Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models [1]

Russell J. Steele
McGill University

Adrian E. Raftery
University of Washington

Technical Report no. 559
Department of Statistics
University of Washington

September 11, 2009

[1]Russell J. Steele is Associate Professor of Mathematics and Statistics, McGill University, 805 Sherbrooke O., Montreal, PQ, Canada H3A 2K6; email: steele@math.mcgill.ca; web: www.math.mcgill.ca/ steele. Adrian E. Raftery is Blumstein-Jordan Professor of Statistics and Sociology, University of Washington, Box 354322, Seattle, WA 98195-4322; email: raftery@stat.washington.edu; web: www.stat.washington.edu/raftery. Raftery's research was supported by NICHD grant HD-054511, NIGMS grant GM-084163 and NSF grant ATM-0724721. The authors are grateful to Matthew Stephens for helpful discussions.

Abstract

Bayesian methods are widely used for selecting the number of components in a mixture models, in part because frequentist methods have difficulty in addressing this problem in general. Here we compare some of the Bayesianly motivated or justifiable methods for choosing the number of components in a one-dimensional Gaussian mixture model: posterior probabilities for a well-known proper prior, BIC, ICL, DIC and AIC. We also introduce a new explicit unit-information prior for mixture models, analogous to the prior to which BIC corresponds in regular statistical models. We base the comparison on a simulation study, designed to reflect published estimates of mixture model parameters from the scientific literature across a range of disciplines. We found that BIC clearly outperformed the five other methods, with the maximum a posteriori estimate from the established proper prior second.

Contents

List of Tables

List of Figures

1 Introduction

Bayesian methods are widely used for selecting the number of components in finite mixture models, particularly since frequentist methods have difficulties for this problem. There is an elegant but rather complicated theory for frequentist testing of one mixture component versus two, or one versus more than one (Lindsay 1995), but we do not know of a fully satisfactory frequentist theory for selecting the number of components more generally.

Here we compare some of the Bayesianly motivated or justifiable methods for choosing the number of components in a mixture. We consider posterior probabilities for a well known proper prior, BIC, ICL, DIC and AIC. We also introduce an explicit unit-information prior for mixture models, analogous to the prior to which BIC corresponds in regular models.

We compare these criteria via a simulation study. The design of the simulation study is critical, as it is easy to design a simulation study to favor one criterion or another. As a result, we based the design on the scientific literature rather than just specifying values ourselves, as is often done. We extracted 43 published estimates of mixture model parameters from the literature across a range of disciplines, achieving broad coverage of the literature prior to 2000. Using cluster analysis after appropriate normalization, we identified six representative sets of parameter values for our study.

The results were perhaps surprising: BIC outperformed the other criteria, including posterior probabilities based on proper priors. This does confirm the informal experience of workers in the area, particularly those using mixture models for clustering, who have been using BIC widely for this purpose for the past decade.

In Section 2 we review the existing Bayesian model selection criteria for mixture models that we include in our comparison. In Section 3 we describe our new unit information prior for mixture models. In Section 4 we give results for a simulated example and a real data example. Then in Section 5 we describe our simulation study and give the results. In Section 6 we discuss some other methods and issues in this area.

2 Bayesian Model Selection for Mixture Models

2.1 Priors for Mixture Models

We consider the univariate Gaussian mixture model with G components where observations $y_1, \ldots, y_n$ are independently and identically distributed with the density

$$p(y_i|\mu, \sigma^2, \lambda) = \sum_{g=1}^{G} \lambda_g f(y_i|\mu_g, \sigma_g^2), \tag{1}$$

where f is the univariate normal density with mean and variance parameters μ_g and σ_g^2 and $\lambda \in (0,1)$ where $\sum_{i=1}^{G} \lambda_i = 1$.

Initially the most used prior was a semi-conjugate specification for the parameters of the mixture model, which was conjugate conditional on the unknown mixture component memberships (West 1992; Diebolt and Robert 1994; Chib 1995). This assumes a Gaussian prior for each of the μ_j with prior mean ξ_j and prior variance $\sigma_j^2 \tau_j$. The priors for the variances are scaled inverse χ^2 random variables, i.e. $\sigma_j^{-2} \sim \frac{1}{2\beta}\chi^2_{2\alpha}$ where α and β are fixed hyperparameters. With this prior, full conditional posterior distribution conditional on the unknown cluster memberships can be found in closed form. Nobile and Fearnside (2007) used a similar structure, but they placed an additional level of hierarchical, uniform prior distributions on τ_j and β_j.

Another commonly used approach is the conditionally semi-conjugate prior of Richardson and Green (1997). This differs from the conditionally conjugate prior in that the priors for the component density means are assumed to be independent of the component density variance parameters. It has been used as for finite Gaussian mixture models (Stephens 2000a; Robert, Rydén, and Titterington 2000), and also more recently for Gaussian hidden Markov Models (Spezia 2009) and image segmentation (Ferreira da Silva 2009).

Richardson and Green (1997), and subsequent authors, also assigned an additional hierarchical prior to the scaling constant of the prior for the variances, assuming that $\beta \sim \frac{1}{2h}\chi^2_{2g}$. The prior proposed by Stephens (2000a) differs slightly from that of Richardson and Green in that κ and ξ are also allowed to vary, namely $\xi \sim \text{Unif}[\infty, \infty]$, and $\kappa^{-1} \sim \frac{1}{l}\chi^2_l$ where $l = 0.0001$. Stephens suggested this prior because he found that the posterior for the number of components G was sensitive to the the prior on μ (and thus to the value of κ).

The choice of prior hyperparameters can have a big effect on the estimation of the mixture parameters (Jasra, Holmes, and Stephens 2005). In general, the hyperparameters for the component density priors have been chosen to be the same for each component density (Frühwirth-Schnatter 2006). Data-dependent choices of the hyperparameters have been proposed by Raftery (1996), Wasserman (2000) and Richardson and Green (1997) to achieve weakly informative priors. Richardson and Green (1997) chose ξ to be the overall mean of the data, the prior variance of the component means to be proportion to the square of the range of the data, and the prior distribution of the variances to have scale parameters proportional to the square of the range, and $\alpha = 2$. Richardson and Green (1997) used a uniform Dirichlet prior for the mixture proportions, λ_g, although this choice can cause difficulties in convergence of Markov Chain Monte Carlo algorithms (Plummer 2008). In our reading of the literature, we have found that these choices have been regularly used in both methodological and applied work.

2.2 Criteria for choosing the number of mixture components

Fully Bayesian MAP estimate: The obvious Bayesian choice of the number of mixture components is the posterior mode, or maximum a posteriori (MAP) estimate of G. This can be evaluated by reversible jump MCMC (Richardson and Green 1997), or by the marked point process method of Stephens (2000b). Here we use Stephen's hierarchical modification of Richardson and Green's prior; his paper provides some evidence that it performs similarly to theirs, but is less sensitive to prior specification. We also use his marked point process algorithm, as implemented in his software, available at
http://www.stat.washington.edu/stephens/papers/software.tar.

