BAYESIAN STATISTICS 7, pp. 249–275
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian Clustering with Variable and Transformation Selections

JUN S. LIU
Harvard University, USA
jliu@stat.harvard.edu

JUNNI L. ZHANG
Peking University, China
zjn@gsm.pku.edu.cn

MICHAEL J. PALUMBO CHARLES E. LAWRENCE
The Wadsworth Center, USA
palumbo@wadsworth.org lawrence@wadsworth.org

SUMMARY

The clustering problem has attracted much attention from both statisticians and computer scientists in the past fifty years. Methods such as hierarchical clustering and the K-means method are convenient and competitive first choices off the shelf for the scientist. Gaussian mixture modeling is another popular but computationally expensive clustering strategy, especially when the data is of high-dimensional. We propose to first conduct a principal component analysis (PCA) or correspondence analysis (CA) for dimension reduction, and then fit Gaussian mixtures to the data projected to the several major PCA or CA directions. Two technical difficulties of this approach are: (a) the selection of a subset of the PCA factors that are informative for clustering, and (b) the selection of a proper transformation for each factor. We propose a Bayesian formulation and Markov chain Monte Carlo strategies that overcome the two difficulties and examine the performances of the new method by both simulation studies and real applications in molecular imaging analysis and DNA microarray analysis.

Keywords: GAUSSIAN MIXTURES, GIBBS SAMPLER, MICROARRAY, SIMULATED TEMPERING.

1. INTRODUCTION

Clustering objects into homogeneous groups is an important step in many scientific investigations. Recently, good clustering techniques are of special interest to biologists because of the availability of large amounts of high dimensional data resulting from the biotechnology revolution. These data include, for example, measurements of mRNA levels in the cell by microarray experiments, single-particle electron micrographs of macromolecules, high-throughput biological sequences of many species, protein-protein interaction data, etc. Although techniques for clustering high-dimensional observations have been subjected to active research for many years, traditional methods such as hierarchical clustering and K-means clustering are still top choices for scientists despite their various limitations in the analysis of complex data.

Due to the recent advances in Markov chain Monte Carlo (MCMC; see Liu (2001) for a recent overview), the Bayesian clustering approach via mixture models has been shown

The work of J. L. Zhang was doen while she was a Ph.D. Student at Harvard University, USA

attractive in many applications (Celeux and Govaert 1995, Fraley and Raftery 1999, Ghosh and Chinnaiyan 2002, Ishwaran *et al.* 2001, Kim *et al.* 2002, McLachlan *et al.* 2002, Moss *et al.* 1999, Richardson and Green 1997, Samsó *et al.* 2002, Yeung *et al.* 2001, etc.). However, the use of Bayesian clustering methods in high-dimensional data has been hindered by the very high computational cost and instability of the generic Gaussian mixture models. To overcome this difficulty, Banfield and Raftery (1993) proposed a general framework for directly modeling/constraining the covariance matrices of the mixture components. Here we recommend to first conduct a principal component analysis (PCA) or correspondence analysis (CA) for data reduction and then fit a mixture model to the factors resulting from these analyses. Indeed, if the original data come from a mixture Gaussian distribution and the estimated PCA or CA directions can be treated as known, then the new data vectors resulting from the projection of the original data onto these directions still follow a mixture Gaussian distribution. A similar approach has been applied to a character recognition problem (Kim *et al.* 2002).

When PCA is used in clustering problems, it can select classification-related directions if these directions are associated with the differneces in the locations of the means of different clusters. It can also pick up some artificial directions resulting from certain unusually noisy components or highly correlated components. Consequently, a potential problem with the PCA-Gaussian-mixture approach is the determination of an appropriate set of the PCA or CA factors useful for clustering. A common practice is to choose the factors corresponding to the few large principal components. But it is not clear where to stop and whether some of these eigen-directions are caused by some artefact or noises in the data unrelated to the clustering task.

In this article, we propose a novel procedure called Bayesian clustering with variable selection (BCVS), which can simultaneously cluster the objects and select "informative" variables, or factors, for the clustering analysis. Since many real data do not fit the multivariate Gaussian or mixture Gaussian models well, it is a common practice to first transform certain variables (using logarithm or some power functions) and then do the model fitting. These transformation steps are often carried out by the investigator based on a certain exploratory pre-processing of the data. We note that if the transformations are indexed as in Box and Cox (1964), each factor can be associated with a transformation variable and a full Bayesian model can be set up to include all the variables. Consequently, the BCVS procedure can be automated to select both informative factors and proper transformations for these factors. The advantage of this type of full Bayesian models is its ability to treat all involved variables in a coherent framework, to combine different sources of information, and to reveal subtle patterns by properly averaging out noise.

Section 2 presents two examples that motivated our development of the method: image clustering and microarray analysis. Section 3 first describes the full Bayesian Gaussian mixture model with variable selection. After the illustration in Section 3.1 of a standard Gibbs sampler for the model, Section 3.2 prescribes a more efficient predictive updating strategy for simultaneous clustering and variable selection. Section 3.3 details a tempering strategy for improving the convergence of the BCVS sampler. Section 4 formulates the transformation selection problem based on the framework of Box and Cox (1964). Section 5 tests the BCVS method on a series of simulated data sets, a micrograph image clustering problem and two microarray studies for cancer patient clustering. Section 6 concludes with a brief discussion.

2. MOTIVATING EXAMPLES

2.1. *Image Analysis for Electron Micrographs*

In single-particle electron micrograph imaging, each macromolecule lies randomly on the

specimen support in a limited number of ways (e.g., onto a few different "faces" of the molecule). A large number of images of the identical molecule are observed, and these images should fall into a few classes corresponding to the different characteristic orientations (Samsó *et al.* 2002). Because particles within a class are still randomly rotated in the plane, they must also be aligned. We focus here only on the classification of previously aligned images. The goals are to identify classes and to infer average images using differences in the appearance of particles. Figure 3 (a) shows images of E. Coli ribosome with four different tilting angles (high-tilt), and (b) shows that with four low-tilt angles. It is seen that after adding noises, the four classes of images are difficult to distinguish. Our goal here is to use a model-based method to automatically cluster these images into different classes.

In most single particle classification techniques, the images are first subject to CA or PCA (Frank and van Heel 1982). The factors produced by CA or PCA are prioritized according to the eigenvalue weights that account for the variance contribution of each factor. Not all factors carry meaningful or signal-related information, since noise or artifacts that are unrelated to the shape of the macromolecule can also contribute to the variance associated with a factor. Although there are some previous researches that address this issue (Frank 1996), these time-consuming methods necessitate an extensive knowledge of the system, are somewhat subjective, and tend to break down when the signal to noise ratio (SNR) is low. It is thus desirable to develop a method that can automatically select the factors to be used in clustering and improve the clustering of the images with low SNR.

2.2. *Patients Clustering Based on Gene Expression Microarrays*

The recent developments in gene chip or microarray technologies allow the scientist to observe simultaneously the expression levels of many genes in a cell at a given time, condition, or developmental stage (Schena *et al.* 1995). Because genes with similar or related functions often behave similarly under various conditions, biologists can discover novel gene-gene relationships and transcriptional regulatory signals by analyzing genes clustered based on the similarities of their expression patterns (Roth *et al.* 1998). It has also been reported that the gene expression profiles can be used to discriminate different cell types and predict patients' responses to certain drug treatment. It is also possible and important to cluster different cell types (or patients) based on their global gene expressions since the resulting gene clusters often correspond to clinically important subgroups. In this latter task, each cell type or patient is associated with the measurements of mRNA levels for thousands to tens of thousands of genes, a very high-dimensional vector, whereas the total number of patients is only in the range of hundreds or fewer. Alizadeh *et al.* (2000) used hierarchical clustering to divide the 96 lymphoma patients into two homogeneous groups based on 4026 expression values of each individual. These two groups correspond to two subgroups of patients who respond differently to the current therapy. Golub *et al.* (1999) used the self-organizing map (SOM) to cluster 38 leukemia samples into two groups based on the microarray values of 6817 genes for each individual. These two clusters again coincide well with the two important subtypes of the leukemia, ALL and AML. We show in the application section that BCVS can be successfully applied to these data to produce as good or better clustering results.

3. VARIABLE SELECTION IN MIXTURE MODELING

It is noticed that if we project a random Gaussian vector to a particular direction $\boldsymbol{v}$, then the resulting random variable also has a Gaussian distribution. Consequently, if we project observations from a mixture Gaussian distribution, the projected vectors should also follow mixture Gaussian. Although the PCA or CA directions need to be estimated from the data in

all applications, it does not seem to make any material difference by treating these directions as given, as long as the number of factors extracted from the data is substantially smaller than the original dimension of the data. Thus, in practice we obtain the first k_0 principal vectors ordered as $V = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_{k_0}^T)$, and project the data onto these directions so as to form n vectors of k_0 dimensions. The data vectors that will be subject to BCVS analysis are $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik_0})$, $i = 1, \ldots, n$, where each $\boldsymbol{x}_i$ is generated by multiplying the ith original data vector by projection matrix V. Each of the $(x_{1j}, x_{2j}, \ldots, x_{nj})^T$ will be called a factor throughout the paper.

It is reasonable to assume that each observation follows a mixture Gaussian distribution, of which each Gaussian component has the mean vector $\boldsymbol{\mu}_j$ and the covariance matrix Σ_j for $j = 1, \ldots, J$. The fractions of each component are $(p_1, \ldots, p_J)$. Notation-wise, we can write

$$\boldsymbol{x}_i \sim p_1 \mathrm{N}(\boldsymbol{\mu}_1, \Sigma_1) + \cdots + p_J \mathrm{N}(\boldsymbol{\mu}_J, \Sigma_J), \quad i = 1, \ldots, n.$$

Without having any constraints, each $\boldsymbol{\mu}_j$ is a k_0-dimensional vector and Σ_j a $k_0 \times k_0$ positive-definite matrix. A membership labeling variable J_i for each observation can be introduced so that

$$\boldsymbol{x}_i \,|\, J_i = j \sim \mathrm{N}(\boldsymbol{\mu}_j, \Sigma_j).$$

For the most part of this article, we assume that J is known in advance. There is a whole body of literature discussing how to choose a proper J in practice, and we will defer this issue to the discussion section.

We further assume that only a subset of the k_0 factors are informative for clustering. In the *anchor mode* model, we assume that this subset consists of the first K factors, where K is a random variable with a prior distribution $K \sim f(k)$. Thus, the data $\boldsymbol{x}_i$ has its first k components to follow a mixture Gaussian and its remaining components to follow a simple Gaussian distribution. Thus,

$$\boldsymbol{x}_i \,|\, J_i = j,\ \boldsymbol{\mu},\ K = k,\ \Sigma \sim N(\boldsymbol{\mu}_j, \Sigma_j) \times N(\boldsymbol{\mu}_0, \Sigma_0),$$

where $\boldsymbol{\mu}_j$ is a vector of length k and Σ_j is a $k \times k$ matrix, both are specific to cluster j; $\boldsymbol{\mu}_0$ is a vector of length $k_0 - k$ and Σ_0 is a $(k_0 - k) \times (k_0 - k)$ covariance matrix, common to all the observations.

Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$, and let $\boldsymbol{J} = (J_1, \ldots, J_n)$. Then

$$P(\boldsymbol{X} \,|\, \boldsymbol{J} = \boldsymbol{j}, \boldsymbol{\mu}, K = k) = \prod_{i=1}^{n} \mathrm{N}(\boldsymbol{x}_{i[1:k]} \,|\, \boldsymbol{\mu}_{j_i}, \Sigma_{j_i}) \prod_{i=1}^{n} \mathrm{N}(\boldsymbol{x}_{i[k+1:k_0]} \,|\, \boldsymbol{\mu}_0, \Sigma_0),$$

where $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \Sigma)$ is the multivariate Gaussian density. The joint posterior distribution is then

$$P(\boldsymbol{J}, K = k, \boldsymbol{\mu}, \Sigma \,|\, \boldsymbol{X}) \propto f(k)\pi(\boldsymbol{\mu}, \Sigma) \prod_{i=1}^{n} \{p_{j_i}\ \mathrm{N}(\boldsymbol{x}_{i[1:k]} \,|\, \boldsymbol{\mu}_{j_i}, \Sigma_{j_i})\ \mathrm{N}(\boldsymbol{x}_{i[k+1:k_0]} \,|\, \boldsymbol{\mu}_0, \Sigma_0)\}.$$

Note that the dimensionality of $\boldsymbol{\mu}_j$, Σ_j, $\boldsymbol{\mu}_0$, and Σ_0 will change when k changes.

We first give some detailed calculation for the derivation of a MCMC algorithm for the anchor mode BCVS. The procedure can be easily generalized to the non-anchor mode, in which BCVS can select any combination of any number of the k_0 factors for the clustering.

3.1. *A Gibbs Sampling Algorithm*

Markov chain Monte Carlo algorithms for the estimation in mixture models have been an active topic in statistical research. Some of the recent articles include Brooks (2001), Diebolt and

Robert (1994), Ishwaran *et al.* (2001), Neal (2000), Richardson and Green (1997), just to start a list. The new feature in our algorithm is its variable/factor selection step.

Assume *a priori* that $[\boldsymbol{\mu}_j \mid \Sigma_j] \sim N(\boldsymbol{x}_0, \Sigma_j/\rho_0)$ and $\Sigma_j \sim \text{Inv-W}_{\nu_0}(S_0^{-1})$ (see the Appendix for its density form). Let $\mathrm{N}(\cdot)$ denote the Gaussian density function. Then

$$\begin{aligned} P(J_i = j \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J, \Sigma; \boldsymbol{X}) &= \frac{p_j \mathrm{N}(\boldsymbol{x}_{i[1:k]} \mid \boldsymbol{\mu}_j, \Sigma_j)\mathrm{N}(\boldsymbol{x}_{i[k+1:k_0]} \mid \boldsymbol{\mu}_0, \Sigma_0)}{\sum_{l=1}^J p_l \mathrm{N}(\boldsymbol{x}_{i[1:k]} \mid \boldsymbol{\mu}_l, \Sigma_l)\mathrm{N}(\boldsymbol{x}_{i[k+1:k_0]} \mid \boldsymbol{\mu}_0, \Sigma_0)} \\ &= \frac{p_j \mathrm{N}(\boldsymbol{x}_{i[1:k]} \mid \boldsymbol{\mu}_j, \Sigma_j)}{\sum_{l=1}^J p_l \mathrm{N}(\boldsymbol{x}_{i[1:k]} \mid \boldsymbol{\mu}_l, \Sigma_l)} \end{aligned}$$

and

$$[\boldsymbol{\mu}_j \mid \Sigma, \boldsymbol{J}, \boldsymbol{X}] \sim N\left(\frac{\rho_0 \boldsymbol{x}_{0[1:k]} + \sum_{i=1}^n \boldsymbol{x}_{i[1:k]} \times I_{\{J_i=j\}}}{\rho_0 + \sum_{i=1}^n I_{\{J_i=j\}}},\ \frac{\Sigma_j}{\rho_0 + \sum_{i=1}^n I_{\{J_i=j\}}}\right);$$

$$[\boldsymbol{\mu}_0 \mid \Sigma, \boldsymbol{X}] \sim N\left(\frac{\rho_0 \boldsymbol{x}_{0[1:k]} + \sum_{i=1}^n \boldsymbol{x}_{i[k+1:k_0]}}{\rho_0 + n},\ \frac{\Sigma_0}{\rho_0 + n}\right).$$

The conditional distribution for Σ_j is then

$$[\Sigma_j \mid \boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{J}, K = k] \propto |\Sigma_j|^{-\frac{\nu_0 + k + n_j + 1}{2}} e^{-\frac{1}{2}\mathrm{tr}[\Sigma_j^{-1}(S_0 + SS_j(\boldsymbol{\mu}_j))]},$$

where n_j is the number of observations in the jth cluster and $SS_j(\boldsymbol{\mu}_j) = \sum_{i:j_i=j}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_j)$ is the mean-corrected sum of squares for the observations in the jth cluster.

A prior $\mathrm{Di}(a_1, \ldots, a_J)$ distribution can be employed for the cluster proportions $(p_1, \ldots, p_J)$. Then, given the clustering indicators $\boldsymbol{J}$, it is straightforward to update the proportion vector by drawing from $\mathrm{Di}(n_1 + a_1, \ldots, n_J + a_J)$, where n_j is the size of the jth cluster.

Updating K in our model setting is not as trivial as the previous steps because once the $\boldsymbol{\mu}_j$ and Σ_j are fixed, the factor number K, which underlies the dimensionality of the mean vectors and covariance matrices, cannot be moved any more. It is possible to propose a change for all the $\boldsymbol{\mu}_j$, the Σ_j and K jointly and use the reversible jumping rule to guide for the acceptance/rejection decision. However, this type of proposals often encounter a high rejection rate, rendering the algorithm inefficient. Here we adopt a more effective alternative: marginalizing the $\boldsymbol{\mu}_j$ and Σ_j analytically. More precisely, using the standard Bayesian Gaussian inference results summarized in the Appendix, we can derive that

$$P(\boldsymbol{X}_{[1:k]} \mid K = k, \boldsymbol{J}) = \prod_{j=1}^J \frac{Z(\nu_0, S_0, k)}{Z(n_j + \nu_0, S_0 + SS_j, k)} (2\pi)^{-\frac{n_j k}{2}} \left(\frac{n_j + \rho_0}{\rho_0}\right)^{-\frac{k}{2}}, \quad (1)$$

where SS_j is defined as in (11) for observations in the jth cluster. This calculation indicates that a more efficient approach, iterative predictive updating (Liu 1994, Chen and Liu 1996), is possible for our MCMC computation.

3.2. *Predictive Updating*

Instead of doing the full Gibbs, a convenient alternative is to iteratively draw the label of each $\boldsymbol{x}_i$ from its predictive distribution conditional on the labels of the remaining observations and then update the component number K conditional on the labels. This strategy not only saves our

effort in updating the mean vectors and covariance matrices, but also improves the convergence rate of the sampler (Liu 1994). More precisely, conditional on K and $\boldsymbol{J}$ and with conjugate priors, we can compute the analytical form of

$$P(\boldsymbol{X} \mid K = k, \boldsymbol{J}) = P(\boldsymbol{X}_{[1:k]} \mid K = k, \boldsymbol{J}) \times P(\boldsymbol{X}_{[k+1:k_0]}), \tag{2}$$

which leads to an iterative sampling of $J_i = j$ with probability proportional to the prior fraction p_j times a multivariate-t density function, and an update of K from $[K = k \mid \boldsymbol{J}, \boldsymbol{X}]$. We can also marginalize p_j if a prior $\mathrm{Di}(a_1, \ldots, a_J)$ has been used for the proportions, in which case we use in the place of p_j,

$$\hat{p}_j = (n'_j + a_j)/(n + a_1 + \cdots + a_J - 1),$$

where n'_j is the total number of objects in the jth cluster excluding the ith observation.

In order to compute the Metropolis ratio for changing the variable component number from k to $k + 1$, we compute the related Bayes factors (normalizing constants) as in the standard Bayesian Gaussian inference (see Appendix). Let the normalizing function $Z(\nu, S, k)$ be defined as in (10). With known clustering information, we can compute (2) in two steps:

- The easier one (the Bayes factor $P(\boldsymbol{X}_{[k+1:k_0]} \mid \boldsymbol{J})$) is

$$P(\boldsymbol{X}_{[k+1:k_0]}) = \frac{Z(\nu_0, S_0, k_0 - k)}{Z(n + \nu_0, S_0 + SS_0, k_0 - k)} (2\pi)^{-\frac{n(k_0-k)}{2}} \left(\frac{n + \rho_0}{\rho_0} \right)^{-\frac{k_0-k}{2}},$$

  where SS_0 is the sum of square matrix (as defined in (11)) computed from $\boldsymbol{X}_{[k+1:k_0]}$. from the prior mean $\boldsymbol{x}_0[k + 1 : k_0]$ of $\boldsymbol{\mu}_{[k+1:k_0]}$. If we model the columns of $\boldsymbol{X}$ from $k + 1$ to k_0 as independent, then the above marginal likelihood can be modified as

$$P(\boldsymbol{X}_{[k+1:k_0]}) = \prod_{j=k+1}^{k_0} \frac{Z(\nu_0, s_0^2, 1)}{Z(n + \nu_0, s_0^2 + s^2(\boldsymbol{x}_j), 1)} (2\pi)^{-n/2} \left(\frac{n + \rho_0}{\rho_0} \right)^{-1/2},$$

  where $s^2(\boldsymbol{x}_j)$ is the modified total sum of squares of the jth column. The prior for the unknown mean and variance of each independent component takes the same form as in the multi-dimensional case (see Appendix), but with s_0^2 replacing S_0.
- Compute the modified sum of square matrix SS_j (by formula (11) for observations in the jth cluster, with data $\{\boldsymbol{x}_{i[1:k]}, i \in \text{Cluster } j\}$. Here n_j is the cluster size. Then

$$P(\boldsymbol{X}_{[1:k]} \mid K = k, \boldsymbol{J}) = \prod_{j=1}^{J} \frac{Z(\nu_0, S_0, k)}{Z(n_j + \nu_0, S_0 + SS_j, k)} (2\pi)^{-\frac{n_j k}{2}} \left(\frac{n_j + \rho_0}{\rho_0} \right)^{-\frac{k}{2}}. \tag{3}$$

  The Metropolis ratio for $k \to k'$ is then

$$r = \min \left\{ 1, \frac{f(k')P(\boldsymbol{X} \mid K = k', \boldsymbol{J})T(k' \to k)}{f(k)P(\boldsymbol{X} \mid K = k, \boldsymbol{J})T(k \to k')} \right\}. \tag{4}$$

  If k_0 is small, we can draw K from its conditional distribution $P(K = k \mid \boldsymbol{J}, \boldsymbol{X})$ directly.

To summarize, we have the following predictive updating iterations to replace the regular Gibbs sampler:

1. Sample the group indicator variable of the ith observation from the t-distribution:

$$[J_i = j \mid \boldsymbol{J}_{[-i]}, K = k, \boldsymbol{X}] \propto \hat{p}_j \; f_t \left(\boldsymbol{x}_{i[1:k]};\; \nu_j, \hat{\boldsymbol{\mu}}_j, \frac{(n'_j + \rho_0 + 1)(S_0 + SS_k)}{(n'_j + \rho_0)(n'_j + \nu_0 - k + 1)} \right)$$

   where $\nu_j = n'_j + \nu_0 - k + 1$, n'_j is the current size of cluster j excluding the ith observation, and $\hat{\boldsymbol{\mu}}_j$ is a weighted combination of prior mean $\boldsymbol{x}_0$ and the sample mean as in (12). The weight for the prior is $\rho_0/(n'_j + \rho_0)$ and for the sample mean is $n'_j/(n'_j + \rho_0)$.
2. Propose a move for k to k' according to a transition rule $T(k \to k')$; accept the move with probability (4).

