Model-Based Clustering With Dissimilarities: A Bayesian Approach

Man-Suk OH and Adrian E. RAFTERY

A Bayesian model-based clustering method is proposed for clustering objects on the basis of dissimilarites. This combines two basic ideas. The first is that the objects have latent positions in a Euclidean space, and that the observed dissimilarities are measurements of the Euclidean distances with error. The second idea is that the latent positions are generated from a mixture of multivariate normal distributions, each one corresponding to a cluster. We estimate the resulting model in a Bayesian way using Markov chain Monte Carlo. The method carries out multidimensional scaling and model-based clustering simultaneously, and yields good object configurations and good clustering results with reasonable measures of clustering uncertainties. In the examples we study, the clustering results based on low-dimensional configurations were almost as good as those based on high-dimensional ones. Thus, the method can be used as a tool for dimension reduction when clustering high-dimensional objects, which may be useful especially for visual inspection of clusters.

We also propose a Bayesian criterion for choosing the dimension of the object configuration and the number of clusters simultaneously. This is easy to compute and works reasonably well in simulations and real examples.

Key Words: Hierarchical model; Markov chain Monte Carlo; Mixture models; Multidimensional scaling.

1. INTRODUCTION

Cluster analysis is the automatic grouping of objects into groups on the basis of numerical data consisting of measures either of properties of the objects, or of the dissimilarities between them. It was developed initially in the 1950s (e.g., Sneath 1957; Sokal and Michener 1958), and the early development was driven by problems of biological taxonomy and market segmentation.

Man-Suk Oh is Professor, Department of Statistics, Ewha Womens University, Seoul 120-750, Korea (E-mail: *msoh@ewha.ac.kr*; Web: *home.ewha.ac.kr/~msoh*). Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322 (E-mail: *raftery@stat.washington.edu*; Web: *www.stat.washington.edu/raftery*.

Journal of Computational and Graphical Statistics, Volume 16, Number 3, Pages 559–585
DOI: 10.1198/106186007X236127

For much of the past half-century, the majority of cluster analyses done in practice have used heuristic methods based on dissimilarities between objects. These include hierarchical agglomerative clustering using various between-cluster dissimilarity measures such as smallest dissimilarity (single link), average dissimilarity (average link) or maximum dissimilarity (complete link). These methods are relatively easy to apply and often give good results. However, they are not based on standard principles of statistical inference, they do not take account of measurement error in the dissimilarities, they do not provide an assessment of clustering uncertainties, and they do not provide a statistically based method for choosing the number of clusters.

Model-based clustering is a framework for putting cluster analysis on a principled statistical footing; for reviews see McLachlan and Peel (2000) and Fraley and Raftery (2002). It is based on probability models in which objects are assumed to follow a finite mixture of probability distributions such that each component distribution represents a cluster. The model-based approach has several advantages over heuristic clustering methods. First, it clusters objects and estimates component parameters simultaneously, avoiding well-known biases that exist when they are done separately. Second, it provides clustering uncertainties, which is important especially for objects close to cluster boundaries. Third, the problems of determining the number of components and the component probability distributions can be recast as statistical model selection problems, for which principled solutions exist. Unlike the previously mentioned heuristic clustering algorithms, however, model-based clustering requires object coordinates rather than dissimilarities between objects as an input. Thus, despite the important advantages of model-based clustering, it can be used only when object coordinates are given, and not when dissimilarities are provided.

In many practical applications in market research, psychology, sociology, environmental research, genomics, and information retrieval for the Web and other document databases, data consist of similarity or dissimilarity measures on each pair of objects (Young 1987; Schutze and Silverstein 1997; Tibshirani et al. 1999; Buttenfield and Reitsma 2002; Condon et al. 2002; Courrieu 2002; Elvevag and Storms 2003; Priem, Love, and Shaffer 2002; Welchew et al. 2002; Ren and Frymier 2003). Examples of such data include the co-purchase of items in a market, disagreements between votes made by pairs of politicians, the number of links between pairs of Web pages, the existence or intensity of social relationships between pairs of families, and the overlap of university applications by high school graduates.

Even when object coordinates are given, visual display of clusters in low-dimensional space is often desired since it may provide useful information about the relationships between the clusters and the underlying data generation process (Hedenfalk et al. 2002; Yin 2002; Nikkila et al. 2002). One way to reduce the dimensionality of objects for visual display in lower dimensional space is multidimensional scaling (MDS). In MDS, objects are placed in a Euclidean space while preserving the distance between objects in the space as well as possible. Reviews of the extensive literature on MDS were provided by Davison (1983), Young (1987), Borg and Groenen (2005), and Cox and Cox (2001).

There are many MDS techniques in the literature. Maximum likelihood MDS methods have been developed by Ramsay (1982), Takane (1982), MacKay (1989) and MacKay

and Zinnes (1986), among others. Least squares MDS methods have been proposed by DeLeeuw and Heiser (1982), Heiser and Groenen (1997), Groenen (1993), and Groenen, Mathar, and Heiser (1995) among others. Recently Oh and Raftery (2001) proposed a Bayesian MDS (BMDS) method using Markov chain Monte Carlo. This provided good estimates of the object configuration in the cases studied, as well as a Bayesian criterion for choosing the object dimension. However, they did not consider clustering and hence clustering has to be done separately with the estimated object configuration from MDS.

In this article, we develop a model-based clustering method for dissimilarity data. We assume that an observed dissimilarity measure is equal to the Euclidean distance between the objects plus a normal measurement error. We model the unobserved object configuration as a realization of a mixture of multivariate normal distributions, each one of which corresponds to a different cluster. We carry out Bayesian inference for the resulting hierarchical model using Markov chain Monte Carlo (MCMC). The resulting method combines MDS and model-based clustering in a coherent framework.

There are three sources of uncertainty in model-based clustering with dissimilarites: (a) measurement errors in the dissimilarities; (b) errors in estimating the object configuration; and (c) clustering uncertainty. Heuristic clustering methods that cluster directly from dissimilarities, such as hierarchical agglomerative clustering and self-organizing maps, do not take account of sources (a) and (c), while source (b) does not arise there. As an alternative, one may consider a two-stage procedure which estimates the object configuration in the first stage using an MDS method, and carries out model-based clustering in the second stage. Sequential application of most MDS methods and model-based clustering takes account of source (c) but not of sources (a) and (b). Sequential application of Bayesian MDS and model-based clustering considers sources (a) and (c), but separately rather than together; it does not consider source (b) when clustering. Heiser and Groenen (1997) proposed cluster difference scaling (CDS) which first finds the clustering and the cluster centers in a low-dimensional space, and then finds the object configuration from the cluster centers and the given dissimilarities. Cluster difference scaling considers sources (a) and (c) simultaneously but it does not consider source (b). In contrast, our approach accounts for all three sources of uncertainty simultaneously. Simultaneous estimation of the errors is important, because errors in the dissimilarity measures and/or the estimated configuration can affect the clustering and the clustering uncertainties, as we will show by example. We summarize the above mentioned clustering techniques and the sources of error they account for in Table 1.

Other important issues are the choice of the number of clusters and of the dimension of the objects. Oh and Raftery (2001) proposed an easily computed Bayesian criterion called MDSIC for choosing object dimension. We extend this to determine the number of clusters as well. The resulting criterion can be computed easily from MCMC output.

Section 2 describes our model for dissimilarities, the mixture model for the object configuration, and the prior distributions we use. Section 3 describes Bayesian estimation for this model using MCMC. The Bayesian criterion for choosing the dimension and the number of clusters is given in Section 4, while the method is applied to several simulated and real datasets in Section 5. Discussions are given in Section 6.

Table 1. Three sources of errors in clustering with dissimilarities.

Error sources	Heuristic clustering	MDS + MBC	BMDS + MBC	CDS	BMCD
Dissimilarity	no	no	yes	yes	yes
Object configuration	not applicable	no	no	no	yes
Clustering	no	yes	yes	yes	yes
Simultaneous consideration	no	no	no	partial	yes

2. MODEL FOR CLUSTERING WITH DISSIMILARITIES

Let δ_{ij} denote the dissimilarity measure between objects i and j, which is assumed to be functionally related to p unobserved attributes of the objects. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object i.

As in Oh and Raftery (2001), we model the true dissimilarity measure δ_{ij} as the distance between objects i and j in a Euclidean space, that is, $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$. The true dissimilarity measure can be different from Euclidean distance in practice, and there can be measurement error in the observations. We therefore assume that the observed dissimilarity measure, d_{ij}, is equal to the true measure, δ_{ij}, plus a Gaussian error. In addition, since dissimilarity measures are typically given as positive values we restrict the observed dissimilarity to be positive. Thus, given the Euclidean distance δ_{ij}, the observed dissimilarity measure d_{ij} is assumed to follow the truncated normal distribution

$$d_{ij} \sim N(\delta_{ij}, \sigma^2)\, I(d_{ij} > 0), \quad i \neq j, i, j = 1, \ldots, n. \tag{2.1}$$

Note that d_{ij} is related to $\mathbf{X} = \{\mathbf{x}_i\}$, called the object configuration, only through δ_{ij}. To represent clustering, we assume that the object configuration is a sample from a mixture of multivariate normal distributions,

$$\mathbf{x}_i \sim \sum_{j=1}^{G} \varepsilon_j\, N(\mu_j, T_j), \tag{2.2}$$

where each component normal distribution represents a cluster.

We use the following priors for the model parameters:

$$\begin{aligned} \sigma^2 &\sim \text{IG}(a, b), \\ (\varepsilon_1, \ldots \varepsilon_g) &\sim \text{Dirichlet}(1, \ldots, 1), \\ \mu_j &\sim N(\mu_{j0}, T_j), \\ T_j &\sim \text{IW}(\alpha, B_j), \end{aligned} \tag{2.3}$$

where $\text{IG}(a, b)$ is the inverse Gamma distribution with mode $b/(a+1)$ and IW is the inverse Wishart distribution.

One may use a more parsimonious covariance structure for T_j than the above unconstrained one. For instance, one may restrict T_j to be a diagonal matrix, or let $T_1 = \cdots =$

T_G, or use some other parsimonious covariance model such as those commonly used in model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 2002). In that case, the priors need to be modified accordingly.

