Bayesian Multidimensional Scaling and Choice of Dimension

Man-Suk Oh and Adrian E. Raftery

Multidimensional scaling is widely used to handle data that consist of similarity or dissimilarity measures between pairs of objects. We deal with two major problems in metric multidimensional scaling—configuration of objects and determination of the dimension of object configuration—within a Bayesian framework. A Markov chain Monte Carlo algorithm is proposed for object configuration, along with a simple Bayesian criterion, called MDSIC, for choosing their dimension. Simulation results are presented, as are real data. Our method provides better results than does classical multidimensional scaling and ALSCAL for object configuration, and MDSIC seems to work well for dimension choice in the examples considered.

KEY WORDS: Clustering; Dimensionality; Dissimilarity; Markov chain Monte Carlo; Metric scaling; Model selection.

1. INTRODUCTION

Multidimensional scaling (MDS) is concerned with data that are given as similarity or dissimilarity measures between pairs of objects. Its goal is to represent the objects by points in a (usually) Euclidean space. MDS has its roots in psychology, specifically psychophysics, as it is based on the analogy between the psychological concept of similarity and the geometrical concept of distance. Subsequently, it has been widely used in other social and behavioral sciences. Recently, interest in MDS has increased further because of its usefulness in some rapidly developed subjects, such as genomics (Tibshirani et al. 1999) and information retrieval for the Web and other document databases (Schutze and Silverstein 1997).

One of the main applications of MDS is visualization, where the user wants to represent a complex set of dissimilarities in a form that is easier to see. One reason for this is to see if visually apparent clusters are present in the data. Another application is exploration, where the user wants to understand the main dimensions underlying the dissimilarities. For example, the objects in MDS might be political candidates, and the data might consist of subjective similarity judgements. MDS might help to suggest which political positions or characteristics are important in forming similarity judgements (e.g., position on Social Security, age, tendency to tell jokes). A third application is hypothesis testing. Monographs on MDS include Davison (1983), Young (1987), Borg and Groenen (1997), and Cox and Cox (2001).

An important issue in MDS is configuration of objects, i.e., estimation of values for attributes of objects. A commonly used MDS method for pairwise dissimilarity data was developed by Torgerson (1952, 1958). Object configurations are easy to compute with this method, now called classical MDS (CMDS). It gives complete recovery (up to location shift) of object configurations when the given dissimilarities are exactly equal to the Euclidean distances and when the dimension is correctly specified.

Another commonly used MDS method is ALSCAL (Takane et al. 1977), which minimizes the sum of the squared differences of squared dissimilarities and squared distances. The attraction of ALSCAL is that it can analyze data in various forms. In many practical situations, however, there are measurement errors in the observed dissimilarities and no clear notion of dimension.

Maximum likelihood MDS methods have been developed for handling measurement errors; see, for example, Ramsay (1982), Takane (1982), Takane and Carroll (1981), MacKay (1989), MacKay and Zinnes (1986), Groenen (1993), and Groenen, Mathar, and Heisser (1995). However, justification of maximum likelihood relies on asymptotic theory, and computation requires nonlinear optimization. The number of parameters to be optimized over typically grows as fast as the number of objects, so that the asymptotic theory may not apply in high dimensions, as pointed out by Cox (1982). Moreover, the likelihood surface will tend to have many more local minima when there are more dimensions, and finding a good initial estimate will be correspondingly more difficult.

Another important issue in MDS is dimensionality, i.e., the number of significant attributes. Despite its importance in many applications, there is no definitive method for determining dimension for dissimilarity data. The most commonly used method is to search for an elbow, that is, a point where a measure of fit or a measure of contribution to the dissimilarity levels off, in a plot of the measure versus dimension (Spence and Graef 1974; Davison 1983; Borg and Groenen 1997). However, it is often difficult to find an elbow—especially when there are significant errors—and visual inspection of a plot may be misleading, because its appearance often depends on the relative scale of the axes.

In this article, we deal with these two important issues in MDS within a Bayesian framework. We use a Euclidean distance model and assume a Gaussian measurement error in the observed dissimilarity. Under the model, we propose a simple Markov chain Monte Carlo (MCMC) algorithm with which to obtain a Bayesian solution for the object configuration. We found that the proposed method, which we call Bayesian MDS (BMDS), provided a much better fit to the data than did CMDS and a moderately better fit than did ALSCAL in all

Man-Suk Oh is Associate Professor of Statistics, Department of Statistics, Ewha Women's University, Seoul 120-750, Korea (E-mail: *msoh@mm.ewha.ac.kr*). Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Seattle 98195-4322 (E-mail: *raftery@stat.washington.edu*). Oh's research was accomplished with a research fund provided by Korean Research Foundation, Support for Faculty Research Abroad. Raftery's research was supported by ONR grant N00014-96-1-1092. The authors thank Chris Fraley for helpful comments and discussions. They thank Kate Stovel for providing the Lloyds Bank data and for helpful discussions about the data.

Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

the examples we tested. Moreover, the improvement in performance of the proposed BMDS scheme relative to CMDS or ALSCAL was more pronounced when there were significant measurement errors in the data or when the Euclidean model assumption was violated or the dimension was misspecified.

On the basis of the BMDS estimate of object configuration over a range of dimensions, we propose a simple Bayesian criterion with which to choose an appropriate dimension. This criterion, called MDSIC, is based on the Bayes factor, or ratio of integrated likelihoods, for the BMDS estimated configuration under one dimension versus a different dimension. In simulated data, we found that the criterion works well for Euclidean models with measurement error that is not too large. In real examples, we found that the criterion gave satisfactory results. We also give an example of cluster analysis on real dissimilarity data in which the BMDS estimates of object configuration are used in conjunction with model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 1998).

In our approach, observed dissimilarities are modeled as equal to Euclidean distances plus measurement error. In this sense, what we do here can be viewed as a Bayesian analysis of metric MDS, and here being Bayesian seems to confer the benefits of yielding good estimated configurations and provides a formal way to choose the dimension and provide measures of uncertainties in estimations. A great deal of MDS research, however, has focused on nonmetric MDS, in which the relationship between dissimilarity and underlying distance is modeled as nonlinear. One could use the basic ideas here to do Bayesian nonmetric MDS, and we suggest some ways of doing this in Section 6.

The article is organized as follows. Classical MDS and ALSCAL are briefly reviewed in Section 2. Bayesian MDS is described in Section 3: Section 3.1 defines the model and the prior, Section 3.2 presents an MCMC algorithm, and Section 3.3 describes the estimation of object configuration from the MCMC output. Based on the BMDS output, a simple Bayesian dimension selection criterion, MDSIC, is described in Section 4. Some simulated and real examples are given in Section 5. We conclude with discussion in Section 6.

2. CLASSICAL MULTIDIMENSIONAL SCALING AND ALSCAL

2.1 Classical Multidimensional Scaling

Let δ_{ij} denote the dissimilarity measure between objects i and j, which are 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.

Torgerson (1952, 1958) developed a technique for multidimensional scaling, now called CMDS. Assume that the dissimilarity measure, δ_{ij}, is the distance between objects i and j in a p-dimensional Euclidean space, i.e.,

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{1}$$

where x_{ik} is the kth element of $\mathbf{x}_i$. The elements x_{ik} are unknown, and the goal of MDS is to recover them from the dissimilarity data.

Construct a double-centered matrix A with elements a_{ij} defined by $a_{ij} = -\frac{1}{2}(\delta_{ij}^2 - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2)$, where $\delta_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2, \delta_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2, \delta_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$. It was shown by Torgerson (1952, 1958) that

$$a_{ij} = \sum_{k=1}^{p} x_{ik} \cdot x_{jk} \quad \text{for all } i, j, \quad \text{i.e., } A = \mathbf{XX}', \tag{2}$$

where $\mathbf{X}$ is the $n \times p$ matrix of object coordinates. The coordinates of $\mathbf{X}$ can be recovered from the spectral decomposition of the matrix A in (2). If the observed dissimilarities, d_{ij}, satisfy the Euclidean distance assumption and there is no measurement error, then the Euclidean distances computed from the matrix $\mathbf{X}$ satisfying (2) will be exactly equal to the given dissimilarities. However, when the model assumption is violated or when there are significant measurement errors in the data, CMDS estimates of object configuration may not be very useful.

Because Euclidean distance is invariant under translation, rotation, and reflection about the origin, the matrix $\mathbf{X}$ can be centered at the origin and rotated to its principal axes orientation (Borg and Groenen 1997, p. 62).

There is no definitive method for choosing the effective dimension of $\mathbf{x}_i$, the number of object attributes that contribute significantly to the dissimilarities. A common way to assess dimension is to look at the eigenvalues of the scalar product matrix A. The kth eigenvalue is a measure of contribution of the kth coordinate of $\mathbf{X}$ to squared distances. Hence, only the first p coordinates of $\mathbf{X}$ corresponding to the first p significantly large eigenvalues suffice to represent the objects. To determine significantly large eigenvalues, one may draw a plot of the ordered eigenvalues versus dimension and look for a dimension at which the sequence of eigenvalues levels off. If each δ_{ij} is equal to a p-dimensional Euclidean distance between objects i and j as given in (1), then the plot should level off precisely at dimension $(p+1)$.