BIC: The BIC (Schwarz 1978) provides a widely used approximation to the integrated likelihood (integrated over the model parameters) for regular models. It was used for mixtures by Roeder and Wasserman (1997) and has been widely used for mixture models since, particularly for clustering (Dasgupta and Raftery 1998; Fraley and Raftery 2002), with good results in practice. It is defined as

$$BIC(G) = 2p(y|\hat{\tau}, G) - d\log(n),$$

where d is the number of free parameters in the mixture model. For regular models, BIC is derived as an approximation to twice the log integrated likelihood using the Laplace method (Tierney and Kadane 1986), but the necessary regularity conditions do not hold for mixture models in general (Aitkin and Rubin 1985). However, Roeder and Wasserman (1997) showed that BIC leads to to a consistent estimator of the mixture density, and Keribin (2000) showed that BIC is consistent for choosing the number of components in a mixture model.

DIC: The DIC (Spiegelhalter, Best, Carlin, and van der Linde 2002) is an AIC-like likelihood penalization criterion, where the number of *effective* model parameters is used instead of the actual number of free parameters in the model. One stated objective of the DIC for model comparison is to minimize predictive error of the selected model. It has the form

$$DIC(G) = -2\log p(y|\hat{\tau}, G) + 2p_d$$

where $\hat{\tau}$ is a "good" estimate of τ with respect to the posterior distribution of the data (often a posterior mean, median or mode) and p_d can be written as:

$$p_D = E_{\tau|y}(\log p(y|\hat{\tau})) - \log p(y|\hat{\tau}).$$

One can thus estimate p_D (and possibly $\hat{\tau}$) via $\tau_t, t = 1, ..., T$ where the τ_t are draws from a posterior sampling algorithm. This gives the following empirical estimate of the effective

number of parameters for a particular model to be used in the above DIC calculation:

$$\hat{p}_D = \frac{1}{T}\sum_{i=1}^{T} \log p(y|\tau_t) - \log p(y|\hat{\tau}).$$

The choice of $\hat{\tau}$ can have a large influence on the estimated number of parameters. Although Spiegelhalter et al. (2002) used the posterior mean in most cases, here we will use the largest posterior mode. The posterior mean (and the posterior median as well) can give poor results (e.g. negative numbers of parameters) because of the multimodality of the mixture density.

Celeux, Forbes, Robert, and Titterington (2006) discussed several alternative choices for $\hat{\tau}$, although they did not come to a definitive conclusion as to which would perform the best for choosing the order of a mixture. Plummer (2008) proposed a modification of the DIC to adjust for the complications with respect to estimating the effective number of parameters derived through a view of the DIC as an approximate penalized loss function. McGrory and Titterington (2007) used a variational Bayes approach to derive a version of the DIC that they found to perform well for choosing the number of components.

ICL: Biernacki, Celeux, and Govaert (1998) suggested an information criterion based on the complete data likelihood. They noted that although integrating over the parameters of the observed is difficult, the following integral

$$p(y, z|G) = \int p(y, z|\tau, G)p(\tau)d\tau$$

is sometimes available in closed form. Even if it is not, one can rewrite the integral as

$$p(y|z, G)p(z|G) = \int p(y|z, \tau, G)p(z|G, \tau)p(\tau)d\tau.$$

Biernacki et al. (1998) noted that Laplace approximations to $p(y|z, \tau, G)$ are often valid and that the remaining integral of $p(z|\tau, G)$ over $p(\tau)$ are often available in closed form, as long as $p(\tau) = p(\tau)p(\lambda)$ and the prior distribution of z is independent of τ. Therefore, they proposed approximating $\log(p(y|z, G))$ using the BIC approximation

$$\log(p(y|z, \hat{\tau}^*, G))) - \frac{d}{2}\log(n),$$

where $\hat{\tau}^* = \text{argmax}_\tau p(y|z, \tau, G)$, which will not necessarily be the same as $\hat{\tau}$ from maximizing the observed data likelihood.

This leads to approximating twice the negative integrated classification likelihood by

$$\text{ICL} = -2 * \log(p(y|\hat{z}', \hat{\tau}^*, G))) + \frac{(d - (G-1))}{G} * \log(n) - 2 * K(\hat{z}')$$

where $K(z) = \int p(z|\lambda, G)p(\lambda|G)$ depends on the prior for λ, d is the number of total free parameters in the mixture model as before, and $\hat{z}'$ is the MAP estimate of z given $\hat{\tau}^*$, i.e.

$$\hat{z}'_{ij} = \begin{cases} 1 \text{ if } \text{argmax}_g \hat{z}^*_{ig} = j \\ 0 \text{ otherwise.} \end{cases}$$

If one specifies a Dirichlet$(\alpha_1, ..., \alpha_G)$ distribution for the mixing parameters, then $K(z)$ can be obtained in closed form as

$$K(z) = \log(\Gamma(\frac{G}{2})) + \sum_{g=1}^{G} \log(\Gamma(n_g(z) + \alpha_g)) - \sum_{g=1}^{G} \log(\Gamma(\alpha_g)) - \log(\Gamma(n + \sum_{g=1}^{G} \alpha_g))$$

where $n(z)_g$ is the number of observations assigned to the g-th group by the allocation matrix z.

Biernacki et al. (1998) noted that if the $n_g(z)$ are far from zero, an additional approximation for $K(z)$ can be made using a Stirling's approximation (dropping $O(1)$ terms) as

$$K(z) \approx n \sum_{g=1}^{G} \hat{\lambda}_g \log(\hat{\lambda}_g) - \frac{1}{2}(G-1) * \log(n)$$

which gives us the following BIC approximation to the integrated complete-data likelihood

$$\text{ICL–BIC} = -2 * \log(p(y, \hat{z}'|\hat{\tau}^*, G))) + d * \log(n).$$

Biernacki, Celeux, and Govaert (2000) presented just the ICL–BIC, with the additional suggestion to use $\hat{\tau}$ instead of $\hat{\tau}^*$. This is the form of ICL that we use in our experiments.

AIC: The best known of the information criteria used for determining the number of components is Akaike's Information Criterion (AIC) . The AIC is calculated for mixtures as:

$$\text{AIC}(G) = -2 \log p(y|\hat{\tau}, G) + 2d$$

where d is the number of free parameters in the mixture (e.g. for one-dimensional normal mixtures with unconstrained variances, $d = 3*G-1$). The theoretical justification for AIC is that choosing the minimum value of the AIC asymptotically minimizes the mean Kullback-Leibler information for discrimination between the proposed distribution and the true distribution, i.e. the model with the minimum value value of the AIC should be asymptotically closest in Kullback-Leibler distance to the true model. However, several studies (Koehler and Murphree 1988; Celeux and Soromenho 1996) have found that the AIC overestimates the number of components for mixtures, most likely due to violations of the regularity conditions required for the approximation to hold. Compared to BIC, the AIC penalizes models with larger numbers of parameters less, leading to the choice of more mixture components.

AIC also has a Bayesian interpretation, leading to the MAP estimate in regular models when the amount of the information in the prior is of the same order as the amount of information in the data (Akaike 1983). This is a highly informative prior, and will not be plausible in most cases, so the Bayesian interpretation of AIC is questionable in many situations.

3 A Unit Information Prior for Mixture Models

Kass and Wasserman (1995) showed that for regular models, the BIC provides an approximation to Bayes factors (posterior odds for one model against another when prior odds are equal to 1, in turn equal to ratios of integrated likelihoods) under a certain prior for the parameters. This is the so-called unit information prior (UIP), equal to a multivariate normal prior distribution with mean equal to the maximum likelihood estimates for the parameters and covariance matrix equal to the inverse information matrix scaled by the number of observations, i.e. the observed information matrix for a single observation. Raftery (1995) extended this result to show that BIC approximates the integrated likelihood for a single model under the UIP.

