum likelihood solutions using the EM algorithm (11), in which the E step involves an expectation with respect to the posterior distribution of the latent variable u. The M-step then has closed form solutions for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, and the value of ν can be found by solution of a simple nonlinear equation. An analogous strategy will be exploited in Section 3 to obtain a tractable Bayesian treatment of Student mixture distributions based on variational inference.

We now consider densities comprising mixtures of Student distributions,

$$p(\mathbf{x}|\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m\}, \boldsymbol{\pi}) = \sum_{m=1}^{M} \pi_m \mathcal{S}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m) \tag{6}$$

where the mixing coefficients $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)^{\mathrm{T}}$ satisfy $\pi_m \geqslant 0$ and $\sum_m \pi_m = 1$.

In order to find a tractable variational treatment of this model, we re-express the mixture density in terms of a marginalization over a binary latent variable $\mathbf{s}$ of dimensionality M having components $\{s_j\}$ such that $s_j = 1$ for $j = m$ and $s_j = 0$ for $j \neq m$, giving

$$p(\mathbf{x}|\mathbf{s}, \{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m\}) = \prod_{m=1}^{M} \mathcal{S}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m)^{s_m} \tag{7}$$

with a corresponding multinomial prior distribution over $\mathbf{s}$ of the form

$$p(\mathbf{s}|\boldsymbol{\pi}) = \prod_{m=1}^{M} \pi_m^{s_m}. \tag{8}$$

It is easily verified that marginalization of the product of (7) and (8) over the latent variable $\mathbf{s}$ recovers the Student mixture (6).

We consider a data set $\mathbf{X}$ comprising observations $\{\mathbf{x}_n\}$, where $n = 1, \ldots, N$, which we shall suppose are drawn independently from the distribution (6). Thus for each data observation $\mathbf{x}_n$ we have corresponding discrete latent variable $\mathbf{s}_n$ specifying which component of the mixture generated that data point, and a continuous

latent variable u_{nm} specifying the scaling of the precision for the corresponding equivalent Gaussian from which the data point was hypothetically generated.

In a Bayesian treatment, we also need priors over the other variables in the model, and again for tractability we choose conjugate priors [1] from the exponential family:

$$p(\boldsymbol{\mu}_m) = \mathcal{N}(\boldsymbol{\mu}_m|\mathbf{m}_0, \rho_0\mathbf{I})$$

as defined in (5),

$$\begin{aligned} p(\boldsymbol{\Lambda}_m) &= \mathcal{W}(\boldsymbol{\Lambda}_m|\mathbf{W}_0, \eta_0) \\ &= C_{\mathcal{W}}(\mathbf{W}_0, \eta_0)\boldsymbol{\Lambda}_m^{(\eta_0-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1}\boldsymbol{\Lambda}_m\right)\right) \end{aligned} \tag{9}$$

which is the Wishart distribution, where the normalisation constant is given by

$$C_{\mathcal{W}}(\mathbf{W}_0, \eta_0) = |\mathbf{W}_0|^{-\eta_0/2}\left(2^{\eta_0 d/2}\pi^{d(d-1)/4}\prod_i^d \Gamma\left(\frac{\eta_0+1-i}{2}\right)\right)^{-1}. \tag{10}$$

Finally,

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = C_{\mathcal{D}}\prod_m^M \pi_m^{\alpha_m-1} \tag{11}$$

which is the Dirichlet distribution, where

$$C_{\mathcal{D}} = \frac{\Gamma(\alpha_0)}{\prod_m^M \Gamma(\alpha_m)} \tag{12}$$

and

$$\alpha_0 = \sum_m^M \alpha_m\ . \tag{13}$$

The parameters of the prior on $\boldsymbol{\mu}_m$ are chosen to give broad distributions, $\mathbf{m}_0 = \mathbf{0}$, $\rho_0 = 10^{-3}$, whereas the prior for $\boldsymbol{\Lambda}_m$ is chosen more moderate, $\mathbf{W}_0 = \mathbf{I}$ and $\eta_0 = 1$. For the prior over $\boldsymbol{\pi}$ we can interpret the parameters $\boldsymbol{\alpha} = \{\alpha_m\}$ as effective numbers of prior observations, which we set to $\alpha_m = 10^{-3}$.

The joint distribution of all random variables can be expressed as a directed graph, as shown in Figure 2. Note that ν is treated as a (non-stochastic) parameter since there is no conjugate prior for ν. However, since there is only one such parameter per mixture component, so we set its value by optimization as part of the variational procedure discussed next.

[1] The fully conjugate prior for unknown mean and precision of a Gaussian is the normal-Wishart. Here we use separate normal and Wishart priors, and this is tractable due to the later assumed factorization, although the analysis is easily extended to the normal-Wishart distribution if desired.