3. POSTERIOR INFERENCE

3.1 MARKOV CHAIN MONTE CARLO

It is well known that inference for mixture models can be simplified with latent variables which indicate the group memberships of objects. We define latent variables K_i such that $P(K_i = j) = \varepsilon_j$ and $\mathbf{x}_i$ belongs to class j if $K_i = j$, so that

$$\mathbf{x}_i | K_i = j \quad \sim \quad N(\mu_j, T_j).$$

From the prior and the model, the full conditional posterior distributions (densities) given all the other unknowns are as follows:

$$\begin{aligned}
\pi(\mathbf{x}_i | K_i = j, \text{ others}) &\propto \exp\left[-1/2(\mathbf{x}_i - \mu_j)' T_j^{-1} (\mathbf{x}_i - \mu_j) - \frac{1}{2\sigma^2} \sum_{i>j} (\delta_{ij} - d_{ij})^2 \right] \\
&\quad \times \prod_{i>j} \Phi(\delta_{ij}/\sigma), \\
\pi(\sigma^2 | \text{ others}) &\propto (\sigma^2)^{-(m/2+a+1)} \exp\left[-\frac{1}{\sigma^2}(\text{SSR}/2 + b) \right. \\
&\quad \left. - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right) \right], \\
(\varepsilon_1, \ldots, \varepsilon_g | \text{others}) &\sim \text{Dirichlet}(n_1 + 1, \ldots, n_g + 1), \\
(\mu_j | \text{ others}) &\sim N\left(\frac{n_j \bar{x}_j + \mu_{j0}}{n_j + 1}, \frac{T_j}{n_j + 1} \right), \\
(T_j | \text{ others}) &\sim \text{IW}(\alpha + n_j/2, B_j + S_j/2), \\
P(K_i = j | \text{ others}) &= \phi(x_i; \mu_j, T_j) / \sum_{j=1}^{g} \phi(x_i; \mu_j, T_j),
\end{aligned}$$

where ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, $\text{SSR} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} (\delta_{ij} - d_{ij})^2$,

$$\begin{aligned}
n_j &= \sum_{i=1}^{j} I(K_i = j), \\
S_j &= \sum_{i=1}^{n} (x_i - \mu_j)(x_i - \mu_j)' I(K_i = j),
\end{aligned}$$

and I is the indicator function.

Iterative simulation of the unknown parameters from their full conditional distributions for a sufficiently long time yields samples of the parameters from the joint posterior distribution, and posterior inference can be done by using the samples. When a simpler covariance structure is used for T_j with appropriate priors, the algorithm can be easily modified since only the generation of T_j needs to be changed.

Simulation of samples from the full conditional distributions of the parameters $\{\varepsilon_j, \mu_j, T_j\}$ is straightforward since the full conditional posterior distributions all have convenient forms. However, the full conditional posterior distributions of $\mathbf{x}_i$ and σ^2 do not have closed forms, and so we apply the Metropolis-Hastings algorithm (Hastings 1970) to generate samples of $\mathbf{x}_i$ and σ^2. Oh and Raftery (2001) suggested an easy random walk Metropolis-Hastings algorithm for generating samples of $\mathbf{x}_i$ and σ^2 when $G = 1$. Given the latent indicator variable K_i, $\mathbf{x}_i$ follows a one-component multivariate normal distribution. Thus, we can easily modify the algorithm of Oh and Raftery (2001) for generating $\mathbf{x}_i$ from its full conditional posterior distribution in the mixture model. Given $\mathbf{X}$, the distribution of σ^2 does not depend on the mixture model parameters, so that the generation of σ^2 is the same in the mixture model as in the one-component model.

To initialize the MCMC algorithm, we first run Oh and Raftery's (2001) BMDS as a preliminary run to obtain an initial guess for $\mathbf{X}$, and then run the model-based clustering software MCLUST (Fraley and Raftery 1999, 2003) with this initial guess.

3.2 NONIDENTIFIABILITY

Euclidean distance is invariant under translation, rotation, and reflection of objects. Thus, the dissimilarity data provide information only about the relative locations of $\mathbf{X}$. In a Bayesian context, $\mathbf{X}$ is identified, strictly speaking, but the absolute location and orientation of $\mathbf{X}$ are defined only by the prior distribution, and in practice are very weakly identified. As a result, the relative positions of the $\mathbf{x}_i$ may have a tight posterior distribution, but their absolute positions will typically have a much more dispersed posterior distribution.

To get around this problem of weak identification, we use a Procrustean similarity transformation (Sibson 1979; Borg and Groenen 2005, chap. 19) which proceeds as follows: (i) Obtain an estimate, $\mathbf{X}^*$, of $\mathbf{X}$ from a preliminary run, for example, the MLE or the posterior mode. (ii) Transform the sample of $\mathbf{X}$ at each iteration of MCMC so that coordinates of $\mathbf{X}$ are as close as possible to the corresponding coordinates of $\mathbf{X}^*$, where the transformation is restricted to be a composition of some or all of a translation, a rotation, and a reflection. See the Appendix (p. 583) for more details. Since the transformation does not change the Euclidean distances between pairs of $\mathbf{x}_i$'s, it does not change the likelihood but it approximately fixes the location and orientation of samples of $\mathbf{X}$ so that $\mathbf{X}$ itself can be stably estimated.

There is another nonidentifiability problem. A mixture of density functions is invariant under arbitrary permutation of component labels. Thus, the posterior density function would be invariant under arbitrary permutation of component labels unless strong prior information is used (Stephens 2000). This may cause label switching during the MCMC iterations, hence typical averages of MCMC samples of the parameters may yield unrea-

sonable estimates of the mixture parameters. To avoid this problem, we adopt the relabeling procedure suggested by Celeux, Hurn, and Robert (2000) at each iteration of MCMC. See the Appendix for details.

By using the two postprocessing steps, Procrustean transformation, and relabeling, we obtain stable samples of the unknown parameters from which posterior estimates can be computed.

4. A BAYESIAN SELECTION CRITERION FOR CONFIGURATION DIMENSION AND THE NUMBER OF CLUSTERS

Posterior inference as described in the previous section presumed that the dimension, p, of the object configuration, and the number of clusters, G, are given. These are typically unknown, however, and we now propose a statistical method for choosing p and G. Oh and Raftery (2001) suggested a dimension selection criterion for MDS, called MDSIC, and found it to work well with Euclidean distance measures with small or moderate error size. In this section, we extend MDSIC and propose a new Bayesian selection criterion, named MIC, for choosing both p and G simultaneously.

We view the overall goal of our analysis as being to choose the best object configuration across the dimension p and the number of clusters G. We therefore base our model selection criteria on $\pi(\mathbf{X}_{pG}, p, G|D)$, the posterior density function of $\mathbf{X}$, p, G, given data D, evaluated at $\mathbf{X} = \mathbf{X}_{pG}$, where $\mathbf{X}_{pG}$ is the best object configuration given p and G.

Note that

$$\pi(\mathbf{X}_{pG}, p, G|D) \propto f(D|\mathbf{X}_{pG}, p, G)\pi(\mathbf{X}_{pG}, p, G),$$

where $f(D|\mathbf{X}_{pG}, p, G) = \int f(D|\mathbf{X}_{pG}, p, G, \sigma^2)\pi(\sigma^2)d\sigma^2$ and $\pi(\mathbf{X}_{pG}, p, G) = \int \pi(\mathbf{X}_{pG}, p, G, \Lambda)d\Lambda$ are the marginal likelihood and the marginal prior of $(\mathbf{X}_{pG}, p, G)$, respectively. Here we use $f(D|\cdots)$ to denote the sampling density of data D given specified parameter values. As p or G increase, the likelihood increases and it can be considered as a measure of fit. However, as p or G increases, the prior density decreases and it can be viewed as a penalty for more complex models.

Under equal prior probabilities for all p in $p_{\min} \le p \le p_{\max}$ and for all G in $G_{\min} \le G \le G_{\max}$,

$$\pi(\mathbf{X}_{pG}, p, G|D) \propto f(D|\mathbf{X}_{pG}, p, G)\pi(\mathbf{X}_{pG}|p, G). \tag{4.1}$$

Thus, one needs only to compute the marginal likelihood and the marginal prior of $\mathbf{X}$ for each p and G. Oh and Raftery (2001) showed that $f(D|\mathbf{X}_{pG}, p, G)$ is approximately proportional to $\mathrm{SSR}_{pG}^{-m/2+1}$, where $\mathrm{SSR}_{pG} = \sum_{i>j}(\delta_{ij}^{(pG)} - d_{ij})^2$ and $\delta_{ij}^{(pG)}$ is the Euclidean distance between the $\mathbf{x}_i$ and $\mathbf{x}_j$ of $\mathbf{X}_{pG}$.

Now consider computation of $\pi(\mathbf{X}_{pG}|p, G)$. We will suppress p, G in (4.1) for simplicity, so that $\pi(\mathbf{X}_{pG}|D) \propto f(D|\mathbf{X}_{pG})\pi(\mathbf{X}_{pG})$. The prior $\pi(\mathbf{X}_{pG})$ is not given in closed form and needs to be estimated. From the relationship

$$\pi(\mathbf{X}_{pG}) = \frac{\pi(\mathbf{X}_{pG}|\Lambda^*)\pi(\Lambda^*)}{\pi(\Lambda^*|\mathbf{X}_{pG})} \tag{4.2}$$

for a fixed value Λ^* of $\Lambda = (\varepsilon, \mu, T)$, $\pi(\mathbf{X}_{pG})$ can be estimated from an estimate of $\pi(\Lambda^*|\mathbf{X}_{pG})$. From Oh (1999) it can be shown that

$$\pi(\Lambda^*|\mathbf{X}_{pG}) = E\left[\pi(\varepsilon^*|K)\prod_{j=1}^{G}\pi(\mu_j^*|K, T_j^*, \mathbf{X}_{pG})\pi(T_j^*|K, \mu_j, \mathbf{X}_{pG})\right], \tag{4.3}$$

where the expectation is with respect to the joint distribution of (K, ε, μ, T) given $(\mathbf{X}_{pG})$. Since the conditional distributions of ε, μ_j, T_j are given in closed forms, $\pi(\Lambda^*|\mathbf{X}_{pG})$ can be easily estimated by using samples of (K, ε, μ, T) generated from the MCMC algorithm. In theory any value of Λ^* can be used, but in practice Λ^* close to the mode of Λ seems to work well from the efficiency point of view.