A measure of fit, called stress, is commonly used to determine the dimensionality. Several definitions of stress have been proposed; the one we use here, and perhaps the mostly commonly used one, is

$$\text{STRESS} = \sqrt{\frac{\sum_{i>j}(d_{ij} - \hat{\delta}_{ij})^2}{\sum_{i>j} d_{ij}^2}},$$

where $\hat{\delta}_{ij}$ is the Euclidean distance obtained from the estimated object configuration (Kruskal 1964). A plot of STRESS versus dimension will level off at the true dimension p if $d_{ij} = \hat{\delta}_{ij}$ and $\hat{\delta}_{ij}$ is given by (1). Note that the squared STRESS is proportional to the sum of squared residuals, $SSR = \sum_{i>j}(d_{ij} - \hat{\delta}_{ij})^2$.

Both methods rely on detecting an elbow in a sequence of values, that is, a point where the plot levels off. However, in real data that do not conform exactly to the model or in which there is a significant amount of measurement or sampling error, an elbow may be difficult to discern.

2.2 ALSCAL

ALSCAL (alternating least squares scaling) was developed by Takane, et al. (1977). It is very general in that it can analyze data given in various forms. In metric MDS, ALSCAL minimizes S-STRESS, where

$$\text{S-STRESS} = \sum_{i>j}(\delta_{ij}^2 - d_{ij}^2)^2.$$

Note that S-STRESS differs from STRESS in that it uses squared distances and dissimilarities, which is done for computational convenience. However, squaring dissimilarities and distances causes S-STRESS to emphasize larger dissimilarities over smaller ones, which may be viewed as a disadvantage of ALSCAL (Borg and Groenen 1997, p. 204). The minimization process can be done by using the Newton–Raphson method, possibly modified. ALSCAL is one of the most commonly used MDS techniques, and it is available in the statistical computer packages SAS and SPSS. For details, see Cox and Cox (2001).

3. BAYESIAN MULTIDIMENSIONAL SCALING

3.1 Model and Prior

Dissimilarity data can be obtained in various forms. However, because Euclidean distance is easy to handle and is relatively insensitive to the choice of dimension compared to other distance measures, it tends to be used in cases in which the dimension is unknown unless there are strong theoretical reasons for preferring a non-Euclidean distance (Davison 1983). Thus, for Bayesian MDS, we model the true dissimilarity measure δ_{ij} as the distance between objects i and j in a Euclidean space, as given in (1).

In practical situations, often there are measurement errors in observations. In addition, dissimilarity measures are typically given as positive values. We therefore assume that the observed dissimilarity measure, d_{ij}, is equal to the true measure, δ_{ij}, plus a Gaussian error, with the restriction that the observed dissimilarity measure is always positive. In other words, given δ_{ij}, the observed dissimilarity measure d_{ij} is assumed to follow the truncated normal distribution

$$d_{ij} \sim N(\delta_{ij}, \sigma^2) I(d_{ij} > 0), \qquad i \neq j, i, j = 1, \ldots, n, \quad (3)$$

where $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$, and x_{ik} are unobserved. From this, the likelihood function of the unknown parameters $\mathbf{X} = \{\mathbf{x}_i\}$ and σ^2 is

$$l(\mathbf{X}, \sigma^2) \propto (\sigma^2)^{-m/2} \exp\left[-\frac{1}{2\sigma^2} SSR - \sum_{i>j} \log \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right], \quad (4)$$

where $m = n(n-1)/2$ is the number of dissimilarities, $SSR = \sum_{i>j}(d_{ij} - \delta_{ij})^2$ is the sum of squared residuals, and $\Phi(\cdot)$ is the standard normal cdf.

For Bayesian analysis of the model, we need to specify priors for $\mathbf{X}$ and σ^2. For the prior distribution of $\mathbf{x}_i$, we use a multivariate normal distribution with mean 0 and a diagonal covariance matrix Λ, i.e., $\mathbf{x}_i \sim N(0, \Lambda)$, independently for $i = 1, \ldots, n$. For the prior of the error variance σ^2, we use a conjugate prior $\sigma^2 \sim IG(a, b)$, the inverse Gamma distribution with mode $b/(a+1)$. For a hyperprior for the elements of $\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_p)$, given dimension p, we also assume a conjugate prior, $\lambda_j \sim IG(\alpha, \beta_j)$, independently for $j = 1, \ldots, p$. We assume prior independence among $\mathbf{X}$, Λ, and σ^2, i.e., $\pi(\mathbf{X}, \sigma^2, \Lambda) = \pi(\mathbf{X})\pi(\sigma^2)\pi(\Lambda)$, where $\pi(\mathbf{X})$, $\pi(\sigma^2)$, and $\pi(\Lambda)$ are the priors given earlier.