However, these results do not hold for mixture models because the required regularity conditions are not satisfied. As a result, we propose here an analogy to the UIP for mixture models. We subsequently use the resulting MAP of G as another Bayesian estimator of G and assess its performance.

Let z be a matrix of indicator variables, such that $z_{ij} = 1$ if observation i is sampled from component j and 0 otherwise. Then one can write:

$$p(y, z|\mu, \sigma^2, \lambda) = \prod_{i=1}^{n} \prod_{g=1}^{G} \lambda_g^{z_{ig}} f(y_i|\mu_g, \sigma_g^2)^{z_{ig}}. \tag{2}$$

Conditional on z, the complete-data information matrix for the parameters μ and σ^2 becomes block diagonal by component:

$$i_C(\mu, \sigma^2) = \begin{array}{c} \\ \mu_1 \\ \sigma_1^2 \\ \vdots \\ \mu_G \\ \sigma_G^2 \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \sigma_1^2 & \vdots & \mu_G & \sigma_G^2 \end{array} \\ \begin{bmatrix} \frac{n(z)_1}{\sigma_1^2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{n(z)_1}{2\sigma_1^4} & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdots & \frac{n(z)_G}{\sigma_G^2} & 0 \\ 0 & 0 & \cdots & 0 & \frac{n(z)_G}{2\sigma_1^4} \end{bmatrix} \end{array},$$

where $n(z)_g$ is equal to the number of observations assigned to group g according to the

matrix z.

We now propose a data-dependent prior for mixture models based on the observed information for the complete data likelihood and $n(z)_g = 1$ for all $g = 1, ..., G$. We use the following conjugate form for the priors for the component parameters:

- $\lambda \sim$ Dirichlet $(\alpha_1, ..., \alpha_G)$
- $\mu_g \sim$ Normal $(\mu_g^0, \sigma_g^2/\kappa_g)$
- $\sigma_g^2 \sim \sigma^{2(0)}$ Inv-$\chi^2_{\nu_g}$

Each of the priors would be conjugate for the likelihoods in question if the group memberships were known and, hence, $p(y, z)$ has an analytic form.

One approach for specifying values for the prior's hyperparameters is to set the prior covariance matrix for the prior parameters equal to the scaled inverse of the observed information matrix for a single observation. Kass and Wasserman (1995) suggested this approach, but in the context of multivariate normal distributions for the parameters, rather than for the conjugate priors above. The prior variance matrix for the parameters μ_g and σ_g^2 is

$$\mathrm{Var}_\pi(\mu_g, \sigma_g^2) = \begin{bmatrix} \frac{\sigma^{2(0)}}{\kappa_g(\nu_g - 2)} & 0 \\ 0 & \frac{2(\sigma_g^{2(0)})^2}{(\nu_g-2)^2(\nu_g-4)} \end{bmatrix}$$

Setting the prior variance matrix equal to $i_C^{-1}(\hat{\mu}_g, \hat{\sigma}_g^2)$, it is necessary to solve a system of two equations in three variables:

$$\begin{aligned} \hat{\sigma}_g^2 &= \frac{\sigma^{2(0)}}{\kappa_g(\nu_g - 2)} \\ 2(\hat{\sigma}_g^2)^2 &= \frac{2(\sigma_g^{2(0)})^2}{(\nu_g - 2)^2(\nu_g - 4)} \end{aligned}$$

which has the following set of solutions:

$$\begin{aligned} (\nu_g - 4) &= \kappa_g^2 \\ (\nu_g - 2)\hat{\sigma}_g^2 &= \sigma_g^{2(0)}. \end{aligned}$$

We choose $\nu_g = 5$ and $\kappa_g = 1$, yielding $\sigma_g^{2(0)} = 3\hat{\sigma}_g^2$, because $\nu_g = 5$ is the smallest integer number of degrees of freedom that guarantees a finite variance for the variance parameters. This also has the very appealing feature of yielding a marginal unit-information prior for the component means. Figures 1 show plots of the priors for the component means for a simulated data set of 400 observations where the true $G = 3$ (a histogram of the data is shown in Figure 2).

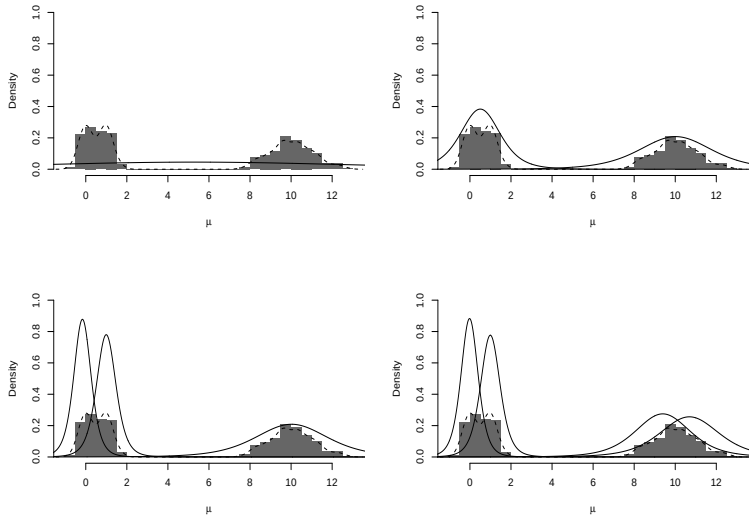


Figure 1: Marginal priors for means for 1-4 components for the trimodal data set using the approximate unit information prior

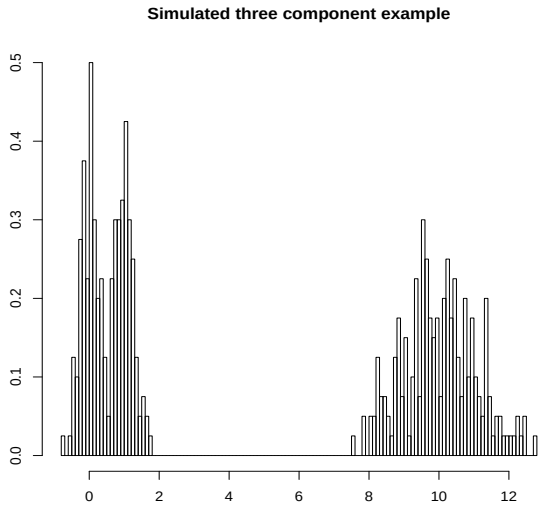


Figure 2: Histogram of 400 simulated data points

Note that we do not assume that the mean and variance parameters have the same prior distribution for each mixture component, and so we impose a prior labeling on the parameters of the mixture. One would need to take this into account when making inference about any of the parameters of the mixture model with the exception of G. In the absence of prior knowledge about the group labelings, the methods of Stephens (2000b) could be applied in order to make inference about all parameters of the distribution.