However, simple comparison of the posterior of $\mathbf{X}_{pG}$ can lead to the choice of large p because of the shrinking effect. As described by Oh and Raftery (2001) there is a shrinking effect as the dimension p increases, that is, the scale (dispersion) of $\mathbf{X}_{pG}$ tends to decrease as p increases without altering the fit. This would yield larger $\pi(\mathbf{X}_{pG})$ for larger p even when the likelihoods are the same and hence would favor larger p. To avoid this shrinking effect, Oh and Raftery (2001) suggested comparing configurations in the same dimensional space. Specifically, denoting $\mathbf{X}_{p1}$ by $\mathbf{X}_p$, they compared dimensions p and $p-1$ through $\mathbf{X}_p$ and $\mathbf{X}_p^* = (\mathbf{X}_{p-1} : 0)$, a $n \times p$ matrix whose first $p-1$ columns are equal to $\mathbf{X}_{p-1}$ and whose last column has all elements equal to 0. Note that $\mathbf{X}_p^*$ yields the same likelihood as $\mathbf{X}_{p-1}$ and it may be considered as an embedding of $\mathbf{X}_{p-1}$ in p-dimensional space.

Consider the case of $G = 1$. Since $f(D|\mathbf{X}_p^*) = f(D|\mathbf{X}_{p-1})$,

$$\begin{aligned}
\frac{\pi(\mathbf{X}_p|D)}{\pi(\mathbf{X}_p^*|D)} &= \frac{f(D|\mathbf{X}_p)\pi(\mathbf{X}_p)}{f(D|\mathbf{X}_p^*)\pi(\mathbf{X}_p^*)} \\
&= \left[\frac{f(D|\mathbf{X}_p)}{f(D|\mathbf{X}_{p-1})}\frac{\pi(\mathbf{X}_p)}{\pi(\mathbf{X}_{p-1})}\right]\left[\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)}\right] \\
&= \left[\frac{\pi(\mathbf{X}_p|D)}{\pi(\mathbf{X}_{p-1}|D)}\right]\left[\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)}\right].
\end{aligned}$$

Thus, $A_p = \frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)}$ is a correction factor to the posterior density ratio of $\mathbf{X}_p$ and $\mathbf{X}_{p-1}$ for the shrinking effect. Assume for a moment that $\mu_{10} = \mathbf{0}$ and $T_1 = \mathrm{diag}(t_1, \ldots, t_p)$, so that μ_1 has a multivariate normal prior with mean at the origin and a diagonal covariance matrix, as in Oh and Raftery (2001). Note that the restriction does not affect the MDS solution and hence does not affect the choice of p with $G = 1$. Then from Oh and Raftery (2001), with a unit information prior IG$(1/2, s_j^{(p)}/2)$ on the jth diagonal element of T_1,

$$-2\log A_p = -2\log\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)} = H_n - n\log\left(\frac{s_p^{(p)}}{n}\right) + \sum_{j=1}^{p-1}\log\left(r_j^{(p)}\right) + (n+1)\frac{\log(n+1)}{(n+r_j^{(p)})},$$

where $s_j^{(p)}$ is the jth diagonal element of $nS^{(p)} = \sum_{i=1}^{n}(\mathbf{x}_i^{(p)} - \bar{\mathbf{x}}^{(p)})(\mathbf{x}_i^{(p)} - \bar{\mathbf{x}}^{(p)})'$; $\mathbf{x}_i^{(p)}$ and $\bar{\mathbf{x}}^{(p)}$ are, respectively, the estimates of $\mathbf{x}_i$ from $\mathbf{X}_p$ and their mean, $r_j^{(p)} = s_j^{(p)}/s_j^{(p-1)}$

and $H_n = -(n+1)\log(\pi) + 2\log(\Gamma((n+1)/2))$. The shrinking effect is related only to p and not to G, so we use A_p for all values of G given the same p.

Now we propose a selection criterion, which we call MIC, as follows. Let

$$\begin{aligned}
\mathrm{MIC}_{1G} &= (m-2)\log \mathrm{SSR}_{1G} - 2\log \pi(\mathbf{X}_{1G}) \\
\mathrm{MIC}_{pG} &= \sum_{q=1}^{p} -2\log \frac{\pi(\mathbf{X}_{qG}|D)}{\pi(\mathbf{X}^*_{qG}|D)} \qquad (4.4) \\
&= \sum_{q=1}^{p} -2\log \frac{l(\mathbf{X}_{qG}|D)}{l(\mathbf{X}_{q-1,G}|D)} \frac{\pi(\mathbf{X}_{qG})}{\pi(\mathbf{X}_{q-1,G})} - 2\log A_q \qquad (4.5) \\
&= (m-2)\log(\mathrm{SSR}_{pG}) - 2\log \pi(\mathbf{X}_{pG}) - 2\sum_{q=1}^{p} \log A_q. \qquad (4.6)
\end{aligned}$$

Note that $(m-2)\log(\mathrm{SSR}_{pG})$ can be considered as a measure of fit, $-2\log \pi(\mathbf{X}_{pG})$ plays the role of a penalty for complexity, and $-2\sum_{q=1}^{p} \log A_q$ is a cumulative correction factor for the shrinking effect. The values of p and G that yield the minimum of MIC_{pG} are viewed as best.

5. EXAMPLES

We now apply the proposed method, which we call Bayesian Model-Based Clustering with Dissimilarities (BMCD), and the model selection criterion MIC, to some simulated and real data sets.

In the simulation and the Lloyds Bank examples given in Sections 5.1 and 5.2, we use a general covariance structure for T_j, and for the Leukemia and Yeast examples in Sections 5.3 and 5.4 we use the same covariance structure for T_j, that is, $T_1 = T_2 = \cdots = T_G$. In all the examples, we use $\bar{\mathbf{x}}$ as the prior mean of μ_j and we let $\alpha = p+4$ and $B_j = (\alpha - p - 1)S_x$ for the hyperparameters of the Inverted Wishart prior of T_j in (2.3), where $\bar{\mathbf{x}}$ and S_x are the average and sample covariance matrix of the initial $\mathbf{x}_i$'s. Thus, we use a common mean for all μ_j and a common vague prior for all T_j, and choose the scale parameter of the inverted Wishart distribution so that the prior mean of T_j is equal to S_x.

5.1 SIMULATION EXAMPLES

Six datasets with 50 objects each were generated from mixtures of bivariate normal distributions with various values of the mixture model parameters, which represent various important cases. Scatterplots of the true objects from the six datasets are given in Figure 1. In all cases the true dimension is $p = 2$, but the true numbers of clusters are different. The first set has a big cluster and a small cluster close to the big one, and the second set has three well-separated clusters. The third set has two big clusters and a small cluster which may be considered as a group of outliers. The fourth set has two clusters with different covariance structures. The fifth set has two big clusters and two small clusters that are symmetrically located. The last set has two close clusters. The true dimension is $p = 2$, but the second

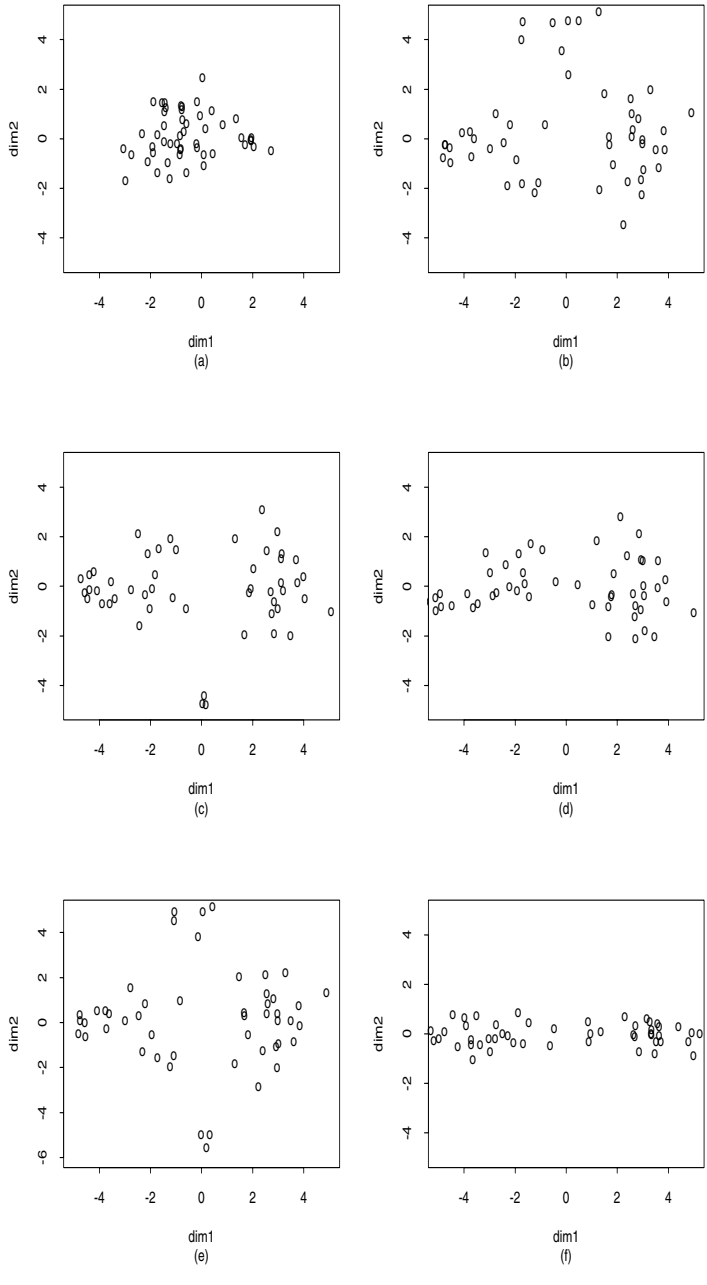

Figure 1. Scatterplots of true objects in the simulation data. There is a big cluster and a small cluster close to the big one in (a), three well-separated clusters in (b), two big clusters and a small cluster of outliers in (c), two clusters with different covariance structures in (d), two big clusters and two small clusters symmetrically arranged in (e), and two close clusters in (f).