When there is little prior information, one may use either the results from a preliminary run or the results from other MDS techniques, such as CMDS or ALSCAL, for parameter specification in the priors. For instance, one may choose a small a for a vague prior of σ^2 and choose b so that the prior mean matches with SSR/m, where SSR is obtained from CMDS or ALSCAL. Similarly, for the hyperprior of λ_j, one may choose a small α and choose β_j so that the prior mean of λ_j matches with the jth diagonal element of the sample covariance matrix $S_x = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i$ of $\mathbf{X}$ obtained from CMDS or ALSCAL. As noted above, the MDS solution can be transformed to have zero sample mean, i.e., $\sum_{i=1}^{n} \mathbf{x}_i = 0$, and a diagonal sample covariance matrix.

Such a prior is mildly data dependent, and it might be argued that this violates the definition of a prior distribution. However, we view this prior as an approximation to the elicited data-independent prior of an analyst who knows a little, but not much, about the problem at hand. Because this prior is diffuse relative to the likelihood, the estimation results are unlikely to be sensitive to its precise specification.

3.2 MCMC

From the likelihood and the prior, the posterior density function of the unknown parameters $(\mathbf{X}, \sigma^2, \Lambda)$ is

$$\begin{aligned}\pi(\mathbf{X}, \sigma^2, \Lambda | D) \propto{}& (\sigma^2)^{-(m/2+a+1)} \prod_{j=1}^{p} \lambda_j^{-n/2} \\ &\times \exp\left[-\frac{1}{2\sigma^2} SSR - \sum_{i>j} \log \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right. \\ &\left. - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i - \frac{b}{\sigma^2} - \sum_{j=1}^{p} \frac{\beta_j}{\lambda_j}\right], \quad (5)\end{aligned}$$

where D is the matrix of observed dissimilarities. Because of the complicated form of the posterior density function (5), numerical integration is required to obtain a Bayes estimate of the parameters. In particular, the posterior is a complicated function of $\mathbf{X}$'s, which in most cases is of high dimension. We therefore use an MCMC algorithm (Gilks, Richardson, and Spiegelhalter 1996) to simulate from the posterior distribution (5). Our algorithm proceeds by iteratively generating new values for each object configuration $\mathbf{x}_i$, the error variance σ^2, and the hyperparameter Λ, given the current values of the other unknowns.

We first suggest initialization strategies for the unknown parameters that are needed for the MCMC algorithm. For initialization of $\mathbf{x}_i$, one may use the output, $\mathbf{x}_i^{(0)}$, of $\mathbf{x}_i$ from CMDS or ALSCAL, because it is easy to obtain. The resulting $\mathbf{X}$ can be centered at the origin and then transformed by using the spectral decomposition so as to have a diagonal sample covariance matrix, thus conforming to the prior. From the adjusted $\mathbf{x}_i^{(0)}$'s, one can compute the sum of squared residuals $SSR^{(0)}$ and $\sigma^{2(0)} = SSR^{(0)}/m$, which can be used as an initial value of σ^2 in the algorithm. In addition, diagonal elements of the adjusted sample covariance matrix of $\mathbf{X}$ can be used as initial values for the λ_j's.

We now describe the details of sample generation in the MCMC algorithm. At each iteration, we simulate a new value of λ_j from its conditional posterior distribution given the other unknowns. From (5), the full conditional posterior distribution of λ_j is the inverse Gamma distribution $IG(\alpha+n/2, \beta_j+s_j/2)$, where s_j/n is the sample variance of the jth coordinates of $\mathbf{x}_i$'s.

We use a random walk Metropolis–Hastings step (Hastings, 1970) to generate $\mathbf{x}_i$ and σ^2 in each iteration of the MCMC algorithm. Specifically, a normal proposal density is used in the random walk Metropolis–Hastings algorithm for generation of $\mathbf{x}_i$. To choose the variance of the normal proposal density, we note that the full conditional posterior density of $\mathbf{x}_i$ is $\pi(\mathbf{x}_i|\cdots) \propto \exp[-\frac{1}{2}(Q_1+Q_2) - \sum_{j\neq i,\, j=1}^{n} \log\Phi(\frac{\delta_{ij}}{\sigma})]$, where $Q_1 = \frac{1}{\sigma^2}\sum_{j\neq i,\, j=1}^{n}(\delta_{ij}-d_{ij})^2$ and $Q_2 = \mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i$. Because $(\delta_{ij}-d_{ij})^2/\sigma^2$ is a quadratic function of $\mathbf{x}_i$ with leading coefficient equal to $1/\sigma^2$, and Q_1 has $n-1$ of this kind whereas Q_2 has only one quadratic term with coefficient Λ, Q_1 would dominate the full conditional posterior density function of $\mathbf{x}_i$ unless there is strong prior information. Thus, we may consider Q_1 only, approximate the full conditional variance of $\mathbf{x}_i$ by $\sigma^2/(n-1)$, and choose the variance of the normal proposal density to be a constant multiple of $\sigma^2/(n-1)$.

