Journal of Classification 24:155-181 (2007)
DOI: 10.1007/s00357-007-0004-z

# Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering

Chris Fraley

University of Washington, Seattle

Adrian E. Raftery

University of Washington, Seattle

**Abstract:** Normal mixture models are widely used for statistical modeling of data, including cluster analysis. However maximum likelihood estimation (MLE) for normal mixtures using the EM algorithm may fail as the result of singularities or degeneracies. To avoid this, we propose replacing the MLE by a maximum a posteriori (MAP) estimator, also found by the EM algorithm. For choosing the number of components and the model parameterization, we propose a modified version of BIC, where the likelihood is evaluated at the MAP instead of the MLE. We use a highly dispersed proper conjugate prior, containing a small fraction of one observation's worth of information. The resulting method avoids degeneracies and singularities, but when these are not present it gives similar results to the standard method using MLE, EM and BIC.

**Keywords:** BIC; EM algorithm; Mixture models; Model-based clustering; Conjugate prior; Posterior mode.

---

Funded by National Institutes of Health grant 8 R01 EB002137–02 and by Office of Naval Research contract N00014-01-10745.

Authors' Address: Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322 USA, e-mail: fraley/raftery@stat.washington.edu

## 1. Introduction

Finite mixture models are an increasingly important tool in multivariate statistics (e.g. McLachlan and Basford 1988; McLachlan and Peel 2000). Approaches to density estimation and clustering based on normal mixture models have shown good performance in many applications (McLachlan and Peel 2000; Fraley and Raftery 2002). Despite this success, there remain issues to be overcome. One is that standard parameter estimation methods can fail due to singularity for some starting values, some models, and some numbers of components. The techniques proposed in this paper are largely able to eliminate these difficulties.

We propose replacing the MLE by a maximum a posteriori (MAP) estimator, for which we use the EM algorithm. For choosing the number of components and the model parameterization, we propose a modified version of BIC, in which the likelihood is evaluated at the MAP instead of the MLE. We use a highly dispersed proper conjugate prior, containing a small fraction of one observation's worth of information. The resulting method avoids degeneracies and singularities, but when these are not present it gives similar results to the standard method that uses MLE, EM and BIC. It also has the effect of smoothing noisy behavior of the BIC, which is often observed in conjunction with instability in estimation.

This paper is organized as follows. In Section 2, we give a brief overview of model-based clustering, which can also be viewed as a proceedure for normal mixture estimation that includes model selection, both in terms of component structure and number of components. In Section 3, we describe our Bayesian regularization method for univariate normal mixtures and we discuss selection of prior hyperparameters appropriate for clustering. In Section 4, we do the same for multivariate normal mixtures. In Section 5, we give examples of mixture estimation with these priors for real data. Other topics treated in this section include alternative priors, extension to other parameterizations of multivariate normal mixtures, and the types of failures that can still occur when a prior is used and how they can be overcome. In Section 6 we discuss our results in the context of other research and further application of this approach.

## 2. Methods

### 2.1 Model-Based Clustering

In model-based clustering, the data $y = (y_1, \ldots, y_n)$ are assumed to be generated by a mixture model with density

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \, f_k(y_i \mid \theta_k), \tag{1}$$

where $f_k(y_i \mid \theta_k)$ is a probability distribution with parameters $\theta_k$, and $\tau_k$ is the probability of belonging to the $k$th component. Most often (and throughout this paper), the $f_k$ are taken to be multivariate normal distributions, parameterized by their means $\mu_k$ and covariances $\Sigma_k$:

$$f_k(y_i \mid \theta_k) = \phi(y_i \mid \mu_k, \Sigma_k) \equiv |2\pi\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1} (y_i - \mu_k) \right\},$$

where $\theta_k = (\mu_k, \Sigma_k)$. The parameters of the model are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997). Each EM iteration consists of two steps, an E-step and an M-step. Given an estimate of the component means $\mu_j$, covariance matrices $\Sigma_j$ and mixing proportions $\tau_j$, the E-step computes the conditional probability that object $i$ belongs to the $k$th component:

$$z_{ik} = \tau_k \phi(y_i|\mu_k, \Sigma_k) / \sum_{j=1}^{G} \tau_j \phi(y_i|\mu_j, \Sigma_j) \; .$$

In the M-step, parameters are estimated from the data given the conditional probabilities $z_{ik}$ (see, e.g., Celeux and Govaert 1995). The E-step and M-step are iterated until convergence, after which an observation can be assigned to the component or cluster corresponding to the highest conditional or posterior probability. The results of EM are highly dependent on the initial values, and model-based hierarchical clustering can be a good source of initial values for datasets that are not too large in size (Banfield and Raftery 1993; Dasgupta and Raftery 1998; Fraley 1998).

The covariance matrices can be either fixed to be the same across all mixture components, or allowed to vary. In general, the multivariate normal density has ellipsoidal contours, and the covariance matrices can also be constrained to make the contours spherical or axis-aligned. Other parameterizations are possible and have been found to be useful; regularization of these is discussed in Section 5.4.

Several measures have been proposed for choosing the clustering model (parameterization and number of clusters); see, e.g., Chapter 6 of McLachlan and Peel (2000). We use the Bayesian Information Criterion (BIC) approximation to the Bayes factor (Schwarz 1978), which adds a penalty to the loglikelihood based on the number of parameters, and has performed well in a number of applications (e.g. Dasgupta and Raftery 1998; Fraley and Raftery 1998, 2002).

The BIC has the form

$$\text{BIC} \equiv 2\,\text{loglik}_{\mathcal{M}}(y, \theta_k^*) - (\#\,\text{params})_{\mathcal{M}}\,\log(n), \tag{2}$$

where $\text{loglik}_{\mathcal{M}}(y, \theta_k^*)$ is the maximized loglikelihood for the model and data, $(\#\,\text{params})_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model $\mathcal{M}$, and $n$ is the number of observations in the data.

The following strategy for model selection has been found to be effective in mixture estimation and clustering:

- Specify a maximum number of components, $G_{\max}$, to consider, and a set of candidate parameterizations of the Gaussian model.
- Estimate parameters via EM for each parameterization and each number of components up to $G_{\max}$. The conditional probabilities corresponding to a classification from model-based hierarchical clustering, or the estimated parameters for a simpler (more parsimonious) model, are good choices for initial values.
- Compute BIC for the mixture likelihood with the optimal parameters from EM for up to $G_{\max}$ clusters.
- Select the model (parameterization / number of components) for which BIC is maximized.

For a review of model-based clustering, see Fraley and Raftery (2002). Efficient implementation for large datasets is discussed in Wehrens, Buydens, Fraley, and Raftery (2004) and Fraley, Raftery, and Wehrens (2005).

The EM algorithm can fail to converge, instead diverging to a point of infinite likelihood. For many mixture models, the likelihood is not bounded, and there are paths in parameter space along which the likelihood tends to infinity (Titterington, Smith and Makov 1985). For example, in the univariate normal mixture model with component-specific variances, where (1) becomes

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \phi(y_i | \mu_k, \sigma_k^2), \tag{3}$$

the likelihood tends to infinity along any path in parameter space along which $\mu_k \longrightarrow y_i$ and $\sigma_k^2 \longrightarrow 0$, for any $i$, if $\tau_k$ is bounded away from zero. While these points of infinite likelihood could technically be viewed as maximum likelihood estimates, they do not possess the usual good properties of MLEs, which do hold for an internal local maximum of the likelihood (Redner and Walker 1984).

In practice, this behavior is due to singularity in the covariance estimate, and arises most often for models in which the covariance is allowed to vary between components, and for models with large numbers of components. It is

natural to wonder whether the best model might be among the cases for which a failure is observed, and to seek to modify the method so as to eliminate convergence failures.

We propose to avoid these problems by replacing the MLE by the maximum a posteriori (MAP) estimate from a Bayesian analysis. We propose a prior distribution on the parameters that eliminates failure due to singularity, while having little effect on stable results obtainable without a prior. The Bayesian predictive density for the data is assumed to be of the form

$$\mathcal{L}(Y \mid \tau_k, \mu_k, \Sigma_k)\ \ \mathcal{P}(\tau_k, \mu_k, \Sigma_k \mid \theta),$$

where $\mathcal{L}$ is the mixture likelihood:

$$\begin{aligned}\mathcal{L}(Y \mid \tau_k, \mu_k, \Sigma_k) &= \prod_{j=1}^{n} \sum_{k=1}^{G} \tau_k \phi(y_j \mid \mu_k, \Sigma_k) \\ &= \prod_{j=1}^{n} \sum_{k=1}^{G} \tau_k \left|2\pi\Sigma_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_j - \mu_k)^T \Sigma_k^{-1}(y_j - \mu_k)\right\},\end{aligned}$$

and $\mathcal{P}$ is a prior distribution on the parameters $\tau_k$, $\mu_k$ and $\Sigma_k$, which includes other parameters denoted by $\theta$. We seek to find a posterior mode or MAP (maxiumum a posteriori) estimate rather than a maximum likelihood estimate for the mixture parameters.

