The sensitivity of the number of clusters in a Gaussian mixture model to prior distributions

William Lima Leão, Cristian Cruz, David Rohde

Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Abstract. Clustering represents an important class of data mining problems of which Bayesian approaches to Gaussian mixture models represent one of the most statistically mature approaches. Bayesian approaches are particularly attractive when the number of clusters and therefore the dimension of the model are unknown as Bayesian model selection techniques can be employed. Two distinct Bayesian approaches have been proposed. The first is to use Bayesian model selection based upon the marginal likelihood, the second is to use an infinite mixture model which 'sides step' model selection. In this study it is empirically demonstrated that in both of these approaches the number of clusters or apparent clusters is prior sensitive. Explanations for the prior sensitivity are given in order to give practitioners guidance in solving this difficult problem. Suggestions are made for testing prior sensitivity by varying the prior over one parameter at a time with conjugate set ups.

Keywords: Bayesian Inference, Clustering, Model Selection, Marginal Likelihood, Gaussian Mixture Model, Markov chain Monte Carlo, Variational Bayes

1 Introduction

The Gaussian mixture model is a powerful modelling for clustering and in its finite form semi-parametric density estimation and in its infinite form non-parametric density estimation. A range of computational methods are available for this model for maximum likelihood including the EM algorithm [5] and for Bayesian inference Markov chain Monte Carlo (MCMC) in particular the Gibbs sampler [8] and variational Bayes [1].

As the number of clusters (K) indexes the dimension of the model inference of K naive model selection concepts such as model fit will always favor more complex models. In contrast Bayesian model selection [2] can be applied to models of differing dimension. Alternatively as Bayesian methods do not overfit model selection can be avoided all together by using infinite mixture models[12] . Together these represent the two Bayesian methods for clustering where the number of clusters are unknown.

In the last 15 years a number of computational methods have been introduced for solving this problem including MCMC [7] such as Gibbs sampling [8] or for computing marginal likelihoods a number of other Monte Carlo algorithms such as annealed importance sampling (AIS) [10]. Another alternative again are variational methods [1] [3].

Despite this algorithmic progress a significant challenge for the practioner is the possible sensitivity of the inference including the number of clusters K to prior distributions. Most machine learning approaches try to avoid prior sensitivity by the use of sophisticated techniques such as hierachichal Bayes or empirical Bayes. On the other hand there is some theoretical work showing that inference of the number of clusters is prior sensitive [9] which suggests that prior judgements can not be completely avoided. Bayesian theory suggests that the value of the marginal likelihood of a model is prior sensitive but the consequence of this in a mixture model setting where the goal is to find K remains unclear.

The contribution of this paper is to study how the choice of K is affected by the prior distributions. Our methodology involves applying state of the art computational techniques AIS and variational Bayes for computing marginal likelihoods and Gibbs sampling for infinite mixtures. In particular the sensitivity of the most likely K or in the case of the infinite mixture effective K are tested with respect to different prior distributions.

The prior distribution for a Gaussian mixture model can be broken into three parts a prior on the coefficients, a prior on the mean and the prior on the variance. The experimental study here shows that the prior on the variance is informative and in cases where the prior permits small variances more clusters often result. The prior on the coefficients has long known to be informative particularly in the infinite mixture model setting. Finally the marginal likelihood of models with different K is relatively insensitive to the prior on the means, yet the infinite mixture models effective K is quite sensitive. An explanation is offered for these observations.

2 Gaussian Mixture Models

2.1 The Model

The finite Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The model can be written in the following way using so called lazy-completion

$$p(y_n|\mu_1,\dots,\mu_K,\pi_1,\dots,\pi_K,\tau_1,\dots,\tau_K) = \sum_{k=1}^{K} \pi_k p_N(y_n|\mu_k,\tau_k^{-1}).$$

The parameters μ_k and τ_k are the mean and the precision (inverse of the covariance matrix) of the normal distribution respectively and π_k are the coefficients and we have the constraint $\sum_{k=1}^{K} \pi_k = 1$ and p_N is the normal distribution.