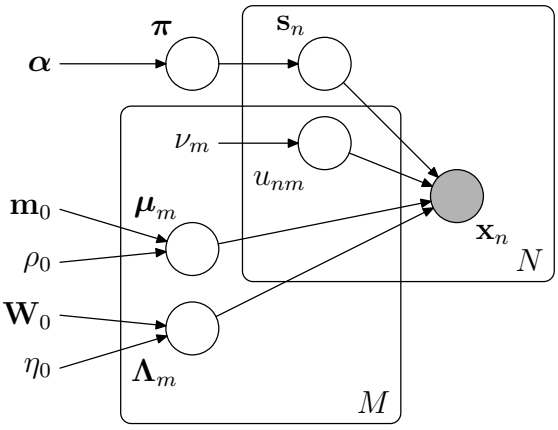

Fig. 2. The diagram shows a representation of the Bayesian Student mixture distribution as a directed graphical model. The boxes denote 'plates', comprising replicas of the entities inside the plates. The M-plate represents the M mixture components and the N-plate the independent identically distributed observations $\mathbf{x}_n$. The circular nodes denote random variables, whereas labels not associated with nodes denote constants (e.g. $\boldsymbol{\alpha}$) or adjustable parameters (ν_m). The shading of the $\mathbf{x}_n$ node indicates that this variable is observed whereas variables denoted by the other (unshaded) nodes are hidden. Note that $\{u_{nm}\}$ belong to both plates, indicating that there are corresponding random variables for each mixture component and each observation.

3 Variational Inference

Exact inference in our Bayesian model is intractable. However, the choice of conjugate-exponential distributions allows us to find an elegant variational framework. To do this we consider the well known equality (10) for the log marginal likelihood given by

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \tag{14}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \left\{ \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}, \tag{15}$$

where in turn $\boldsymbol{\theta}$ denotes the set of all unobserved stochastic variables, which for our model comprise $\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}, \boldsymbol{\pi}, \{\mathbf{s}_n, \mathbf{u}_n\}$. Here $q(\boldsymbol{\theta})$ denotes a variational posterior distribution, and $p(\mathbf{X}, \boldsymbol{\theta})$ is the joint distribution over all hidden and observed variables, as defined by the model in Figure 2. The second term on the right hand side of (14) is the Kullback-Leibler (KL) divergence between the variational posterior distribution $q(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta}|\mathbf{X})$,

$$\mathrm{KL}(q\|p) = -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|\mathbf{X})}{q(\boldsymbol{\theta})} \, d\boldsymbol{\theta}.$$

Since the KL divergence is non-negative, $\mathcal{L}(q)$ is a lower bound of the log marginal likelihood. This bound would become exact if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{X})$, in which case the KL-divergence would vanish, but if so, we would not have had to resort to approximate inference in the first place.

In order to make progress, we choose a constrained family of distributions for $q(\boldsymbol{\theta})$ such that the evaluation of the lower bound $\mathcal{L}(q)$ becomes tractable. Here we assume a factorized variational distribution of the form

$$q(\boldsymbol{\theta}) = q(\{\boldsymbol{\mu}_m\})q(\{\boldsymbol{\Lambda}_m\})q(\boldsymbol{\pi})q(\{\mathbf{s}_n\})q(\{\mathbf{u}_n\}) \tag{16}$$

where we are using a generic notation $q(\cdot)$ for all of the factors, since the particular distribution involved can be determined from its argument. This fully factorized approximation is often referred to as the 'mean field approximation' and has its roots in statistical physics (e.g. (6)). It can be seen as one particular instance from the set of variational approximations (10). A consequence of this choice is that our variational approximation will not capture correlations among the factors, which in turn will lead to an under-estimation of the variance in the posterior distribution, the consequences of which we will see in Section 4.

With a chosen family of approximating distributions, we can now search for the optimal member of this family by maximization of $\mathcal{L}(q)$, which is equivalent to minimization of the KL divergence. This is achieved by optimizing with respect to each factor in turn, holding the others fixed. It is easily shown that the log of the optimum solution for a particular factor is obtained, up to an additive constant, by taking the log of the joint distribution for the model and averaging with respect to variational distributions of all the remaining factors. Since this represents a coupled solution, it is necessary to cycle through the factors in turn in an iterative manner. As a consequence of the conjugate-exponential structure of the model, the resulting optimal factors take the same functional form as the corresponding conditional distributions comprising $p(\mathbf{X}, \boldsymbol{\theta})$ (2; 16).