In the case of G, relabeling is not necessary, as the suggested prior yields posterior inference about G equivalent to the posterior inference using an exchangeable mixture prior with $G!$ components with each component corresponding to a different labeling for the components of the prior above. This is shown by

Theorem 1 *Posterior inference for the number of components based on a non-exchangeable prior distribution for the component parameters is equivalent to inference based on an exchangeable mixture prior that contains all $G!$ label-permuted versions of the non-exchangeable prior.*

Proof: Let $p_1(\tau)$ be the non-exchangeable prior distribution of interest. Let $p_s(\tau), s = 1, ..., G!$ be the the $G!$ label-permuted versions of the non-exchangeable prior. Let $\pi^*(\tau)$ be the exchangeable mixture of $p_s(\tau)$. Then $p(y|G) = \int p(y|\tau)p_i(\tau)d\tau$ is the same for $i = 1, ..., G!$ because the likelihood is symmetric with respect to permutations of the labels. So,

$$\begin{aligned} p_1(y|G) &= \frac{1}{G!}G! \int p(y|\tau, G)p_1(\tau|G)d\tau \\ &= \frac{1}{G!}\sum_{s=1}^{G!} \int p(y|\tau, G)p_s(\tau|G)d\tau \\ &= \int p(y|\tau, G)(\frac{1}{G!}\sum_{s=1}^{G!} p_s(\tau|G))d\tau \\ &= \int p(y|\tau, G)\pi^*(\tau)d\tau \\ &= p_*(y|G) \end{aligned}$$

Because posterior inference about G depends on τ only through $p(y|G)$, any posterior inference about G using $p_1(\tau)$ as a prior will be equivalent to inference using $\pi^*(\tau)$ as a prior. □

Theorem 1 shows that the simpler non-exchangeable prior provides the same posterior inference about G as the computationally more expensive exchangeable mixture prior with $G!$ components. The exchangeable prior has the appealing characteristic that it does not assign substantial prior mass to values of the μ_g's far from the likelihood modes.

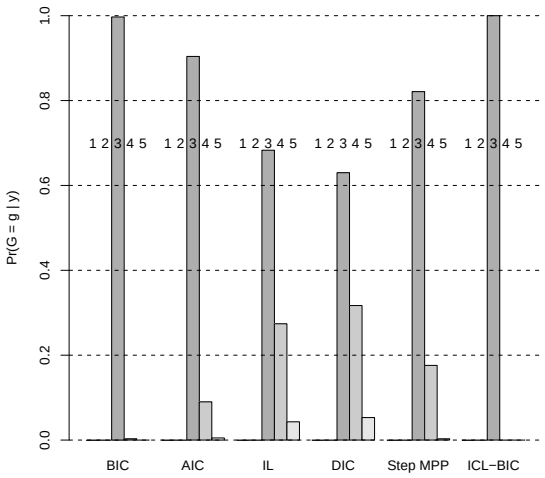


Figure 3: Posterior probabilities of G for trimodal data in Figure 2. IL = posterior probabilities from the unit information prior of Section 3. StepMPP = posterior probabilities using the prior and methods of Stephens (2000a).

We compute the posterior probabilities of different values of G with this prior using the incremental mixture importance sampling (IMIS) method of Steele, Raftery, and Emond (2006).

4 Examples

We now compare the results of choosing the number of mixture components from the methods described above for one simulated example and one real data example.

4.1 Simulated Example

Figure 3 shows the posterior probabilities of different values of G for the trimodal dataset in Figure 2 using the different methods discussed above. For AIC, the posterior probabilities are derived from the Bayesian interpretation of AIC as minus twice the log integrated likelihood for a highly informative prior, as $p(G|y) \propto \exp(-\mathrm{AIC}/2)$. DIC is computed using the unit information prior in Section 3 and is also viewed as an approximation to minus twice the log integrated likelihood. The prior for G was uniform over the integers $1, 2, \ldots, 7$.

The true number of components in this case is $G = 3$, and BIC and ICL both put almost all the posterior mass on this value. The other methods put some posterior mass on bigger

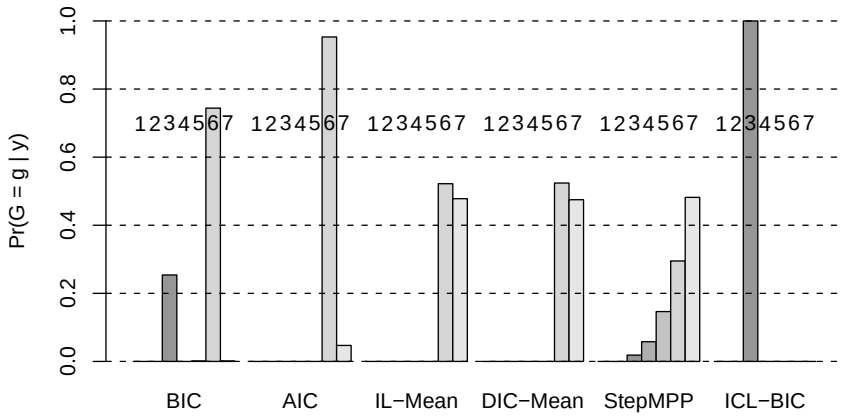


Figure 4: Posterior probabilities of G for galaxy example. IL-Mean = posterior probabilities from the unit information prior of Section 3. DIC-Mean = posterior probabilities based on DIC.

values.

4.2 Galaxy Data

Figure 4 shows the posterior probabilities of G for the canonical galaxy velocity data set (Postman, Huchra, and Geller 1986) analyzed by Roeder (1990) and several others since.

The methods give very different results. The fully Bayesian approach of Stephens (2000a) (which is similar to that of Richardson and Green 1997) puts the most mass on the largest value considered, $G = 7$. AIC, DIC and the Bayesian approach using our unit information prior put most of the mass on $G = 6$ and $G = 7$. BIC shares the mass between $G = 3$ (25%) and $G = 6$ (75%), while ICL puts almost all the mass on $G = 3$.

We do not know the "correct" answer for this famous dataset, and so we cannot say which methods are "right" or "wrong". However, exploratory analysis of this dataset by Fraley and Raftery (2007) sheds some light. If one looks at the full dataset using standard goodness-of-fit assessment methods (the empirical cumulative distribution function and Kolmogorov-Smirnov tests), it is clear that it is not well fit by a single normal. This is in line with the fact that all methods give essentially zero posterior probability to $G = 1$.

There are two clear small "clusters", one at the left and one at the right of the dataset. When a mixture model with $G = 3$ is fit, these are clearly separated out, with one component

corresponding to each of these small clusters, and the remaining component corresponding to the majority of points in the middle (72 of the 82 points).

When one looks at this majority group of points in the middle, there are no strong deviations from normality apparent (Fraley and Raftery 2007, Figure 4), and standard goodness of fit tests do not reject normality. For example, the P-value from a Kolmogorov-Smirnov test of normal is 0.254.

This suggests that no reasonable statistical analysis based on the data alone should categorically exclude $G = 3$ in favor of larger values of G. Yet, AIC, DIC and the fully Bayesian analyses based on both priors considered do just that. ICL goes to the other extreme: it overwhelmingly favors $G = 3$ over all other possibilities. That leaves BIC, which gives mass to both $G = 3$ and $G = 6$, which seems scientifically reasonable.

We attach two caveats to these results. First, the DIC is based on the unit information prior in Section 3, and the results might be different for other priors, including the (very similar) priors of Richardson and Green (1997) and Stephens (2000a). Second, only BIC, the fully Bayesian methods and, arguably, AIC, can be interpreted as yielding posterior probabilities, but we put DIC and ICL on the same probabilistic evidence scale for comparison purposes.