From a preliminary numerical study, we found that the full conditional density function of σ^2 is well approximated by the density function of $IG(m/2+a, SSR/2+b)$. When the number of dissimilarities, $m = n(n-1)/2$, is large, which is often the case because m is a quadratic function of n, the inverse Gamma density function is well approximated by a normal density. Thus, we propose a random walk Metropolis–Hastings algorithm with a normal proposal density with variance proportional to the variance of $IG(m/2+a, SSR/2+b)$ distribution.

3.3 Posterior Inference

Iterative generation of $\{\mathbf{x}_i\}$, σ^2, and $\{\lambda_j\}$ for a sufficiently long time provides a sample from the posterior distribution of the unknown parameters, and Bayesian estimation of the parameters can be obtained from the samples. However, because the model assumes a Euclidean distance for the dissimilarity measure δ_{ij}, the posterior samples of $\{\mathbf{x}_i\}$ would be invariant under translation, rotation, and reflection about the origin, as in other MDS, unless there is strong prior information to the contrary. We can retrieve only the relative locations of the objects from the data, not their absolute locations. Hence, the convergence of δ_{ij} rather than $\mathbf{X}$ needs to be checked to verify the convergence of MCMC. The near lack of identifiability in $\mathbf{X}$ also suggests the use of sample averages as Bayes estimates to be inadvisable, because the MCMC samples of $\mathbf{X}$ may be unstable even when the distances δ_{ij} are stable. Thus, we take an approximate posterior mode of $\mathbf{X}$ as a Bayes estimate of $\mathbf{X}$, i.e., the BMDS solution of the object configuration. The posterior mode provides relative positions of $\mathbf{x}_i$'s corresponding to the maximum posterior density. A meaningful absolute position of $\mathbf{X}$ may be obtained from an appropriate transformation, if desired.

To obtain the posterior mode of $\mathbf{X}$, one may compute the posterior density for each MCMC sample. However, this can be time consuming, because the posterior is complicated. However, we observed that the likelihood dominates the prior, and that in the likelihood (4), the term involving SSR is dominant, so that the posterior mode of $\mathbf{X}$ can be approximated by the value of $\mathbf{X}$ that minimizes the sum of squared residuals SSR.

Because the center and direction of $\mathbf{X}$ can be arbitrary, we postprocess the MCMC sample of $\mathbf{X}$ at each iteration by using the transformation $\mathbf{x}_i = D_x'(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is the sample mean of $\mathbf{x}_i$'s and D_x is the matrix whose columns are the eigenvectors of the sample covariance matrix $S_x = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})'(\mathbf{x}_i-\bar{\mathbf{x}})$ of $\mathbf{x}_i$'s. This transformation does not solve the nonidentifiability problem, but the new $\mathbf{x}_i$'s have sample mean 0 and a diagonal covariance matrix to correspond to the prior specification.

Although samples of $\mathbf{x}_i$'s are unstable because of lack of identifiability, samples of δ_{ij}'s are stable after convergence, and hence they can be used to measure uncertainties in quantities that are functions of δ_{ij}'s. For example, one may be interested in whether object i is more closely related to object j than to object k, which can be answered by looking at the posterior distribution of $\delta_{ij} - \delta_{ik}$, as approximated by the MCMC sample.

4. CHOICE OF DIMENSION

As described in the previous section, BMDS gives object configurations in a given dimensional Euclidean space. In most cases, the dimension of the objects (the number of significant attributes) is unknown. In this section, we propose a simple Bayesian dimension selection criterion based on the BMDS object configurations.

Consider the dimension p as an unknown variable, and assume equal prior probability for all values of p up to some maximum value, $p_{\max}$. Then the posterior is given by

$$\begin{aligned}
&\pi(\mathbf{X}, \sigma^2, \Lambda, p|D)\\
&\quad\propto l(\mathbf{X}, \sigma^2, p|D)\pi(\mathbf{X}|\Lambda, p)\pi(\sigma^2)\pi(\Lambda|p)\\
&\quad= (2\pi)^{-m/2}\sigma^{-m}\exp\left[-\frac{1}{2\sigma^2}SSR - \sum_{i>j}\log\Phi(\delta_{ij}/\sigma)\right]\\
&\qquad\times(2\pi)^{-np/2}\prod_{j=1}^{p}\lambda_j^{-n/2}\exp\left[-\sum_{j=1}^{p}\frac{1}{2\lambda_j}s_j\right]\\
&\qquad\times\Gamma(a)^{-1}b^a(\sigma^2)^{-(a+1)}\exp[-b/\sigma^2]\\
&\qquad\times\Gamma(\alpha)^{-p}\prod_{j=1}^{p}\beta_j^{\alpha}\lambda_j^{-(\alpha+1)}\exp[-\beta_j/\lambda_j]\\
&\quad= A(p)\cdot h(\sigma^2, \mathbf{X})\cdot g(\Lambda, \mathbf{X}, p),
\end{aligned}$$