For instance, for the variational posterior distribution $q(\{\boldsymbol{\Lambda}_m\})$ this yields

$$\ln q(\boldsymbol{\Lambda}_m) = \langle \ln p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{m}_0, \rho_0, \mathbf{W}_0, \eta_0, \nu, \boldsymbol{\alpha}) \rangle_{\boldsymbol{\mu},\mathbf{u},\mathbf{s}} + \text{const.} \tag{17}$$

$$= \frac{N}{2} \ln |\boldsymbol{\Lambda}_m| - \frac{1}{2} \text{Tr} \left[\sum_n^N \left\langle (\mathbf{x}_n - \boldsymbol{\mu}_m)(\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}} u_{nm} s_{nm} \right\rangle_{\boldsymbol{\mu},\mathbf{u},\mathbf{s}} \boldsymbol{\Lambda}_m \right]$$
$$+ \frac{\eta_0 - d - 1}{2} \ln |\boldsymbol{\Lambda}_m| - \frac{1}{2} \text{Tr}[\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_m] + \text{const.} \tag{18}$$

where in (17) we have used Bayes' theorem and the last term summarizes terms independent of $\boldsymbol{\Lambda}_m$. Here $\langle \cdot \rangle_{\boldsymbol{\mu},\mathbf{u},\mathbf{s}}$ denotes expectations with respect to the corresponding $q(\cdot)$.

From (9) and (18) we see that

$$q(\boldsymbol{\Lambda}_m) = \mathcal{W}(\boldsymbol{\Lambda}_m | \mathbf{W}_m, \eta_m) \tag{19}$$

where

$$\begin{aligned}\mathbf{W}_m^{-1} &= \mathbf{W}_0^{-1} + \sum_n^N \left\langle (\mathbf{x}_n - \boldsymbol{\mu}_m)(\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}} u_{nm} s_{nm} \right\rangle_{\boldsymbol{\mu},\mathbf{u},\mathbf{s}} \\ &= \mathbf{W}_0^{-1} + \sum_n^N \langle u_{nm} \rangle \langle s_{nm} \rangle \left(\mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} - 2\mathbf{x}_n \langle \boldsymbol{\mu}_m \rangle^{\mathrm{T}} + \langle \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\mathrm{T}} \rangle \right) \end{aligned} \tag{20}$$

and

$$\eta_m = \eta_0 + \sum_n^N \langle s_{nm} \rangle. \tag{21}$$

Thus the optimal solution for the factor $q(\boldsymbol{\Lambda}_m)$ depends on moments evaluated with respect to other factors in the variational posterior, in this case $\langle s_{nm} \rangle$, $\langle u_{nm} \rangle$, $\langle \boldsymbol{\mu}_m \rangle$ and $\langle \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\mathrm{T}} \rangle$. Analogous results are obtained for the other factors in (16), and the details of these, together with formulae for the necessary moments can be found in Appendix A. Since the solutions for the variational factors are all coupled we solve them iteratively by first initializing the distributions and then cycling round each factor in turn and replacing its current estimate by its optimal solution given the current estimates for the other factors. We also update ν_m for each mixture component, replacing it with its log-marginal maximum likelihood estimate, by setting the corresponding gradient to zero and solving the resulting (independent) non-linear equations [2]

$$1 + \frac{1}{\hat{s}_m} \sum_n^N \langle s_{nm} \rangle \left[\langle \ln u_{nm} \rangle - \langle u_{nm} \rangle \right] + \ln \frac{\nu_m}{2} - \Psi\left(\frac{\nu_m}{2} \right) = 0$$

where

$$\hat{s}_m = \sum_n^N \langle s_{nm} \rangle$$

and

$$\Psi(a) = \frac{d \ln \Gamma(a)}{da} \tag{22}$$

commonly known as the di-gamma function.

Note that the only assumption made about $q(\boldsymbol{\theta})$ is that it factorizes as in (16). It is thus interesting to note that the optimal solutions for the factors in the variational posterior distribution exhibit additional factorizations arising from interactions between the assumed form of (16) and the conditional independence properties of the joint distribution $p(D, \boldsymbol{\theta})$ (2). For instance, in the optimal solution for $q(\{\boldsymbol{\Lambda}_m\})$ given by (19) there is an additional factorization with respect to the different components $\boldsymbol{\Lambda}_m$. Similar 'consequential' factorizations with respect to m are observed in $q(\{\boldsymbol{\mu}_m\})$, and with respect to n in $q(\{\mathbf{u}_n\})$ and $q(\{\mathbf{s}_n\})$.

[2] We are using the `fzero` function in Matlab ® (`http://www.mathworks.com/matlab`) to do this.