The lazy-completion form of the model is intuitive and is useful for generic algorithms such as Metropolis Hastings and it is used in this study for annealed importance sampling, however by augmenting the observations with latent data the model is put into complete data exponential family form which enables a number of efficient algorithms. In order to do this a data point y_n is augmented with a discrete latent variable which takes values from 1 to K and identifies which cluster y_n belongs to, the model then has the following form

$$p(y_n, z_n|\mu_1, .., \mu_K, \pi_1, ..., \pi_K, \tau_1, ..., \tau_K) = \pi_{z_n} p_N(y_n|\mu_{z_n}, \tau_{z_n}^{-1}).$$

In a Bayesian setting an infinite mixture model can also be derived in which case $K \to \infty$ and the coefficients π are given a Dirichlet process prior distribution. The usefulness of an infinite parameter model is of course dependent on there being tractable algorithms available, in a later section following [12] it is demonstrated that a Gibbs sampler for this model can be derived as a limiting case of the standard Gibbs sampler.

2.2 The Prior

The Gaussian mixture model has three sets of parameters π which is a vector on a $K-1$ simplex, $\mu_1, ..., \mu_K$ each of which is a vector with the same dimension as the dataset and $\tau_1, ..., \tau_K$ each of which is a semi-definite square precision matrix of the same dimension as the data. Here the standard conjugate prior distribution is analyzed with particular interest in both the tractability and interpretability of the model. An important aspect of interpretability is that it is possible to modify the prior over τ without modifying the prior over μ, within some formulations this is easier and more flexible than others.

Indpendant prior over expectation and precision An interpretable prior distribution is the following

$$\tau_k \sim \mathcal{W}(\nu_0, \omega_0), \qquad \mu_k \sim \mathcal{N}(\rho_0, \Psi_0^{-1})$$

$$\pi \sim \text{Dirichlet}(\alpha_0/K)$$

An advantage of this set up is that the priors are specified independently for all the parameters. This prior distribution results in a complete data exponential family representation which enables the Gibbs sampler, EM algorithm and the variational EM algorithm (although we are not aware of any implementation).

A disadvantage of this setup is that updates for μ must be computed conditional on τ and updates for τ must be computed conditional on μ.

Joint prior over expectation and precision The following prior involves a joint specification for μ and τ and similarly enables joint inference for μ and τ.

$$\tau_k \sim \mathcal{W}(\nu_0, \omega_0), \qquad \mu_k|\tau_k \sim \mathcal{N}(\rho_0, (\beta_0\tau_k)^{-1})$$

$$\pi \sim \text{Dirichlet}(\alpha_0/K)$$

The advantage of this setup is computational, it is more efficient to infer μ and τ jointly than independently.

The disadvantage is that the prior is specified for μ and τ jointly and it becomes difficult to modify the prior for τ while keeping the prior for μ constant, which is useful for understanding prior sensitivity to τ alone. While it is difficult to make such a modification it is possible by modifying V and β_0 in the following way. If one model has parameters $\Theta_1 = [\nu_0, \omega_0, \rho_0, \beta_0, \alpha_0]$, a different model can be set up with the same priors over all parameters except τ with $\Theta_2 = [\nu_0, \omega_0/h, \rho_0, h\beta_0, \alpha_0]$ where h is a parameter adjusting the prior on τ, the ν_0 hyper-parameter must be common to both models which is the main limitation of this set up. The only variational Bayes methods for mixtures that we are aware of use this set up.

Hierachichal models Another approach to setting priors is to put a hierachichal model over some or all the hyper-parameters e.g. [12]. In these models the prior for $\mu_1, ..., \mu_K$ and $\tau_1, ..., \tau_K$ are not independent but rather exchangeable. The affect of this is to change the way the model operates, in particular the model will include the ability to "borrow strength" i.e. the top layer of the model will have a (possibly weak) ability to make all the clusters similar in mean and precision/variance.

These models replace the need to set hyper-parameters with the need to set hyper-hyper parameters. These hyper-hyper parameters must be set either using prior knowledge or in a (perhaps weak) violation of Bayesian principles using empirical Bayes. While these models are very interesting claims that these models avoid prior sensitivity should therefore be treated with skepticism.