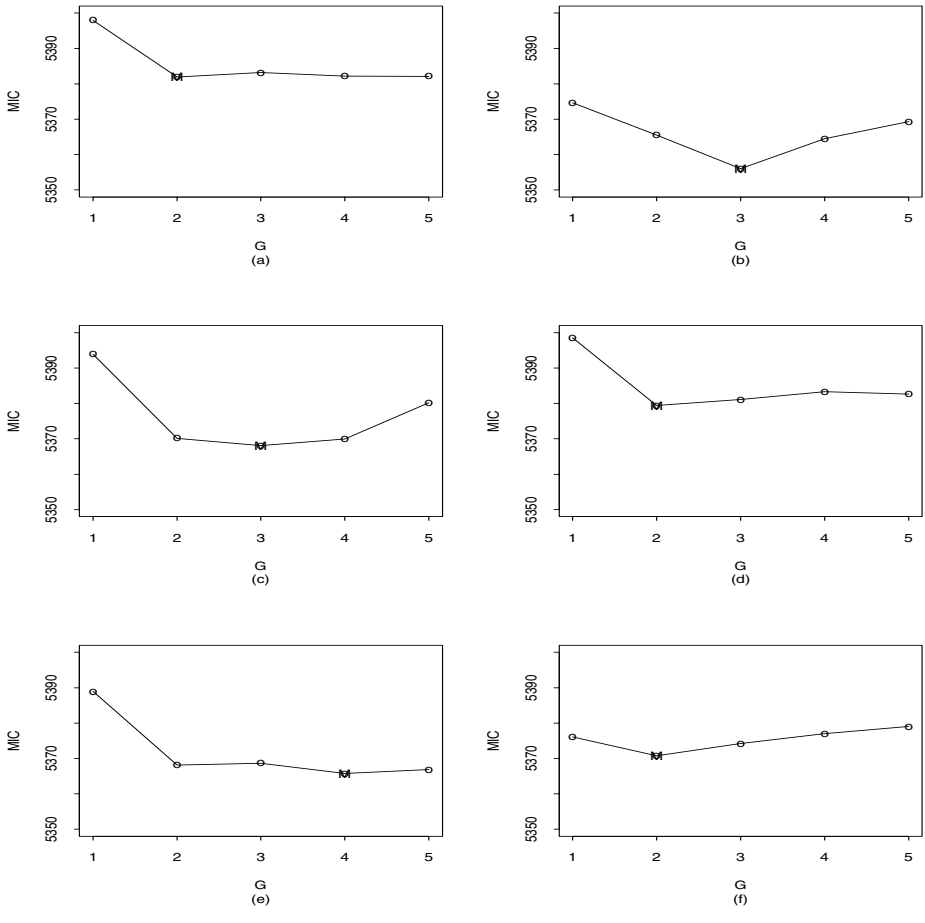

Figure 2. Plots of MIC for $p = 2$ in the six simulated datasets when $\sigma = 0.3$. The best (minimum) value is indicated by the symbol M on the plot. MIC chooses the correct number of clusters in each case.

coordinates of the objects are close to zero, resulting in a slim scatterplot. For each dataset, Euclidean distances δ_{ij} between pairs of objects were computed and a 50×50 matrix of observed dissimilarity measures d_{ij} was obtained by generating the d_{ij} from a normal distribution with mean δ_{ij} and standard deviation 0.3, restricted to be positive.

For each dataset, we applied BMCD with 20,000 MCMC iterations, of which the first 5,000 were discarded for burn-in, and we computed MIC for various values of p and G. In all cases, the MIC values for p other than 2 were much larger than the MIC value for $p = 2$ for every value of G considered, so MIC correctly identifed the true dimension $p = 2$. Figure 2 shows plots of MIC for various G when $p = 2$ for the six datasets. In all cases, MIC selected the correct number of clusters, though MIC's preference for the selected G was not as strong as for the selected p. Also, the estimated object configurations from BMCD were good, as can be seen in Figure 3.

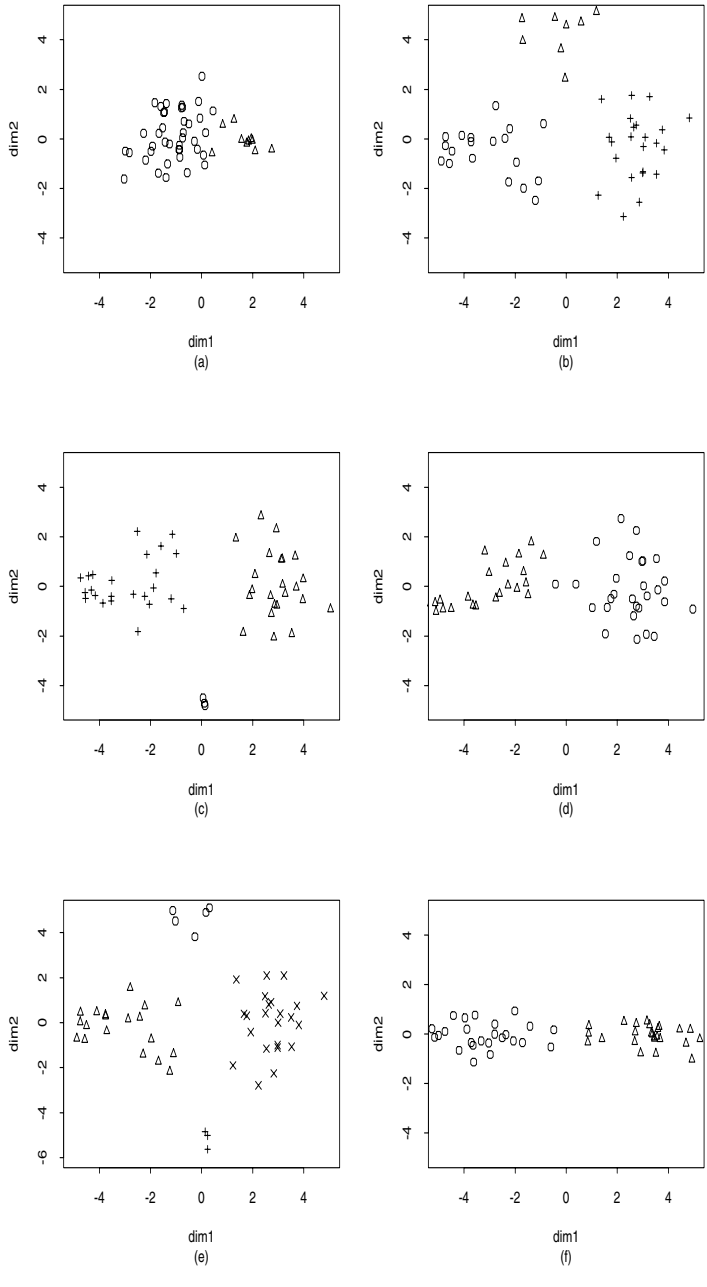

Figure 3. Scatterplots of the estimated object configuration and the classification from BMCD with the optimal p and G in the six simulated datasets when $\sigma = 0.3$. Different symbols represent different clusters.

Table 2. The optimal (p, G) values chosen by MIC for $\sigma = 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 3.0$ for datasets (a)–(f).

σ	0.6	0.9	1.2	1.5	1.8	2.1	3.0
(a)	(2,2)	(2,2)	(2,2)	(2,3)	(2,2)	(2,2)	(2,2)
(b)	(2,2)	(2,2)	(2,2)	(2,2)	(2,3)	(2,2)	(2,2)
(c)	(2,2)	(2,2)	(2,2)	(2,2)	(2,3)	(2,1)	(1,1)
(d)	(2,2)	(2,2)	(2,2)	(2,2)	(2,2)	(1,2)	(1,1)
(e)	(2,2)	(2,3)	(2,2)	(2,2)	(2,3)	(2,2)	(2,4)
(f)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)

To assess the effect of having a much greater measurement error variance for the dissimilarities, σ^2, we repeated the same experiment with larger standard deviations, namely $\sigma = 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 3.0$. The optimal values of (p, G) for each σ are summarized in Table 2. In choosing the optimal (p, G), sometimes MIC values are not much different for various values of G. Since there is estimation error in MIC, we ignored a small difference in MIC and chose the smallest G among those with MIC values within ± 5 from the minimum MIC value. The estimated object configurations and their clusterings for $\sigma = 1.8$ are presented in Figure 4 for illustration. Table 2 shows that as σ gets larger MIC tends to choose smaller p and G, implying that BMCD tends to put the object configurations in a smaller dimension and cluster the objects into a small number of clusters as the measurement error in dissimilarity gets larger.

Next, to see the effect of an asymmetric error distribution for the dissimilarities, we simulated $d_{ij} = \delta_{ij} + \sigma(\chi_{ij} - 3)/\sqrt{6}$, for the six datasets, where χ_{ij} is a random number from a chi-squared distribution with three degrees of freedom. Thus, the error term $\sigma(\chi_{ij} - 3)/\sqrt{6}$ has mean 0 and variance σ^2 but the distribution is skewed rather than symmetric. We applied BMCD to these new dissimilarity datasets with various values of σ, and observed that the clustering and the object configuration were almost the same as when the error term was normally distributed. This suggests that the proposed BMCD method may be applied to a moderately skewed error distribution.

We also tried some three-dimensional examples with clustering structures similar to those in Figure 1. MIC chose the correct dimension in all cases. In fact, MIC never missed the correct dimension in all the cases we studied. The clustering results in three dimensions are similar to the results in two dimensions. From these observations and from the results for the real data examples we now turn to, there seems to be little or no interaction between p and G, and the choice of G seems to depend not on p but on the clustering structure of the data.

5.2 LLOYDS BANK DATA

We now consider dissimilarity measures between the careers of 80 employees of Lloyds Bank in England during the period 1905–1950 (Stovel, Savage, and Bearman 1996), computed using the optimal alignment method of Sankoff and Kruskal (1983), as applied to

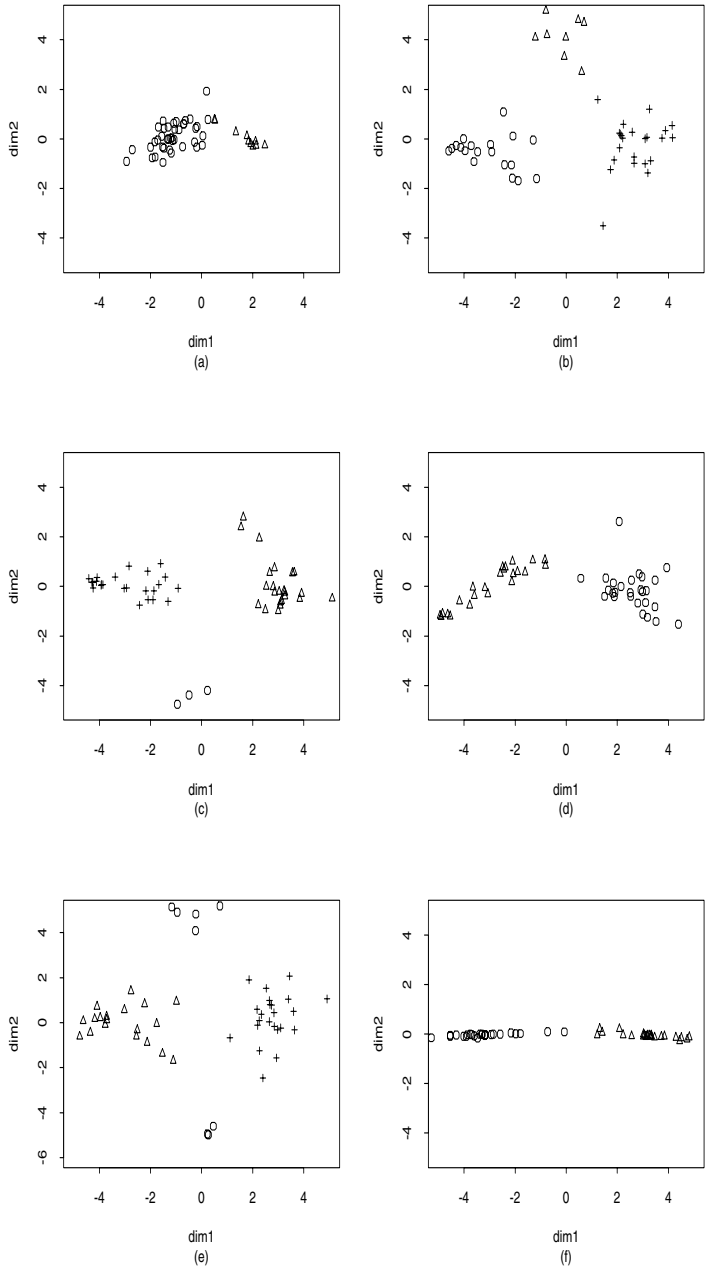

Figure 4. Scatterplots of the estimated object configuration and the classification from BMCD with $p = 2$ and the optimal G in the six simulated data sets when $\sigma = 1.8$. Different symbols represent different clusters.

