ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

Robust Bayesian Mixture Modelling

Christopher M. Bishop Markus Svensén

Microsoft Research
7 J J Thomson Avenue, Cambridge, CB3 0FB, U.K.
http://research.microsoft.com/∼{cmbishop,markussv}

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. This can lead to excessive sensitivity to small numbers of data points and consequent overestimates 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.

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.

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. This problem can be addressed 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 [1] and more recently using variational methods [2, 3].

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

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

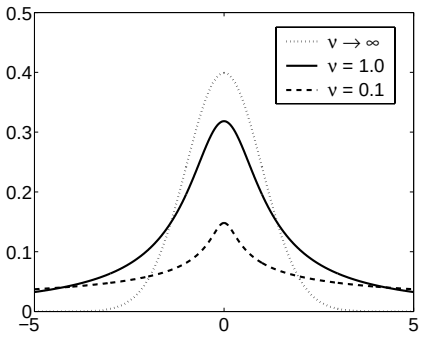

Figure 1: The plots shows the univariate Student distribution $\mathcal{S}(x|\mu, \lambda, \nu)$ for various values of ν, with μ and λ fixed. The limit $\nu \to \infty$ corresponds to a Gaussian, while for finite values of ν, this distribution has heavier tails than a Gaussian having the same μ and λ.

least-squares solution, whose lack of robustness is well known. In the Bayesian model selection context, the presence of outliers often increase the number of mixture components employed in the model.

In this paper we develop a Bayesian treatment of mixture models based on components having a Student distribution [4] which has heavier tails compared to the exponentially decaying tails of a Gaussian. In order to obtain a tractable variational solution for this model we express the Student distribution as an infinite sum of scaled Gaussians through the introduction of additional latent variables. Results on real data sets demonstrate a worthwhile improvement in robustness compared with Gaussian mixtures.

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

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

is the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$. The Student distribution represents a generalization of the Gaussian, 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 as an infinite mixture of scaled Gaussians. In particular we can write the Student distribution in the form

$$\mathcal{S}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, u\boldsymbol{\Lambda}) \mathcal{G}(u|\nu/2, \nu/2)\, du \tag{3}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$, and $\mathcal{G}(u|a, b)$ is the Gamma distribution. For each observation of $\mathbf{x}$ there is a corresponding posterior distribution over the latent variable u, and this can be exploited to find maximum likelihood solutions using the EM algorithm [5].

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

We now consider densities comprising mixtures of Student distributions,

$$p(\mathbf{x}|\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m\}, \boldsymbol{\pi}) = \sum_{m=1}^{M} \pi_m \mathcal{S}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m), \tag{4}$$

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 dimension M having components $\{s_j\}$ such that $s_j = 1$ for $j = m$ and $s_j = 0$ for $j \neq m$, giving

$$p(\mathbf{x}|\mathbf{s}, \{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m\}) = \prod_{m=1}^{M} \mathcal{S}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m)^{s_m} \tag{5}$$

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

$$p(\mathbf{s}|\boldsymbol{\pi}) = \prod_{m=1}^{M} \pi_m^{s_m}. \tag{6}$$

It is easily verified that marginalization of the product of (5) and (6) over the latent variable $\mathbf{s}$ recovers the Student mixture (4).

We consider a data set $\mathbf{X}$ comprising N observations $\mathbf{x}_n$ where $n = 1, \ldots, N$ which we shall suppose are drawn independently from the distribution (4). 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.

Finally, in a Bayesian treatment we need priors over the variables in the model, and again for tractability we choose conjugate priors from the exponential family, of the form

$$p(\boldsymbol{\mu}_m) = \mathcal{N}(\boldsymbol{\mu}_m|\mathbf{m}, \rho\mathbf{I}), \quad p(\boldsymbol{\Lambda}_m) = \mathcal{W}(\boldsymbol{\Lambda}_m|\mathbf{W}_0, \eta_0) \quad \text{and} \quad p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{7}$$

where $\mathcal{W}(\boldsymbol{\Lambda}|\cdot,\cdot)$ denotes the Wishart distribution and $\mathcal{D}(\boldsymbol{\pi}|\cdot)$ denotes the Dirichlet. The parameters of the priors on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are chosen to give broad distributions, in particular $\mathbf{m} = \mathbf{0}$, $\rho = 10^{-3}$, $\mathbf{W}_0 = \mathbf{I}$ 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 $\alpha_m = 10^{-3}$.

Note that there is no conjugate distribution for the parameter ν. However, there is only one such parameter per mixture component, and so we set its value by optimization as part of the variational procedure discussed in Section 3.

3 Variational Inference

Exact inference in our Bayesian model is intractable. However, given our choice of model and prior distributions over the variables in the model, we can make use of

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

established methods for variational inference [6]. This form of approximate inference maximizes a lower bound of the (intractable) log-likelihood of the data with respect to a chosen, parameterised, *variational distribution* over the variables in the model. If we chose this distribution to factorise over the different variables in our model, $\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}$, $\boldsymbol{\pi}$, and $\{\mathbf{s}_n, \mathbf{u}_n\}$, it turns out that we can obtain a closed form fomula for the marginal distribution governing any particular variable. Moreover, this distribution will have the same form as the corresponding prior in (3), (6) or (7), and its parameters will depend on the parameters of the prior as well as statistics calculated under the current fixed variational distributions of the other variables in the model.