3 Algorithms

3.1 Gibbs sampling for individual of $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$

The individual Gibbs sampler for μ and τ is given by

$$\tau_k | \boldsymbol{y}, \boldsymbol{z}, \mu_k \sim \mathcal{W}(\nu_0 + N_k, (\omega_0 + \sum_{n:z_n=k} (y_n - \rho_k)(y_n - \rho_k)^T)^{-1})$$

$$\mu_k | \boldsymbol{y}, \boldsymbol{z}, \tau_k \sim \mathcal{N}((\Psi_0 + N_k\tau_k)^{-1}(\Psi_0\rho_0\tau_k\bar{y}_k), (\Psi_0 + N_k\tau_k)^{-1})$$

where N_k, is the number of observations belonging to class k and $\bar{y}_k = \frac{1}{N_k}\sum_{n:z_k=k} y_n$.

3.2 Gibbs sampling for joint sampling of $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$

The conditional posterior distributions for the means and the precision are

$$\tau_k|\boldsymbol{y},\boldsymbol{z} \sim \mathcal{W}\left(\nu_0 + N_k, \omega_k^*\right)$$

$$\omega_K^* = \omega_0 + (\sum_{n:z_n=k} (y_n - \bar{y}_k)(y_n - \bar{y}_k)^T) + \frac{\nu_0 N_k}{\nu_0 + N_k}(\rho_k - \bar{y}_k)(\rho_k - \bar{y}_k)^T$$

$$\mu_k|\boldsymbol{y},\boldsymbol{z},\tau_k \sim \mathcal{N}(\rho_k^*, (\beta_k^*\tau_k)^{-1})$$

where the occupation number, N_k, is the number of observations belonging to class k, and $\bar{y}_k$ is the mean of these observations. where

$$\rho_k^* = \frac{\nu_0\rho_0 + N_k\bar{y}_k}{\nu_0 + N_k}, \qquad \beta_k^* = \beta_0 + N_k.$$

3.3 Sampling of $\boldsymbol{Z}$ and $\boldsymbol{\pi}$

There are two approaches to sampling Z and π a standard Gibbs sampling approach and a Rao-Blackwelized approach that has the advantage that it enables a Gibbs sampler for the infinite mixture model.

Standard Gibbs Setup The conditional posterior distributions for the weights are

$$\pi_1, \ldots, \pi_K|\boldsymbol{z} \sim \text{Dirichlet}(\alpha/K + N_1, \ldots, \alpha/K + N_K).$$

The latent variables distribution is given by

$$p(z_n = k|y_n, \mu_k, \tau_k, \pi_k) \propto \pi_k p_N(y_i|\mu_k, \tau_k^{-1}).$$

Rao-Blackwelized Sampler In order to be able to Gibbs sampling for the (discrete) indicators, z_i, we will consider the probability of one indicator variable conditional on all the others with π marginalized out

$$p(z_n = k|\boldsymbol{z}_{n-}) = \frac{N_{n-,k} + \alpha/K}{N - 1 + \alpha}$$

where the subscript $n-$ indicates all indexes except n and $N_{n-,k}$ is the number of observations, excluding y_n, that are associated with component k.

The main advantage of the Rao-Blackwelized sampler is that it is still valid when $K \to \infty$ and results in a sampler based upon the Chinese restaurant process [13]. Allowing $K \to \infty$ the conditional prior becomes

$$p(z_n = k|\boldsymbol{z}_{k-}) = \frac{N_{n-,k}}{N-1-\alpha}, \text{ if } N_{n-,k} > 0,$$
$$p(z_n \neq z_{n'} \ \forall n' \neq n|\boldsymbol{z}_{n-}, \alpha) = \frac{\alpha}{N-1+\alpha}, \text{ all other components combined..}$$

In each individual sample of the indicator variables a datapoint is assigned either to one of the existing clusters or alternatively to a new cluster, the parameters for the new cluster are drawn from the prior distribution.

3.4 Variational Bayes

The variational Bayes for mixtures is given in detail in [3] and exploits exponential family form in a way that is analogous to the Gibbs sampler for joint sampling of μ and τ, as complete data exponential family form.

3.5 Annealed Importance Sampling (AIS)

Standard Monte Carlo algorithms sample from the posterior as predictive quantities of interest can be approximated from posterior samples, this is however not true for the computation of the marginal likelihood which requires more sophisticated Monte Carlo methods. Annealed Importance sampling is one possible method.