We continue to use BIC for model selection, but in a slightly modified form. We replace the first term on the right-hand side of (2), equal to twice the maximized log-likelihood, by twice the log-likelihood evaluated at the MAP or posterior mode.

## 3. Bayesian Regularization for Univariate Normal Mixture Models

For one-dimensional data, we use a normal prior on the mean (conditional on the variance):

$$\begin{aligned}\mu \mid \sigma^2 &\sim \mathcal{N}(\mu_{\mathcal{P}}, \sigma^2/\kappa_{\mathcal{P}}) \\ &\propto \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\kappa_{\mathcal{P}}}{2\sigma^2}\left(\mu - \mu_{\mathcal{P}}\right)^2\right\}\end{aligned} \tag{4}$$

and an inverse gamma prior on the variance:

$$\begin{aligned}\sigma^2 &\sim \text{inverseGamma}(\nu_{\mathcal{P}}/2, \varsigma_{\mathcal{P}}^2/2) \\ &\propto \left(\sigma^2\right)^{-\frac{\nu_{\mathcal{P}}+2}{2}} \exp\left\{-\frac{\varsigma_{\mathcal{P}}^2}{2\sigma^2}\right\}.\end{aligned} \tag{5}$$

This is called a *conjugate prior* for a univariate normal distribution because the posterior can also be expressed as the product of a normal distribution and an inverse gamma distribution. The hyperparameters $\mu_{\mathcal{P}}$, $\kappa_{\mathcal{P}}$, $\nu_{\mathcal{P}}$, and $\varsigma^2_{\mathcal{P}}$ are called the *mean*, *shrinkage*, *degrees of freedom* and *scale*, respectively, of the prior distribution. The values of the mean and variance at the posterior mode for a univariate normal under this prior are:

$$\hat{\mu} = \frac{n\bar{y} + \kappa_{\mathcal{P}}\mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n}$$

and

$$\hat{\sigma}^2 = \frac{\varsigma^2_{\mathcal{P}} + \frac{\kappa_{\mathcal{P}} n}{\kappa_{\mathcal{P}}+n}(\bar{y} - \mu_{\mathcal{P}})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2}{\nu_{\mathcal{P}} + n + 3}.$$

For derivations, see the appendix of Fraley and Raftery (2005).

For univariate mixtures, it is usually assumed that either all of the components share a common variance $\sigma^2$, or else that the variance is allowed to vary freely among the components. Constraining the variance to be equal is a form of regularization, and singularities are not typically observed when this is done. Singularities often arise, however, when the variance is unconstrained.

We use the normal inverse gamma conjugate prior (4) and (5), and take the prior distribution of the vector of component proportion $(\tau_1, \ldots, \tau_G)$ to be uniform on the simplex. Then the M-step estimators are given in Table 1. The prior hyperparameters $(\mu_{\mathcal{P}}, \kappa_{\mathcal{P}}, \nu_{\mathcal{P}}, \varsigma^2_{\mathcal{P}})$ are assumed to be the same for all components. For derivations, see the appendix of Fraley and Raftery (2005).

We make the following choices for the prior hyperparameters:

$\mu_{\mathcal{P}}$: the mean of the data.

$\kappa_{\mathcal{P}}$: .01

The posterior mean $\frac{n_k\bar{y}_k + \kappa_{\mathcal{P}}\mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n_k}$ can be viewed as adding $\kappa_{\mathcal{P}}$ observations with value $\mu_{\mathcal{P}}$ to each group in the data. The value we used was determined by experimentation; values close to and bigger than 1 caused large perturbations in the modeling in cases where there were no missing BIC values without the prior. The value .01 resulted in BIC curves that appeared to be smooth extensions of their counterparts without the prior.

$\nu_{\mathcal{P}}$: $d + 2 = 3$

The marginal prior distribution of $\mu$ is a Student's $t$ distribution centered at $\mu_{\mathcal{P}}$ with $\nu_{\mathcal{P}} - d + 1$ degrees of freedom. The mean of this distribution is $\mu_{\mathcal{P}}$ provided that $\nu_{\mathcal{P}} > d$, and it has a finite variance provided that $\nu_{\mathcal{P}} > d + 1$ (see, e.g. Schafer 1997). We chose the smallest integer value for the degrees of freedom that gives a finite variance.

$\varsigma_{\mathcal{P}}^2$: $\frac{\text{var(data)}}{G^2}$ (The empirical variance of the data divided by the square of the number of components.) The resulting prior mean of the precision corresponds to a standard deviation is one $G$th that of all of the data, where $G$ is the number of components. This is roughly equivalent to partitioning the range of the data into $G$ intervals of fairly equal size.

## 4. Bayesian Regularization for Multivariate Normal Mixtures

For multivariate data, we use a normal prior on the mean (conditional on the covariance matrix):

$$\begin{aligned} \mu \mid \Sigma &\sim \mathcal{N}(\mu_{\mathcal{P}}, \Sigma/\kappa_{\mathcal{P}}) \\ &\propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{\kappa_{\mathcal{P}}}{2}\text{trace}\left[(\mu-\mu_{\mathcal{P}})^T \Sigma^{-1}(\mu-\mu_{\mathcal{P}})\right]\right\}, \end{aligned} \tag{6}$$

and an inverse Wishart prior on the covariance matrix:

$$\begin{aligned} \Sigma &\sim \text{inverseWishart}(\nu_{\mathcal{P}}, \Lambda_{\mathcal{P}}) \\ &\propto |\Sigma|^{-\frac{\nu_{\mathcal{P}}+d+1}{2}} \exp\left\{-\frac{1}{2}\text{trace}\left[\Sigma^{-1}\Lambda_{\mathcal{P}}^{-1}\right]\right\}. \end{aligned} \tag{7}$$

As in the univariate case, the hyperparameters $\mu_{\mathcal{P}}$, $\kappa_{\mathcal{P}}$ and $\nu_{\mathcal{P}}$ are called the *mean*, *shrinkage* and *degrees of freedom* respectively, of the prior distribution. The hyperparameter $\Lambda_{\mathcal{P}}$, which is a matrix, is called the *scale* of the inverse Wishart prior. This is a *conjugate prior* for a multivariate normal distribution because the posterior can also be expressed as the product of a normal distribution and an inverse Wishart distribution. Under this prior, the posterior means of the mean vector and the covariance matrix are:

$$\hat{\mu} = \frac{n\bar{y} + \kappa_{\mathcal{P}}\mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n}$$

and

$$\hat{\Sigma} = \frac{\Lambda_{\mathcal{P}}^{-1} + \left(\frac{\kappa_{\mathcal{P}} n}{\kappa_{\mathcal{P}}+n}\right)(\bar{y}-\mu_{\mathcal{P}})(\bar{y}-\mu_{\mathcal{P}})^T + \sum_{j=1}^{n}(y_j-\bar{y})(y_j-\bar{y})^T}{\tilde{\nu}_{\mathcal{P}} + d + 2}.$$

The normal inverted Wishart prior and its conjugacy to the multivariate normal are discussed in e.g. Gelman, Carlin, Stern, and Rubin (1995) and Schafer (1997). The derivations leading to these results are given in the appendix of Fraley and Raftery (2005).