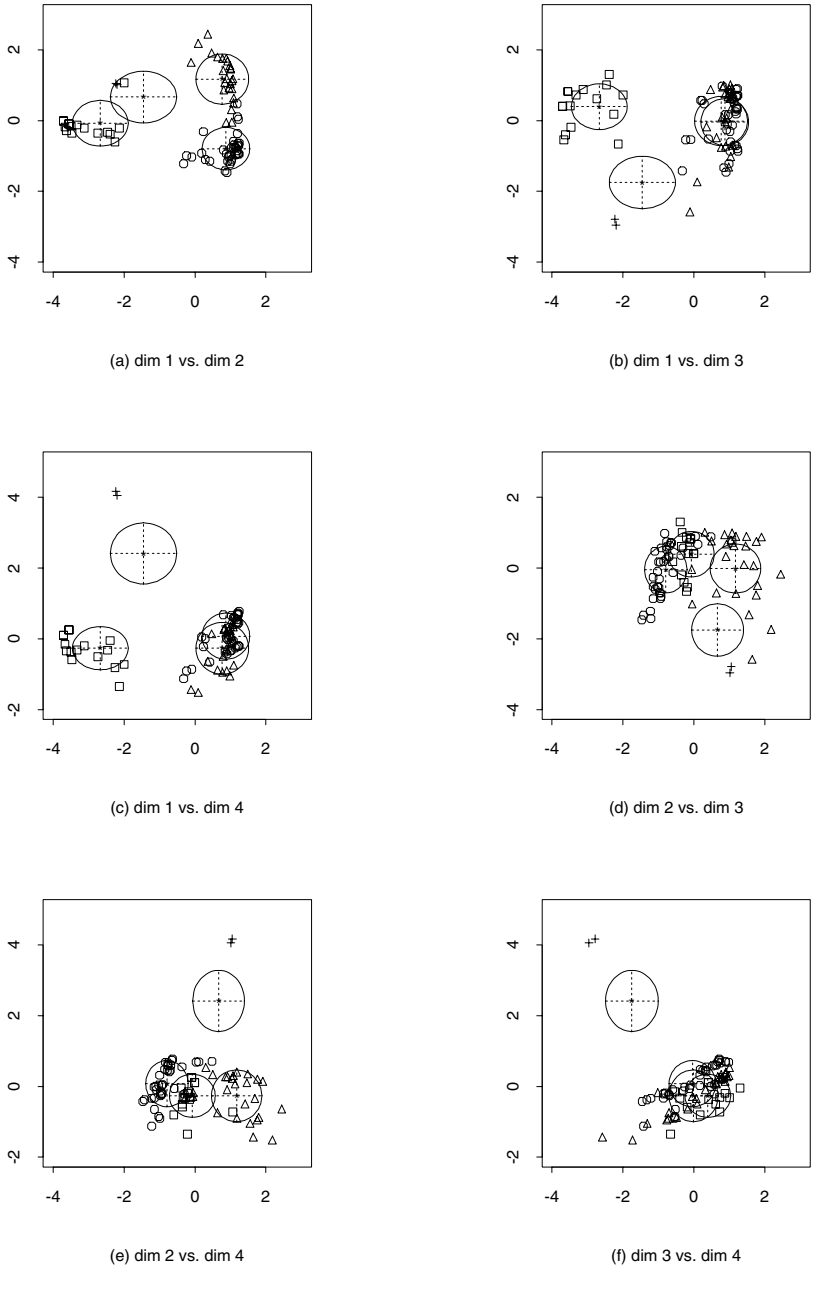

Figure 5. Scatterplots of the estimated object configuration and the component density functions from BMCD for the Lloyds Bank data.

career data by Abbott and Hrycak (1990). Note that the dissimilarity measures in this dataset are not Euclidean distances and may not satisfy some properties of typical metric distances. Oh and Raftery (2001) analyzed the data and MIC chose $p = 8$ as the optimal dimension. After removing two outlying employees who had extremely short careers at the bank, MCLUST was applied to the estimate of $\mathbf{X}$ from BMDS. It chose $G = 3$ and yielded a reasonable classification of objects. The first group consisted of 16 employees who had short careers at the bank and spent all or most of their careers at the lowest clerk rank. The second group consisted of 30 employees who had long careers at the bank and spent all or most of their careers at the lowest clerk rank. The third group consisted of 32 employees, 24 of whom were promoted to managers, and 8 of whom had medium length careers and ended at the clerk level.

We applied BMCD to the data with 40,000 MCMC iterations, of which the first 10,000 were discarded as burn-in. MIC chose $p = 8$ and $G = 4$, which coincides with the results from the previous analysis. The first group was identical to the first group found in the previous analysis. The second and the third groups were almost identical to those found in the previous study except that the eight employees who had medium length careers and ended at the clerk level were clustered together with those who had long careers and ended at the clerk level in BMCD, rather than with the managers. This is more satisfactory, substantively. The fourth group consisted of the two outliers which were removed before clustering in the previous analysis. Thus, BMCD picked up the outliers during the process and yielded a more sensible clustering than the previous analysis. Figure 5 shows pairwise scatterplots of object configurations and the estimated component densities for the first four coordinates. Note that in Figure 5, the component density for the outliers has mean close to zero and a large covariance since there are only two objects in the group and the effect of the prior is dominant for this component. From Figure 5, it seems that BMCD gives clear separation between clusters and takes care of the outliers.

5.3 LEUKEMIA GENE EXPRESSION DATA

Golub et al. (1999) used gene expression data on 50 genes and 72 acute leukemia patients to classify the patients into different types of leukemia. The 50 genes are believed to be informative about the distinction between acute myeloid leukema (AML) and acute lymphoblastic leukemia (ALL) in the known samples.

We follow the standardization process given by Getz, Levine, and Domang (2000) and compute the Euclidean distance between genes, yielding a dissimilarity matrix similar to a correlation matrix. Due to the standardization, the mean of each gene is set to zero and this reduces the true dimension of the objects from 50 to 49.

Investigation of plots of dissimilarity measures between pairs of individuals shows that there are two big groups and most dissimilarity measures between individuals in the same group are small while those between individuals in different groups are large. Figure 6(a) and (b) are typical plots for individuals in the first and the second groups, respectively. However, there are some individuals, such as numbers 12 and 55, who seem to be misclassified since they have smaller dissimilarities with those in the different group and larger

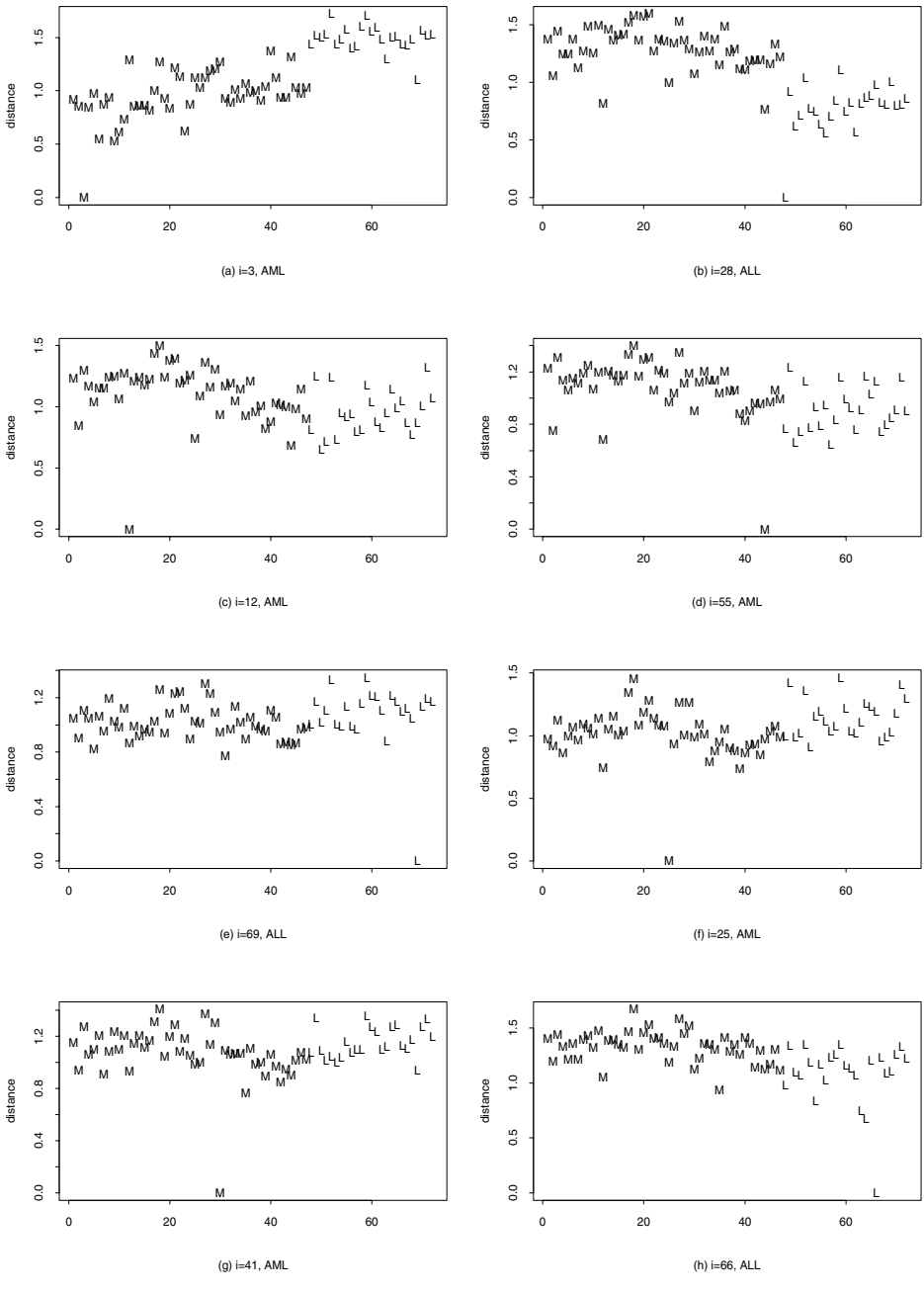

Figure 6. Distances for some selected individuals in Leukemia data (id numbers and the type of Leukemia is given at the bottom of each figure). Typical distances for individuals in AML and ALL groups are shown in (a) and (b), respectively. Figures (c)–(d) presents distances for individuals who have smaller distances with those in the different group and larger distances with those in the same group. Figures (e)–(h) presents distances for individuals whose distances do not clearly show their closeness to either group.