This algorithm can be easily modified to accommodate the non-anchor mode. Without the restriction of having to include the first k factors, we can choose a factor at random and ask whether we should treat it as an informative factor or not, i.e., turn it on or not. Conditional on each factor's on-off states, $\boldsymbol{O}$, we can compute $P(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{Y})$ the same way as in (3). The Metropolis-ratio similar to (4) can be used to guide the transition from $\boldsymbol{O}$ to a new vector of the on-off states, $\boldsymbol{O}'$.

3.3. *Parallel Tempering*

It has been observed that the MCMC samplers for fitting a mixture model tend to be very "sticky". With the inclusion of the variable selection and transformation indicators, the MCMC algorithm tends to perform even worse. Parallel tempering (Geyer 1991) seems to be an effective means to improve the mixing of a MCMC sampler.

To implement a tempering sampler conditional on $K = k$, we define the target distribution as

$$\pi(\boldsymbol{J}) \propto P(\boldsymbol{Y} \mid K = k, \boldsymbol{J}) P(\boldsymbol{J}),$$

where $P(\boldsymbol{J})$ is the prior distribution for $\boldsymbol{J}$. Then $\pi(\boldsymbol{J})$ can be evaluated as described in Section 3.2, up to a normalizing constant. In order to carry out the tempering idea, we construct a temperature ladder, $1 = t_1 < t_2 < \cdots < t_L$, and define $\pi_l(\boldsymbol{J}) \propto \{\pi(\boldsymbol{J})\}^{\frac{1}{t_l}}$.

The sample space of the tempering sampler is the product space of the $\boldsymbol{J}$. In other words, the new target distribution is

$$\Pi(\boldsymbol{J}_1, \ldots, \boldsymbol{J}_L) = \pi_1(\boldsymbol{J}_1) \times \cdots \times \pi_L(\boldsymbol{J}_L),$$

for which our tempering sampler will converge to. Here we let $\boldsymbol{J}_l = (J_{l,1}, \ldots, J_{l,n})$. The tempering process can be implemented as follows:

Tempering Sampler

1. Iterative classification (independently) at each level. For levels $l = 1, 2, \ldots, L$:
   - For $i = 1, \ldots, n$, compute

$$[J_{l,i} = j \mid \boldsymbol{J}_{l,[-i]}] \propto \left\{ p_j \; f_t \left(\boldsymbol{y}_{i[1:k]};\; \nu_j, \hat{\boldsymbol{\mu}}^*_j, \frac{(n_j + \rho_0 + 1)(S_0 + SS_k)}{(n_j + \rho_0)(n_j + \nu_0 - k + 1)} \right) \right\}^{1/t_k}$$

     for all possible j (in our example, $j = 1, 2, 3, 4$).
   - Update $J_{l,i}$ by a random draw from the above distribution.
2. For every N_0 (say, 10) cycles of iterative updating, we conduct one cycle of level exchange, starting from the highest-temperature configuration. For $k = 1, \cdots, L - 1$:

- compute the ratio

$$r = \frac{\pi_{L-k}(\boldsymbol{J}_{L-k+1})\pi_{L-k+1}(\boldsymbol{J}_{L-k})}{\pi_{L-k}(\boldsymbol{J}_{L-k})\pi_{L-k+1}(\boldsymbol{J}_{L-k+1})} = \left\{\frac{\pi(\boldsymbol{J}_{L-k+1})}{\pi(\boldsymbol{J}_{L-k})}\right\}^{\frac{1}{t_{L-k}}-\frac{1}{t_{L-k+1}}}$$

- exchange $\boldsymbol{J}_{L-k}$ and $\boldsymbol{J}_{L-k+1}$ with probability $\min\{1, r\}$.

3. Go back to step 1.

4. SELECTING A PROPER TRANSFORMATION

For ease of presentation, we give only the details for deciding whether logarithm transformations should be applied to certain factors. The formulas for the more general selection from a continuum of transformations using the Cox-Box formulation is presented with details omitted. We first consider the univariate case, and then generalize to consider several variables.

Suppose we have a set of univariate observations $x_1, \ldots, x_n$. Suppose the prior distribution for μ and Σ is as given in (8) and (9), we have

$$P(x_1, \ldots, x_n \,|\, \nu_0, s_0, x_0, \rho_0) = \frac{Z(\nu_0, s_0, 1)}{Z(n+\nu_0, SS+s_0, 1)}(2\pi)^{-n/2}\left(\frac{n+\rho_0}{\rho_0}\right)^{-1/2},$$

where $\sigma^2 \sim$ Inv-W$_{\nu_0}(s_0^{-1})$, $\mu \,|\, \sigma^2 \sim N(x_0, \sigma^2/\rho_0)$, and SS is as defined in (11). If the data needs to be log-transformed, then the likelihood is

$$P_l(x_1, \ldots, x_n \,|\, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(\log x_i - \mu)^2}{2\sigma^2}\right\}\prod_{i=1}^n x_i^{-1}.$$

Thus, under the log-transformation model and a set of different prior parameters (indicated by "$*$"), we have

$$P_l(x_1, \ldots, x_n \,|\, \nu_0^*, s_0^*, x_0^*, \rho_0^*) = \frac{Z(\nu_0^*, s_0^*, 1)}{Z(n+\nu_0^*, SS^*+s_0^*, 1)}(2\pi)^{-n/2}\left(\frac{n+\rho_0^*}{\rho_0^*}\right)^{-1/2}\prod_{i=1}^n x_i^{-1},$$

where SS^* is the corresponding residual sum of squares for the $\log(x_i)$.

More generally, suppose we have n iid observations, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n$, from a k-dimensional distribution. Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$. We notice what when considering the logarithm transformation for one of the variables, x_l, say, the computation of the data likelihood is almost unchanged. More precisely, we have

$$P_l(\boldsymbol{X} \,|\, \nu_0^*, S_0^*, \boldsymbol{x}_0^*, \rho_0^*) = \frac{Z(\nu_0^*, S_0^*, k)}{Z(n+\nu_0^*, SS^*+S_0^*, k)}(2\pi)^{-nk/2}\left(\frac{n+\rho_0^*}{\rho_0^*}\right)^{-k/2}\prod_{i=1}^n x_{il}^{-1},$$

where SS is the sum of square matrix of the original data as defined in (11), and SS^* is the sum of square matrix of the transformed data.

It is not very sensible to use the same prior for the original and the transformed data, but it is also difficult to decide what corresponding priors should be applied to the transformed data. Thus, noninformative priors seem to be desirable, although they will typically result in improper (infinite) Bayes factors. We note, however, that if we let $\rho_0 = \rho_0^*$, $\nu_0 = \nu_0^*$, $S_0 = S_0^*$, and let all of them converge to zero, we still have a proper ratio:

$$\frac{P(\boldsymbol{X})}{P_l(\boldsymbol{X})} = \frac{Z(n, SS^*, k)}{Z(n, SS, k)}\prod_{i=1}^n x_{il} = \left(\frac{|SS^*|}{|SS|}\right)^{n/2}\prod_{i=1}^n x_{il}. \tag{5}$$

For a transformation of the Box-Cox form $f_\alpha(x) = (x^\alpha - 1)/\alpha$, $\alpha > 0$, the ratio is

$$\frac{P(\boldsymbol{X})}{P_\alpha(\boldsymbol{X})} = \left(\frac{|SS^*|}{|SS|}\right)^{n/2} \prod_{i=1}^{n} x_{il}^{1-\alpha}. \tag{6}$$

Note that $\alpha = 0$ corresponds to the logarithm transformation. To insert the variable selection step into the predictive updates, we sample conditional on $\boldsymbol{J}$ for each variable whether a logarithm transformation should be applied, according to the ratio (5).

When the observations involve negative values (or are very large), it is sometimes useful to consider the transformation of the form $f_{\alpha,m} = [(x+m)^\alpha - 1]/\alpha$, where m is some constant to be added to the observation and can be estimated as well. We see from the foregoing derivations that the Bayes ratio should take a similar form:

$$\frac{P(\boldsymbol{x})}{P_{\alpha,m}(\boldsymbol{x})} = \left(\frac{|SS^*|}{|SS|}\right)^{n/2} \prod_{i=1}^{n} (x_{il} + m)^{1-\alpha}. \tag{7}$$

In practice, we may start m with $1 - \min\{x_{1i}, i = 1, \ldots n\}$ and update m along with the MCMC iterations.

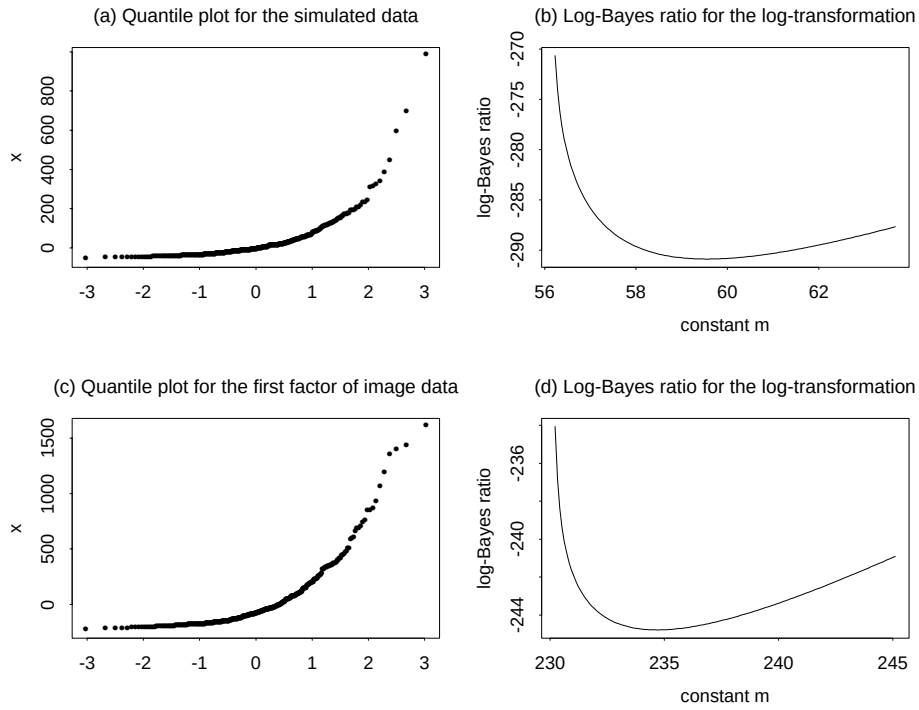

Figure 1. *Deciding whether a logarithm transformation should be used for (a) a simulated dataset and (c) the first factor of a micrograph image dataset. (b) and (d): the logarithm of the Bayes ratio as computed by (7) with* $\alpha = 0$.

We simulated four hundred observations from $\exp(Z + 4) - 60$, where Z is the standard Gaussian random variable. Its qq-plot is seen in Figure 1 (a). The log-Bayes ratio (the untransformed likelihood versus the logarithm transformation) as shown in Figure 1 (b) clearly indicates that a constant around 50 to 60 should be added and the logarithm transformation is necessary. The qq-plot in Figure 1 (c) shows the long-tailness of the first factor in the micrograph image example. By applying the transformation-selection procedure just described, we see that from Figure 1 (d) that a number around 235 should be added before the logarithm transformation.

The Bayes ratio strongly indicates the use of log-transformation for a wide range of constant m. As shown later, this transformation significantly improved the clustering result for a molecular micrograph application.