Table 1. M-step Estimators for the Mean and Variance of Univariate Mixture Models Under the Normal Inverse Gamma Conjugate Prior. The two rows for the variance correspond to the assumptions of equal or unequal variance across components. Here $z_{jk}$ is the conditional probability that observation $j$ belongs to the $k$th component, $n_k \equiv \sum_{j=1}^{n} z_{jk}$ and $\bar{y}_k \equiv \sum_{j=1}^{n} z_{jk} y_j / n_k$.

| Parameter | Without Prior | With Prior |
|---|---|---|
| $\hat{\mu}_k$ | $\bar{y}_k$ | $\dfrac{n_k \bar{y}_k + \kappa_{\mathcal{P}} \mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n_k}$ |
| $\hat{\sigma}^2$ | $\dfrac{\sum_{k=1}^{G} \sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)^2}{n}$ | $\dfrac{\varsigma_{\mathcal{P}}^2 + \sum_{k=1}^{G} \left[ \frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)} (\bar{y}_k - \mu_{\mathcal{P}})^2 + \sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)^2 \right]}{\nu_{\mathcal{P}} + n + G + 2}$ |
| $\hat{\sigma}_k^2$ | $\dfrac{\sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)^2}{n_k}$ | $\dfrac{\varsigma_{\mathcal{P}}^2 + \frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)} (\bar{y}_k - \mu_{\mathcal{P}})^2 + \sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)^2}{\nu_{\mathcal{P}} + n_k + 3}$ |

### 4.1 Multivariate Mixture Models

For multivariate normal mixtures, the contours of the component densities are ellipsoidal, and the component covariance matrices can be constrained so that the contours are spherical (proportional to the identity matrix), or axis-aligned (diagonal). It is usually assumed either that all the components share a common covariance matrix, or that the covariance matrix can vary freely between the components. As in the univariate case, constraining the variance to be equal is a form of regularization, and failures in estimation are not typically observed when this is done except when large numbers of mixture components are involved, causing the mixing proportions of some components to shrink to zero. Constraining the covariance matrix to be diagonal or spherical is also a form of regularization for multivariate data. Although singularities may arise when the covariance is restricted to one of these forms but otherwise allowed to vary among components, they occur less frequently than when the covariance matrix is unconstrained.

Under the conjugate prior with the inverse gamma prior (5) on the variance components for the diagonal and spherical models, the inverse Wishart prior (7) on the covariance for the ellipsoidal models, and the normal prior (6) on the mean, the M-step estimators are given in Table 2. The prior hyperparameters $(\kappa_{\mathcal{P}}, \nu_{\mathcal{P}}, \Lambda_{\mathcal{P}}, \varsigma^2_{\mathcal{P}})$ are assumed to be the same for all components. For derivations, see the appendix of Fraley and Raftery (2005).

### 4.2 Multivariate Prior Hyperparameters

We make the following choices for the prior hyperparameters for mulivariate mixtures.

$\mu_{\mathcal{P}}$: the mean of the data.

$\kappa_{\mathcal{P}}$: .01: the same reasoning applies as in the univariate case.

$\nu_{\mathcal{P}}$: $d + 2$
As in the univariate case, the marginal prior distribution of $\mu$ is multivariate $t$ centered at $\mu_{\mathcal{P}}$ with $\nu_{\mathcal{P}} - d + 1$ degrees of freedom. The mean of this distribution is $\mu_{\mathcal{P}}$ provided that $\nu_{\mathcal{P}} > d$, and it has a finite covariance matrix provided $\nu_{\mathcal{P}} > d+1$ (see, e. g. Schafer 1997). We chose the smallest integer value for the degrees of freedom that gives a finite covariance matrix.

$\varsigma^2_{\mathcal{P}}$: $\frac{\text{sum(diag(var(data)))}/d}{G^{2/d}}$ (For spherical and diagonal models.) The average of the diagonal elements of the empirical covariance matrix of the data divided by the square of the number of components to the $1/d$ power.

Table 2. M-step Estimators for the Mean and Variance of Multvariate Mixture Models under the Normal Inverse Gamma and Normal Inverse Wishart Conjugate Priors. The rows for the variance correspond to the assumptions of equal or unequal spherical variance across components, and equal or unequal ellipsoidal variance across components. Here $z_{jk}$ is the conditional probability that observation $j$ belongs to the $k$th component, $n_k \equiv \sum_{j=1}^{n} z_{jk}$, $\bar{y}_k \equiv \sum_{j=1}^{n} z_{jk} y_j / n_k$, and $W_k \equiv \sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)(y_j - \bar{y}_k)^T$.

| Parameter | Without Prior | With Prior |
|---|---|---|
| $\hat{\mu}_k$ | $\bar{y}_k$ | $\frac{n_k \bar{y}_k + \kappa_{\mathcal{P}} \mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n_k}$ |
| $\hat{\sigma}^2$ | $\frac{\sum_{k=1}^{G} \text{trace}\,(W_k)}{nd}$ | $\frac{\varsigma_{\mathcal{P}}^2 + \sum_{k=1}^{G} \text{trace}\left[\frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k\right]}{\nu_{\mathcal{P}} + (n+G)d + 2}$ |
| $\hat{\sigma}_k^2$ | $\frac{\text{trace}\,(W_k)}{n_k d}$ | $\frac{\varsigma_{\mathcal{P}}^2 + \text{trace}\left[\frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k\right]}{\nu_{\mathcal{P}} + n_k d + d + 2}$ |
| $\text{diag}(\hat{\delta}_i^2)$ | $\frac{\text{diag}\left(\sum_{k=1}^{G} W_k\right)}{n}$ | $\frac{\text{diag}\left(\varsigma_{\mathcal{P}}^2 I + \sum_{k=1}^{G}\left[\frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k\right]\right)}{\nu_{\mathcal{P}} + n + 2}$ |
| $\text{diag}(\hat{\delta}_{ik}^2)$ | $\frac{\text{diag}\,(W_k)}{n_k}$ | $\frac{\text{diag}\left(\varsigma_{\mathcal{P}}^2 I + \left[\frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k\right]\right)}{\nu_{\mathcal{P}} + n_k + 2}$ |
| $\hat{\Sigma}$ | $\frac{\sum_{k=1}^{G} W_k}{n}$ | $\frac{\Lambda_{\mathcal{P}} + \sum_{k=1}^{G}\left[\frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k\right]}{\nu_{\mathcal{P}} + n + G + d + 1}$ |
| $\hat{\Sigma}_k$ | $\frac{W_k}{n_k}$ | $\frac{\Lambda_{\mathcal{P}} + \frac{\kappa_{\mathcal{P}} n_k}{(\kappa_{\mathcal{P}} + n_k)}(\bar{y}_k - \mu_{\mathcal{P}})(\bar{y}_k - \mu_{\mathcal{P}})^T + W_k}{\nu_{\mathcal{P}} + n_k d + d + 2}$ |

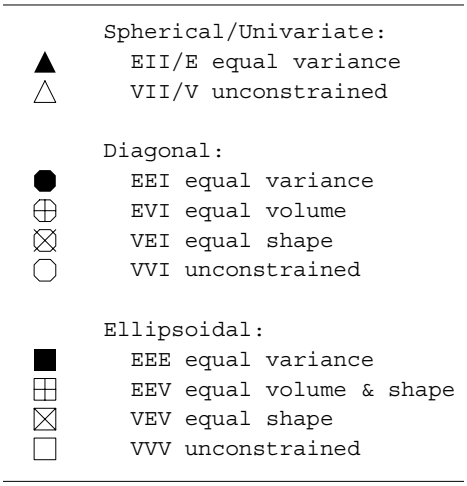


Figure 1. Legend for BIC Plots. Different symbols correspond to different model parameterizations. The three letter codes are those used to designate equal (E) or varying (V) shape, volume, and orientation, respectively, in the MCLUST software (Fraley and Raftery 1999, 2003, 2006). The letter I designates a spherical shape or an axis-aligned orientation.

$\Lambda_{\mathcal{P}}$: $\frac{\text{var(data)}}{G^{2/d}}$ (For ellipsoidal models.) The empirical covariance matrix of the data divided by the square of the number of components to the $1/d$ power.

The volume of the ellipsoid defined by $\varsigma^2_{\mathcal{P}}$ or $\Lambda_{\mathcal{P}}$ is one $G^2$th of the volume of the ellipsoid defined by the empirical covariance matrix of all of the data, where $G$ is the number of components.

## 5. Examples

Figure 1 displays the symbols used in the BIC plots throughout this section along with their associated model parameterization.