Table 3. MIC values for Leukemia data for $p = 2, 3, 49$

G	1	2	3	4
$p = 2$	14444	14425	14428	14440
$p = 3$	12986	12970	12973	12979
$p = 49$	−11058	−22457	−22432	−22320

dissimilarities with those in the same group as shown in Figure 6(c)–(d). Also there are a few individuals whose distances do not clearly show their closeness to either group as shown in Figure 6(e)–(h).

In all the examples we analyzed, MIC showed a sharp drop at the same optimal value of p for all values of G, so that the choice of G does not affect the choice of dimension. This is because the object configurations are not much different for different G's when the same p is used. Thus, one may choose optimal dimension p with $G = 1$ and then choose optimal G with the selected p. In other words, one may apply BMDS to choose the dimension and then apply BMCD with the optimal dimension. This would reduce computation time significantly. Note that although p and G are chosen sequentially, the object configuraton and the clustering are still estimated simultaneously.

Following the above suggestion, BMDS (i.e., BMCD with $G = 1$) was applied to the data and MDSIC clearly chose $p = 49$, which is the true dimension, as shown in Figure 7. We used 18,000 MCMC iterations, of which 3,000 were discarded as burn-in.

We next applied the method with $p = 49$ for various values of G, and the MIC values are shown in Table 3. MIC clearly chooses $G = 2$, which is viewed as the correct number of groups for this dataset. Three individuals, numbers 12, 55, and 69, are misclassified, so the misclassification rate is $3/72 = 4.2\%$. The 41st individual is classified into the correct AML group, but his or her posterior membership probability for AML is only 0.51,

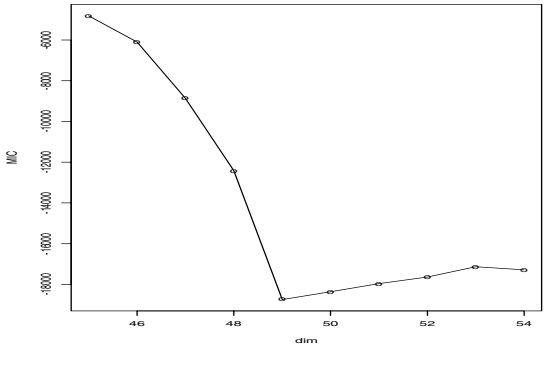

Figure 7. MICs for Leukemia data with $G = 1$.

showing large uncertainty. This result makes sense in light of the dissimilarity measures in Figure 6.

For visual display of the clusters and objects in low-dimensional space, we applied the method with $p = 2, 3$ for various values of G and the results are also shown in Table 3. MIC chose $G = 2$ groups both when $p = 2$ and when $p = 3$. The classification results from $p = 2, 3$ were the same as those from $p = 49$ with $G = 2$ except for the 41st and the 69th individuals. When $p = 2$ and 3, the 41st individual was misclassified into the ALL group but with a membership probability of only 0.51, and the 69th individual was correctly classified into the AML group when $p = 2, 3$. However, the membership probability for the 69th individual was about 0.56 when $p = 2, 3$ while it was close to 1.0 when $p = 49$, showing significant uncertainty when $p = 2, 3$. Thus, in terms of clustering, 2 or 3 dimensions from BMCD did as well as the true 49 dimensions for all the individuals except for the 69th.

Figure 8 shows estimated object configurations with their classifications when $p = 3$ and $G = 2$. It is interesting to observe that the clusters can be well identified in the plot of the first two coordinates from BMCD. And it is hard to see clusters in the plots of the other coordinates. We have observed this in most of the datasets we have analyzed, suggesting that in some cases only a few coordinates from BMCD would do as well as all the coordinates, in terms of clustering, at least for most of the data points. One can use our method to obtain a hard classification by classifying each patient into their cluster of highest posterior probability. The patients misclassified by this were in the middle of the plot of the first two coordinates, and had the highest estimated uncertainty. This suggests that our method is indeed capturing uncertainty fairly well, and that there is benefit to reporting the posterior probabilities, or at least an overall measure of uncertainty for each patient, rather than a hard classification. One such measure of uncertainty is one minus the posterior probability of being in the most likely group for that individual (Bensmail et al. 1997).

To assess the benefits from simultaneous rather than separate estimation of object configuration and clustering, we compared BMCD with a two-stage scheme, which estimates object configuration and then applies model-based clustering, when $p = 3$ and $G = 2$. For a fair comparison, we used the same priors, and we applied the same MCMC procedures for posterior estimation of the parameters and used the estimated object configuration from BMCD as input data in the model-based clustering of the two-stage scheme. Note that the only difference between the two methods lies in the randomness of the object configuration. In most cases the estimated membership probabilities were more extreme in the two-stage method. This may be because it is more likely to assign each object to a certain group when $\mathbf{X}$ is fixed than when it is random. More extreme probabilities yield smaller posterior standard deviations for the probabilities, suggesting that sequential application of MDS and model-based clustering can underestimate the clustering uncertainties.

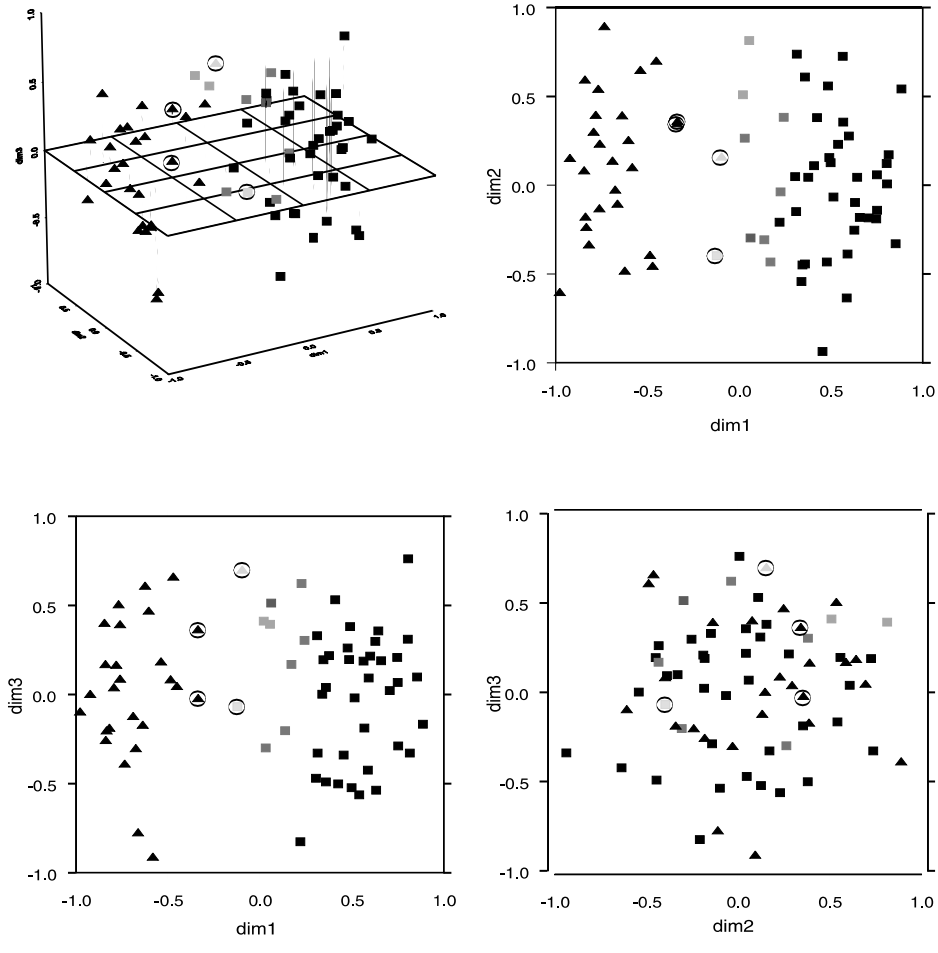

Figure 8. Three-dimensional and pairwise scatter plots for the object configuration and their classification from BMCD for the Leukemia gene expression data when $p = 3$ dimensions and $G = 2$ groups. The membership probabilities are represented by the darkness of the symbols, (black for probability 1 and gray for probability 0.5). Misclassified objects are marked by circles.

5.4 YEAST CELL CYCLE DATA

The Yeast cell cycle data (Cho et al. 1998) consist of gene expression levels of approximately 6,000 genes over 17 time points. Yeung et al. (2001) used a subset of this data consisting of 384 genes whose expression levels peak at different time points corresponding to five known phases of the cell cycle (Cho et al. 1998).

We normalized the data analyzed by Yeung et al. (2001) using a typical standardization method, subtracting the mean and dividing by the standard deviation, for each gene. We then computed the Euclidean distance for each pair of genes and used the Euclidean distances as dissimilarity measures. Due to the normalization, we lose one dimension and hence the true dimension of object configuration from the dissimilarity matrix is 16.

Table 4. MIC values for Yeast data for $p = 2, 3, 16$.