5. PERFORMANCE EVALUATION OF THE BCVS

5.1. *A Simulation Study*

In order to investigate the usefulness of variable selection in clustering analysis, we simulated observations from a mixture of two bivariate Gaussian distributions,

$$\alpha \mathrm{N}(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)}) + (1-\alpha)\mathrm{N}(\boldsymbol{\mu}^{(2)}, \Sigma^{(2)}),$$

then we add some factors which are just independent Gaussians. Thus, in these data, only the first two factors are "informative," and the other dimensions are noises. We compare the results of the following: clustering using only the first two factors; using BCVS with the anchored mode, using BCVS with non-anchored mode, and using all the factors indiscriminately.

We first simulated 100 data sets with four factors, for each data set, the parameters for the bivariate mixture Gaussian distribution were drawn according to

$$\begin{aligned}
&\alpha \sim \mathrm{Un}(0.3, 0.7),\\
\sigma_{11}^{(1)} \sim Inv-\chi^2(2, 4.0), \quad &\sigma_{22}^{(1)} \sim Inv-\chi^2(8, 5.0), \quad \rho^{(1)} \sim \mathrm{Un}(0, 0.6),\\
&\mu_1^{(1)}|\sigma_{11}^{(1)} \sim \mathrm{N}(3, \sigma_{11}^{(1)}), \quad \mu_2^{(1)}|\sigma_{22}^{(1)} \sim \mathrm{N}(1, \sigma_{22}^{(1)}),\\
\sigma_{11}^{(2)} \sim Inv-\chi^2(8, 5.0), \quad &\sigma_{22}^{(2)} \sim Inv-\chi^2(2, 4.0), \quad \rho^{(2)} \sim \mathrm{Un}(0, 0.6),\\
&\mu_1^{(2)}|\sigma_{11}^{(2)} \sim \mathrm{N}(0, \sigma_{11}^{(2)}), \quad \mu_2^{(2)}|\sigma_{22}^{(2)} \sim \mathrm{N}(0, \sigma_{22}^{(2)})
\end{aligned}$$

where $\sigma_{ll}^{(j)}$ is the variance of the lth variable and $\rho^{(j)}$ the correlation coefficient for cluster j. Each noise factor has mean 0 and variance τ^2 drawn from Inv-$\chi^2(8, 5.0)$.

After the parameters were drawn, 200 observations were then simulated with these parameters. This setup resulted in a wide range of data sets, some of which had well-separated clusters, and some had close clusters. We also generated 100 data sets of size 200 with nine factors (seven noisy factors). For each data set, the parameters for the bivariate mixture Gaussian distribution and those for each of the seven noise factors were drawn similarly as described above. In each dimension setting (4 and 9, respectively), we stratified the 100 simulated datasets into 3 groups (easy, median, and difficult) of about equal sizes based on the performances of the "gold standard", i.e., the clustering algorithm using only the two informative factors. Figure 2 compares the performances of the three clustering approaches, BCVS with anchor mode, BCVS with non-anchor mode, and clustering using all the factors, with the "gold standard", in each data set group as well as all data sets.

In the MCMC implementations, we assumed *a priori* that for each cluster j, $[\boldsymbol{\mu}_j \mid \Sigma_j] \sim \mathrm{N}(\boldsymbol{x}_0, \Sigma_j/\rho_0)$ and $\Sigma_j \sim$ Inv-W$_{\nu_0}(S_0^{-1})$, where $\boldsymbol{x}_0$ is the vector of sample means of the k factors, $\rho_0 = 0.01$, $v_0 = k$, and S_0 is a diagonal matrix with sample variances of the k factors as the diagonal elements. We also assumed uniform prior for the number of factors used for clustering in anchor mode and for the combination of factors used for clustering in non-anchor mode.

Figure 2 showed the differences in the number of correctly clustered objects for each method. Suppose the true clustering answer groups the observations into $\boldsymbol{T} = (T_1, T_2)$. Then the number of correctly clustered objects of a clustering result $\boldsymbol{C} = (C_1, C_2)$ is defined as the number of matched objects for the best match, i.e.,

$$N(\boldsymbol{C}) = max\{|C_1 \cap T_1| + |C_2 \cap T_2|, |C_2 \cap T_1| + |C_1 \cap T_2|\}.$$

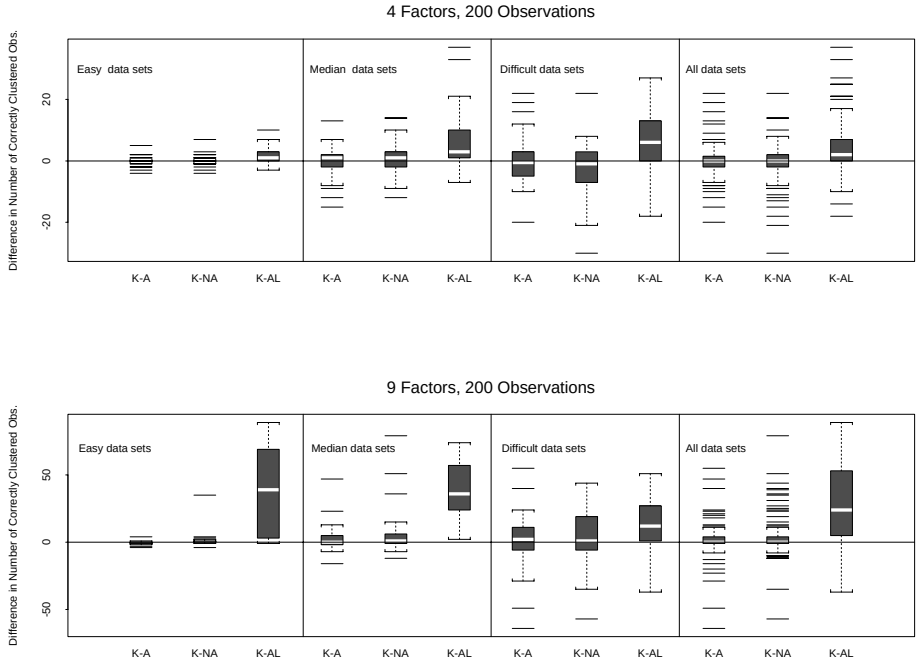

Figure 2. *The number of correctly clustered objects of the three clustering strategies (anchor, non-anchor, and all factors) were each compared with that of the clustering using the two known factors. A: BCVS with anchor mode; NA: non-anchor mode; K: clustering using only the first two factors; AL: using all the factors. The 100 simulated datasets for each setting are stratified according to the clustering performances when the two informative factors are known (K).*

A completely random assignment will result in a $N(\boldsymbol{C})$ slightly greater than $n/2$.

It is interesting to note from the boxplots that for the datasets with 4 factors the BCVS with both the anchor and non-anchor modes performed almost as good as the method with the two informative factors known, and all the three are better than the method using all the 4 factors indiscriminately. The datasets with 9 factors showed more striking differences. Note that when all the 9 factors are used in fitting a mixture of two Gaussians, we have to entertain 109 free parameters: two nine-by-nine dimensional covariance matrices, two mean vectors, and the mixture proportion. It is not surprising that such a method performed poorly. It is somewhat surprising, however, that BCVS performed so much better, suggesting that we do not lose much by not knowing which factors to use. A careful examination shows that the BCVS with anchor mode performed slightly better than that with non-anchor mode.

5.2. *Classification of Images from Electron Micrographs*

In a recent study (Samsó *et al.* 2002), we applied the BCVS to classify the electron micrograph images and compared its performances with the popular hierarchical clustering (HAC) method. The signal components of these images were generated by projecting a volume of the 50S ribosomal subunit from *Escherichia coli* reconstructed from electron micrographs of a negatively stained sample (Radermacher *et al.* 1987). A set of four images, as shown in Figure 3 (a), was created by tilting four projections by different angles. A set of lower-tilt images were also produced for closer comparisons but is not shown here.

One hundred copies of each projection were generated and noise was added to each of them, yielding four tilt groups of 100 images each. Two different sources of noise were used. In the first series, Gaussian noise with standard deviation from 1 to 20 was added to each projection.

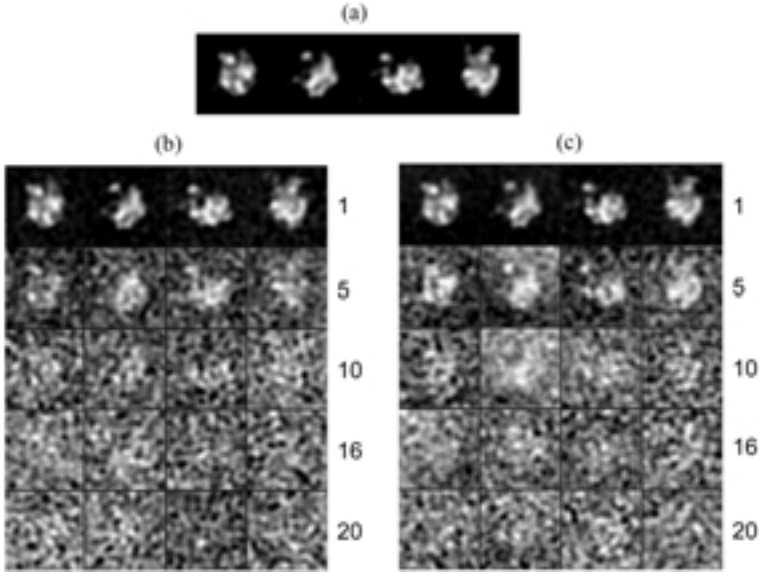

Figure 3. *(a) The images of the 50S ribosomal subunit from E. coli at four tilting angels. (b) Images with Gaussian noises at different SNR levels. (c) Images with real noises.*

In a second series, the "real noise" obtained by windowing 400 regions of an image of carbon film obtained in the electron microscope was added. The real-noise windows were also scaled to obtain a wide range of SNR. Examples of the resulting individual images from selected data sets are shown in Figure 3 (b) and (c). These data sets differ from those obtained in real experiments where particles are found in random orientations. However, these partially simulated data allow the comparison of different classification methods without confounding the comparison with alignment issues.

Each image data were submitted to either CA or PCA dimension reduction, and the first eight factors were saved. These eight-dimensional data sets were classified by both the BCVS and HAC with "complete-link" merging criterion. Since methods for factor selections in HAC are subjective, all eight factors were employed with this procedure. The HAC dendrogram tree was cut at a threshold giving four classes.

Results. The BCVS yielded a smooth sigmoidal decline of the success rate of classification with increased noise (Figure 4). HAC showed a more erratic pattern, e.g., from noise 4 to 5 there is a sudden decrease of classification (from 100% to 75.5%, Figure 4 (a), resulting from the merging of nearly all of the observations of two tilt groups (94% and 96%) into a single class, leaving a fourth class almost empty. At this noise level, BCVS classified 100% correctly. Up to and including noise level 14, the BCVS anchor and non-anchor modes produced similar results. At noise levels greater than 14, the anchor mode gave the best results.

For real-noise data sets, at high-tilt, both BCVS and HAC produce similar results through noise level 10 (Figure 4 (b)). At noise level 12, HAC decreases its score suddenly (from 92% to 59%), whereas BCVS continues to return favorable scores through noise level 14, at which a score of over 79% is seen. The BCVS outperformed HAC more significantly in all low-tilt datasets (e.g., 81% *vs* 48% at noise level 4, as shown by dashed lines in Figure 4 (c)).