### 5.1 Univariate Examples

Figure 2 shows histograms and model-based clustering results for the three univariate datasets analyzed in Richardson and Green (1997). The data are described in Table 3. The equal-variance model is started with the outcome of model-based hierarchical clustering for that model, and the unequal variance model is started with the result of the equal variance model[1]. Note that without

1. This initialization differs from the default in the MCLUST software for the unequal variance case.

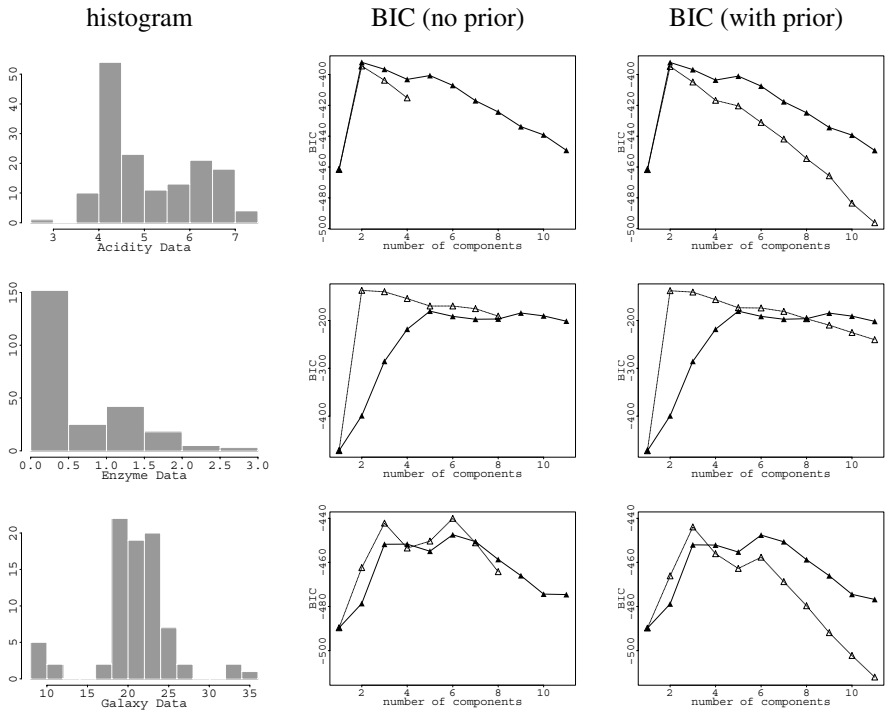


Figure 2. BIC for the Univariate Acidity, Enzyme and Galaxy Datasets. The histogram for the dataset is given in the left column, while plots of the number of components versus BIC values are shown for the model-based clustering without (center) and with (right). The equal variance model is indicated by filled symbols, while the model in which the variance is allowed to vary across components is indicated by open symbols.

the prior, there are no results available for the unconstrained variance model when the number of components is sufficiently large. The reason is that parameters could not be estimated due to singularities.

For the acidity data, the standard BIC values based on the MLE are not available for the unequal variance model with five or more mixture components. In the five-component case, the EM algorithm hits a path around the 250th iteration along which the variance for the first component tends rapidly to zero and the likelihood diverges to infinity (see Table 4). With the prior, BIC values are available for all models and numbers of groups. The results are similar with and without the prior in cases where both are available, and the overall conclusion is unchanged.

For the enzyme data, including the prior allows us to assess solutions with more than eight components, but does not otherwise change the analysis.

Table 3. Source and Description for the Three Univariate Data Sets Used in the Examples.

| dataset | # observations | reference |
|---|---|---|
| Acidity | 155 | Crawford et al. 1992 |
| Enzyme | 245 | Bechtel et al. 1993 |
| Galaxy | 82 | Roeder 1990 |

Table 4. Values of $\sigma_k^2$ from the M-step of EM for the Acidity Data Under the Five-Component, Unconstrained Variance Model Without the Prior. The variance for one of the components falls below the threshold for failure due to singularity.

| iteration | $\sigma_k^2,\ k = 1, \ldots, 5$ | | | | |
|---|---|---|---|---|---|
| 1 | 0.046731 | 0.018704 | 0.071418 | 0.058296 | 0.024956 |
| 2 | 0.056739 | 0.025452 | 0.085215 | 0.070723 | 0.030543 |
| 3 | 0.065152 | 0.031345 | 0.092778 | 0.077127 | 0.033145 |
| 4 | 0.072763 | 0.036744 | 0.097979 | 0.080919 | 0.034621 |
| 5 | 0.079982 | 0.041453 | 0.102011 | 0.08329 | 0.035558 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 246 | 0.22418 | 0.049378 | 0.182963 | 0.063843 | 0.044054 |
| 247 | 0.171405 | 0.049469 | 0.183083 | 0.063836 | 0.044056 |
| 248 | 0.108567 | 0.049606 | 0.183261 | 0.063829 | 0.044057 |
| 249 | 0.038607 | 0.049819 | 0.183493 | 0.063823 | 0.044058 |
| 250 | 0.000307 | 0.050004 | 0.183625 | 0.063815 | 0.04406 |

For the galaxy data, BIC without the prior chooses six components with unequal variances by a small margin over the three-components model, while with the prior it chooses three components fairly clearly. This dataset has been extensively analyzed in the literature, including a number of studies using mixtures of normals.

Roeder and Wasserman (1997) chose three components using an approach similar to ours based on MLE via EM and BIC; their choice of three rather than six components seems to be due to their using a different (and lower) local mode of the likelihood for the six-component model. Richardson and Green (1997) did a Bayesian analysis using reversible jump MCMC, and reported a posterior distribution with both mode and median at six components, although they indicated later that convergence may not have been achieved (Richardson and Green 1997, p. 789).

Figure 3 shows the classifications and the estimated densities for these two competing models. They agree that the seven smallest observations form a group, as do the largest three. They disagree about whether the middle 72

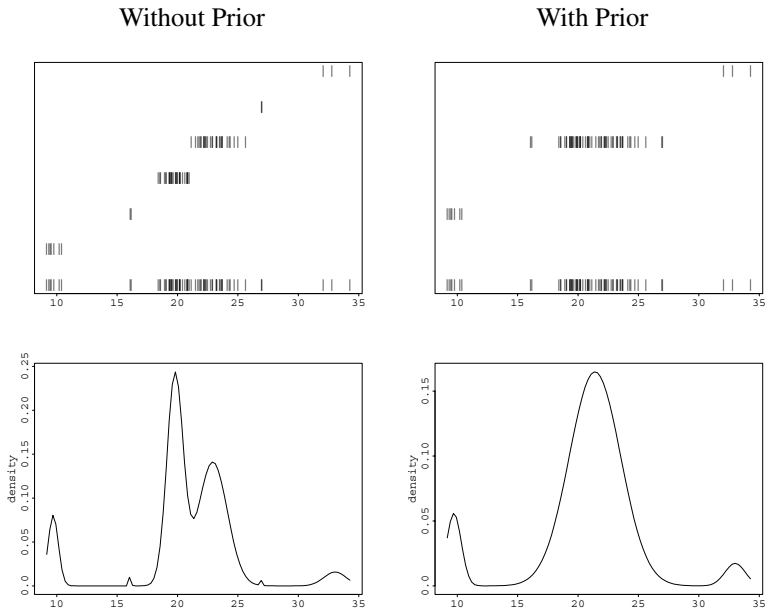


Figure 3. Classifications (top) and Densities Corresponding to the Mixture Models Fitted to the Univariate Galaxy Data Without and With the Prior. In the classification plots, all of the data is shown at the bottom, while the different classes are separated on the lines above.

observations can be adequately described by one normal density, or whether four normal components are needed. The figure suggests that several of the groups in the six-group solution may be spurious.

One can also shed some light on this issue by assessing the fit of a normal distribution to the middle 72 observations. Figure 4(a) shows the cumulative distribution function (CDF) for the full galaxy dataset, together with the CDFs for 99 datasets of the same size simulated from the fitted normal distribution. Thirteen of the 82 observations lie outside the pointwise band, in line with the well-accepted conclusion that one normal does not fit the entire dataset. Figure 4(b) shows the same thing for the middle 72 observations; the entire empirical CDF lies inside the band, suggesting that one normal distribution is adequate for this group. Kolmogorov-Smirnov test statistics for testing a normal distribution tell a similar story: the a $P$ value for the full dataset is 0.005, while that for the middle 72 observations is 0.254. Note that these results are based on estimated parameters, which is anti-conservative. If account were taken of parameter uncertainty, the tests would be less likely to reject the null hypothesis of normality, and so the conclusion for the middle 72 observations would be unchanged. Overall, this analysis provides support for the three-group solution.

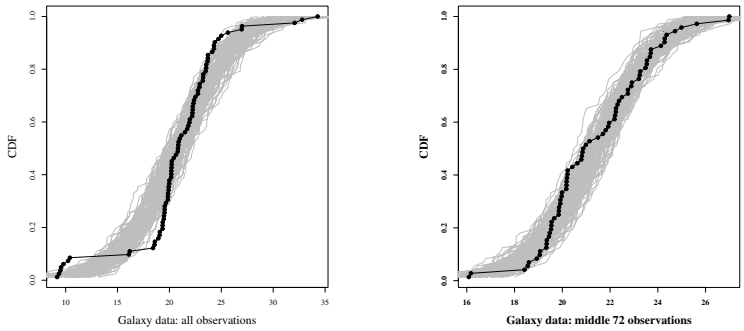


Figure 4. The Empirical CDF (black) Together with Empirical CDFs for 99 Datasets of the Same Size Simulated from the Fitted Normal Distribution (gray) for the Full Galaxy Dataset (left), and the Middle 72 Observations (right).