where

$$s_j = \sum_{i=1}^{n} x_{ij}^2, \tag{6}$$

$$A(p) = (2\pi)^{-(m+np)/2}\Gamma(a)^{-1}b^a\Gamma(\alpha)^{-p}\prod_{j=1}^{p}\beta_j^{\alpha}, \tag{7}$$

$$h(\sigma^2,\mathbf{X}) = (\sigma^2)^{-(m/2+a+1)} \times \exp\left[-(SSR/2+b)\Big/\sigma^2 - \sum_{i>j}\log\Phi(\delta_{ij}/\sigma)\right], \quad (8)$$

$$g(\Lambda,\mathbf{X},p) = \prod_{j=1}^{p}\lambda_j^{-(n/2+\alpha+1)}\exp[-(s_j/2+\beta_j)/\lambda_j]. \quad (9)$$

Note that, because of the postprocessing described in Section 3.3, the $\mathbf{x}_i$'s have sample mean 0 and a diagonal sample covariance matrix.

We adopt a Bayesian approach to choosing the dimension. We view the overall task to be that of choosing the best configuration, and hence we view the choice between dimension p and dimension p' as being between the estimated configuration with dimension p and the estimated configuration with dimension p'. Thus, we consider the marginal posterior, $\pi(\mathbf{X},p|D)$, of $(\mathbf{X},p)$ with $\mathbf{X}$ equal to the BMDS solution and choose the value of p that gives the largest value of $\pi(\mathbf{X},p|D)$. Choosing between dimension p and dimension p' is based on the posterior odds for the estimated configuration of dimension p versus that of dimension p'.

Now, note that $\pi(\mathbf{X},p|D) = c\int l(\mathbf{X},\sigma^2,p|D)\pi(\sigma^2)d\sigma^2 \cdot \int\pi(\mathbf{X},\Lambda,p)d\Lambda \approx c\cdot l(\mathbf{X},p|D)\pi(\mathbf{X},p)$, where c is a constant independent of $\mathbf{X}$ and $p, l(\mathbf{X},p|D)$ is the marginal likelihood of $(\mathbf{X},p)$, and $\pi(\mathbf{X},p)$ is the marginal prior of $(\mathbf{X},p)$. The marginalized likelihood term increases as p increases. However, the marginal prior term decreases as p increases, because we are using a diffuse (but proper) prior, and so this term penalizes more complex models. The approach has the simplicity of a maximum likelihood method as well as the advantage of a Bayesian method in penalizing more complex models.

Integrating the function $g(\Lambda,\mathbf{X},p)$ given in (9) with respect to Λ gives $\int g(\Lambda,\mathbf{X},p)d\Lambda = \Gamma^p(n/2+\alpha)\prod_{j=1}^{p}(s_j/2+\beta_j)^{-(n/2+\alpha)}$. The integral of the function $h(\sigma^2,\mathbf{X})$ given in (8) with respect to σ^2 is approximately equal to

$$\int h(\sigma^2,\mathbf{X})d\sigma^2 \approx (2\pi)^{1/2}(m/2)^{-1/2} \times (SSR/m)^{-m/2+1}\exp[-m/2]. \quad (10)$$

This formula is justified in the Appendix. From these,

$$\begin{aligned}\pi(\mathbf{X},p|D) &= c\cdot A(p)\cdot\int h(\sigma^2,\mathbf{X})d\sigma^2\cdot\int g(\Lambda,\mathbf{X},p)d\Lambda \\ &= c\cdot A^*(p)\cdot(SSR/m)^{-m/2+1} \\ &\quad \times\prod_{j=1}^{p}(s_j/2+\beta_j)^{-(n/2+\alpha)}, \qquad (11)\end{aligned}$$

where

$$A^*(p) = A(p)\cdot(2\pi)^{1/2}\Gamma^p(n/2+\alpha)(m/2)^{-1/2}\exp[-m/2].$$

To clarify the dependence of $\mathbf{X}$ on p, let $\mathbf{X}^{(p)}$ denote the BMDS solution of $\mathbf{X}$ when the dimension is p. There is a difficulty in directly comparing $(\mathbf{X}^{(p)},p)$ and $(\mathbf{X}^{(p+1)},p+1)$. The marginal posterior $\pi(\mathbf{X},p|D)$ is dependent on the scale of $\mathbf{X}$, because it includes the term $\prod_{j=1}^{p}(s_j/2+\beta_j)^{-(n/2+\alpha)}$. Note that s_j/n is the sample variance of the jth coordinate of $\mathbf{X}$. However, without improvement in the fit, the scale of $\mathbf{X}$ may change with the dimension p. Given the same Euclidean distances, the coordinates of $\mathbf{X}$ would get closer to the origin as p increases, unless all the extra coordinates are equal to 0. For instance, the Euclidean distance between -1 and 1 in one-dimensional space is equal to the Euclidean distance between $(1/\sqrt{2},1/\sqrt{2})$ and $(-1/\sqrt{2},-1/\sqrt{2})$ in two-dimensional space, and hence the variance in each coordinate is smaller in two-dimensional space. This would give a smaller s_j and hence a larger $\pi(\mathbf{X},p|D)$ in a higher dimension, although there is no change in the distance and the fit.