If we initialise the parameters of the variational distribution to reasonable values, we can subsequently circulate through the model variables in an iterative fashion, updating the corresponding marginal variational distribution of each variable. This process will converge to a stable posterior variational distribution over all the variables.

We can also evaluate the lower bound in terms of moments of the variational distribution and use this value as a surrogate for the log-marginal likelihood of the data for the purpose of model selection. Note, however, that 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. This procedure uses the entire training set in a single pass for each random start and does not require cross-validation, such as would be needed with maximum-likelihood.

4 Experimental Results

We now present the results of applying the Student mixture model to three real data sets. First, however, we note that if a model having an excess of components is used, then in our Bayesian treatment the unwanted components simply revert to their broad prior distributions, and do not interact with the data. The corresponding terms in the lower bound cancel out, and such components are effectively pruned out of the model. We say that the *effective* number of components is the number of components for which there exists at least one data point in which the posterior probability that the component generated this data point is numerically greater than zero.

We fitted Gaussian mixture models (GMMs) and Student mixture models (SMMs) having between 1 and 6 mixture components to three real data sets, with and without 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. The data sets were the univariate Enzyme, Acidity and Galaxy data used by Richardson and Green [1]. 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 $[-10, 10]$, and added after the normalization.

The results are shown in Figure 2. Let us start with the Enzyme data set (top). Without the outliers added, both the GMM and SMM give similar results, favouring models having two effective components. With the outliers, however, the best per-

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

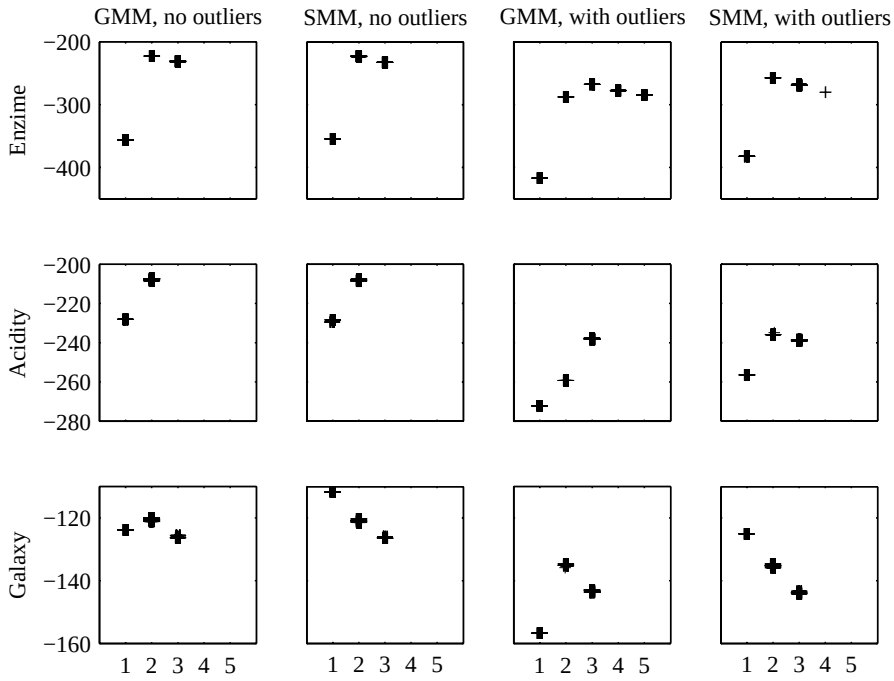

Figure 2: Comparison of Gaussian (GMM) and Student (SMM) mixture models in their robustness to outliers on three different data sets, showing plots of the lower bound of the fitted model versus the effective number of components. All plots share the same horizontal range of [0, 6] and all plots on the same row have the same scale on the vertical axis. Each solution is plotted as a '+'-sign and a small amount of uniform noise has been added to its horizontal position, in order to make the results visually clearer.

forming GMM favours three components, whereas the SMM continues to favour two components. Results from the Acidity data set (middle), show the same pattern.

For the Galaxy data set (bottom), we see a rather different behaviour. Before adding outliers we see that the GMM has a clear preference for two components whereas the SMM strongly favours just one component. It turns out, the two component Gaussian mixture essentially mimics a Student distribution, with components having similar values for their means but very different variances; the resulting mixture distribution has a sharp peak and heavy tails. It is thus not surprising that the addition of artificial outliers leaves this situation unchanged.

It is worth noting that, for all three 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 [1]. This is unsurprising since the factorized variational distribution tends to under-estimate the variance of the posterior distribution, leading to an under-estimate of the model evidence, and this effect be-

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 69-74

comes 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 [7]

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. One of the most common motivations for using Student distributions is to obtain robust estimates for the mean of a set of data points.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.