### 5.2 Butterfly Example

In this section we consider observations from the `butterfly` dataset (Celeux and Robert, 1993), consisting of the measurements of the widths of the upper and lower wings for 23 butterflies, shown in Figure 5. The original goal was to ascertain how many species were present in the sample and classify the butterflies into them.

Figure 6 shows the BIC for equal variance and unconstrained variance models, assuming spherical, diagonal, and elliptical shapes (the models for which priors were discussed in Section 4). For all models, EM was started using the results from model-based hierarchical clustering on the unconstrained ellipsoidal variance model. The model and classification chosen according to the BIC are the same regardless of whether or not the prior is imposed, but failures due to singularity for the unconstrained models are eliminated with the prior.

The four-component unconstrained model without the prior fails at the outset. The hierarchical clustering result based on the unconstrained model used for initialization assigns a single observation to one of the groups in this case. The Bayesian regularization allows the identification of a group with a single member while allowing the covariance matrix to vary between clusters, which is not possible without the prior.

### 5.3 Trees Example

In this section we analyze the `trees` dataset (Ryan, Joiner, and Ryan 1976) included in the `R` language (`www.r-project.org`). A pairs plot of the data is shown in Figure 7. There are 31 observations of 3 variables.

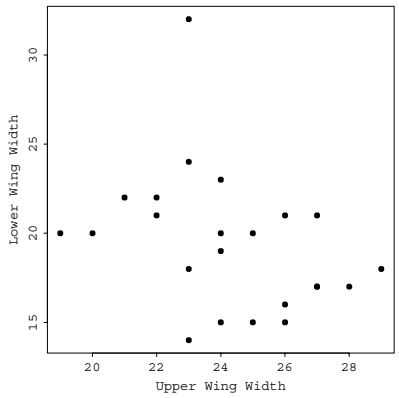


Figure 5. Wing widths from the `butterfly` dataset. There are 23 observations.

Figure 8 shows the BIC for the trees data for the equal variance and unconstrained variance models, assuming spherical, diagonal, and ellipsoidal shapes. For the equal variance models, EM was started using the results from model-based hierarchical clustering assuming no constraints on the variance. For the unconstrained variance model, EM was started using the results from the equal variance model.

Figure 9 shows the 3 and 5 group classifications where the BIC has peaks without the prior, and the 2 group classification corresponding to the maximum BIC with the prior. The six-component unconstrained model fails to converge without the prior in this case because one of the covariances becomes singular as the EM iterations progress, as shown in Figure 10.

### 5.4 Other Parameterizations and Priors

Banfield and Raftery (1993) expressed the covariance matrix for the $k$-th component of a multivariate normal mixture model in the form

$$\boldsymbol{\Sigma}_k = \lambda_k D_k A_k D_k^T, \tag{8}$$

where $D_k$ is the matrix of eigenvectors determining the orientation, $A_k$ is a diagonal matrix proportional to the eigenvalues determining the shape, and $\lambda_k$ is a scalar determining the volume of the cluster. They used this formulation to define a class of hierarchical clustering methods based on cross-cluster geometry, in which mixture components may share a common shape, volume, and/or orientation. This approach generalizes a number of existing clustering methods. For example if the clusters are restricted to be spherical and identical in volume, the clustering criterion is the same as that which underlies Ward's method

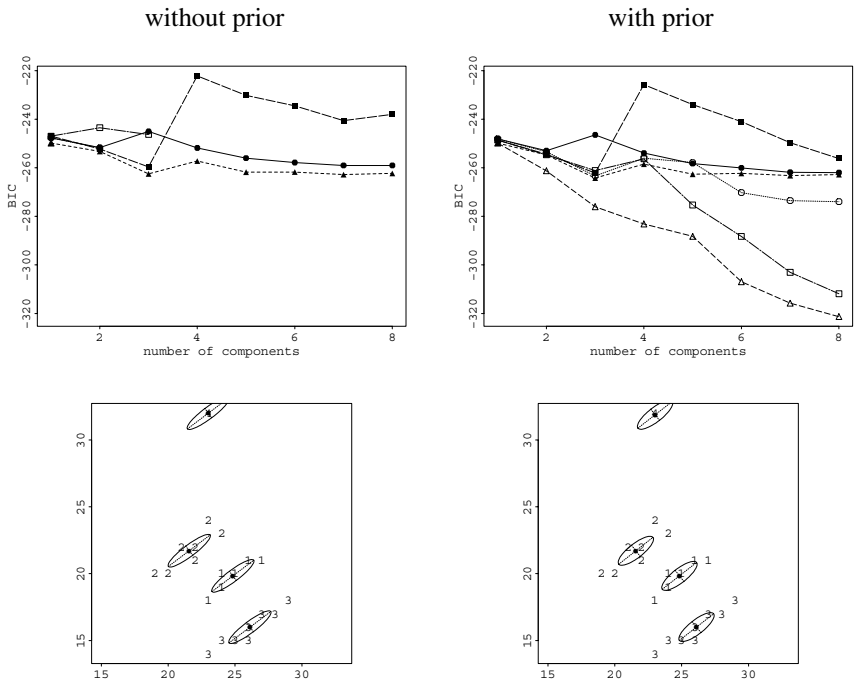


Figure 6. The top row gives the BIC for the six models for variables 3 and 4 the butterfly dataset. Refer to Figure 1 for the correspondence between symbols and models. The bottom row displays a projection of the data showing the classification corresponding to the maximum BIC. Failures due to singularities no longer occur when the prior is included in the model, although the BIC selects a model mapping to the same four group classfication regardless of whether or not the prior is imposed.

(Ward 1963) and $k$-means clustering (MacQueen 1967). Banfield and Raftery (1993) developed this class of models in the context of hierarchical clustering estimated using the classification likelihood, but the same parameterizations can also be used with the mixture likelihood. A detailed description of 14 different models that are possible under this scheme can be found in Celeux and Govaert (1995).

Many of these models can be estimated by the `MCLUST` software (Fraley and Raftery 1999, 2003, 2006), where they are designated by a three-letter symbol indicating volume, shape and orientation, respectively. The letter `E` indicates cross-cluster equality, while `V` denotes freedom to vary across clusters, and the letter `I` designates a spherical shape or an axis-aligned orientation.

It is possible to formulate priors that eliminate singularity in a manner similar to that used to derive the priors in Table 2 for most cases of the parame-

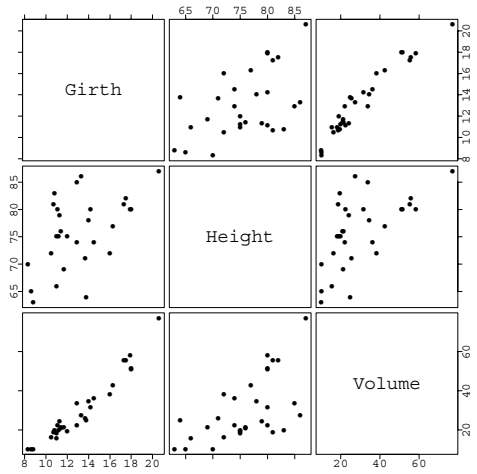


Figure 7. Pairs plot of the `trees` dataset. There are 31 observations.

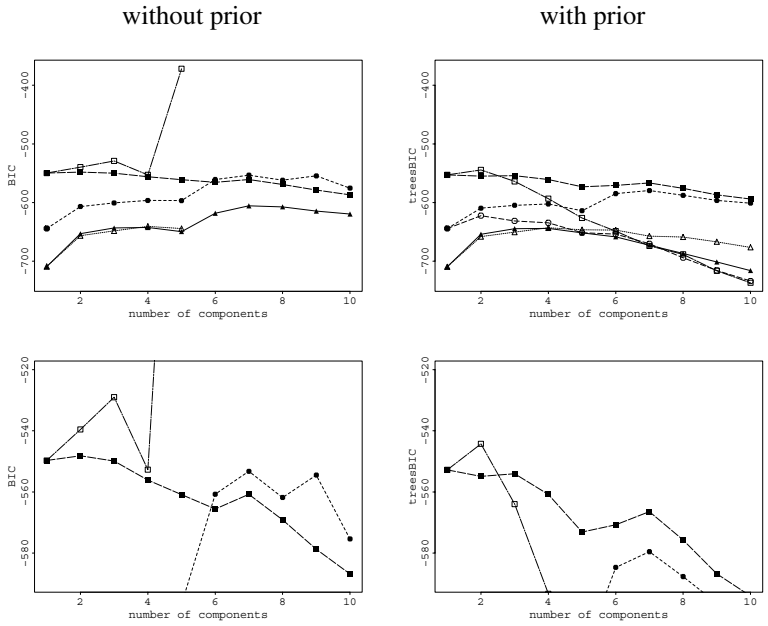


Figure 8. The top row gives the BIC for the six models for the trees dataset, while the bottom row shows details of the BIC near the maximum. Refer to Figure 1 for the correspondence between symbols and models. Two groups are favored over five when the prior is imposed.