3.2 The Lower Bound

The lower bound (15) can be re-written as:

$$\begin{aligned}\mathcal{L}(q) = &\langle \ln p(\mathbf{x}|\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}, \mathbf{u}, \mathbf{s})\rangle + \sum_m^M \langle \ln p(\boldsymbol{\mu}_m|\mathbf{m}_0, \rho_0)\rangle \\ &+ \sum_m^M \langle \ln p(\boldsymbol{\Lambda}_m|\mathbf{W}_0, \eta_0)\rangle + \langle \ln p(\mathbf{u}|\nu)\rangle + \langle \ln p(\boldsymbol{\pi}|\hat{\boldsymbol{\alpha}})\rangle + \langle \ln p(\mathbf{s}|\boldsymbol{\pi})\rangle \\ &- \sum_m^M \langle \ln q(\boldsymbol{\mu}_m)\rangle - \sum_m^M \langle \ln q(\boldsymbol{\Lambda}_m)\rangle - \langle \ln q(\mathbf{u})\rangle - \langle \ln q(\boldsymbol{\pi})\rangle - \langle \ln q(\mathbf{s})\rangle. \quad (23)\end{aligned}$$

Thus, it can be evaluated in terms of moments of the variational posterior factors, which typically already have been computed as part of the inference procedure, and can therefore be efficiently computed. Details are given in Appendix B. The value of the bound can be monitored during the optimization and can be used to set a convergence criterion.

Evaluation of the lower bound plays a useful role in checking the correctness of the variational formulae, as well as their software implementation, since at every update the value of the bound must not decrease. We have taken this a stage further and used numerical central differences to evaluate the derivatives of the bound with respect to the parameters of each factor in the variational distribution immediately after the corresponding factor has been updated. The central differences must take account of constraints such as positivity, symmetry, or summation to unity and this is done through appropriate changes of variables. For instance the variational posterior distribution $q(\boldsymbol{\pi})$ is a Dirichlet with parameters $\widetilde{\alpha}_m$ and by expressing these using a softmax (normalized exponential) transformation applied to unconstrained variables γ_m we can make small variations in the γ_m and ensure that the positivity and summation constraints on the $\widetilde{\alpha}_m$ remain satisfied. The central difference evaluation of the derivatives should give zero immediately after each update, and this provides a powerful check on the correctness of the software implementation. This diagnostic is switched off once the correctness of the code is confirmed, in order to save computation.

The lower bound is a non-convex function of the variational posterior distribution, and so there will in general exist multiple maxima, and the resulting solution will depend on the initialization. We address this by performing multiple optimizations from random starts, and retaining the solution giving the largest value of the resulting bound $\mathcal{L}(q)$. Note that, as a consequence of adopting a Bayesian approach, this procedure can make use of the entire training set in a single pass of training and does not require cross-validation.

4 Experimental Results

We now present the results of applying the Student mixture model to four real data sets. First, however, we note that if a model with an excess of components is used, then in our Bayesian treatment the unwanted components simply revert to their prior distributions, and do not interact with the data. The corresponding prior and posterior terms in the lower bound cancel out, and there is no contribution to the predictive distribution, so that such components are effectively pruned out of the model. This arises because any component whose parameters deviate from their prior distributions will incur a penalty, the significance of which will be determined by the deviation and the broadness of the prior (12). We say that the *effective* number of components is the number of components for which there exists at least one data point for which the posterior probability, or 'responsibility', that the component generated this data point is numerically greater than zero.

We have compared mixture models with Gaussian (GMM) and Student (SMM) components on four real data sets, to which we optionally add outliers. Specifically, we have fitted GMMs and SMMs having between 1 and 6 mixture components to these data sets, with and without the added outliers, and then compared them in terms of the resulting bounds as well as the effective number of mixture components used in the fitted models. For each model we used 50 different random initializations to handle the non-convexity of the lower bound. In maximum likelihood approaches it is common to use cross-validation against an independent data set to select an appropriate model complexity. However, in our Bayesian approach we can simply use the value of the bound $\mathcal{L}$ at its maximum to select the best model, since this approximates the log marginal likelihood for the model. Not only does this save on expensive cross-validation, but it allows all of the available data to be used for training without running into over-fitting problems.

The data sets were the Enzyme, Acidity and Galaxy data used by (15) and the Old Faithful data (9). These data sets were also used by (7) to demonstrate a similar method for model complexity selection for GMMs. All data sets were normalized to zero mean and unit variance. Outliers, numbering 2% of the size of the original data set, were drawn from a uniform distribution on $[-10, 10]$ along each dimensions, and added after the normalization.

The results are shown in Figure 3. We first note that although all models initially had six components, the maximum number of effective components is never higher than five. Moving on to the individual data sets, let us start with the top row of plots, which corresponds to the Enzyme data set. Without the outliers added (columns one and two), the GMM and SMM give similar results, in each case strongly favouring models having two effective components. With the addition of outliers, however, the best performing GMM (column three) now favours three components, whereas the SMM (column four) continues to favour two components. Results from the