5 Simulation Study

We now describe our simulation study to compare the different methods for choosing the number of components.

5.1 Study Design

The results of a simulation study for comparing methods can depend critically on the study design. Often the simulation study is designed by the researcher in a fairly arbitrary way. Given this, we tried to ground our study explicitly on the published experience with mixture models.

We searched the literature for published parameter estimates to be used as the basis for our design. In order to choose examples in a fair, comprehensive, but reasonable fashion, we looked at all papers published before July 1999 listed in the Science Citation Index/Web of Science that referenced one of three textbooks on mixture models:

- Everitt and Hand, Finite Mixture Distributions (1981)
- Titterington, Smith, and Makov, Statistical Analysis of Mixture Distributions (1985)

Table 1: Parameters for the simulation study, as suggested by the 43 parameter estimates from the literature

G=1: (Experiments 1-5) n=400, 300, 200, 150, 100 $\mu_1 = 0, \sigma_1 = 1$, for all experiments				
G=2: $\mu_1 = 0, \sigma_1 = 1$ for all experiments				
Experiment	μ_2	σ_2	λ_2	n
6	1.0	0.5	.25	400
7	2.0	2.0	.25	300
8	3.0	1.0	.25	200
9	4.0	2.5	.25	150
10	6.0	1.5	.25	100

- McClachlan and Basford, Mixture Models: Inference and Applications to Clustering (1987)

We also included a set of examples that were listed in the Titterington, et al. book. In all, we looked at over 200 papers published between 1936 and 1999. Of these, we found 22 papers that listed 43 parameter estimates for two-component models. (There were very few reported estimates for mixture models with more than two components.)

The 43 two-component examples and references are listed in Tables 4 and 5, along with the mixture parameter estimates. We standardized each example's parameters such that the group with the larger estimated component membership had mean 0 and variance 1. The smaller group's mean was then normalized to $|\mu_2 - \mu_1|$ and the variance to $\frac{\sigma_2^2}{\sigma_1^2}$. The standardized values are shown in Table 6.

From an experimental design perspective, one would like the simulation study to "cover" the space of the mixture examples in the literature. Using the model-based clustering R package `mclust` (Fraley and Raftery 1998), we fit a mixture of multivariate normal distributions to the literature parameter estimates, restricting the component densities to have equal, diagonal covariance matrices, using the BIC to choose the "optimal" number of components. Similar experimental design approaches for computer experiments can be found in texts on number theoretic methods (see for example Fang and Wang 1994).

The BIC suggested 5 components, and we therefore used 5 settings of the mixture parameters for the two-component mixture model in our experiment. For balance we also included 5 experiments with one component, and different sample sizes. The final simulation study design is shown in Table 1, and the mixture densities used are shown in Figure 5.

We generated 50 data sets for each of the 10 experiments. For each of the 500 simulated datasets, we then found the selected number of components for each of the six selection methods compared, as described in Section 4.

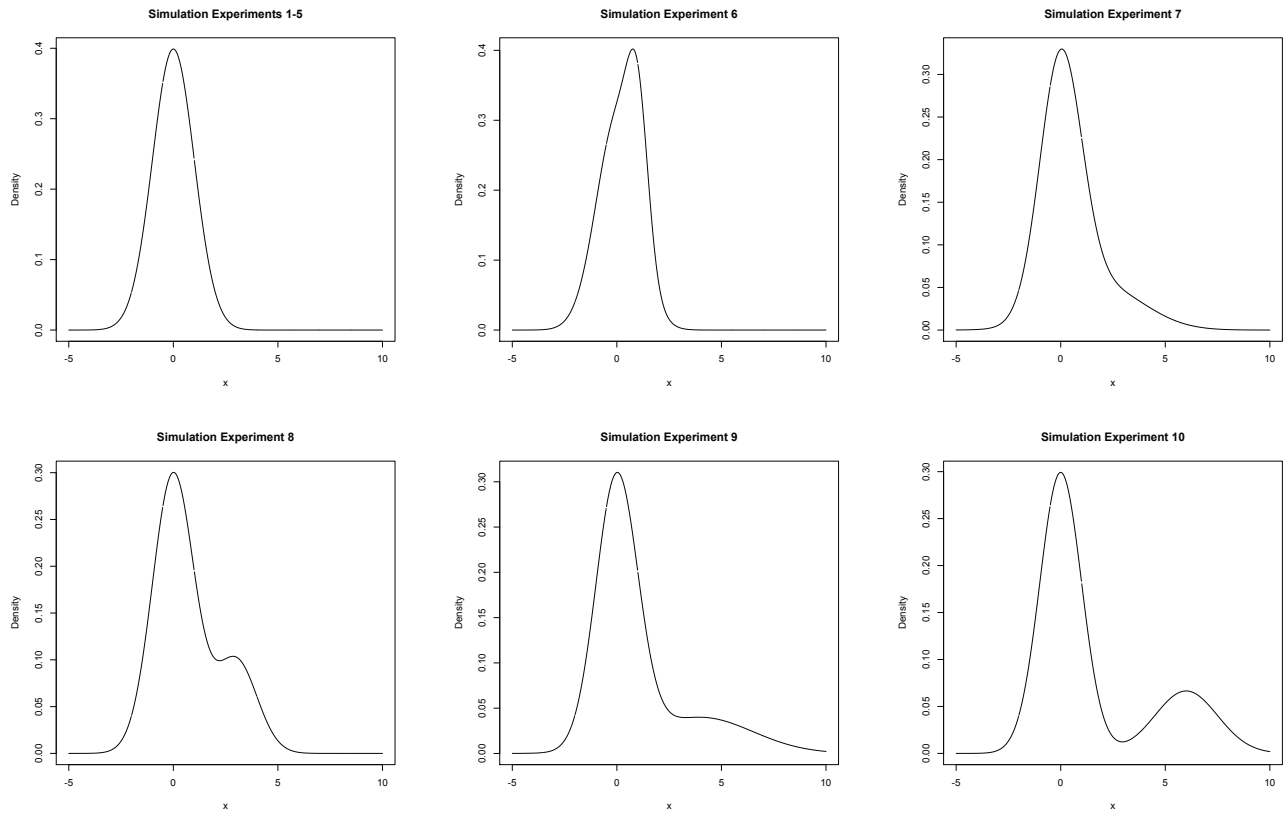
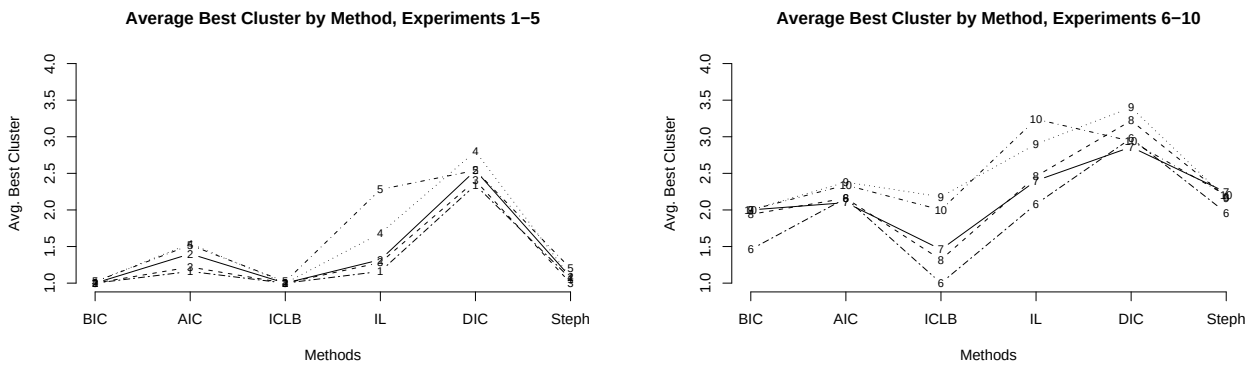


Figure 5: Simulation study: The densities used. The top left plot shows the one-component normal density which was the basis for experiments 1–5. The remaining plots show the two-component mixture of normals used for experiments 6–10.