Table 5. Parameterizations of the Covariance Matrix via Eigenvalue Decomposition, with the associated prior. $\lambda$ is a scalar controling volume, $A$ a diagonal matrix controling shape, and $D$ an orthogonal matrix corresponding to the eigenvectors of the covariance matrix controling orientation. The subscript $k$ indicates variation across components in the mixture model.

| MCLUST symbol | parameterization | prior | applied to |
|---|---|---|---|
| EII | $\lambda I$ | inverse gamma | $\lambda$ |
| VII | $\lambda_k I$ | inverse gamma | $\lambda_k$ |
| EEI | $\lambda A$ | inverse gamma | each diagonal element of $\lambda A$ |
| VEI | $\lambda_k A$ | | |
| EVI | $\lambda A_k$ | | |
| VVI | $\lambda_k A_k$ | inverse gamma | each diagonal element of $\lambda_k A_k$ |
| EEE | $\lambda DAD^T$ | inverse Wishart | $\Sigma = \lambda DAD^T$ |
| VEE | $\lambda_k DAD^T$ | inverse gamma | $\lambda_k$ |
| | | inverse Wishart | $\tilde{\Sigma} = DAD^T$ |
| EVE | $\lambda DA_k D^T$ | | |
| VVE | $\lambda_k DA_k D^T$ | inverse gamma | each diagonal element of $\lambda_k A_k$ |
| EEV | $\lambda D_k A D_k^T$ | inverse gamma | each diagonal element of $\lambda A$ |
| VEV | $\lambda_k D_k A D_k^T$ | | |
| EVV | $\lambda D_k A_k D_k^T$ | inverse gamma | $\lambda$ |
| | | inverse Wishart | $\tilde{\Sigma}_k = D_k A_k D_k^T$ |
| VVV | $\lambda_k D_k A_k D_k^T$ | inverse Wishart | $\Sigma_k = \lambda_k D_k A_k D_k^T$ |

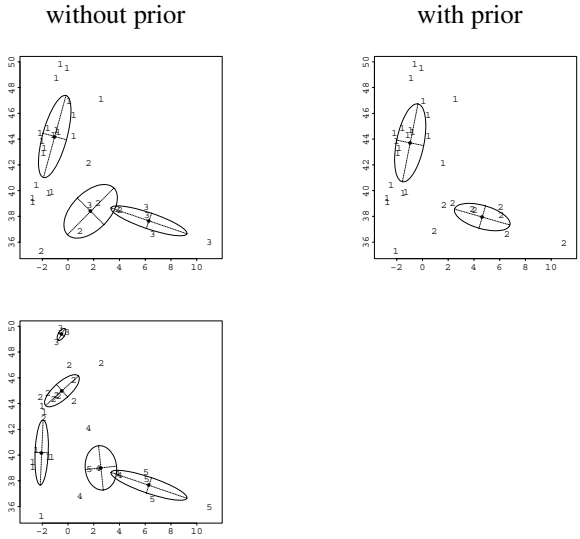


Figure 9. A projection of the trees data showing the 3 and 5 group classifications, where the BIC has peaks without the prior, and the 2 group classification where the BIC peaks with the prior.

terization (8). These priors are summarized in Table 5. The cases for which no prior is given are those for which neither the volume and shape nor the shape and orientation can be treated as a unit across components. For other models that are not covered in Table 2 of Section 4, the M-step computations would need to be derived in the manner described in Celeux and Govaert (1995) once the prior terms have been added to the complete data loglikelihood.

We have obtained good results with an alternative heuristic that can be applied to all parameterizations based on the decomposition (8): noting that the computations involved in the M-step without the prior involve the weighted sums of squares and products matrix

$$W_k \equiv \sum_{j=1}^{n} z_{jk}(y_j - \bar{y}_k)(y_j - \bar{y}_k)^T$$

(see Celeux and Govaert 1995), we replace $W_k$ in the M-step formulas with their analogs from Table 2.

Figures 11 and 12 show results for the trees dataset for ten models, including four (`VEI`, `EVI`, `EEV`, and `VEV`) for which we use the heuristic just described. In Figure 11, all models are initialized with the result from model-based hierarchical clustering using the unconstrained model, so there are some differences with Figure 8 in those instances where the models are the same. Without a prior, the model fragments the data into several small, highly ellipsoidal components (see Figure 12). There is one component to which only one observation would be assigned according to its highest conditional probability. With the prior, a model with fewer components is selected.

### 5.5 Vanishing Components

Of note in Figure 11 is that the estimation for the trees data fails for some models, even when the prior is imposed. For this example, all failures without the prior were due to singularity in the estimated covariance matrices, while all failures with the prior were due to the mixing proportion of one or more components shrinking to zero. In this latter case, it is still possible to estimate the BIC, although no associated parameter or loglikelihood values are available. This is accomplished by adding the appropriate terms penalizing the number of parameters (formula (2) in Section 2.) to the loglikehood available for the largest number of components with the same parameterization scheme. Figure 13 plots the BIC for mixture estimation with the prior for the trees data, with the values for models that include vanishing components shown.

## 6. Discussion

We have proposed a Bayesian regularization method for avoiding the singularities and degeneracies that can arise in estimation for model-based clus-

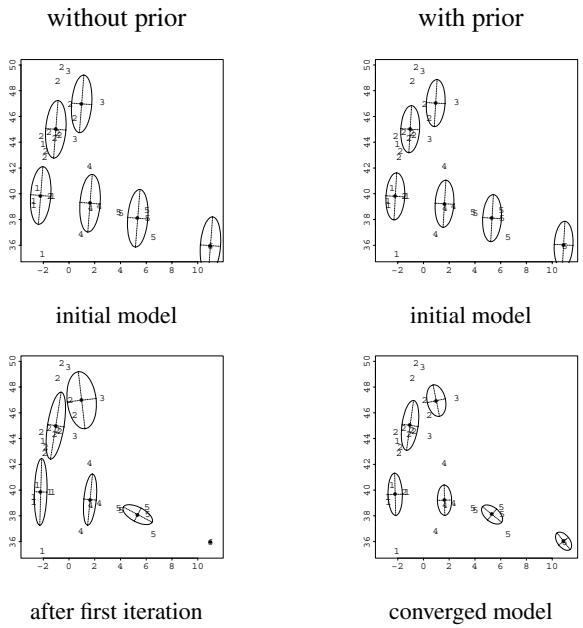


Figure 10. For a projection of the trees data, the top row shows the initial equal-variance models fitting a six-component unconstrained model without and with the prior. The bottom left figure shows the covariance for component 6 of the six-component model collapsing to near singularity after the first iteration without the prior. At the bottom right is the six-component unconstrained model converged fit with the prior.

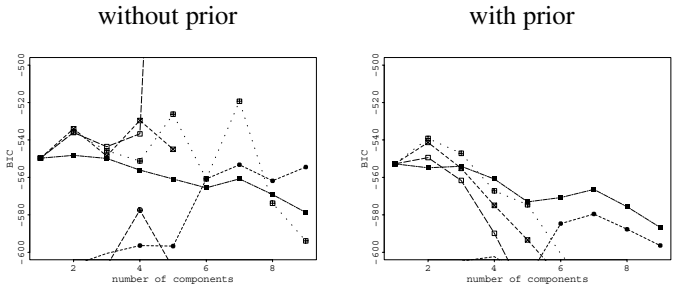


Figure 11. BIC for Ten Models for the Trees Dataset. Refer to Figure 1 for the correspondence between symbols and models. Without the prior, the BIC is maximized with the same five-component unconstrained model obtained in Section 5., with the steep increase and cutoff in BIC suggesting near singularity. With the prior it is maximized at a two-component model in which the volume and shape of the covariances of the components in the model are the same, but their orientation may vary. Compare with Figure 8, which shows the BIC for the restricted set of six models for which an explicit prior is available.