In a Bayesian application the AIS operates on a hierarchy of distributions of size M that interpolate smoothly between the prior f_m and the unnormalized posterior f_0. An annealing parameter controls the rate at which the interpolation occurs. An expression for the mth hierarchy of the distribution is given by

$$f_m(\Theta) = f_0(\Theta)^{\kappa_j} f_m(\Theta)^{1-\kappa_j}$$

where $1 = \kappa_0 > \kappa_1 > \cdots > \kappa_M = 0$, and the values of κ control the annealing schedule.

The algorithm operates by constructing a complex proposal distribution on this hierarchy of distributions including the target distribution (the posterior). The proposal distribution consists of drawing from the prior and then applying a standard MCMC kernel $T()$ to each of the $M-1$ hierarchies of the distribution after $M-1$ steps applying the kernel to the posterior.

$$\begin{array}{l} \text{Generate } \Theta_{(M-1)} \text{ from } f_M. \\ \text{Generate } \Theta_{(M-2)} \text{ from } \Theta_{(M-1)} \text{ using } T() \\ \qquad\qquad \vdots \\ \text{Generate } \Theta_{(0)} \text{ from } \Theta_{(1)} \text{ using } T(). \end{array}$$

A single AIS sample involves drawing from the prior and applying an MCMC kernel $M-1$ times to annealed sequence of distributions, if M is large and the Markov chain and the annealing heuristic are operating well a sample from $\Theta_{(0)}$ might be close to a sample from the posterior. After the annealing is applied M steps the algorithm remains a standard importance sampling algorithm consisting of samples from a (complex) proposal distribution and associated importance weights.

$$w^{(i)} = \frac{f_{M-1}(\Theta_{(M-1)})}{f_M(\Theta_{(M-1)})} \frac{f_{M-2}(\Theta_{(M-2)})}{f_{M-1)}(\Theta_{(M-2)})} \cdots \frac{f_1(\Theta_{(1)})}{f_2(\Theta_{(1)})} \frac{f_0(\Theta_{(0)})}{f_1(\Theta_{(0))})}$$

The mean of the weights converges to the marginal likelihood as with any importance sampling method.

AIS is vulnerable to catastrophic failure if the proposal distribution is not capable of finding areas of high probability e.g. in the case that it fails to find important modes in the he posterior in order to avoid this the Markov chain length (m) must be long and the annealing heuristic satisfactory. Like other Monte Carlo approaches it can be difficult to diagnose catastrophic failure.

4 Results

All simulations are applied to the Galaxy velocity dataset which is a classic dataset in the context of mixtures e.g. [4]. This dataset is one dimensional and as such all of the Wishart distributions could be replaced with Gamma distributions.

4.1 Finite Mixture Models Independent Prior Specification

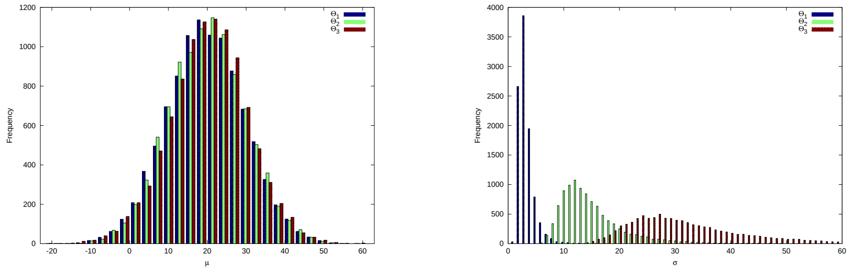

Fig. 1. Prior distribution for $\omega_0 = 1/40$, $\omega_0 = 1/800$, $\omega_0 = 1/4000$

In the case of specifying the priors for a finite mixture model with independent specification for the mean and the variance it becomes straight forward