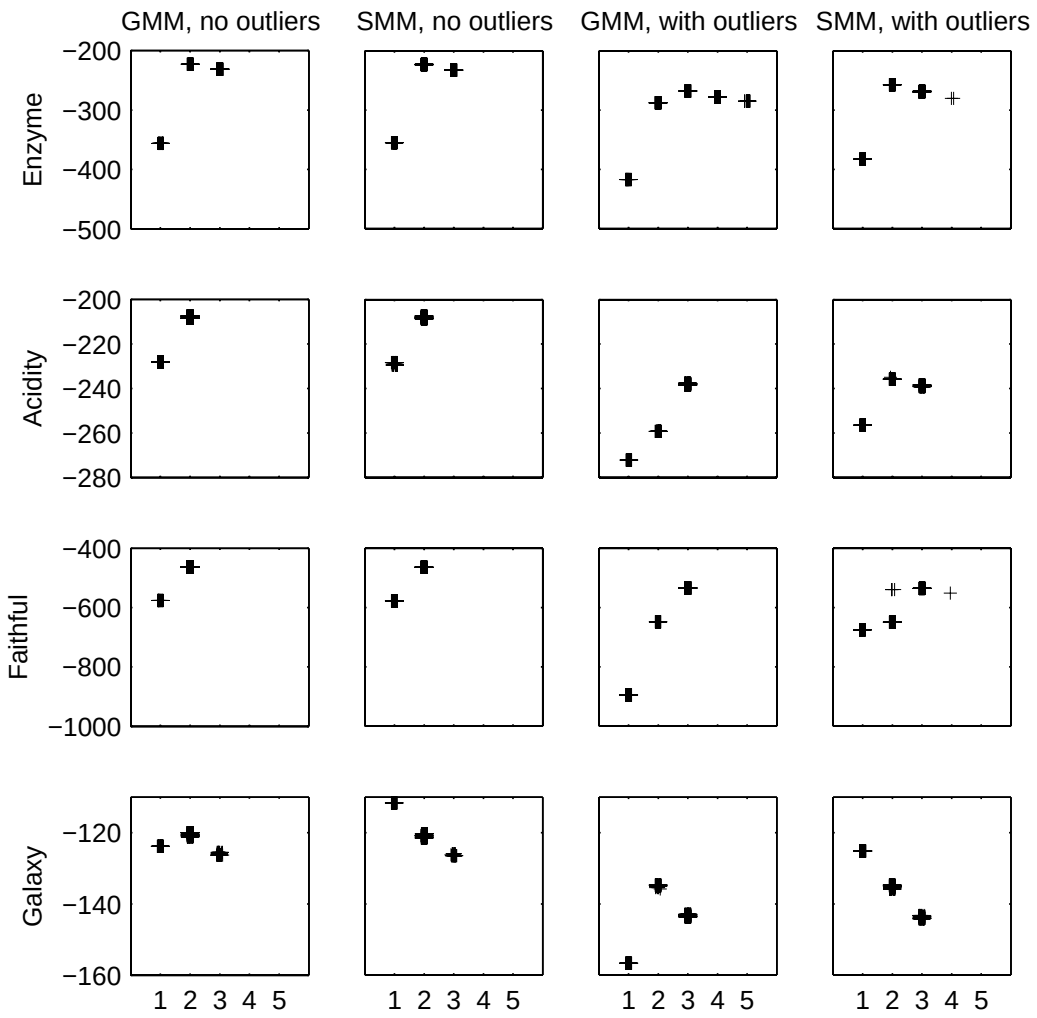

Fig. 3. Comparison of Gaussian (GMM) and Student (SMM) mixture models in their robustness to outliers, showing plots of the lower bound $\mathcal{L}(q)$ of the fitted model versus the number of effective components. Each row corresponds to one data set, and all plots on the same row have the same scale on the vertical axis. The first two columns show the results of the GMM and SMM, respectively, on the original data sets, whereas columns three and four show the corresponding results after the addition of outliers. All plots share the same horizontal range of [0, 6]. In the plots, a small amount of uniform noise has been added to the horizontal position of the points, in order to obtain a better visualization of the results. Each plot contains a total of 300 points, corresponding to the 50 random initializations of each model in which the initial number of components is varied in the range 1 to 6.

Acidity data set, corresponding to the second row of plots, show the same pattern.

In the case of the Old Faithful data set, corresponding to row three, both the GMM and the SMM consistently use two components without outliers. Once outliers are added the GMM strongly prefers three components, whereas the SMM now has solutions for both two and three components which have almost identical values of the bound.

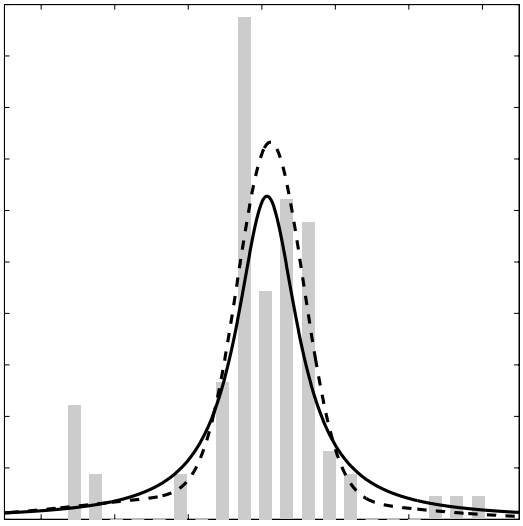

Fig. 4. Histogram plot of the Galaxy data together with densities of the optimal single component SMM (solid) and the optimal dual component GMM (dashed). The density curves have been computed using the expected values of $\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \boldsymbol{\pi}_m\}$ together with the optimal value of ν.