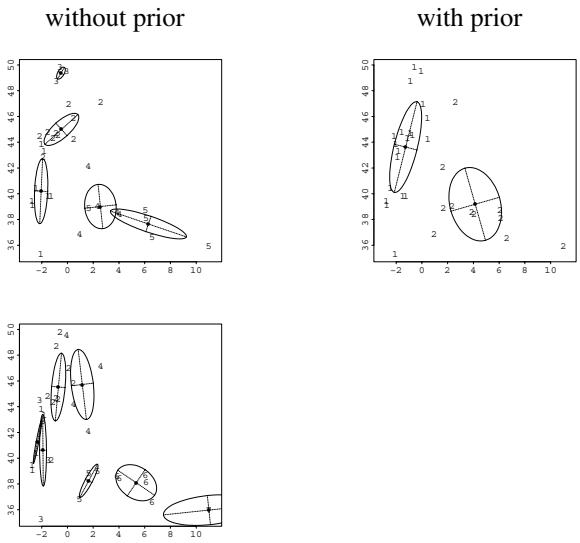


Figure 12. LEFT: A projection of the trees data showing the model and classification for the 5-component unconstrained model (highest BIC), and the 7-component constant volume and shape model (second highest BIC) without the prior. RIGHT: A projection of the trees data showing the model and classification corresponding to the two-component constant volume and shape model with the prior (highest BIC). Compare with Figure 9, which shows the choice from a restricted set of six models using an explicit prior.

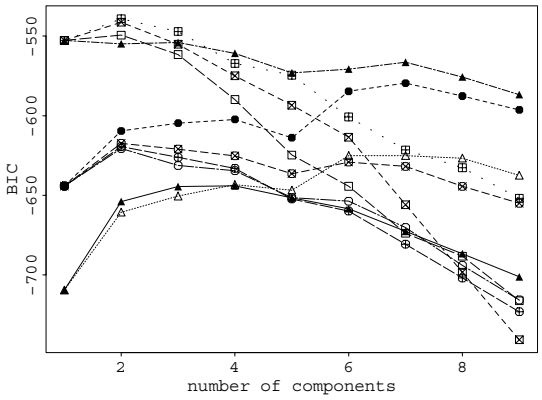


Figure 13. BIC plot for mixture estimation with the prior for the trees data, with values for models that include vanishing components shown. Refer to Figure 1 for the correspondence between symbols and models.

tering using the EM algorithm. The method involves a dispersed but proper conjugate prior, and uses the EM algorithm to find the posterior mode, or MAP estimator. For model selection it uses a version of BIC that is slightly modified by replacing the maximized likelihood by the likelihood evaluated at the MAP. In application to a range of datasets, the method did eliminate singularities and degeneracies observed in maximum likelihood methods, while having little effect on stable results. When the number of observations is small, the choice of prior can influence the modeling outcome even when no singularities are observed.

In model-based clustering, parameterization through eigenvalue decomposition with explicit eigenvector normalization (8) allows cross-component constraints on the geometry (shape, volume and orientation) of normal mixture components, and constitutes a form of regularization that has been exploited both in clustering (Banfield and Raftery 1993; Celeux and Govaert 1995; Fraley and Raftery 1998, 2002) and in discriminant analysis (Flury 1988; Bensmail and Celeux 1996; Fraley and Raftery 2002). With this scheme models can, for example, be constrained to be either spherical or diagonal and have either fixed or varying covariance across components; there are a number of intermediate possibilities as well. Failures due to singularity are less likely to occur with models having less variation across components, although with a corresponding loss of modeling flexibility.

In this paper we have limited our treatment to cases in which there are more observations than variables. However, there are many instances, such as gene expression data (e.g. Quackenbush 2001), where the reverse is true. While the methods given here will not fail due to singular covariance estimates when the prior on the covariance is nonsingular (for example, $\frac{\text{diag(var(data))}}{G^{2/d}}$ could be used if $n \leq d$) more sophisticated techniques are usually needed to obtain useful clustering. One approach is to adopt a mixture of factor analyzers model (Ghahramani and Hinton 1997; McLachlan, Peel, and Bean 2003), in which the covariance matrix has the form

$$\Sigma_k = D_k + B_k B_k^T,$$

where $D_k$ is diagonal and $B_k$ is a $d \times m$ matrix of factor loadings, with $m \ll d$. Some regularization is still required to prevent the elements of $D_k$ from vanishing during estimation; this could be done, for example, by imposing a common value of $D_k$ across components.

Another approach to mixture modeling of high-dimensional data is variable selection, which arises as an important consideration in many types of modeling, even when there are more observations than attributes or variables. Usually variable selection is accomplished by a separate (often heuristic) procedure prior to analysis. Raftery and Dean (2006) developed a model-based

clustering that incorporates variable selection as an integral part of the procedure. It would be possible to include a prior distribution in their method as we have done here.

The presence of noise in the data can have a significant effect on density estimation and clustering. Spurious singularities can be introduced due to noise when fitting Gaussian mixtures. In mixture modeling, one approach is to add a term with a first-order Poisson distribution to the mixture model to account for noise (Banfield and Raftery 1993; Dasgupta and Raftery 1998); our Bayesian regularization approach could also be applied to these models. An alternative is to model with mixtures of $t$ distributions, which have broader tails than normals (McLachlan and Peel 1998; McLachlan, Peel, Basford, and Adams 1999).

Mkhadri, Celeux, and Nasroallah (1997) reviewed regularization techniques in discriminant analysis, including parameterization of normal mixture models by eigenvalue decomposition (8) and Bayesian estimation using conjugate priors analogous to those we have used here in the case of unconstrained covariance. These methods were extended to cases with mixed discrete and continuous variables by Merbouha and Mkhadri (2004).

Several studies have used the EM algorithm to estimate the posterior mode in a Bayesian approach for mixture models. Roberts, Husmeier, Rezek, and Penny (1998) used a Dirichlet prior on the mixing proportions and a noninformative prior on the elements of the means and covariances, while Figueiredo and Jain (2002) used noninformative priors on all of the parameters to be estimated. Brand (1999) proposed an entropic prior on the mixing proportions, and applied his method to hidden Markov models as well as Gaussian mixtures. These methods work by starting with more components than necessary, and then pruning those for which mixing proportions are considered negligible.

Bayesian estimation for mixture models can be done via Markov chain Monte Carlo simulation, using priors on the means and covariances similar to the ones we have used here (e.g. Lavine and West 1992, Diebolt and Robert 1994, Crawford 1994, Bensmail, Celeux, Raftery, and Robert 1997, Richardson and Green 1997, Dellaportas 1998, Bensmail and Meulman 2003, Zhang, Chan, Wu, and Chen 2004, Bensmail, Golek, Moody, Semmes, and Haoudi 2005). We have shown that Bayesian estimation using the posterior mode from EM is straightforward for many models, and that these results can be used for approximate Bayesian estimation in other models (Section 5). Thus, for many applications it is not clear that the use of MCMC for mixture estimation and model-based clustering is needed given its computational demands.

## References

BANFIELD, J.D. and RAFTERY, A.E. (1993), "Model-based Gaussian and Non-Gaussian Clustering", *Biometrics 49*, 803–821.

BECHTEL, Y.C., BONAÏTI-PELLIÉ, C., POISSON, N., MAGNETTE, J., and BECHTEL, P.R. (1993), "A Population and Family Study of $n$-acetyltransferase Using Caffeine Urinary Metabolites", *Clinical Pharmacology and Therapeutics 54*, 134–141.

BENSMAIL, H., and CELEUX, G. (1996), "Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition", *Journal of the American Statistical Association 91*, 1743–1748.

BENSMAIL, H., CELEUX, G., RAFTERY, A.E., and ROBERT, C.P. (1997), "Inference in Model-based Cluster Analysis", *Statistics and Computing 7*, 1–10.

BENSMAIL, H., GOLEK, J., MOODY, M.M., SEMMES, J.O., and HAOUDI, A. (2005), "A Novel Approach to Clustering Proteomics Data Using Bayesian Fast Fourier Transform", *Bioinformatics 21*, 2210–2224.

BENSMAIL, H., and MEULMAN, J.J. (2003), "Model-based Clustering with Noise: Bayesian Inference and Estimation", *Journal of Classification 20*, 49–76.

BRAND, M. (1999), "Structure Discovery in Conditional Probability Models via an Entropic Prior and Parameter Extinction", *Neural Computation 11*, 1155–1182.