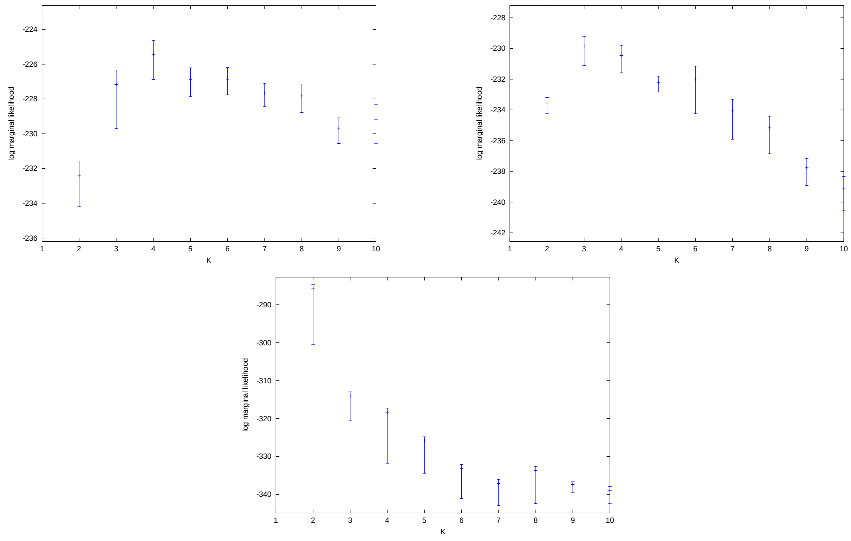

Fig. 2. Marginal Likelihoods approximated for different values of K and different priors left, Θ_1, centre Θ_2 and right Θ_3. Approximations computed with 100 samples of AIS with a 95% confidence interval.

to test the sensitivity to one parameter while leaving the other fixed. As annealed importance sampling does not make use of exponential form, there is no computational cost for considering arbitrary independent priors. The marginal likelihoods was approximated with 100 samples. For each sample an annealing schedule $\kappa_0 = 1 > \ldots > \kappa_n = 0$ was used with $\kappa_n = e^{-n/500}$ for $n = 1, \ldots, 3500$. The 95% confidence interval is produced in order to assess the Monte Carlo error using 10000 bootstrap samples [6].

Two random walk normal proposal distributions were used, one with small variance equal to 0.005 for π and τ, and set 0.25 to μ; and other with large variance equal to 0.03 for π and τ, and set 1.5 to μ. This gave reasonable acceptance throught the annealing schedule.

For all experiments the prior on the mean was fixed to $\mu_k \sim \mathcal{N}(20, 10^2)$, the prior on $\pi \sim$ Dirichlet(1) and variations on the prior on τ_k were considered with Θ_1 being $\tau_k \sim \mathcal{W}(6, 1/40)$, Θ_1 being $\tau_k \sim \mathcal{W}(6, 1/800)$ and Θ_1 being $\tau_k \sim \mathcal{W}(6, 1/4000)$. In order to visualise the prior plots of the more intuitive $\sigma = \frac{1}{\sqrt{\tau_k}}$ are used see Figure 1. In order to consider the implications of the prior we first note that the standard deviation of the data is ≈ 4, this suggests that if there are multiple clusters they each must have σ smaller than 4. We note that the prior for Θ_1 favors small values of σ and the prior for Θ_2 less so and Θ_3 much less so again. This gives a clear interpretation of the marginal likelihood selection of K under each of the three priors shown in Figure 2 where Θ_1 favours 4 clusters (and does so strongly), Θ_2 also favours 4 clusters (but 3 clusters has similar marginal likelihood) and Θ_3 which favours 2 clusters.

4.2 Infinite Mixture Models Independent Prior Specification

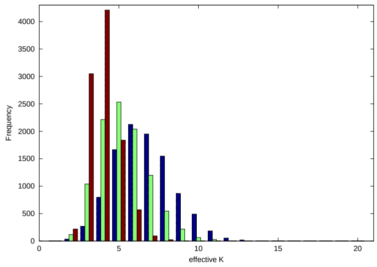

Fig. 3. Samples from the posterior of the effective K for three different priors Θ_1, centre Θ_2 and right Θ_3.

In order to test the sensitivity to the effective K in an infinite mixture model the following experiment was carried out. The same priors for Θ_1, Θ_2 and Θ_3 for the components were used again here. The prior for the coefficients was a Dirichlet process with hyper parameter $\alpha_0 = 1$. The Gibbs sampler was run for 10000 samples in order to find the distribution over the effective K i.e. the number of clusters within the finite dataset.

Applying the same logic that prior distributions favouring small values of σ will result in more clusters it is possible to explain the posterior on K shown in Figure 3. Analogous to the finite mixture model prior belief about scale is an important factor in the effective K in an infinite mixture model.