(a) (b)

Figure 6: Average selected number of mixture components by method and experiment. The experiments are listed in Table 1. ICLB refers to the ICL-BIC criterion, IL is the integrated from the unit information prior in Section 3, and Steph refers to the MAP estimate from Stephens's fully Bayesian method. (a) Experiments 1–5, for which the true $G = 1$. (b) Experiments 6–10, for which the true $G = 2$.

5.2 Results

The average number of components selected by each method in each experiment is shown in Figure 6. For experiments 1–5 for which the true number of components was 1, BIC and ICL were very accurate, Stephens's fully Bayesian method was almost as accurate, while AIC, our new UIP, and, most strikingly, DIC, overestimated the number of components considerably.

For experiments 6–10, for which the true number of components was 2, BIC was highly accurate on average for experiments 7–10 but less so for experiment 6. Stephens's method was also accurate on average. AIC overestimated the number somewhat on average for experiments 9 and 10. The other methods, ICL, the UIP and DIC, were much more variable and inaccurate.

The number of times each method chose the correct number of components is shown in Table 2. BIC performed best overall, achieving almost perfect accuracy for each experiment except experiment 6. This may be because experiment 6, while a two-component mixture, is very close to a single Gaussian, and BIC chose $G = 1$ a relatively high proportion of the time. Stephens's method was second best overall, clearly outperforming the other methods. It also did worse for experiment 6 than for the other experiments.

DIC performed particularly poorly, uniformly across all experiments. This may be due to the fact that it used our unit information prior, but the fully Bayesian method using the UIP performed much better, so this is doubtful.

Mixture modeling is often done to estimate the density rather than to assess the number

Table 2: Simulation study: The number of times each of the six model selection criteria chose the correct number of components for the ten experiments in Table 1. UIP refers to the MAP estimate from the fully Bayesian analysis with the unit information prior in Section 3.

Expt.	BIC	Stephens	AIC	ICL	UIP	DIC
1	50	49	45	50	44	20
2	50	48	38	50	39	17
3	50	50	42	50	40	22
4	49	48	34	50	30	14
5	49	46	33	49	19	16
6	23	29	35	0	40	20
7	50	42	46	19	34	23
8	47	45	45	16	33	14
9	50	41	37	39	22	10
10	50	43	39	50	7	20
Total	468	441	394	373	308	176
% Correct	94	88	79	75	62	35

of components. We therefore compared the performance of the six methods for density estimation. To do this in a comparable way, we chose the value of G selected by each method and then found the MLE of the parameters for that value of G and the corresponding density, and computed its mean integrated squared error.

The results are shown in Table 3. Again, BIC did best, followed by the fully Bayesian Stephens's method, and DIC did worst by far.

6 Discussion

We have considered five established Bayesianly motivated methods for choosing the number of components in a mixture model, and introduced an additional one based on a new unit information prior for mixture models. We compared all six methods using a simulation study whose design was based on a broad survey of mixture model parameter estimates published in the literature before 2000. BIC performed best, quite decisively, with the fully Bayesian method of Stephens (2000a) (similar to that of Richardson and Green 1997) outperforming the other methods. AIC, ICL and DIC performed poorly. So did our own new proposed prior in this context, unfortunately.

There are other Bayesian approaches to choosing the number of components that we have not considered here. Steele (2002) proposed a modification of BIC, defined as follows:

$$\mathrm{BIC}_2 = 2\log(p(y|\hat{\tau}, G) + 2\sum_{g=1}^{G}\log(\hat{n}_g + 0.0001) + (G-1)\log(n). \tag{3}$$

Table 3: The average Mean Integrated Squared Error (MISE) for the 10 experiments in Table 1. The values in the table are multiplied by 1000.

Expt.	BIC	Stephens	AIC	ICL	UIP	DIC
1	0.19	0.21	0.22	0.19	0.23	0.67
2	0.21	0.24	0.33	0.21	0.31	0.65
3	0.35	0.35	0.41	0.35	0.50	1.32
4	0.48	0.51	1.30	0.48	1.35	2.24
5	0.60	1.00	1.58	0.60	2.75	3.20
6	1.53	1.13	0.86	2.31	0.77	0.76
7	0.23	0.24	0.23	2.18	0.25	0.28
8	0.55	0.39	0.37	2.45	0.42	0.61
9	0.37	0.75	0.47	0.61	0.58	0.77
10	0.34	0.44	0.39	0.34	0.75	0.58
Mean	0.48	0.53	0.62	0.97	0.79	1.11

BIC_2 penalizes the parameters of the mixture components by the logarithm of the estimated sample size of their component rather than of the entire sample size, which seems more in line with the derivation of BIC for regular models. However, BIC_2 performed less well than the raw BIC in Steele's (2002, chapter 3) simulation study, and so we did not include it here.

Fraley and Raftery (2007) proposed a regularized version of BIC. This addressed the problem with BIC that it depends on the maximum likelihood estimator, which can sometimes be degerate in mixture models, particularly if there are ties in the data. It replaced the MLE by the posterior mode with a very weakly informative prior, crafted so that the posterior mode hardly differs from the MLE except when the MLE is degenerate. This is implemented in the `mclust` R package. We did not include this in the current comparison, and we conjecture that it would perform similarly to BIC, but perhaps slightly better because of the regularization.

We have addressed only one aspect of the prior specification, namely the prior for the component density parameters. Analysts also have the flexibility to specify different priors on the mixing parameters (for example, a choice between the often used uniform Dirichlet and the Jeffrey's prior (Robert 1994)), but this has been relatively unexplored in the mixture literature. Stephens (2000a) uses a Poisson prior on the number of components, and it would be interesting to examine how the interplay between the three priors (on the component densities, mixing parameters, and the number of components, respectively) affects inference and can be used by analysts in a constructive way.

We did not consider the class of Dirichlet process mixture priors (Escobar and West 1995), mostly for reasons of brevity and relevance. The motivation behind the use of the Dirichlet process mixture approach to finite mixture models is mostly to model the underlying

population density, rather than to make inference about the number of components. For further discussion of this issue, see Green and Richardson (2001) on the choice between parametric and non-parametric modeling for mixtures and the effect on choosing the number of components.