PCA versus CA. CA is historically preferred to PCA in image processing because PCA often results in factors whose scales differ in magnitudes, rendering the Euclidean distance-based HAC algorithm difficult to produce meaningfully clusters. However, since the BCVS adaptively adjusts scales of factors through inferences of variances, we examined its application under PCA

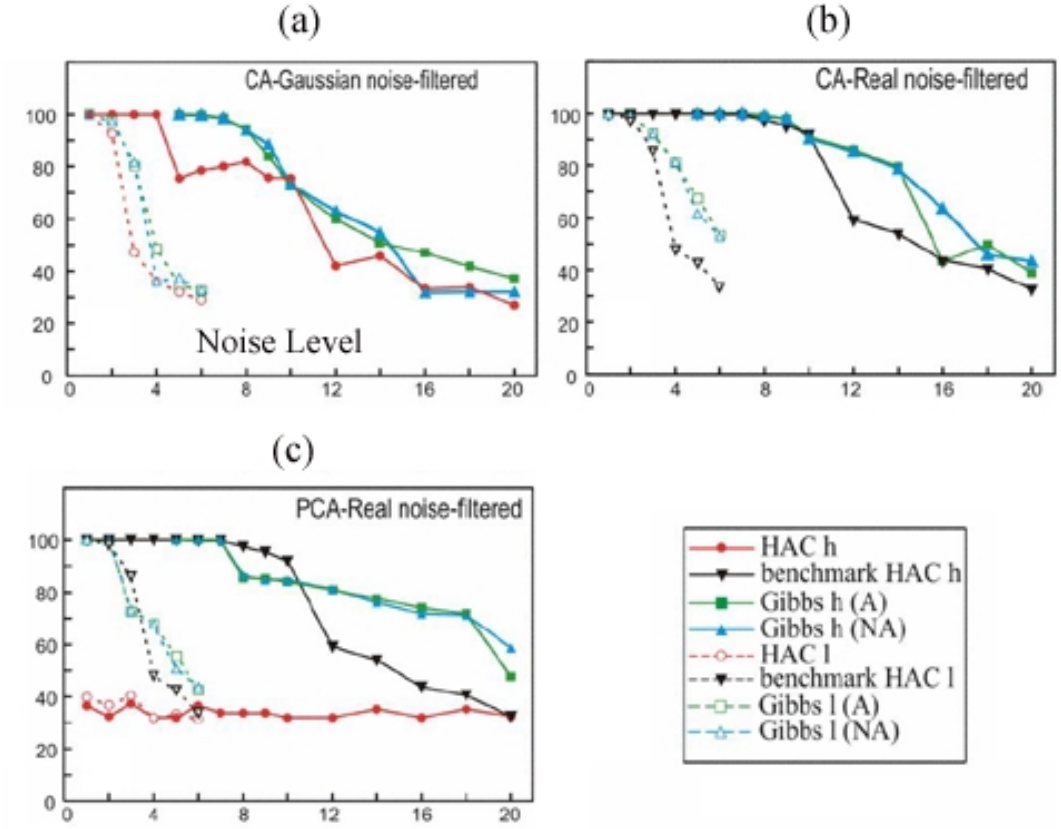

Figure 4. *The comparison results for the micrograph image data. Solid lines are for high-tilt datasets and the dashed lines for low-tilt datasets.*

data reduction. With well-behaved Gaussian noise we found that in all cases both BCVS and HAC perform better on PCA data than on CA data (Figure not shown).

HAC's results on real-noise data sets using PCA show very low scores (below 40%) for all noise levels (Figure 4 (c)). Scores in this range indicate that the classification is not much better than a random assignment of particles into classes. Because HAC uses Euclidean distances, it is not well suited for cases in which the variance (i.e., the spread) of a factor differs by an order of magnitude or more from the variance of other factors. These results support the accepted practice of using CA data reduction for HAC. In contrast, the BCVS algorithm is not hindered by data with such characteristics and can perform better than that using the CA.

Logarithm transformation of factors. We observed that, under PCA reduction, the first factor had a much larger variance than the other factors and the degree of this effect varies from class to class. The q-q plot for the first factor is displayed in Figure 1 (c), which shows a clear sign of long-tailness. We considered the transformation of the form $\log(x + m)$. As discussed in Section 4, the Bayes ratio of the original model versus the log-transformation data model for this factor can be computed. Figure 1 (d) shows the logarithm of the ratio for a range of m values, which is overwhelmingly in favor of transformation (a value of -100 means that the logarithm transformation is e^{100} times more favorable than no-transformation).

With this factor log-transformed, the rate of correct classification is significantly improved for BCVS on real-noise, PCA-reduced data. For example, at the noise level 8, the correct-classification rate increases from 83.50 to 94.25 for the anchor mode, and from 82.75 to 94.25 for the non-anchor mode, and such a high success rate of more than 80% is maintained through noise level 18.

5.3. *Clustering of Microarray Data*

Lymphoma Data. Alizadeh *et al.* (2000) reported a microarray gene expression study on diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgekin's lymphoma. It was known that this DLBCL is clinically heterogeneous with about 40% of patients responding well to the current therapy and the remaining 60% succumbing to the disease. The microarray studies of the tumor cells in these two types of patients revealed the molecular heterogeneity of the two types of DLBCL. It is of interest here to see if a clustering technique, without using any knowledge about the identities of the tumor cells (i.e., whether they come from a responding patient or not), can identify the two distinct groups.

A complementary DNA (cDNA) microarray chip (Schena *et al.* 1995) was constructed by Alizadeh *et al.* (2000), who selected genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in cancer or immune systems. The final "Lymphochip" consists of 18,000 cDNA clones. About 1.8 million measurements of gene expression were made in 96 normal and malignant lymphocyte samples using 128 Lymphochip microarrays. After some preprocessing of the raw data, the authors discarded about 75% of the mRNA measurements of all the patients and made available a 4026×96 matrix corresponding to the mRNA expression measurements of 4026 genes in the 96 patients.

We first submitted the data to Splus software package to conduct the PCA and return the projections of the data to the first 16 eigen directions. Then we asked the BCVS to cluster the objects into two groups based on the 16 factors. The algorithm typically picked 7 to 8 factors in the clustering analysis. When BCVS was asked to produce 8 clusters with non-anchor mode, it resulted in very similar clusters as that reported in Alizadeh *et al.* (2000) who used HAC. On average 11 factors were selected, 1-9, 11 and 15. When four clusters were asked for, BCVS chose to use ten factors, 1-9 and 12, and produced a group (47 members) with mostly DLCL types, a group that mixes CLL and FL types (24), a group of Blood B cell types (13), and a group of Blood T cell types (12). This result is in close agreement with Alizadeh *et al.* (2000). A small difference is that the two germinal center B cells were grouped with the CLL+FL cluster and DLCL-0009 and SUDHL5 were put into the DLCL cluster.

Leukemia data. Golub *et al.* (1999) applied gene expression microarray techniques to study human acute leukemias and discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Distinguishing ALL from AML is crucial for successful treatment, since chemotherapy regimens for ALL can be harmful for AML patients. Their results demonstrated the feasibility of cancer classification based solely on gene expression monitoring and suggested a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge. They first constructed a classifier that can distinguish the two types of cancers using an initial collection of samples belonging to known classes. As a second task, they applied a two-cluster self-organizing map (SOM) to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 gene. Their SOM results matched the known classes closely: Class A1 contained mostly ALL (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). So SOM clustered 3 ALL samples with the AML class and 1 AML sample with ALL. A drawback of the SOM, however, is its lack of statistical interpretation and the uncertainty measurement of the results.

The dataset contains the expression levels of 6817 genes for 38 patients. There are actually 7129 probe sets - controls and gene redundancies bringing the 6817 up to 7129. The patient samples are known to come from two distinct classes of leukemia: 27 are ALL and 11 are AML. We submitted all the data (controls & redundancies) to PCA using the S-Plus function. Recognizing that the total number of observations is very small, we chose to output the first

eight eigenvectors from the PCA and use the BCVS with the anchor mode. The BCVS chose the first 7 eigenvectors to include in the mixture modeling, and reported two classes: Class 1 contained 27 samples, 26 of which were ALL samples and Class 2 contained 11 samples, 10 of which were AML samples. This result is slightly better than that obtained by Golub *et al.* (1999) using SOM, illustrating that PCA plus BCVS produces competitive results for clustering. More importantly, in comparison with SOM, the PCA and Gaussian mixture models possess clear statistical interpretations and well-established mathematical properties, enabling the investigator to think further about the modeling issues.

6. DISCUSSIONS

Based on Gaussian mixture models, we propose a novel Bayesian method BCVS for clustering high-dimensional data. The new method has the following features: (a) factors informative to the clustering model are automatically selected; (b) transformations of the factors can be selected in an automatic and principled way; and (c) the new method combines the PCA with the formal Bayesian modeling. We have shown by simulation that the variable selection step in BCVS can significantly improve the clustering result especially when the number of factors in consideration is high. We have also shown by a few real applications that the BCVS produced as good or better results than the popular hierarchical clustering method.

What is lacking in our current method is a way to determine the total number of clusters for the mixture model. Conceptually, one can just give a prior for the clustering variables (total cluster number and memberships) and then proceed with the MCMC machinery. For example, a popular prior for clustering indicators is that derived from the Dirichlet process (Neal 2000), which is most conveniently described conditionally: given the clustering of the first i observations, the $i+1$st observation can join an existing cluster of size n_j with probability $n_j/(q+i)$, and form a new cluster of its own with probability $q/(q+i)$. The prior expectation of the total number of clusters in this model is $O(log(n))$, which may not be desirable in practice. Qin *et al.* (2002) proposed a modification: the $i+1$st observation joins an existing cluster with probability $1/(q+c_i)$ and forms a new cluster with probability $q/(q+c_i)$, where c_i is the total number of clusters formed by the first i observations. This prior gives an expected number of clusters of $O(\sqrt{n})$. One can also prescribe a prior for the clustering variable, as suggested by Richardson and Green (1997), by giving first a distribution on the number of clusters and then a distribution for the memberships of the n objects. A potential problem with this line of approach is the additional computation cost.

Another possible avenue to achieve the automatic selection of cluster number is to treat it as a model selection problem and use the Bayesian information criterion (BIC) to determine the number of Gaussian components (Yeung *et al.* 2001). Although this approach is not directly based on a model, it is much cheaper computationally and gives satisfactory result in general. However, with the additions of variable and transformation selection variables, the BIC in Yeung *et al.* (2001) needs to be revised to suit for the new variable selection task.

ACKNOWLEDGEMENTS

This research was supported partly by NSF Grants DMS-0094613 and DMS-0204674 to JSL and the NIH Grant R21RR14036 and NSF Grant DBI-9515518 to CEL. We thank Epaminondas Sourlas in Harvard statistics department for some computational assistance.

REFERENCES

Alizadeh, A. A. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

Banfield J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations *J. Roy. Statist. Soc. B* **26**, 211–246, (with discussion).

Brooks, S. P. (2001). On Bayesian analyses and finite mixtures for proportions. *Statist. Comput.* **11**, 179–190.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recogn.* **28**, 781–793.

Chen, R. and Liu, J.S. (1996). Predictive updating methods with applications in Bayesian classification. *J. Roy. Statist. Soc. B* **58**, 397–415.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56** 363–375.

Fraley, C. and Raftery, A.E. (1999). Mclust: software for model-based cluster analysis. *J. Classification* **16**, 297–306.

Frank, J. (1996). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies.* New York: Academic Press.

Frank, J. and van Heel, M. (1982). Correspondence analysis of aligned images of biological particles. *J. Mol. Biol.* **161**, 134–137.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: The 23rd symposium on the interface* (E. Keramigas, ed.). Fairfax: Interface Foundation, 156–163.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275–286.

Golub, T. R., Slonim, D. K., Tamayo, P. *et al.* (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**, 531–537.

Ishwaran, H., James, L. F., and Sun J. Y. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96**, 1316–1332.

Kim, H. C., Kim, D. J., and Bang, S. Y. (2002). A numeral character recognition using the PCA mixture model. *Pattern Recogn. Lett.* **23**, 103–111.