Finally, for the Galaxy data set, corresponding to the fourth row in Figure 3, we see a rather different picture. Before adding outliers, the GMM has a clear preference for two components whereas the SMM strongly favours just one component. Insight into this result is obtained by plotting a histogram of the data, along with the best solutions from the GMM and from the SMM, as shown in Figure 4. It turns out that the two components of the Gaussian mixture model are almost concentric but have significantly different variances. Thus the GMM is effectively modelling the data set as a single cluster but with heavy tails, and in this sense is approximating the Student distribution (which is an infinite mixture of concentric Gaussians) with a mixture of just two Gaussian components. The addition of artificial outliers (columns three and four) leaves this situation unchanged, with the GMM still favouring two components and the SMM strongly preferring just one component. In effect the data set already has outliers and the addition of further artificial outliers has no qualitative influence on the clustering algorithms.

It is worth noting that, for the Enzyme, Acidity and Galaxy data sets, the GMM models preferred by the variational bound have fewer components than those preferred under the MCMC selection scheme used by Richardson and Green (15). This is unsurprising since the factorized variational distribution tends to under-estimate the variance of the posterior distribution, leading in turn to an under-estimate of the model evidence, and this effect becomes more pronounced as the number of hidden variables increases. However, the advantage of a variational approach compared with MCMC is its applicability to large scale applications without incurring high computational cost (5).

5 Conclusions

In this paper we have developed a novel approach to Bayesian mixture modelling which includes Gaussian mixture models as a special case, but which is more robust to non-Gaussianity in the data. Singularities of the kind associated with maximum likelihood are absent, and surplus components revert to the prior distribution and play no role in the predictive density.

It should be emphasized that our approach involves only a small computational overhead compared to the use of maximum likelihood techniques, since the dominant computational costs arise from the evaluation and inversion of weighted empirical precision matrices, which is also the dominant cost in maximum likelihood EM.

A further advantage of our approach is that the inference of the *mean* of a cluster of data points is also less sensitive to outliers when a heavy tailed Student distribution is used in place of a Gaussian. In fact one of the most common motivations for using Student distributions is to obtain robust estimates for the mean of a set of data points.

A Variational Distributions

Here we provide the formulae for the variational distributions over the random variables in our model, $\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}, \boldsymbol{\pi}, \{\mathbf{s}_n, \mathbf{u}_n\}$, and the necessary moments of these. The formulae are all derived in the same fashion as $q(\boldsymbol{\Lambda}_m)$ in section 3.1, i.e., we equate the logarithm of the variational posterior distribution of interest with the logarithm of the corresponding true posterior and then take the expectations of both sides with respect to all other factors under their respective current variational posterior (see equation (18)). Dropping terms independent of the factor of interest, we arrive at familiar distributions for all variables.

A.1 $q(\mathbf{s})$

$$\ln q(\mathbf{s}) \propto \sum_{n,m}^{N,M} s_{nm} \ln r_{nm},$$

where

$$r_{nm} = \exp\left(\langle \ln \pi_m \rangle + \frac{1}{2}\langle \ln |\mathbf{\Lambda}_m| \rangle + \frac{d}{2}\langle \ln u_{nm} \rangle - \frac{\langle u_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} - \frac{d}{2}\ln 2\pi\right), \quad \text{(A.1)}$$

where in turn (cf. (3))

$$\begin{aligned}\left\langle \Delta_{nm}^2 \right\rangle &= \left\langle (\mathbf{x}_n - \boldsymbol{\mu}_m)^{\mathrm{T}} \mathbf{\Lambda}_m (\mathbf{x}_n - \boldsymbol{\mu}_m) \right\rangle_{\boldsymbol{\mu}_m, \mathbf{\Lambda}_m} \\ &= \mathbf{x}_n^{\mathrm{T}} \langle \mathbf{\Lambda}_m \rangle \mathbf{x}_n - 2\mathbf{x}_n^{\mathrm{T}} \langle \mathbf{\Lambda}_m \rangle \langle \boldsymbol{\mu}_m \rangle + \mathrm{Tr}\left[\left\langle \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\mathrm{T}} \right\rangle \langle \mathbf{\Lambda}_m \rangle\right] \end{aligned} \quad \text{(A.2)}$$