Gaussian mixture models are often used for clustering, particularly in more than one dimension. Choosing the number of mixture components is not necessarily the same as choosing the number of clusters. This is because a cluster may arise from a non-Gaussian distribution, which is itself approximated by a mixture of Gaussians. Methods for combining mixture components to form clusters have been proposed by Wang and Raftery (2002, Section 4.5), Tantrum, Murua, and Stuetzle (2003), Li (2005), Jörnsten and Keleş (2008) and Baudry et al. (2008).

References

Aitkin, M. and D. B. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B, Methodological 47*, 67–75.

Aitkin, M. and G. T. Wilson (1980). Mixture models, outliers, and the EM algorithm (Corr: V23 p427). *Technometrics 22*, 325–331.

Akaike, H. (1983). Information measures and model selection. In *Proceedings of the International Statistical Institute 44th Session*, pp. 277–291. The Hague, Netherlands: International Statistical Institute.

Ashman, K. M., C. M. Bird, and S. E. Zepf (1994). Detecting bimodality in astronomical datasets. *The Astronomical Journal 108*, 2348–2361.

Baudry, J. P., A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo (2008). Combining mixture components for clustering. Technical Report 540, Department of Statistics, University of Washington, Seattle, Wash.

Biernacki, C., G. Celeux, and G. Govaert (1998). Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report No. 3521, Institut National de Recherche en informatique et en automatique.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Brown, J., R. Moulton, S. Konasiewicz, and A. Baker (1998). Cerebral oxidative metabolism and evoked potential deterioration after severe brain injury: new evdience of early posttraumatic ischemia. *Neurosurgery 42*, 1057–1063.

Bull, W. and M. Brandon (1998). Lichen dating of earthquake-generated regional rockfall events, Southern Alps, New Zealand. *Geological Society of America Bulletin 110*, 60–84.

Cassie, R. (1954). Some uses of probability paper for the graphical analysis of polymodal frequency distributions. *Australian Journal of Marine and Freshwater Resources 5*, 513–22.

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington (2006). Deviance information criteria for missing data models (Pkg: P651-706). *Bayesian Analysis 1*, 651–674.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*, 195–212.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Clark, M. (1977). Gethen: A computer program for the decomposition of mixtures of two normal distributions by the method of moments. *Computers and Geosciences 3*, 257–267.

Clark, V. A., J. M. Chapman, A. H. Coulson, and V. Hasselblad (1968). Dividing the blood pressures from the Los Angeles heart study into two normal distributions. *The Johns Hopkins Medical Journal 122*, 77–83.

Cohen, J. G., J. P. Blakeslee, and A. Ryzhov (1998). The ages and abundances of a large sample of M87 globular clusters. *The Astrophysical Journal 496*, 808–826.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

Dernburg, A., K. Browman, J. Fung, W. Marshall, J. Philips, D. Agard, and J. Sedat (1996). Perturbation of nuclear architecture by long-distance chromosome interactions. *Cell 85*, 745–759.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, Methodological 56*, 363–375.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Everitt, B. S. and D. J. Hand (1981). *Finite mixture distributions*. Chapman and Hall Ltd.

Fang, K.-T. and Y. Wang (1994). *Number-theoretic methods in statistics*. Chapman and Hall Ltd.

Ferreira da Silva, A. R. (2009, APR). Bayesian mixture models of variable dimension for image segmentation. *Computer Methods and Programs in Biomedicine 94*(1), 1–14.

Flury, B. D., J. P. Airoldi, and J. P. Biber (1992). Gender identification of water pipits (Anthus spinoletta) using mixtures of distributions. *Journal of Theoretical Biology 158*, 465–480.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*, 155–181.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Gage, T. B. and G. Therriault (1998). Variability of birth-weight distributions by sex and ethnicity: analysis using mixture models. *Human Biology 70*, 517–534.

Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics 28*, 355–375.

Gutierrez, R. G., R. J. Carroll, N. Wang, G.-H. Lee, and B. H. Taylor (1995). Analysis of tomato root initiation using a normal mixture distribution. *Biometrics 51*, 1461–1468.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions (Com: p445-446). *Technometrics 8*, 431–444.

Jasra, A., C. Holmes, and D. Stephens (2005, FEB). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science 20*(1), 50–67.

Jörnsten, R. and S. Keleş (2008). Mixture models with multiple levels, with application to the analysis of multifactor gene expression data. *Biostatistics 9*, 540–554.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya, Series A, Indian Journal of Statistics 62*, 49–66.

Koehler, A. B. and E. S. Murphree (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics 37*, 187–195.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics 14*, 547–568.

Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications*. Hayward, Calif.: Insitute of Mathematical Statistics.

Livezey, B. (1993). An ecomorphological review of the dodo (Raphus cucullatus) and solitaire (Pezophaps solitaria), flightless Colombiformes of the Mascarene Islands. *Journal of Zoology London 230*, 247–292.

Martin, E. S. (1936). A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing. *Biometrika 28*, 149–178.

McGrory, C. A. and D. M. Titterington (2007, JUL 15). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis 51*(11), 5352–5367.

McLachlan, G. J. and K. E. Basford (1987). *Mixture models: Inference and applications to clustering*. Marcel Dekker Inc.

McLaren, C., V. Gordeuk, A. Looker, V. Hasselblad, C. Edwards, L. Griffen, and G. Brittenham (1995). Prevalence of heterozygotes for hemochromatosis in the white population of the United States. *Blood 86*, 2021–2027.

Monetti, A., G. Versini, G. Dalpias, and F. Reniero (1996). Sugar adulterations control in concentrated rectified gradpe musts by finite mixture distribution analysis of the myo- and scyllo-inositol content andthe D/H methyl ratio of fermentative ethanol. *Journal of Agricultural and Food Chemistry 44*, 2194–2201.

Murray, G. D. and D. M. Titterington (1978). Estimation problems with data from a mixture. *Applied Statistics 27*, 325–334.

Nemec, J. and A. Nemec (1991). Mixture models for studying stellar populations. i. Univariate mixture models, parameter estimation, and the number of discrete population components. *Publications of the Astronomical Society of the Pacific 103*, 95–121.

Nobile, A. and A. T. Fearnside (2007, JUN). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing 17*(2), 147–162.

Plummer, M. (2008, JUL). Penalized loss functions for Bayesian model comparison. *BIOSTATISTICS 9*(3), 523–539.

Postman, M., J. Huchra, and M. Geller (1986). Probes of large-scale structure in the corona borealis region. *The Astronomical Journal 92*, 1238–47.

Raftery, A. E. (1995). Bayesian model selection for social research (with discussion). *Sociological Methodology 25*, 111–196.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, D. J. Spiegelhalter, and S. Richardson (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 163–188. London: Chapman and Hall.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society - Series B 59*, 731–792.

Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation*. Springer-Verlag Inc.

Robert, C. P., T. Rydén, and D. M. Titterington (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 62*, 57–75.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association 85*, 617–624.

Roeder, K. and L. Wasserman (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*, 894–902.

Rushforth, N., P. Bennett, A. Steinberg, T. Burch, and M. Miller (1971). Diabetes in the Pima indians. *Diabetes 20*, 756–765.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Spezia, L. (2009, JUL 1). Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference 139*(7), 2305–2315.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (Pkg: P538-639). *Journal of the Royal Statistical Society, Series B: Statistical Methodology 64*, 583–616.