Liu, J. S. (1994). The collapsed Gibbs sampler with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89**, 958–966.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing.* New York: Springer.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

Moss, S., Wilson, R.C., and Hancock, E.R. (1999). A mixture model for pose clustering. *Pattern Recogn. Lett.* **20**, 1093–1101.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models J. Comp. Graph. Statis. **9**, 249–265.

Qin, Z. S., McCue, L. A., Thompson, W., Mayerhofer, L., Lawrence, C.E. and Liu, J. S. (2002). Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Tech. Rep.*, Harvard University, USA.

Radermacher, M., Wagenknecht, T., Verschoor, A., and Frank J. (1987). Three-dimensional reconstruction from a single-exposure random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. *J. Microsc.* **146**, 113–136.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–758.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945.

Samsó, M., Palumbo, M. J., Radermacher, M., Liu, J. S., and Lawrence, C. E. (2002). A Bayesian Method for Classification of Images from Electron Micrographs. *J. Struct. Biol.*, (to appear).

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Yeung, K. Y., Fraley, C., Murua, A. *et al.* (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

APPENDIX: BAYESIAN MULTIVARIATE GAUSSIAN INFERENCE

For the consistency of notations and the self-containedness of the article, we here present the standard conjugate Bayesian analysis with Gaussian observations. More details can be found in, for example, Gelman *et al.* (1995). Suppose n iid realizations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are observed from the k-dimensional Gaussian distribution $N(\boldsymbol{\mu}, \Sigma)$. It is of interest to make inference on $\boldsymbol{\mu}$, Σ, and a future observation $\boldsymbol{x}_{n+1}$.

The prior distribution for Σ is the Inverse-Wishart distribution, Inv-W$_{\nu_0}(S_0^{-1})$, which is of the form:

$$p_0(\Sigma) = c_0 |\Sigma|^{-\frac{\nu_0+k+1}{2}} e^{-\frac{1}{2}tr(\Sigma^{-1}S_0)}, \tag{8}$$

where c_0 is the normalizing constant, k is the dimensionality, and ν_0 and S_0 are two hyper-parameters to be given by the user. Conditional on Σ, the prior of $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} \,|\, \Sigma \sim N(\boldsymbol{x}_0, \Sigma/\rho_0). \tag{9}$$

To make our later analysis more convenient, we define the normalizing function

$$Z(\nu, S, k) = |S|^{\frac{\nu}{2}} \left\{ 2^{\frac{\nu k}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^{k} \Gamma\left(\frac{\nu+1-i}{2}\right) \right\}^{-1}. \tag{10}$$

Then $c_0 = Z(\nu_0, S_0, k)$.

In order to make p_0 a proper distribution, we need to have $\nu_0 > k-1$ and $|S_0| > 0$. With this distribution and $\nu_0 > k+1$, we have $\mathrm{E}(\Sigma) = S_0/(\nu_0 - k - 1)$.

If we have n iid observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from $N(\boldsymbol{\mu}, \Sigma)$, then data-parameter joint distribution is

$$\begin{aligned}
P(\boldsymbol{X}, \boldsymbol{\mu}, \Sigma) &= c_0 (2\pi)^{-\frac{(n+1)k}{2}} |\Sigma|^{-\frac{n+k+\nu_0+2}{2}} \rho_0^{\frac{k}{2}} \\
&\times \exp\left[-\frac{1}{2} \{ tr(\Sigma^{-1}S_0) + \rho_0 (\boldsymbol{x}_0 - \boldsymbol{\mu}) \Sigma^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu})^T + \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}) \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \} \right] \\
&= c_0 (2\pi)^{-\frac{(n+1)k}{2}} |\Sigma|^{-\frac{n+k+\nu_0+2}{2}} \rho_0^{\frac{k}{2}} \\
&\times \exp\left[-\frac{1}{2} \{ tr\{\Sigma^{-1}(S_0 + SS)\} + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^*) \left(\frac{\Sigma}{n+\rho_0} \right)^{-1} (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^*)^T \} \right]
\end{aligned}$$

where

$$SS = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) + \frac{n\rho_0}{n+\rho_0} (\bar{\boldsymbol{x}} - \boldsymbol{x}_0)^T (\bar{\boldsymbol{x}} - \boldsymbol{x}_0) \tag{11}$$

and

$$\bar{\boldsymbol{x}}^* = \frac{n}{n+\rho_0} \bar{\boldsymbol{x}} + \frac{\rho_0}{n+\rho_0} \boldsymbol{x}_0. \tag{12}$$

Thus, after integrating out $\boldsymbol{\mu}$, we have

$$\begin{aligned}
P(\boldsymbol{X}, \Sigma) &= c_0 (2\pi)^{-\frac{nk}{2}} \rho_0^{\frac{k}{2}} (n+\rho_0)^{-\frac{k}{2}} |\Sigma|^{-\frac{n+k+\nu_0+1}{2}} \exp\left[-\frac{1}{2} tr\{\Sigma^{-1}(S_0 + SS)\} \right] \\
&= c_0 (2\pi)^{-\frac{nk}{2}} \rho_0^{\frac{k}{2}} (n+\rho_0)^{-\frac{k}{2}} c_1^{-1} \text{Inv-W}_{n+\nu_0}\{(S_0 + SS)^{-1}\}
\end{aligned}$$

where

$$c_1 = Z(n + \nu_0, S_0 + SS, k).$$

Hence, the model likelihood $P(\boldsymbol{X})$ is of the form

$$P(\boldsymbol{X} \mid n, k, \lambda_0) = \frac{Z(\nu_0, S_0, k)}{Z(n + \nu_0, S_0 + SS, k)} (2\pi)^{-\frac{nk}{2}} \left(\frac{n + \rho_0}{\rho_0} \right)^{-\frac{k}{2}}. \tag{13}$$

Marginally, the distribution of $\boldsymbol{\mu}$ is

$$[\boldsymbol{\mu} \mid \boldsymbol{x}_1, \ldots \boldsymbol{x}_n] \sim t_{n+\nu_0-k+1} \left(\bar{\boldsymbol{x}}^*, \frac{S_0 + SS}{(n + \rho_0)(n + \nu_0 - k + 1)} \right).$$

The predictive distribution of a new observation $\boldsymbol{x}$ is

$$[\boldsymbol{x} \mid \boldsymbol{x}_1, \ldots \boldsymbol{x}_n] \sim t_{n+\nu_0-k+1} \left(\bar{\boldsymbol{x}}^*, \frac{(n + \rho_0 + 1)(S_0 + SS)}{(n + \rho_0)(n + \nu_0 - k + 1)} \right). \tag{14}$$

DISCUSSION

ADRIAN E. RAFTERY *((University of Washington, USA)*

1. Introduction. It is a pleasure to congratulate the authors on a paper that proposes promising solutions to several outstanding problems in model-based clustering. I prefer the term *model-based clustering* to *Bayesian clustering*, because many of the references that the authors cite as examples of Bayesian clustering in their Section 1 are not specifically Bayesian, but are model-based, and because most clustering done in practice is heuristic rather than model-based. That is the big dichotomy in the clustering literature, rather than the Bayesian-frequentist one. For recent reviews of the literature on this topic, see Fraley and Raftery (2002) and McLachlan and Peel (2000).

I will start by summarizing the main features of the paper and highlighting what I see as some of its most important contributions. I will then address the issue of whether dimension reduction and clustering should be done separately (as the authors do), or together, and suggest that simultaneous solutions are possible. I will also discuss the nature of the hierarchical agglomerative clustering that the authors used as a comparison method. Finally, I will discuss the relative advantages and disadvantages of maximum likelihood via the EM algorithm, and fully Bayesian estimation via MCMC, for model-based clustering.

2. Highlights of the Paper. The paper addresses the important problem of clustering high-dimensional data, i.e. $n \times p$ data matrices containing J groups, where the dimension, p, is large relative to the number of data points, n. The work is motivated by three examples: clustering of electron microscope images, where $n = 400$, p is in the tens of thousands, and $J = 4$; lymphoma gene expression data, with $n = 96$, $p = 4,026$ and there are 2 groups; and leukemia gene expression data, with $n = 38$, $p = 6,817$ and there are 2 groups.

The method proposed is as follows, in brief. The number of groups, J, is taken to be known.

1. Extract the first k_0 principal components. In the examples, $k_0 = 8$ or 16.
2. Fit a mixture of multivariate normal distributions to a subset of the k_0 principal components. The covariance matrices are taken to be unconstrained.
3. The subset of the principal components to be used is treated as a parameter, and estimated using the Markov chain Monte Carlo model composition (MC^3) algorithm of Madigan and York (1995). This is a Metropolis-Hastings algorithm over the discrete space of possible

subsets of the principal components, in which the mixture model parameters are integrated out analytically.

4. Taking the first k principal components, rather than an arbitrary subset, is an option, called *anchor mode*.
5. Bayes factors can be used to choose a Box-Cox transformation of the variables.

A simulation study was carried out, in which $n = 200$, $p = 9$ and there were two groups. This gave good results, but in a situation very different from the motivating examples. The method was applied to the motivating data sets, with encouraging results.

The two biggest contributions of the paper are methods for variable selection and for the selection of transformations in model-based clustering. The results show clear gains when variable selection is used in clustering, over the approach when all the variables are used. This is similar to the situation in regression, where it has been shown that one gains by doing variable selection or model averaging in a Bayesian context (e.g. Hoeting *et al.* 1999; Clyde 1999). The method proposed for selecting transformations makes a lot of sense.

3. Dimension Reduction and Clustering: Together or Separate?. The authors first perform principal component analysis on the full data set. They then select the first k_0 principal components, where k_0 is very small relative to the total number of principal components, and they then do clustering on the resulting reduced data set. This is simple and appealing, and gives good results in their simulations and examples. It is worth pointing out that this is actually a very well known general approach in the clustering literature and goes back a long way (e.g. Chien 1978; Everitt 1974; Schnell 1970; Tyron and Bailey 1970). The originality of the present approach is that after the first k_0 principal components have been extracted, further variable selection is carried out simultaneously with the clustering.

However, Chang (1983) has shown that the practice of reducing the data to the first principal components before clustering is not justified in general. He showed that the principal components with the larger eigenvalues do not necessarily contain the most information about the cluster structure, and that taking a subset of principal components can lead to a major loss of information about the groups in the data. Chang demonstrated this theoretically, by simulations, and also in applications to real data. Similar results have been found by other researchers, including Yeung and Ruzzo (2001) for clustering gene expression data, and Green and Krieger (1995) for market segmentation.

This point is illustrated in Figure 5. This is a simulated two-dimensional data set in which there are two clear groups. The first principal component is the diagonal line with equation roughly $y = -x$ that separates the two groups. This first principal component accounts for about 90% of the variance, so consideration of the eigenvalues of the covariance matrix would often lead to collapsing the data to the first principal component only. But if this were done, it is clear that all the cluster information would be lost. In fact, all the cluster information is contained in the second principal component, which accounts for only a small proportion of the variance.

Of course, Liu *et al.* do not use only one principal component; they use $k_0 = 8$ or 16. But they are reducing data with thousands of dimensions to 8 or 16; how can we be sure that something similar to Figure 1 is not happening, on a much larger scale? In their examples, this does not seem to be the case, but the other papers I have cited provide evidence that this can happen in practice.

This also suggests that the choice of k_0 is crucial in practice, and I wonder how the authors propose choosing it so as to avoid problems such as those I have mentioned.