Taking into account that the probability distribution $q(\mathbf{s}_n)$ must be normalised for each data point $\mathbf{x}_n$, we see that

$$q(\mathbf{s}) = \prod_{n,m}^{N,M} p_{nm}^{s_{nm}} \quad \text{(A.3)}$$

which is a multinomial distribution, where

$$p_{nm} = \frac{r_{nm}}{\sum_{m'}^{M} r_{nm'}}. \quad \text{(A.4)}$$

although the last term in the argument for the exponential in (A.1) will cancel out in (A.4). From (A.3), we see directly that

$$\langle s_{nm} \rangle = p_{nm}.$$

A.2 $q(\boldsymbol{\pi})$

$$\ln q(\boldsymbol{\pi}) \propto \sum_{n,m}^{N,M} \left(\langle s_{nm} \rangle + (\alpha_m - 1)\right) \ln \pi_m$$

from which we see that

$$q(\boldsymbol{\pi}) = \mathcal{D}\left(\boldsymbol{\pi} | \hat{\boldsymbol{\alpha}}\right) \quad \text{(A.5)}$$

which is Dirichlet distribution, as defined in (11), with parameter

$$\hat{\alpha}_m = \sum_{n}^{N} \langle s_{nm} \rangle + \alpha_m$$

From (A.5) and (11)–(13), we can deduce that

$$\langle \ln \pi_m \rangle = \Psi(\hat{\alpha}_m) - \Psi(\hat{\alpha}_0),$$

where $\Psi(\cdot)$ is defined in (22).

A.3 $q(\mathbf{\Lambda}_m)$

The formulae for $q(\mathbf{\Lambda}_m)$ were given in (19)–(21). The required moments under this posterior are

$$\langle \mathbf{\Lambda}_m \rangle = \eta_m \mathbf{W}_m$$

and

$$\langle \ln |\mathbf{\Lambda}_m| \rangle = d \ln 2 - \ln |\mathbf{W}_m| + \sum_i^d \Psi \left(\frac{\eta_m + 1 - i}{2} \right).$$

A.4 $q(\boldsymbol{\mu}_m)$

For $q(\boldsymbol{\mu}_m)$, we obtain

$$q(\boldsymbol{\mu}_m) = \mathcal{N}\left(\boldsymbol{\mu}_m | \mathbf{m}_m, \mathbf{R}_m\right)$$

where

$$\mathbf{R}_m = \langle \mathbf{\Lambda}_m \rangle \sum_n^N \langle w_{nm} \rangle + \rho_0 \mathbf{I},$$

$$\mathbf{m}_m = \mathbf{R}_m^{-1} \left(\langle \mathbf{\Lambda}_m \rangle \sum_n^N \langle w_{nm} \rangle \mathbf{x}_n + \rho_0 \mathbf{m}_0 \right)$$

and

$$\langle w_{nm} \rangle = \langle s_{nm} \rangle \langle u_{nm} \rangle.$$

We have

$$\langle \boldsymbol{\mu}_m \rangle = \mathbf{m}_m$$

and

$$\langle \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\mathrm{T}} \rangle = \mathbf{m}_m \mathbf{m}_m^{\mathrm{T}} + \mathbf{R}_m.$$

A.5 $q(\mathbf{u})$

Finally, for $q(\mathbf{u})$, we get

$$q(u_{nm}) = \mathcal{G}\left(u_{nm} | a_{nm}, b_{nm}\right)$$

where

$$a_{nm} = \frac{\nu_m + \langle s_{nm} \rangle d}{2} \tag{A.6}$$

and

$$b_{nm} = \frac{\nu_m + \langle s_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} \tag{A.7}$$

where $\langle \Delta_{nm}^2 \rangle$ is defined as in (A.2). The required moments are

$$\langle u_{nm} \rangle = \frac{a_{nm}}{b_{nm}}$$

and

$$\langle \ln u_{nm} \rangle = \Psi(a_{nm}) - \ln b_{nm}.$$

B The Lower Bound

Given values for $q(\{\boldsymbol{\mu}_m\})$, $q(\{\boldsymbol{\Lambda}_m\})$, $q(\mathbf{s})$, $q(\mathbf{u})$, $q(\boldsymbol{\pi})$ and $\{\eta_m\}$, we can evaluate the lower bound of the log-marginal likelihood, (23). This is useful for several purposes, as discussed in section 3.2. We evaluate the terms of (23) separately:

$$\begin{aligned}
&\langle \ln p(\mathbf{X}|\{\boldsymbol{\mu}_m \boldsymbol{\Lambda}_m\}, \mathbf{u}, \mathbf{s}) \rangle = \\
&\frac{1}{2} \sum_{n,m}^{N,M} \langle s_{nm} \rangle \left(\langle \ln |\boldsymbol{\Lambda}_m| \rangle - d \ln(2\pi) + d \langle \ln u_{nm} \rangle - \langle u_{nm} \rangle \left\langle \Delta_{nm}^2 \right\rangle \right)
\end{aligned}$$

where we have used (A.2).