Steele, R. J. (2002). *Importance sampling methods for inference in mixture models and missing data*. Ph. D. thesis, Department of Statistics, University of Washington, Seattle, Wash.

Steele, R. J., A. E. Raftery, and M. J. Emond (2006, SEP). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS).

Journal of Computational and Graphical Statistics 15(3), 712–734.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *The Annals of Statistics 28*, 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, Methodological 62*, 795–809.

Stricker, C., S. Redman, and D. Daley (1994). Statistical analysis of synaptic transmission: model discrimination adn confidence limits. *Biophysical Journal 67*, 532–547.

Tantrum, J., A. Murua, and W. Stuetzle (2003). Assessment and pruning of hierarchical model based clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 197–205. Association for Computing Machinery.

Thalange, N., P. Foster, M. Gill, D. Price, and P. Clayton (1996). Model of normal prepubertal growth. *Archives of Disease in Childhood 75*, 427–431.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. John Wiley and Sons.

Wang, N. and A. E. Raftery (2002). Nearest neighbor variance estimation (NNVE): Robust covariance estimation via nearest neighbor cleaning (with discussion). *Journal of the American Statistical Association 97*, 994–1019.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B, Methodological 62*, 159–180.

West, M. (1992). Modelling with mixtures (Disc: P519-524). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. . M. Smith (Eds.), *Bayesian Statistics 4. Proceedings of the Fourth Valencia Inter national Meeting*, pp. 503–519. Clarendon Press [Oxford University Press].

Appendix

Table 4: The first 21 of 43 parameter estimates obtained from the literature

Author(s)	Year	μ_1	μ_2	σ_1^2	σ_2^2	n	λ_1
Murray, Titterington	1978	98.940	118.680	13.330	9.970	166	0.3310
Martin	1936	116.840	110.540	5.510	5.730	337	.4860
Cassie	1962	-0.330	6.506	1.262	1.193	100	0.1500
Cassie	1962	19.967	26.175	2.158	2.772	1000	0.6540
Cassie	1962	0.814	2.584	1.547	0.699	64651	0.8140
Aitken, Tunnicliffe-Wilson	1980	-57.350	33.800	19.630	19.630	15	0.1300
Hasselblad	1966	3.210	7.339	1.000	2.220	430	0.2162
Gutierrez	1995	0.520	0.820	0.141	0.063	40	0.9100
Thalange, et al.	1996	0.002	0.046	0.032	0.155	3674	0.4230
Stricker, et al.	1994	-0.200	1.350	0.760	0.460	589	0.5600
McLaren, et al.	1995	29.500	46.500	7.000	7.000	1375	0.8400
McLaren, et al.	1995	26.700	43.400	7.000	7.000	1547	0.8500
Dernburg, et al.	1996	0.410	0.140	0.200	0.070	450	0.5700
Dernburg, et al.	1996	0.480	0.190	0.190	0.080	450	0.2500
Gage, Therriault	1998	3545.000	3050.000	475.000	1110.000	70428	0.9260
Gage, Therriault	1998	3215.000	2634.000	453.000	1042.000	14954	0.8400
Brown, et al.	1998	0.470	-28.870	4.440	6.190	88	0.8300
Clark, et al.	1968	118.300	147.600	14.660	26.340	1148	0.7400
Clark, et al.	1968	65.700	78.400	10.100	13.000	1146	0.7200
Clark, et al.	1968	116.100	145.900	12.610	23.490	210	0.6500
Clark, et al.	1968	71.000	94.900	10.090	7.340	208	0.8500

Table 5: The last 22 of 43 parameter estimates obtained from the literature

Author(s)	Year	μ_1	μ_2	σ_1^2	σ_2^2	n	λ_1
Rushforth, et al.	1971	1.987	2.527	0.109	0.245	245	0.9750
Rushforth, et al.	1971	2.178	2.538	0.130	0.148	52	0.6450
Bull, Brandon	1998	22.100	22.100	0.710	4.700	40	0.3200
Bull, Brandon	1998	22.000	22.000	3.200	23.300	175	0.6400
Bull, Brandon	1998	19.300	19.300	3.800	7.000	430	0.4600
Bull, Brandon	1998	14.800	14.800	2.500	15.200	108	0.6800
Flury, et al	1992	85.210	92.120	1.200	1.230	36	0.5280
Flury, et al.	1992	86.170	92.350	1.530	1.530	381	0.4900
Flury, et al.	1992	86.830	93.620	1.310	1.870	123	0.3710
Flury, et al.	1992	85.970	92.350	1.530	1.530	298	0.5750
Flury, et al.	1992	473.430	428.160	22.670	22.670	288	0.5080
Livezey	1993	131.600	125.100	2.600	2.200	137	0.5200
Livezey	1993	262.500	220.400	10.400	5.300	19	0.2100
Clark	1977	3.670	1.930	1.020	0.404	229	0.7312
Ashman, et al.	1994	1.530	1.950	0.150	0.150	60	0.5500
Ashman, et al.	1994	-629.000	771.000	22.700	22.700	133	0.6300
Ashman, et al.	1994	-0.346	1.462	0.492	0.492	223	0.2800
Cohen, et al.	1998	1.460	2.730	0.600	0.600	129	0.5500
Nemec, Nemec	1991	-1.560	-0.540	0.330	0.220	119	0.7700
Monetti, et al.	1996	95.190	102.180	1.470	1.318	354	0.2572
Monetti, et al.	1996	6.130	7.250	0.528	0.239	354	0.2644
Monetti, et al.	1996	4.270	5.440	0.854	0.270	354	0.3007

Table 6: Standardized means and variances of the components with largest estimated group membership for each of the 43 published estimates from the literature. The estimated mixing parameters for the larger estimated group and the total number of observations are also shown.

μ_2^*	σ_2^{2*}	λ_1	n	μ_2^*	σ_2^{2*}	λ_1	n
1.98	0.75	0.66	166	4.95	0.44	0.97	245
1.09	1.03	0.51	337	2.76	0.87	0.64	52
5.73	0.94	0.85	100	0.00	6.61	0.68	40
2.87	0.77	0.65	1000	0.00	0.13	0.64	175
1.14	2.21	0.81	64651	0.00	1.84	0.54	430
4.64	1.00	0.87	15	0.00	0.16	0.68	108
1.85	2.22	0.78	430	5.75	0.97	0.52	36
2.12	2.23	0.91	40	4.03	1.00	0.51	381
0.28	4.84	0.57	3674	3.63	1.42	0.62	123
2.03	1.65	0.56	589	4.16	1.00	0.57	298
2.42	1.00	0.84	1375	1.99	1.00	0.50	288
2.38	1.00	0.85	1547	2.50	1.18	0.52	137
1.35	2.85	0.57	450	7.94	0.50	0.79	19
3.62	0.42	0.75	450	1.70	2.52	0.73	229
1.04	0.42	0.92	70428	2.80	1.00	0.55	60
1.28	0.43	0.84	14954	1.67	1.00	0.63	133
6.60	0.71	0.83	88	3.67	1.00	0.72	223
1.99	0.55	0.74	1148	2.11	1.00	0.55	129
1.25	0.77	0.72	1146	3.09	1.50	0.77	119
2.36	0.53	0.65	210	5.30	0.89	0.74	354
2.36	1.37	0.85	208	4.68	0.45	0.73	354
4.33	0.31	0.69	354				