4. Simultaneous Parameter Reduction and Clustering: Is it Possible? It is easy to criticize strategies such as those of the authors, but dimension reduction of some sort is clearly necessary

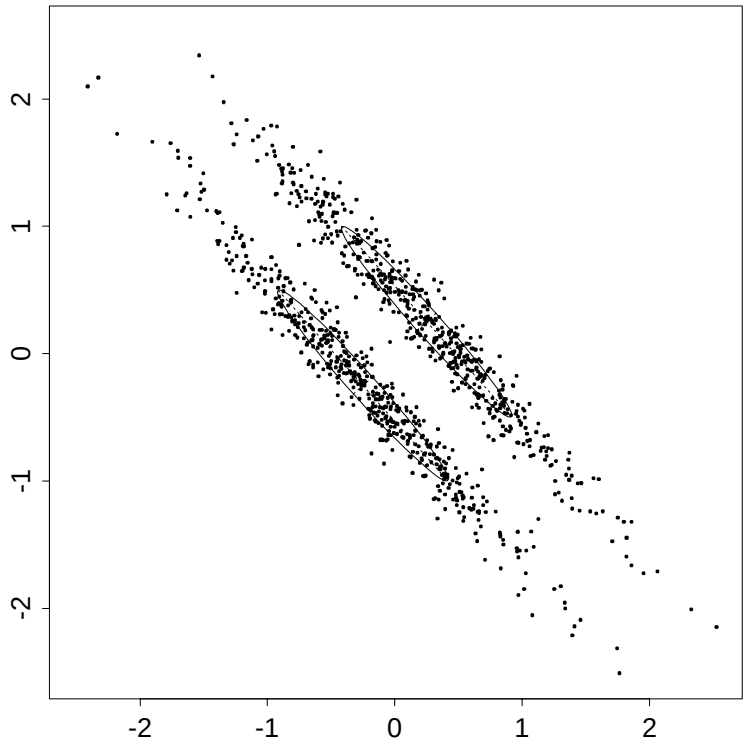

Figure 5. *Two Groups in Two Dimensions. All cluster information would be lost by collapsing to the first principal component. The principal ellipses of the two groups are shown as solid curves.*

in these very high-dimensional problems. The authors' strategy of reducing the data to the first few principal components is commonly used in high-dimensional clustering problems. So the question is, what else can be done? Directly clustering data of such high dimension seems hopeless.

Some initial dimension reduction may be necessary and easy to do in problems like this. For example, in gene expression data, eliminating the nonexpressed genes is standard practice, and this can reduce the number of dimensions from around 6,000 to a few hundred in typical examples. In medical image data, many of the pixels are of no interest because they belong to the background or to broad features, and eliminating them can reduce the effective dimension of the problem by 90% or more.

A general alternative strategy is to keep all the remaining dimensions in the analysis, but to use more parsimonious models for the covariance matrices in the multivariate normal mixture model. Various such parsimonious models have been proposed, and they allow surprising flexibility, as well as huge reductions in the number of parameters needed.

Perhaps the simplest such model is the diagonal covariance, or naive Bayes model, in which $\Sigma_j = \text{diag}(v_{1j}, \ldots, v_{pj})$. This is quite parsimonious relative to the Liu et al model. For example, for a 400-dimensional data set (typical of gene expression microarrays when nonexpressed genes have been removed), it has about the same number of parameters as the Liu et al model with $k_0 = 18$ principal components. It is also quite flexible: it allows different volumes and shapes for each cluster, and also different orientations. The main restriction is that the orientations are axis-aligned. The big advantage here is that clustering is done simultaneously with parameter reduction. This model has performed well for clustering gene expression data (Yeung et al, 2001).

A more general, but still parsimonious set of models arises from the volume-shape-orientation decomposition of Banfield and Raftery (1993), in which $\Sigma_j = \lambda_j D_j A_j D_j^T$. Here λ_j is a scalar that determines the volume of the jth cluster, A_j is a diagonal matrix of scaled eigenvalues determining its shape, and D_j is an orthogonal matrix determining its orientation. Each of the volume, shape and orientation may be constant across clusters, or allowed to vary between clusters.

If the orientation is the same across clusters, but the volume and shape are allowed to vary between clusters, a parsimonious but flexible model results. This generalizes the diagonal or naive Bayes model: the cluster orientations are still axis-aligned, but the axes are rotated, and the rotation is determined simultaneously with the clustering. The diagonal model would not work well in the situation of Figure 1, but this generalization would work very well.

Another parsimonious principal component-based approach is the mixture of probabilistic principal component analyzers of Tipping and Bishop (1999). This allows a different set of principal components for each group, and these are estimated simultaneously with the clustering. This seems likely to achieve much of the parsimony of the Liu et al approach, but without the potential disadvantages pointed out by Chang (1983) and others. This is related to the mixture of factor analyzers approach of Ghahramani and Hinton (1997); see McLachlan and Peel (2000) for a review of these approaches.

Finally, I note that using a time series covariance structure for Σ_j may be very parsimonious and quite appropriate for some high dimensional situations that have a sequential structure, such as chemical spectra. The obvious covariance structures of this kind are those that arise from autoregressive-moving average (ARMA) models, and these are highly parsimonious. Model-based clustering has been applied with some success to the clustering of chemical spectra by Wehrens, Simonetti and Buydens (2002), but without using the specifically sequential nature of the data. It seems possible that exploiting this aspect of the data more fully might lead to more efficient and even more successful clustering in such applications.

5. The Hierarchical Agglomerative Clustering Straw Man. Liu *et al.* compare their methods with complete link hierarchical agglomerative clustering based on Euclidean distances. They refer to this as hierarchical agglomerative clustering, or HAC, but it is worth noting that this is only one of many possible kinds of HAC. It is related, although vaguely, to model-based clustering with the the model $\Sigma_j = \lambda I$, i.e. spherical, equal-volume clusters. This is likely often be to an inappropriate model for the data.

It is possible to carry out *model-based* hierarchical agglomerative clustering. One just uses the likelihood as merging criterion at each stage (Banfield and Raftery 1993). This has been used recently with considerable success in the high-dimensional situation of text classification (Tantrum, Murua and Stuetzle 2002).

Thus BCVS's advantage over HAC in the authors' simulation study may be due either to (i) being Bayes, or (ii) using a better model (the unconstrained covariance model) for the components. Which of (i) or (ii) is correct is an empirical question, and is not answered by the paper. It could be addressed by carrying out model-based HAC, and comparing the result of this with BCVS. This can be done, for example, using the MCLUST software available at `www.stat.washington.edu/mclust`.

6. Label-Switching. The authors have laid out a Bayesian approach to the estimation of a mixture model using MCMC. As such, it would seem to be subject to the label-switching problem discussed, for example, by Richardson and Green (1997). This arises because one can change the labeling of the mixture components without changing the likelihood. Because there are J! labelings, it follows that there are J! components of the posterior distribution, which are identical except for the labeling, if the prior is symmetric with respect to labelings. This

has various perverse consequences: for example, if the MCMC sampler does truly explore the posterior distribution, the posterior means of the means of the mixtures components will all be the same. This problem is not easily diagnosed; for example, label switches do not necessarily correspond to sudden big jumps in the MCMC chain.

In the vanilla mixture model, various solutions to the label-switching problem have been proposed (Celeux, Hurn and Robert 2000; Stephens 2000), and these seem to work quite well.

I would be interested in the authors' comments on this problem. On the face of it, it seems as if it invalidates their method in its current form. I wonder if the authors feel that the approach could be rescued using some of the proposed solutions to the label-switching problem; could these be adapted to BCVS?

7. MLE via EM, or Bayes via MCMC?. Several of the groups active in research on model-based clustering now focus more on maximum likelihood estimation via the EM algorithm than on Bayesian estimation via MCMC. These include the Grenoble group (e.g. Celeux, Chaveau and Diebolt 1996), the Queensland group (e.g. McLachlan and Peel 2000), the Microsoft Research group (e.g. Cadez et al 2000), and our own group at the University of Washington. This is surprising, because of the appeal of the Bayesian framework and because several of these researchers had earlier adopted Bayesian MCMC approaches to the model-based clustering problem (e.g. Bensmail et al 1997). How can we explain this?

For estimation, the two approaches give similar results in many situations. Both have strengths and weaknesses. The EM approach is often simpler to implement, particularly when the components of the mixture are complex. The Bayesian approach requires the additional work involved in prior specification and the assessment of prior sensitivity, which may not seem very rewarding if similar results are obtained without using a prior. For example, Liu et al use the prior

$$\mu_j \sim N(x_0, 100\Sigma_j).$$

This seems extremely spread out. How sensitive is the posterior for the principal components chosen to this choice, for example to the choice of the constant 100? The Bayesian approach also suffers from the label-switching problem mentioned earlier.

The Bayesian approach also has advantages. In model-based clustering, the MLE of Σ_j can be degenerate, with zero or near-zero eigenvalues or diagonal elements, yielding infinite or near-infinite likelihoods, corresponding to small and/or highly linear clusters. By specifying a prior for Σ_k, a Bayesian approach can alleviate this by effectively smoothing the likelihood so that its many uninteresting infinite spikes are removed. Similarly, the Bayes estimates of posterior cluster membership probabilities, $p(J_j|x_i)$, take account of parameter uncertainty, and so are more accurate and less extreme towards 0 or 1.

But the biggest advantage of the Bayesian approach lies in model selection, model averaging and hypothesis testing, rather than estimation. Liu et al have shown this convincingly in their treatment of the selection of the principal components to be used. In this regard, the good performance of the anchor mode is striking. Do the authors have any thoughts on how general this is?

Throughout, the authors have taken the number of clusters to be known. They point out that a Bayesian approach could solve this also, and mention the use of BIC (Fraley and Raftery 2002; Yeung et al 2001) as a possibility. As they say, this is cheap computationally and has given good results in many applications. It could be adapted to variable and transformation selection as well. Its use for transformation selection might require the calculation of a more exact Bayes factor.

Steele (2002) has derived a unit information prior for model selection in mixtures, and in simulation studies he found that Bayes factors based on it gave similar performance to BIC. This

provides both some further justification for BIC, and a basis for possible further refinements of model selection methods in model-based clustering.

P. L. IGLESIAS and F. A. QUINTANA *(Pontificia Universidad Católica, Chile)*

This paper proposes a Bayesian clustering procedure after transforming the data via PCA. The number of factors to be included is considered random and therefore part of the inference problem. The goal is to search for homogeneous groups. Under BCVS the transformed vector via PCA is modelled as a mixture of gaussian distributions with Dirichlet-distributed weights and conjugate style priors for mean vectors and covariance matrices. We congratulate the authors for an excellent piece of work, which not only contains novel material, but also opens the door to a number of potentially fruitful research topics.

BCVS involves highly dimensional parameters and many variables. It is then natural to wonder about sensitivity on prior specification and robustness of the clustering method to departures from the gaussian mixture model. Concretely,

1. How would the clustering structure change under different choices of the prior $f(k)$ for the number of components? Would it be possible to elicit such prior in a fully Bayesian way, that is, without carrying out a pre-exploratory data analysis? Perhaps a natural alternative is to perform a reference analysis, which is not necessarily equivalent to a uniform prior on k.
2. Much of the structure in the gaussian mixture model is preserved for some elements in the class of multivariate elliptical distributions, without the need of using Box-Cox transformations. In this case, we anticipate some changes in the clustering structure. For instance, with heavy-tailed distributions we would expect some points to shift to different clusters, and maybe even less clusters than in the gaussian case.
3. An alternative way to produce clusters consists of using product partition models (PPM) after the k principal components are chosen. An advantage of this method is that the number of clusters J does not need to be known in advance.