G	3	4	5	6	7	8
$p = 2$	844334	844232	844207	844185	844173	844201
$p = 3$	772569	772459	772402	772423	772407	772396
$p = 16$	−1346244	−1346602	−1346788	−1346521	−1347010	−1346648

Yeung et al. (2001) applied MCLUST to the 384×17 dataset and found that the "EEE" model, which assumes that the mixture components have the same covariance matrix, gave a better fit than other models, so we made this assumption in applying BMCD to this data set. We used 18,000 MCMC iterations, of which we discarded 3,000 as burn-in.

First, to choose the optimal dimension of objects, BMDS was applied for $p = 1$ to 18. It indicated clear evidence for $p = 16$, which is the correct dimension in this example. Next, BMCD was applied with $p = 16$ for $G = 3$ to 8. The values of MIC are given in Table 4. MIC reaches a minimum at $G = 7$ but has a first local minimum at $G = 5$, and here we consider $G = 5$ and $G = 7$ as possible optimal numbers of groups.

We applied BMCD with $p = 2, 3$ for visualization and parsimony. Values of MIC are given in Table 4. When $p = 2$, MIC chooses $G = 7$ and when $p = 3$ it chooses $G = 8$, but MIC has about the same value at $G = 5$ and $G = 7$. Figure 9(a) shows the two-dimensional object configuration from BMCD with the actual five known phases. There are significant overlaps between the actual clusters. Figure 9(b) shows the estimated object configuraton and classification results from BMCD with $p = 2$ and $G = 5$. It can be seen that BMCD yields a reasonable clustering of the objects.

We compared the clustering results for $p = 2$, $p = 3$, and $p = 16$, when $G = 5$. There are 20 mismatches between $p = 2$ and $p = 16$, and 22 mismatches between $p = 3$ and

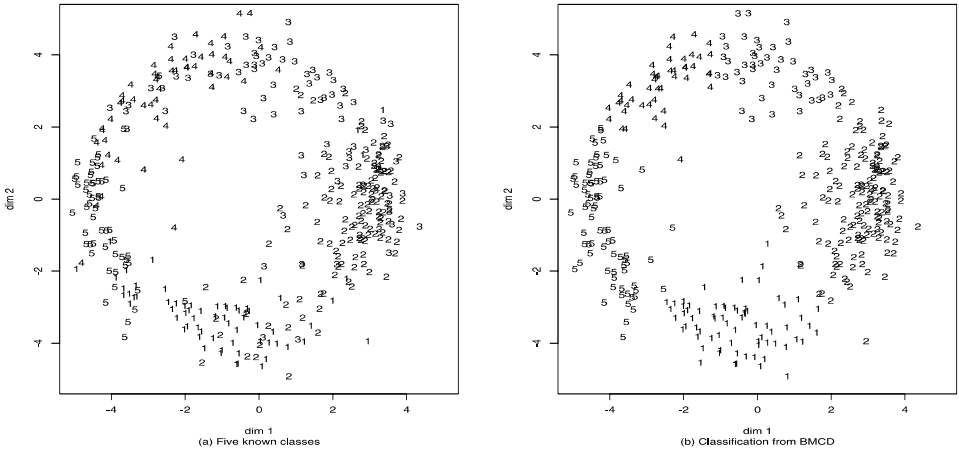

Figure 9. Scatterplots of the object configuration from BMCD with $p = 2$ in the Yeast cell cycle data and their classifications: (a) the actual five clusters; (b) the classification from BMCD with $G = 5$.

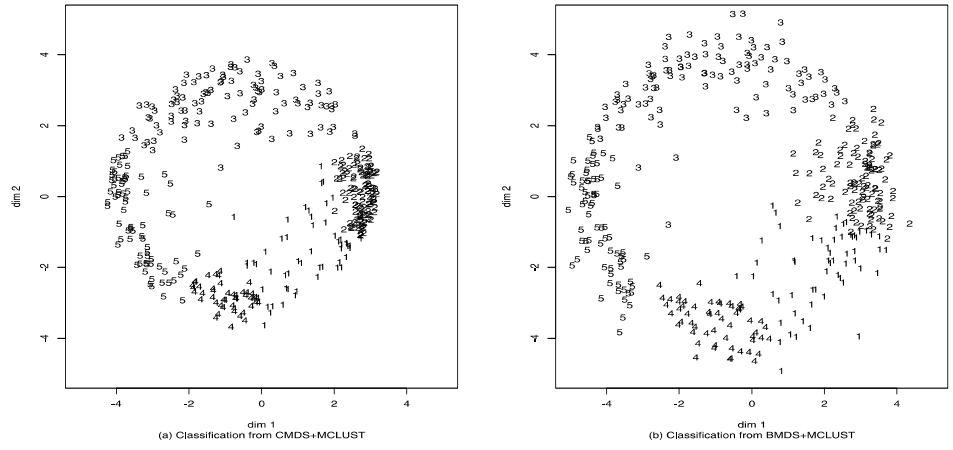

Figure 10. Scatterplots of object configuration from sequential application of MDS and MCLUST with $p = 2$ and $G = 5$ in Yeast data and their classifications: (a) presents classification from the classical MDS followed by MCLUST and (b) presents classification from the Bayesian MDS followed by MCLUST.

$p = 16$. Thus, the proportion of mismatches is less than 6% between the low dimensional clustering and the clustering with the true dimension. Many of the mismatched genes show significant clustering uncertainties in low dimensions, suggesting that the assessment of uncertainty is accurate in these cases. We next computed the proportion of mismatches between clustering from BMCD and the actual five clusters. These are 0.268, 0.266, and 0.279 for $p = 16$, $p = 3$, and $p = 2$, respectively, indicating that there is almost no difference in clustering quality between the different dimensions.

For comparison, we tried classical MDS followed by MCLUST and Bayesian MDS followed by MCLUST. Figure 10 shows the two-dimensional object configuration and their clustering in each case. As shown by Oh and Raftery (2001), the object configurations from classical MDS and Bayesian MDS are quite different. In addition, the clustering results from MCLUST (Figure 10) and those from the proposed BMCD (Figure 9) are significantly different even when Bayesian MDS results are used in MCLUST. It seems that the first cluster in BMCD is split into two clusters in MCLUST and the third and the fourth clusters in BMCD are combined into one cluster in MCLUST. The proportion of mismatches between clustering from MCLUST (with both CMDS and BMDS) and the actual five clusters is about 0.476.

We also performed the two-stage scheme of MDS plus model-based clustering with $p = 2$ and $G = 5$, and compared the resulting membership probabilities with those from BMCD. The main practical difference between the two methods lies in whether $\mathbf{X}$ is fixed or randomly generated at each MCMC iteration. In almost all cases the two-stage scheme yields more extreme membership probabilities. This again suggests that first estimating object configuration and then clustering (as opposed to doing both simultaneously as in BMCD) does not take into account the variation in object configuration when clustering and that it may underestimate the clustering uncertainties.

6. DISCUSSION

We have proposed a model-based clustering method for the situation where the data consist of dissimilarity measures between pairs of objects. It is also useful for clustering objects in low-dimensional space for visual display and parsimony even when the object coordinates are given, but are high-dimensional.

Hierarchical models are used to represent the possible sources of error, namely measurement errors in the dissimilarities, errors in estimating object configuration, and errors in clustering the objects. A probabilistic model is used for the observed dissimilarities and a mixture model is used for the unobserved latent object configuration. The object configuration, the mixture model parameters, and the objects' group memberships are estimated simultaneously via Bayesian inference using MCMC. The object configuration can be used for display of objects and the mixture parameters can be used for clustering objects.

When the estimated dimension is high, we have compared Bayesian MDS with the selected dimension with Bayesian MDS with low dimension (2 or 3). We found that the clustering results were very similar, and that those misclassified in the low-dimensional analysis had high clustering uncertainties, which is good. Thus, in practice low-dimensional configurations from BMDS may be good enough for many purposes, especially if it is followed up with more intensive investigation of objects with high clustering uncertainty.

We have proposed a Bayesian criterion, MIC, for simultaneously selecting the object dimension and the number of clusters, which is easy to compute from MCMC output. In our simulations and in real examples, it worked reasonably well in all cases. MIC varied more between dimensions than between numbers of clusters, and the choice of dimension was not affected by the choice of the number of clusters. Thus, as an approximation, we suggest selecting the dimension assuming one cluster (i.e., using BMDS), and then choosing the number of clusters given the selected dimension. This greatly reduces computation time.

In general, MIC chooses the correct number of clusters G when the clusters are reasonably well-separated or when the clusters overlap but the entire scatterplot is not ellipsoidal. However, when there are small clusters with very few objects (e.g., outliers), MIC sometimes fails to detect the correct G. This is because there are not enough observations relative to the number of parameters in the multivariate normal distribution and hence the vague prior dominates the posterior. One way to improve the ability of MIC to detect a small number of outliers could be to use a simpler covariance structure, such as a spherical covariance $\gamma^2 I$.

Our BMCD method allows various covariance structures, including unrestricted, equal covariance for all the clusters, spherical, and others. For parsimony, one can choose a simpler covariance matrix when there is available information, or when there is a need for compromise between the dimension and the covariance structure. Allowing unrestricted covariance structure in a high-dimensional space would bring too many parameters to estimate, which may be a problem when there is not a sufficient number of objects. Taking these considerations into account, we imposed no restriction for the simulation and the bank data but we did impose some restrictions for the Leukemia and the Yeast cell cycle

data, based on preliminary studies using the model-based clustering software MCLUST. The development of more formal methods for making this decision is an appropriate goal for further research.

The proposed method may be computationally expensive when both the number of objects, n, and the dimension, p, are large, since it estimates an $n \times p$ matrix of object configurations. However, dimension reduction is a main purpose of MDS, and in the examples we showed that the method provided a similar clustering in a low dimensional space. Thus, the method could be used for clustering a not-too-large number of objects in a low-dimensional space.

In this article, we assumed that the observed dissimilarities follow a truncated normal distribution with constant variance. Thus, the error in the dissimilarity measure is symmetrically distributed apart from the truncation at zero, and the error does not depend on the size of the dissimilarity. In some cases, however, the error distribution may be skewed or the error may depend on the size of the dissimilarity measure (e.g., the errors may be small for close objects and large for distant objects). In such cases, one may incorporate this feature by using an appropriate distribution for the error distribution, such as a gamma distribution or a normal distribution with a nonconstant variance. The proposed algorithm can be easily modified according to the different modeling of the error since we use a Metropolis-Hastings algorithm for generating $\mathbf{X}$. Note that the error distribution affects only the distribution of $\mathbf{X}$, not those of ε, μ_j, T_j.