To circumvent this scale dependency, a dimension selection criterion should compare $\mathbf{X}$'s in the same dimension. For this, let $\mathbf{X}^{*(p+1)} = (\mathbf{X}^{(p)}:\mathbf{0})$ in $(p+1)$-dimensional space, which has the first p coordinates equal to $\mathbf{X}^{(p)}$ and the last coordinates all equal to 0. Then $\mathbf{X}^{*(p+1)}$ provides the same Euclidean distances and the same fit as $\mathbf{X}^{(p)}$ and may be considered an implantation of $\mathbf{X}^{(p)}$ in $(p+1)$-dimensional space. Ideally, if p is the correct dimension, then the optimal solution $\mathbf{X}^{(p+1)}$ in $(p+1)$-dimensional space would be $\mathbf{X}^{*(p+1)}$. Thus, we compare $\mathbf{X}^{*(p+1)}$ and $\mathbf{X}^{(p+1)}$ and choose p to be the dimension if $\mathbf{X}^{*(p+1)}$ has a larger marginal posterior density than $\mathbf{X}^{(p+1)}$.

From (11), the ratio of the marginal posteriors of $\mathbf{X}^{*(p+1)}$ and $\mathbf{X}^{(p+1)}$ is

$$\begin{aligned}R_p &\equiv \frac{\pi(\mathbf{X}^{(p+1)},p+1|D)}{\pi(\mathbf{X}^{*(p+1)},p+1|D)} \\ &= \left(\frac{SSR_{p+1}}{SSR^*_{p+1}}\right)^{-m/2+1}\left(\prod_{j=1}^{p+1}\frac{s_j/2+\beta_j}{s_j^*/2+\beta_j}\right)^{-(n/2+\alpha)} \\ &= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-m/2+1}\left(\prod_{j=1}^{p}\frac{s_j^{(p+1)}/2+\beta_j}{s_j^{(p)}/2+\beta_j}\right)^{-(n/2+\alpha)} \\ &\quad \times\left(\frac{s_{p+1}^{(p+1)}/2+\beta_{p+1}}{\beta_{p+1}}\right)^{-(n/2+\alpha)},\end{aligned}$$

where $s_j^{(p)}$ is s_j given in (6), computed from $\mathbf{X}^{(p)}$. Clearly, the ratio R_p depends on the choice of the hyperparameters α and β_j of Λ.

When there is no strong prior information, a reasonable choice for (α,β_j) in $(p+1)$-dimensional space might be $\alpha = \frac{1}{2}$ and $\beta_j = \frac{1}{2}s_j^{(p+1)}/n$, so that the prior information roughly corresponds to the information from one observation. This is close to the unit information prior, which was observed by Kass and Wasserman (1995) to correspond to the Bayesian information criterion (BIC) approximation to the Bayes factor (Schwarz 1978) and by Raftery (1995) to correspond to a similar approximation to the integrated likelihood. Raftery (1999) argued that this is a reasonable proper prior for approximating the situation where the amount of prior information is small.

This yields the ratio

$$R_p = \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-m/2+1}\left(\prod_{j=1}^{p}\frac{r_j^{(p+1)}(n+1)}{(n+r_j^{(p+1)})}\right)^{-(n+1)/2} \times (n+1)^{-(n+1)/2},$$

where $r_j^{(p+1)} = s_j^{(p+1)}/s_j^{(p)}$. Taking minus twice the logarithm of the ratio gives

$$\begin{aligned} LR_p &\equiv -2 \log R_p \\ &= (m-2)\log(SSR_{p+1}/SSR_p) \qquad (12) \\ &\quad + \left\{ (n+1)\sum_{j=1}^{p} \log\left[\frac{r_j^{(p+1)}(n+1)}{(n+r_j^{(p+1)})} \right] \right. \\ &\quad \left. + (n+1)\log(n+1) \right\}. \qquad (13) \end{aligned}$$

Note that the term (12) in LR_p is roughly the log-likelihood ratio and would be negative because higher dimension results in a smaller SSR. The term (13) plays the role of penalty on the increase of dimension by 1 and would be positive if $\prod_{j=1}^{p}(n/r_j^{(p+1)}+1) < (n+1)^{p+1}$. When there is no significant change in $\mathbf{X}$ between p- and $(p+1)$-dimensional spaces, then $r_j^{(p+1)} \approx 1$, and the penalty term is approximately $(n+1)\log(n+1)$.