DANIEL PEÑA *(Universidad Carlos III de Madrid, Spain)*

This is a very interesting paper with many insights and I would regret that the method proposed in the paper may fail because the initial step of the procedure. Working with the first k_0 principal components of the data may lead to loose very useful information, as the directions of maximum variability need not be the directions most useful for clustering.

An interesting direction for clustering is one in which the projected points cluster around well separated different means. Note that a univariate sample of zero-mean variables of size n will have maximum separation in two groups when it is composed of $n/2$ points equal to $-a$ and $n/2$ points equal to a, and then the kurtosis coefficient of the sample will take its minimum value, equal to one. Thus, directions of minimum kurtosis coefficient seems useful to show possible clusters.

Another interesting direction for clustering will be the one in which the majority of the points cluster around a common mean, but there are some small groups of points at both sides of the main group. In particular we may have a central group and some outliers. In this case, the kurtosis coefficient of the distribution of the data will be large. Thus, interesting directions for clustering are those in which the projected data have either a small or large kurtosis coefficient and a powerful cluster method for high dimensional data can be obtained by projecting the data in the directions with maximum or minimum kurtosis coefficient (see Peña and Prieto, 2001). It can be shown that this method has some optimal properties when the data have been generated by mixtures of elliptical distributions. For instance, when the data are generated by a mixture

of two normal distributions with the same covariance matrix the direction which minimizes the kurtosis of the projection is the Fisher linear discriminant function (Peña and Prieto, 2000).

Given these results, I would suggest to the authors that instead of selecting the first k_0 principal components select those components in which the projected data have the largest or the smallest kurtosis coefficient. Alternatively, instead of working with principal components, they can directly take as variables the projections of the original data on the $k_0/2$ orthogonal directions of maximum and minimum kurtosis coefficient.

CHRISTIAN P. ROBERT *(CEREMADE, Université Paris Dauphine,France)*

It is quite exciting to track the growing importance of mixture modelling in the Bayesian literature and, in particular, at the *Valencia 7* meeting. This paper is an exemplary illustration of the versatility of mixture modelling, since this modelling applies to many areas from clustering to nonparametric settings. My comments will mainly be limited to relevant works in the area, although I first want to point out that the computation of the Bayes factor (5) using improper priors is invalid, *stricto sensu* (Robert, 2000).

First, the authors rightly perceive the relevance of using *Principal Components* (PC) to reduce the dimension of the model and they introduce the compelling idea of *anchor*. To my opinion, using PC is an almost compulsory step as fitting a mixture model to a high dimensional dataset will almost certainly lead to a large number of components, unless some strong and well-documented structure is available. During the talk, I was wondering, however, how the exploratory technique of PC (and the concept of anchor) could be embedded within a more Bayesian Decision Theory framework.

Second, the authors note that integrating the parameters out lead to improved performances of the sampler. There is more to this point: we showed in Casella, Robert and Wells (2000) that this integration provides an easier exploration of the space of the missing variables (or of the corresponding sufficient statistics), and exhibited a quite peculiar phenomenon of *concentration*. In the cases we studied, it indeed appears that the marginal posterior distribution of the sufficient statistics, $(\sum_i I_j(J_i), \sum_i I_j(J_i)\, x_i, \sum_i I_j(J_i)\, x_i x_i^T)$ say, is highly concentrated on a few values, and thus that partitionned sampling is very helpful in uncovering the important parts of the missing variables space.

Third, Celeux, Hurn and Robert (2000) also implemented tempering in this setting, following Neal (1996). The algorithm in §3.2 of our paper is essentially the same as the *Tempering Sampler* of Liu *et al.* , with the difference that the level exchange is operated at every step of our algorithm. The motivation for using tempering there was to evacuate the *label switching* problem that plagues mixture posterior simulation, particularly in higher dimensions.

REPLY TO THE DISCUSSION

We thank all the discussants for their exciting ideas, sharp questions, and knowledges in the area of model-based clustering. As Raftery rightly pointed out, many of the methods we labeled as "Bayesian" are in fact EM-based likelihood method based on a complete statistical model. To us, there is no major distinction between the two approaches except that the complete Bayesian approach often wins in giving one the full inference. Besides providing a principled approach to model selection, the Bayesian procedure is also more flexible in incorporating complex structures, such as the one for Box-Cox transformations.

Although the EM-based approach appears to have avoided the troubles of prior specifications and "label-switching" nuisance, it does so by sacrificing its inference power — now the inference has to be based on certain asymptotic results which may not hold particularly well for mixture model fitting. Being able to assess the sensitivity to prior specifications should be seen as a

virtue of the Bayes approach, although it might give the researchers (us) some additional hassles. In other words, the EM approach in a way only hides the problems, not really solves them. If the Bayesians are willing to take an asymptotic approximation in the place of the (perhaps) more accurate MCMC computations, they will not have any label-switching problem at all — asymptotically the posterior modes corresponding to different labeling can never communicate. It is indeed true that in the finite-sample case the label-switching problem causes the Bayesians some headache, but we have not yet seen a problem where this difficulty causes any material damage. Furthermore, a cheap and effective way to get around the issue is to artificially impose an order among the clusters (e.g., based on the ordering of the first component of the mean vector of each cluster).

Professor Robert objected to our use of improper priors in transformation selections. Although what Robert pointed out is a well-known and general phenomenon for Bayesian model selections, our situation is slightly different. As explained in Section 4, the ratio of the Bayes factors needed for the transformation selection is perfectly proper, as long as the priors for the parameters in the transformed and untransformed models are "equally" noninformative. One can probably also calibrate these priors to make them "equally" informative and proper. Using noninformative priors here, however, makes practical sense: if we are not sure whether we should impose a Gaussian model for the original data or the logarithm of the data, it is very likely that we are completely ignorant about the mean and variance parameters of the corresponding model.

Both Raftery and Robert pointed out the "incompatibility" and potential dangers of using the exploratory PCA method together with the general Bayesian framework. Although we agree with Robert that a coherent Bayesian model encompassing both PCA and mixture modeling would be a more aesthetic thing to do, we are skeptical whether such a full Bayesian paradigm exists that can completely dominate the approach we took. Regarding the counter-example shown by Raftery, we feel that if distinctions among different clusters cannot be discerned with the first 10 to 15 principle components, it is perhaps more fruitful for us to re-examine the underlying science and revise the model accordingly than to search for a more omnipotent model. Nevertheless, the suggestions made by Professors Peña and Raftery are very interesting and worth further exploration. In particular, Peña's idea of choosing judiciously additional "special" directions according to either kurtosis or other statistics can be an important addition to the current BCVS framework.

Microarray analysis provides for statisticians an excellent entry point to bioinformatics/ computational biology. We would like to take this opportunity to mention a few other important bioinformatics problems in which statistics is likely to play an important role. These include, by no means exclusively, the protein folding prediction, multiple sequence alignments, transcription factor binding-sites identification, gene regulatory network constructions, evolutionary analysis, and the analyses of single nucleotide polymorphisms (SNPs) in the human genome. Some related references can be found in Liu (2002).

The prediction of protein tertiary structures is of great importance to drug designs and the basic biochemistry. Although many proteins' structures have been worked out by X-ray crystallographers, these only account for a small part of the protein universe. Multiple sequence alignment is still the main tool for protein or DNA sequence analysis, which has been at the center of computational biology for about 30 years. With the completion of the human genome and genomes of many other species, the task of organizing and understanding the generated sequence data through multiple alignment becomes even more pressing and challenging. The control of genes' expressions is fundamental to cell survival, growth, and differentiation. An important form of control is exerted by interfering with genes' transcriptions by specialized

proteins called the transcription factors, which recognizes short DNA segments in front of genes. Predicting novel transcription factor binding sites and deciphering genetic networks are all important steps towards the general goal of understanding gene regulation. As the great evolutionist T. Dobzhansky pointed out: nothing in biology made sense except in the context of evolution. Evolution study can not only help us understand where and how we come from, but also shed light on protein functions and cellular processes. The SNPs refer to frequently occurred (one in 1000 bases) single-base variations among the genomes of different individuals. Because of the availability of high through-put SNPs detection and analysis tools and their great potential in mapping genes responsible for complex diseases, the SNPs have recently attracted much attention from scientists. Statistical modeling and computation are crucial to fully realize the power of SNPs.

Now is clearly an exciting time for statisticians, especially Bayesian statisticians. We are challenged by many important biological and other scientific problems through massive amount of data, which are often associated with in-depth subject knowledges; yet, we are equipped with the powerful Bayesian modeling machine and unprecedented computational tools, such as MCMC and EM algorithms, and computer power. We hope that our paper and the discussions can get some of the readers interested in looking into a broader array of bioinformatics problems.

ADDITIONAL REFERENCES IN THE DISCUSSION

Bensmail, H., Celeux, G., Raftery, A.E. and Robert, C.P. (1997). Inference in model-based cluster analysis. *Statistics and Computing* **7**, 1–10.

Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. (2000). Visualization of navigation patterns on a web site using model-based clustering. *Tech. Rep.*, Microsoft Research, Redmond, USA.

Casella, G., Robert, C.P. and Wells, M.T. (2000). Mixture Models, Latent Variables and Partitioned Importance Sampling. *Tech. Rep.*, CREST, France.

Celeux, G., Chaveau, D. and Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation* **55**, 287–314.

Celeux, G., Hurn, M. and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–970.

Chang, W.-C. (1983). On using principal components before separating a mixture of tow multivariate normal distributions. *Appl. Statist.* **32**, 267–275.

Chien, Y.Y. (1978). *Interactive Pattern Recognition.* New York: Marcel Dekker.

Clyde, M. (1999). Bayesian model averaging and model search strategies (with discussion). *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 157–185.

Everitt, B.S. (1974). *Cluster Analysis*. New York: Wiley.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611–631.

Ghahramani, Z. and Hinton, G.E. (1997). The EM algorithm for factor analyzers. *Tech. Rep.*, University of Toronto, Canada.

Green, P.E. and Krieger, A.M. (1995). Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society* **37**, 221–239.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statist. Sci.* **14**, 382–417. Corrected version at
`http://www.stat.washington.edu/www/research/online/hoeting1999.pdf`.

Liu, J.S. (2002). Bioinformatics: Microarrays Analyses and Beyond. *Amstat News*, September, 59–67.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

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

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statist. Computing* **4**, 353–366.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–758.

Robert, C.P. (2000). *The Bayesian Choice*. New York: Springer.

Peña, D. and Prieto, F.J. (2000). The kurtosis coefficient and the linear discriminant function. *Statistics & Probability Letters* **49**, 257–261.

Peña, D. and Prieto, F.J. (2001). Clustering by projections. *J. Amer. Statist. Assoc.* **96**, 1433–1445.

Schnell, G.D. (1970). A phenetic study of the suborder *Lari (Aves)* I. Methods and results of principal components analysis. *Systematic Zoology* **19**, 35–57.

Steele, R. (2002). *Practical Importance Sampling Methods for Mixture Models and Missing Data Problems*. Ph.D. Thesis, University of Washington, Seattle, USA.

Stephens, M. (2000). Dealing with label-switching in mixture models. *J. Roy. Statist. Soc. B* **62**, 795–809.

Tantrum, J.M., Murua, A. and Stuetzle, W. (2002). Hierarchical model-based clustering of large data sets through fractionation and refractionation. *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD02)*, 183–190.

Tipping, M.E. and Bishop, C.M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation* **11**, 443–482.

Tyron, R.C. and Bailey, D.E. (1970). *Cluster Analysis*. New York: McGraw-Hill.

Wehrens, R., Simonetti, A.W. and Buydens, L.M.C. (2002). Mixture modelling of medical magnetic resonance data. *Journal of Chemometrics* **16**, 274–282.

Yeung, K.Y. and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.