In MDS it is popular to transform the dissimilarities, for example, by an optimal ordinal or spline transformation; this is often called nonmetric scaling. Our method could be generalized to allow for this. One way of doing it would be to generalize Equation (2.1) to

$$d_{ij} \sim N(g(\delta_{ij}), \sigma^2) I(d_{ij} > 0),$$

where $g(\cdot)$ is a function to be estimated from the data along with the other parameters and the configuration. If g is a parametric function, the methodology could be extended to handle this in a conceptually straightforward way, by including the parameters specifying g in the MCMC algorithm. In nonmetric scaling, g is often nonparametric, and this could be included in the current framework by allowing g to be a spline function. In that case, g is still specified by a finite number of parameters, and these could be included in the MCMC algorithm.

One important area where data come in the form of measures on pairs of objects is social networks, where data consist of the presence or absence (or in some cases the intensity) of ties between actors. Hoff, Raftery, and Handcock (2002) used ideas similar to those of Oh and Raftery (2001) to represent actors in a social network by positions in a Euclidean latent space and estimate the positions. The model used was the same as that of Oh and Raftery (2001), except that the conditional distribution of "dissimilarities" (in the social network case, presence or absence of ties) given distances was taken to be binary with a conditional probability specified by logistic regression, rather than truncated normal. The analysis of social network data is often motivated by questions about the presence and nature of clusters in the network, and these are often answered fairly heuristically. It would seem straightforward to extend the present approach to social network data, again

modeling presence or absence of ties as conditionally binary with a probability depending on distance in a logistic regression manner. This could provide a more formal way of answering questions about clustering in social networks.

APPENDIXES

A. PROCRUSTEAN TRANSFORMATION

- Step 0: Let $\mathbf{J}$ be the centering matrix, that is, $\mathbf{J} = \mathbf{I} - 1/n\mathbf{1}\mathbf{1}'$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is the vector of all 1's.
- Step 1: Compute $\mathbf{C} = \mathbf{X}^{*\prime}\mathbf{J}\mathbf{X}$.
- Step 2: Compute the singular value decomposition of $\mathbf{C}$, that is, $\mathbf{C} = PDQ'$, where P and Q are orthogonal matrices and D is a diagonal matrix.
- Step 3: Let $\mathbf{T} = QP'$.
- Step 4: Let $\mathbf{t} = 1/n(\mathbf{X}^* - \mathbf{X}\mathbf{T})'\mathbf{1}$.
- Step 5: Transform $\mathbf{X}$ by $\mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{1}\mathbf{t}'$.

B. RELABELING PROCEDURE

Let $\boldsymbol{\theta}$ be a d dimensional vector of all the parameters in the mixture distribution, and J be the number of components in the mixture.

- Step 0. Estimate elements of $\boldsymbol{\theta}$ and their variances using samples taken before the first label switching. Let $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$ be the above estimates. Permute the labeling of the latent classes and deduce $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{J!-1}$ and $\mathbf{s}^1, \ldots, \mathbf{s}^{J!-1}$.
- Step 1. For each sample of $\boldsymbol{\theta}$, do :

  (1) Get l^* which minimizes the squared-distances

  $$||\boldsymbol{\theta} - \boldsymbol{\theta}^l||^2 = \sum_{i=1}^{d} \frac{(\theta_i - \theta_i^l)^2}{s_i},$$

  for $l = 0, \ldots, J! - 1$, where θ_i and s_i are the ith coordinate of $\boldsymbol{\theta}$ and $\mathbf{s}$, respectively. Allocate $\boldsymbol{\theta}$ to the label given by l^*. Switch permutations l^* and 0.

  (2) Update $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$ and derive $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{J!-1}$ and $\mathbf{s}^1, \ldots, \mathbf{s}^{J!-1}$ by permutation.

ACKNOWLEDGMENTS

The authors thank two referees and an associate editor for helpful comments. Man-Suk Oh's research was supported by research funds from KOSEF R04-2002-00-0046-0, and Adrian Raftery's research was supported by NIH Grant 1R01CA094212-01 and ONR Grant N00014-01-10745.

[Received September 2003. Revised December 2006.]

REFERENCES

Abbott, A., and Hrycak, A. (1990), "Measuring Sequence Resemblance," *American Journal of Sociology*, 96, 144–185.

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. (1997), "Inference in Model-Based Cluster Analysis," *Statistics and Computing*, 7, 1–10.

Borg, I., and Groenen, P. (2005), *Modern Multidimensional Scaling* (2nd ed.), Berlin: Springer-Verlag.

Buttenfield, B., and Reitsma, R. F. (2002), "Loglinear and Multidimensional Scaling Models of Digital Library Navigation," *Internation Journal of Human-Computer Studies*, 57, 101–119.

Celeux, G., Hurn, M., and Robert C. P. (2000), "Computational and Inferential Difficulties with Mixture Posterior Distribution," *Journal of the American Statistical Association*, 95, 957–970.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D. J., Lockhart, D. J., and Davis, R.W. (1998), "A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, 2, 65–73.

Condon, E., Golden, B., Lele, S., Raghavan, S., and Wasil, E. (2002), "A Visualization Model Based on Adjacency Data," *Decision Support Systems*, 33, 349–362.

Courrieu, P. (2002), "Straight Monotonic Embedding of Data Sets in Euclidean Spaces," *Neural Networks*, 10, 1185–1196.

Cox, T. F., and Cox, M. A. A. (2001), *Multidimensional Scaling*, London: Chapman & Hall.

Davison, M. L. (1983), *Multidimensional Scaling*, New York: Wiley.

De Leeuw, J., and Heiser, W.J. (1982), "Theory of Multidimensional Scaling," in *Handbook of Statistics*, eds. P. R. Krishnaiah and L. N. Kanal, Amsterdam: North Holland.

Elvevag, B., and Storms, G. (2003), "Scaling and Clustering in the Study of Semantic Disruptions in Patients With Schizophrenia: A Re-evaluation," *Schizophrenia Research*, 63, 237.

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

——— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

——— (2003), "Enhanced Model-Based Clustering, Density Estimation and Discriminant Analysis Software: MCLUST," *Journal of Classification*, 20, 263–286.

Getz, G., Levine, E., and Domany, E. (2000), "Coupled Two-way Clustering Analysis of Gene Microarray Data," *Proceedings of National Academy of Sciences*, 97, 12079–12084.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, M. L., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Groenen, P. J. F. (1993), *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, Lieden, The Netherlands: DSWO Press.

Groenen, P. J. F., Mathar, R., and Heisser, W. J. (1995), "The Majorization Approach to Multidimensional Scaling for Minkowski Distances," *Journal of Classification*, 12, 3–19.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods using Markov Chains and their Applications," *Biometrika*, 57, 97–109.

Hedenfalk, I. A., Ringer, M., Trent, J., and Borg, A. (2002), "Gene Expression in Inherited Breast Cancer," *Advances in Cancer Research*, 84, 1–34.

Heiser, W. J., and Groenen, P.J.F. (1997), "Cluster Difference Scaling with Within-Cluster Loss Component and a Fuzzy Successive Approximation Strategy to Avoid Local Minima," *Psychometrika*, 62, 63–83.

Hoff, P., Raftery, A. E., and Handcock, M. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

MacKay, D. (1989), "Probabilistic Multidimensional Scaling: An Anisotropic Model for Distance Judgements," *Journal of Mathematical Psychology*, 33, 187–205.

MacKay, D., and Zinnes, J. L. (1986), "A Probabilistic Model for the Multidimensional Scaling of Proximity and Preference Data," *Marketing Sciences*, 5, 325–334.

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

Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., and Wong, G. (2002), "Analysis and Visualization of Gene Expression Data using Self-Organizing Maps," *Neural Networks*, 15, 953–966.

Oh, M-S. (1999), "Estimation of Posterior Density Functions from a Posterior Sample," *Computational Statistics & Data Analysis*, 29, 411–427.

Oh, M-S., and Raftery, A. (2001), "Bayesian Multidimensional Scaling and Choice of Dimension," *Journal of the American Statistical Association*, 28, 259–271.

Priem, R.L., Love, L., and Shaffer, M.A. (2002), "Executives' Perceptions of Uncertainty Sources: A Numerical Taxonomy and Underlying Dimensions," *Journal of Management*, 28, 725–746.

Ramsay, J. O. (1982), "Some Statistical Approaches to Multidimensional Scaling," *Journal of the Royal Statistical Society*, Series A, 145, 285–312.

Ren, S., and Frymier, P.D. (2003), "Use of Multidimensional Scaling in the Selection of Wastewater Toxicity Test Battery Components," *Water Research*, 37, 1655–1661.

Sankoff, D., and Kruskal, J.B. (1983), *Time Warps, String Edits and Macromolecules,* Reading, MA: Addison-Wesley.

Schutze, H., and Silverstein, C. (1997), "Projections for Efficient Document Clustering," *ACM SIGIR 97*, pp. 74–81.

Sibson, R. (1979), "Studies in the Robustness of Multidimensional Scaling: Perturbation Analysis of Classical Scaling," *Journal of the Royal Statistical Society*, Series B, 41, 217–229.

Sneath, P. H. A. (1957), "The Application of Computers to Taxonomy," *Journal of General Microbiology*, 17, 201–206.

Sokal, R. R., and Michener, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Scientific Bulletin*, 38, 1409–1438.

Stephens, M. (2000), "Dealing with Label-Switching in Mixture Models," *Journal of the Royal Statistical Society*, Series B, 62, 795–809.

Stovel, K., Savage, M., and Bearman, P. (1996), "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890–1970," *American Journal of Sociology*, 102, 358–399.

Takane, Y. (1982), "The Method of Triadic Combinations: A New Treatment and Its Applications," *Behaviormetrika*, 11, 37–48.

Tibshirani, R., Lazzeroni, L., Hastie, T., Olshen, A., and Cox, D. (1999), "The Global Pairwise Approach to Radiation Hybrid Mapping," Technical Report, Department of Statistics, Stanford University.

Welchew, D.E., Honey, G.D., Sharma, T., Robins, T.W., and Bullmore, E.T. (2002), "Multidimensional Scaling of Integrated Neurocognitive Function and Schizophrenia as a Disconnexion Disorder," *NeuroImage*, 17, 1227–1239.

Yeung, K., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001), "Model-Based Clustering and Data Transformation for Gene Expression Data," *Bioinformatics*, 17, 977–987.

Yin, H. (2002), "Data Visualization and Manifold Mapping using the ViSOM," *Neural Networks*, 15, 1005–1016.

Young, F.W. (1987), *Multidimensional Scaling, History, Theory, and Applications*, ed. R. M. Hammer, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.