$$\langle \ln p(\boldsymbol{\mu}_m | \mathbf{m}_0, \rho_0) \rangle = \frac{d}{2} \ln \frac{\rho_0}{2\pi} - \frac{\rho_0}{2} \left\langle \|\boldsymbol{\mu}_m - \mathbf{m}_0\|^2 \right\rangle$$

$$\langle \ln p(\boldsymbol{\Lambda}_m | \mathbf{W}_0, \eta_0) \rangle = \ln C_{\mathcal{W}}(\mathbf{W}_0, \eta_0) + \frac{\eta_0 - d - 1}{2} \langle \ln |\boldsymbol{\Lambda}_m| \rangle - \frac{1}{2} \mathrm{Tr}[\mathbf{W}_0^{-1} \langle \boldsymbol{\Lambda}_m \rangle],$$

where $C_{\mathcal{W}}(\cdot)$ is defined in (10).

$$\begin{aligned}
\langle \ln p(\mathbf{u}|\{\nu_m\}) \rangle = \sum_m^M \Bigg(& N \left(\frac{\nu_m}{2} \ln \left(\frac{\nu_m}{2} \right) - \ln \Gamma \left(\frac{\nu_m}{2} \right) \right) \\
& + \sum_n^N \left(\left(\frac{\nu_m}{2} - 1 \right) \langle \ln u_{nm} \rangle - \frac{\nu_m}{2} \langle u_{nm} \rangle \right) \Bigg)
\end{aligned}$$

$$\langle \ln p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \rangle = \ln \Gamma(\alpha_0) + \sum_m^M \left((\alpha_m - 1) \langle \ln \pi_m \rangle - \ln \Gamma(\alpha_m) \right)$$

$$\langle \ln p(\mathbf{s}|\boldsymbol{\pi}) \rangle = \sum_{n,m}^{N,M} \langle s_{nm} \rangle \langle \ln \pi_m \rangle.$$

Note that the last five terms of (23) are simply the entropies of the corresponding variational distributions.

$$\langle \ln q(\boldsymbol{\mu}_m) \rangle = \frac{1}{2} \ln |\mathbf{R}_m| - \frac{d}{2}(1 + \ln(2\pi))$$

$$\langle \ln q(\boldsymbol{\Lambda}_m) \rangle = \ln C_{\mathcal{W}}(\mathbf{W}_m, \eta_m) + \frac{\eta_m - d - 1}{2} \langle \ln |\boldsymbol{\Lambda}_m| \rangle - \frac{\eta_m d}{2}$$

where $C_{\mathcal{W}}(\cdot)$ is defined in (10).

$$\langle \ln q(\mathbf{u}) \rangle = \sum_{n,m}^{N,M} ((a_{nm} - 1)\Psi(a_{nm}) - a_{nm} - \ln \Gamma(a_{nm}) + \ln b_{nm}) ,$$

where we have used (A.6) and (A.7).

$$\langle \ln q(\boldsymbol{\pi}) \rangle = \ln \Gamma(\hat{\alpha}_0) + \sum_{m}^{M} ((\hat{\alpha}_m - 1)\langle \ln \pi_m \rangle - \ln \Gamma(\hat{\alpha}_m))$$

$$\langle \ln q(\mathbf{s}) \rangle = \sum_{n,m}^{N,M} \langle s_{nm} \rangle \ln \langle s_{nm} \rangle .$$

References

[1] H. Attias. Learning parameters and structure of latent variables by variational Bayes. In K. B. Laskey and H. Prade, editors, *Proceedings of the Fifthteenth Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

[2] C. M. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems*, volume 15, pages 793–800, Cambridge, MA, 2002. MIT Press.

[3] C. M. Bishop and M. Svensén. Robust Bayesian mixture modelling. In *Proceedings of ESANN 2004*, pages 69–74, Bruges, Belgium, April 2004.

[4] C. M. Bishop and J. Winn. Non-linear Bayesian image modelling. In *Proceedings of the Sixth European Conference on Computer Vision, Dublin*, volume 1, pages 3–17. Springer, 2000.

[5] D. M. Blei, M. I. Jordan, and A. Y. Ng. Hierarchical Bayesian models for applications in information retrieval. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Bayesian Statistics*, volume 7, pages 25–43. Oxford University Press, 2003.

[6] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987.

[7] A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola, editors, *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.

[8] J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, B*, 56:363–375, 1994.

[9] W. Härdle. *Smoothing techniques with implementation in S*. Springer, New York, 1991.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, 1998.

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

[12] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[13] G. J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. *Lecture Notes in Computer Science*, 1451:658–666, 1998.

[14] C. E. Rasmussen. The infinite gaussian mixture model. In Todd K. Leen Sara A. Solla and Klaus-Robert Mller, editors, *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

[15] S. Richardson and P. J. Green. On bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.

[16] John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 2004. Accepted for publication.