4.3 Finite Mixture Models Joint Prior Specification

In the context of the finite mixture model employing a joint specification on the mean and the variance we compare the marginal likelihood for the following prior distributions Θ_1 with $\omega_0 = \frac{1}{50}$, $\nu_0 = 30$, $\beta_0 = 1$ and $\rho_0 = 20$ and Θ_2 with $\omega_0 = \frac{1}{1000}$, $\nu_0 = 30$, $\beta_0 = 20$ and $\rho_0 = 20$. By considering the plots in Figure 4 it is clear that the prior for the two models is identical except for the difference in the prior on $\sigma = \frac{1}{\sqrt{\tau}}$ which for Θ_1 favours low values of σ and which for Θ_2 favors values around 6. Recall that the standard deviation of the galaxy dataset is ≈ 4 so that if there are multiple components in the mixture it makes sense for σ to have values smaller than 4. Consistent with this argument it is seen that the marginal likelihood in Figure 5 approximated with AIS and the bound approximated with variational Bayes selects 5 clusters with Θ_1 and just one cluster with Θ_2. The AIS marginal likelihood is approximated using the previously described Markov chain set up and 260 samples per model (per value of K). The variational Bayes bound is the largest computed after 5 tries of the algorithm it is seen that this is consistently a lower bound. As the variational Bayes algorithm fits a single mode of the posterior a plot of the variational bound multiplied by the minimum $K!$ modes expected due to symmetry alone, this is also observed to be much lower than the marginal likelihood suggesting that the posterior also contains modes not due to symmetry.

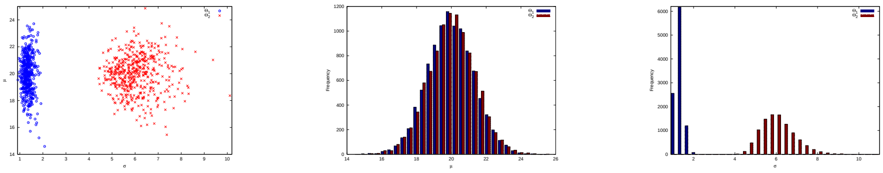

Fig. 4. Two prior distributions Θ_1 with $\omega_0 = \frac{1}{50}$, $\nu_0 = 30$, $\beta_0 = 1$ and $\rho_0 = 20$ and Θ_2 with $\omega_0 = \frac{1}{1000}$, $\nu_0 = 30$, $\beta_0 = 20$ and $\rho_0 = 20$.

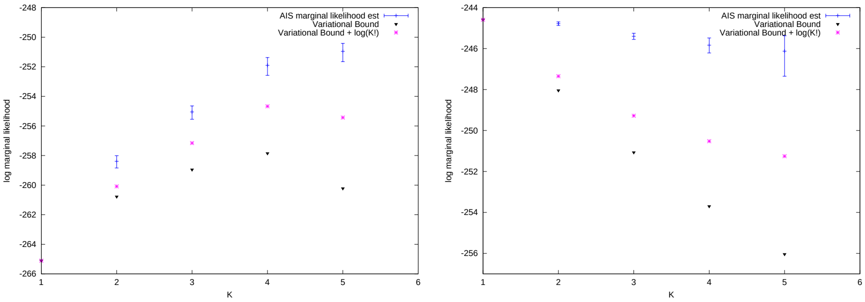

Fig. 5. Approximated Marginal likelihood as a function of K computed using AIS and a bound computed with variational Bayes for two different prior distributions left Θ_1 and right Θ_2.

4.4 Infinite Mixture Models Joint Prior Specification

In this experiment three different prior distributions (Θ_1, Θ_2 and Θ_3) were considered all of which had $\rho_0 = 20, \nu_0 = 6$ and $\alpha = 1$, the priors for Θ_1 and Θ_2 both had $\beta_0 = 1$ but Θ_1 had $\omega = 1/40$ and Θ_2 had $\omega = 1/400$. Θ_3 had $\beta = 0.1$ and $\omega = 1/400$. The three joint priors are illustrated in Fig 6. It is noteworthy that all three have different prior (marginal) distributions on μ, but the prior for Θ_2 and Θ_3 are the same on $\sigma = \frac{1}{\sqrt{\tau}}$. When we consider the posterior distribution on the effective K in Figure 7 two observations are striking. Firstly the posterior effective K while the prior on τ for Θ_2 and Θ_3 is the same the posterior effective K is different this suggests that the effective number of clusters is sensitive to the prior on the mean. As the creation of a new cluster is dependent on draws from the prior distribution this is perhaps unsurprising