CAMPBELL, J.G., FRALEY, C., MURTAGH, F., and RAFTERY, A.E. (1997), "Linear Flaw Detection in Woven Textiles Using Model-based Clustering", *Pattern Recognition Letters 18*, 1539–1548.

CAMPBELL, J.G.,FRALEY, C., STANFORD, F. MURTAGH, and A. E. RAFTERY (1999), "Model-Based Methods for Real-Time Textile Fault Detection", *International Journal of Imaging Systems and Technology 10*, 339–346.

CELEUX, G., and GOVAERT, G. (1995), "Gaussian Parsimonious Clustering Models", *Pattern Recognition 28*, 781–793.

CELEUX, G., and ROBERT, C.P. (1993), "Une Histoire de Discrétisation (avec Commentaires)", *La Revue de Modulad 11*, 7–44.

CRAWFORD, S.L. (1994), "An Application of the Laplace Method to Finite Mixture Distributions", *Journal of the American Statistical Association 89*, 259–267.

CRAWFORD, S.L., DEGROOT, M.H., KADANE, J.B., and SMALL, M.J. (1992), "Modeling Lake Chemistry Distributions: Approximate Bayesian Methods for Estimating a Finite Mixture Model", *Technometrics 34*, 441–453.

DASGUPTA, A., and RAFTERY, A.E. (1998), "Detecting Features in Spatial Point Processes with Clutter via Model-based Clustering", *Journal of the American Statistical Association 93*, 294–302.

DELLAPORTAS, P. (1998), "Bayesian Classification of Neolithic Tools", *Applied Statistics 47*, 279–297.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm (with Discussion)", *Journal of the Royal Statistical Society, Series B 39*, 1–38.

DIEBOLT, J., and ROBERT, C. (1994), "Estimation of Finite Mixtures Through Bayesian Sampling", *Journal of the Royal Statistical Society, Series B 56*, 363–375.

FIGUEIREDO, M.A.T., and JAIN, A.K. (2002), "Unsupervised Learning of Finite Mixture Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 381–396.

FLURY, B.W. (1988), *Common Principal Components and Related Multivariate Models*, New York: Wiley.

FRALEY, C. (1998), "Algorithms for Model-based Gaussian Hierarchical Clustering", *SIAM Journal on Scientific Computing 20*, 270–281.

FRALEY, C., and RAFTERY, A.E. (1998), "How Many Clusters? Which Clustering Method? - Answers via Model-based Cluster Analysis. *The Computer Journal 41*, 578–588.

FRALEY, C., and RAFTERY, A.E. (1999), "MCLUST: Software for Model-based Cluster Analysis", *Journal of Classification 16*, 297–306.

FRALEY, C., and RAFTERY, A.E. (2002), "Model-based Clustering, Discriminant Analysis and Density Estimation", *Journal of the American Statistical Association 97*, 611–631.

FRALEY, C., and RAFTERY, A.E. (2003), "Enhanced Software for Model-based Clustering, Density Estimation, and Discriminant Analysis: MCLUST", *Journal of Classification 20*, 263–286.

FRALEY, C., and RAFTERY, A.E. (2005, August), "Bayesian Regularization for Normal Mixture Estimation and Model-based Clustering", Technical Report 486, University of Washington, Department of Statistics.

FRALEY, C., and RAFTERY, A.E. (2006, September), "`MCLUST` Version 3 for `R`: Normal Mixture Modeling and Model-based Clustering", Technical Report 504, University of Washington, Department of Statistics.

FRALEY, C., RAFTERY, A.E., and WEHRENS, R. (2005), "Incremental Model-based Clustering for Large Datasets with Small Clusters", *Journal of Computational and Graphical Statistics 14*, 1–18.

GELMAN, A., CARLIN, J.B., STERN, H.S., and RUBIN, D.B. (1995), "*Bayesian Data Analysis*", London: Chapman and Hall.

GHAHRAMANI, Z., and HINTON, G.E. (1997), "The EM Algorithm for Mixtures of Factor Analyzers", Technical Report CRG-TR-96-1, Toronto: University of Toronto, Department of Computer Science (revised).

LAVINE, M., and WEST, M. (1992), "A Bayesian Method for Classification and Discrimination", *Canadian Journal of Statistics 20*, 451–461.

MACQQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Eds. L.M.L. Cam and J. Neyman, University of California Press, Volume 1, pp. 281–297.

MCLACHLAN, G.J., and BASFORD, K.E. (1988), *Mixture Models : Inference and Applications to Clustering*, New York: Marcel Dekker.

MCLACHLAN, G.J., and KRISHNAN, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

MCLACHLAN, G.J. and PEEL, D. (1998), "Robust Cluster Analysis via Mixtures of Multivariate $t$-distributions", in *Lecture Notes in Computer Science*, Eds. A. Amin, D. Dori, P. Pudil, and H. Freeman, Springer, Volume 1451, pp. 658–666.

MCLACHLAN, G. J. and D. PEEL (2000), *Finite Mixture Models*, New York: Wiley.

MCLACHLAN, G.J., PEEL, D., BASFORD, K.E., and ADAMS, P. (1999), "The EMMIX Software for the Fitting of Mixtures of Normal $t$-components", *Journal of Statistical Software 4* (on-line publication www.jstatsoft.org).

MCLACHLAN, G.J., PEEL, D., and BEAN, R.W. (2003), "Modelling High-dimensional Data by Mixtures of Factor Analyzers", *Computational Statistics and Data Analysis 41*, 379–388.

MERBOUHA, A., and MKHADRI, A. (2001), "Regularization of the Location Model in Discrimination with Mixed Discrete and Continuous Variables", *Computational Statistics and Data Analysis 45*, 563–576.

MKHADRI, A., CELEUX, G., and NASROALLAH, A. (1997), "Regularization in Discriminant Analysis: An Overview", *Computational Statistics and Data Analysis 23*, 403–423.

MUKHERJEE, S., FEIGELSON, E.D., BABU, G.J., MURTAGH, F., FRALEY, C., and RAFTERY, A.E. (1998), "Three Types of Gamma Ray Bursts", *The Astrophysical Journal 508*, 314–327.

QUACKENBUSH, J. (2001), "Computational Analysis of Microarray Data", *Nature Reviews Genetics 2*, 418–427.

RAFTERY, A.E., and DEAN, N. (2006), "Variable Selection for Model-based Clustering", *Journal of the American Statistical Association 101*, 168–178.

REDNER, R.A., and WALKER, H.F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm", *SIAM Review 26*, 195–239.

RICHARDSON, S., and GREEN, P.J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components", *Journal of the Royal Statistical Society, Series B 59*, 731–758.

ROBERTS, S., HUSMEIER, D., REZEK, I., and PENNY, W. (1998), "Bayesian Approaches to Gaussian Mixture Modeling", *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 1133–1142.

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 WASSERMAN, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals", *Journal of the American Statistical Association 92*, 894–902.

RYAN, T.A., JOINER, B.L., and RYAN, B.F. (1976), *The MINITAB Student Handbook*, North Scituate, Massachusetts: Duxbury.

SCHAFER, J.L. (1997), *Analysis of Incomplete Multivariate Data by Simulation*, London: Chapman and Hall.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model", *The Annals of Statistics 6*, 461–464.

TITTERINGTON, D.M., SMITH, A.F., and MAKOV, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

WANG, N., and RAFTERY, A.E. (2002), "Nearest Neighbor Variance Estimation (NNVE): Robust Covarince Estimation via Nearest Neighbor Cleaning (with Discussion)", *Journal of the American Statistical Association 97*, 994–1019.

WARD, J.H. (1963), "Hierarchical Groupings to Optimize an Objective Function", *Journal of the American Statistical Association 58*, 234–244.

WEHRENS, R., BUYDENS, L., FRALEY, C., and RAFTERY, A.E. (2004), "Model-based Clustering for Image Segmentation and Large Datasets via Sampling", *Journal of Classification 21*, 231–253.

WEHRENS, R., SIMONETTI, A., and BUYDENS, L. (2002), "Mixture-modeling of Medical Magnetic Resonance Data", *Journal of Chemometrics 16*, 1–10.

YEUNG, K.Y., FRALEY, C., MURUA, A., RAFTERY, A.E., and RUZZO, W.L. (2001), "Model-based Clustering and Data Transformation for Gene Expression Data", *Bioinformatics 17*, 977–987.

ZHANG, Z., CHAN, K.L., WU, Y., and CHEN, C. (2004), "Learning a Multivariate Gaussian Mixture Model with the Reversible Jump MCMC Algorithm", *Statistics and Computing 14*, 343–355.