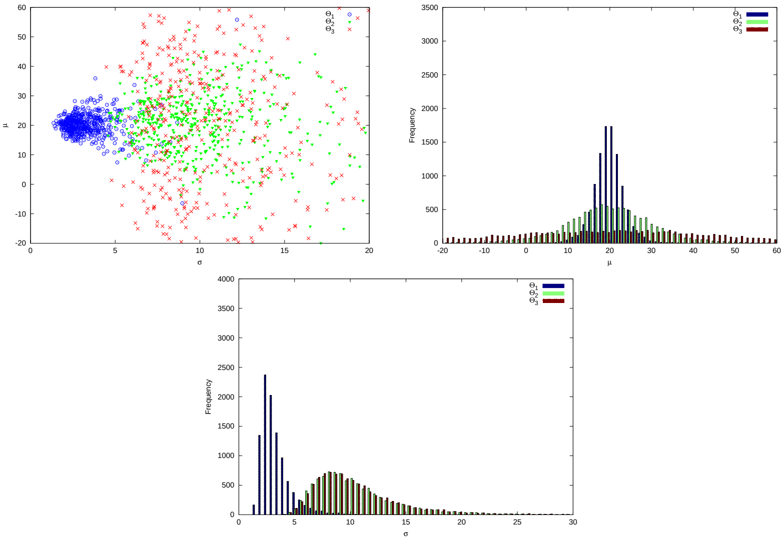

Fig. 6. Prior samples from the joint model for Θ_1, Θ_2 and Θ_3.

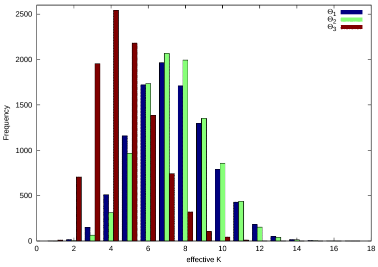

Fig. 7. Histograms of Effective K.

and priors which are constantrated near the data should result in more clusters. A second striking observation is that the posterior effective K on Θ_1 and Θ_2 are very similar despite the fact that they have very different joint priors on the mean and the variance. The effect on the posterior K on the change in the prior on the mean in the case seems to be offset by a change in the prior on σ. This example illustrates the difficult in conducting sensitivity analysis when the prior to both the mean and the variance are changed simultaneously, this is a particular difficulty in the joint prior set up (although see Section 2.2 for ways of independently changing priors within this setup). Our simulations also suggest that prior sensitivity to the prior on the mean seems to be much more important when considering the posterior effective K of the infinite model.

5 Conclusion

The number of clusters within a mixture model whether selected using Bayesian model comparison with finite mixture models or considering the effective number of clusters in an infinite model is prior sensitive. In particular it is demonstrated that more clusters result if the prior favours small variances. It is recommended that sensitivity analysis be carried out by fixing the prior on the mean and changing the prior on the variance, this is easier and more flexible with independent prior distributions but is possible with the joint prior.

This observed prior sensitivity has wider implications for Bayesian machine learning where an engineering approach to problem solving may in some situations need to be augmented with subjective Bayesian tools which may help in setting prior distributions through introspection.

References

1. H. Attias. A variational bayesian framework for graphical models. In *In Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.
2. J. Bernardo and A. F. M Smith. *Bayesian Theory.* John Wiley, Chichester, 1994.
3. C. M. Bishop. *Pattern recognition and machine learning.* Springer, 1st ed. 2006. corr. 2nd printing edition, 2006.
4. S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.
5. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.
6. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* Chapman and Hall, New York, 1993.
7. D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall, 1997.
8. J. M. Marin, K. Mengersen, and C. P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics 25*, 2006.
9. P. Mccullagh and J. Yang. How many clusters? *Bayesian Analysis*, 3(1), 2008.
10. R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 1998.
11. R.M. Neal. Bayesian mixture modelling by Monte Carlo simulation. Technical Report Dept. of Computer Science, University of Toronto, Technical Report CRG–TR–91–2, 1991.
12. C. E. Rasmussen. The infinite Gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.
13. Y. W. Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2010.