A positive LR_p would prefer the dimension p to $(p+1)$ and a negative value would prefer the dimension $(p+1)$ to p, and hence one can select the dimension where the value of LR_p turns positive. Alternatively, if we define MDSIC as

$$\begin{aligned} \mathrm{MDSIC}_1 &= (m-2)\log SSR_1, \\ \mathrm{MDSIC}_p &= \mathrm{MDSIC}_1 + \sum_{j=1}^{p-1} LR_j, \qquad (14) \end{aligned}$$

then the optimal dimension is the one that achieves the minimum of $MDSIC_p$.

5. EXAMPLES

BMDS requires that prior parameters be specified. For all the examples given in this section, we chose 5 as the degrees of freedom a for the prior of σ^2, and we chose b to match the prior mean of σ^2 with SSR/m obtained from the CMDS or ALSCAL. Note that a smaller a would not make much difference because $m = n(n-1)/2$ is large. For the hyperprior of λ_j, we chose $\alpha = 1/2$ and $\beta_j = \frac{1}{2}s_j^{(0)}/n$, where $s_j^{(0)}/n$ is the sample variance of the jth coordinate of $\mathbf{X}$ obtained from CMDS or ALSCAL, which roughly corresponds to information from one observation as described in Section 4.

For the multiplicity constant of the variance of the normal proposal density in the Metropolis–Hastings algorithms for generating $\mathbf{x}_i$ and σ^2, we chose 2.38^2 for both $\mathbf{x}_i$ and σ^2 as suggested by Gelman, Roberts, and Gilks (1996). We found reasonably fast mixing in MCMC with this choice of multiplicity constant.

In all the examples, we ran 13,000 iterations of MCMC and observed very quick convergence in σ^2 and the δ_{ij}'s. Here we are interested in an approximate posterior mode of $\mathbf{X}$ rather than its full posterior distribution, so that convergence requirements are less stringent than if one seeks the full posterior distribution, and all the iterations may be used for the purpose of obtaining the $\mathbf{x}_i$'s that give minimum STRESS.

5.1 A Simulation

As an illustrative example, we generated 50 random samples of $\mathbf{x}_i$ from a 10-dimensional multivariate normal distribution with mean 0 and variance I, the identity matrix. We used the Euclidean distances between pairs $(\mathbf{x}_i, \mathbf{x}_j)$ as dissimilarities δ_{ij}. Given these δ_{ij}'s, we generated the observed distances d_{ij} from a normal distribution with mean δ_{ij} and standard deviation .3, truncated at 0. Thus, the data consist of a 50×50 symmetric matrix of dissimilarities computed from Euclidean distances with Gaussian errors.

Using the results from ALSCAL for initialization, BMDS as described in Section 3 was applied for various values of the dimension p. With minimum STRESS and $\mathbf{x}_i$ obtained from BMDS, we applied MDSIC described in Section 4 to select the dimension of $\mathbf{x}_i$. The results are summarized in Table 1.

The table presents values of STRESS from CMDS, ALSCAL, and BMDS. It also presents the likelihood ratio term of (12), the penalty term of (13), and the MDSIC given in (14). It can be observed that BMDS shows significant improvement over CMDS, providing a smaller STRESS, and a moderate improvement over ALSCAL when the dimension is low. When the dimension is close to the true dimension 10, BMDS shows a moderate improvement over CMDS and the same results as ALSCAL. It is interesting to observe that the better performance of BMDS is more pronounced when the dimension is low because, for visualization purposes, dimension $p = 2$ is often chosen.

The table shows that the log-likelihood ratios computed from the BMDS solution for $\mathbf{X}$ decrease monotonically as p increases up to 10, that there is no significant change after dimension 10, and that the penalties stay about the same for various p. Moreover, the MDSIC assumes a minimum at the correct dimension, namely, 10.

We also applied BMDS with initial values obtained from the CMDS results. It provided about the same results as before except that it gives slightly larger STRESS (with the difference less than .01) when the dimension is larger than or equal to 5. MDSIC with the BMDS results chose the same dimension, 10, as in the previous case.

Table 1. Analysis of Simulation Data in Example 1, $\mathbf{x}_i \sim N_{10}(0, I)$

dim	CMDS STRESS	ALSCAL STRESS	BMDS STRESS	LRT	Penalty	MDSIC
1	.6622	.5079	.4813	−1105.8	173.0	10647
2	.4943	.3198	.3063	−828.7	170.3	9714
3	.3720	.2250	.2182	−695.5	169.5	9056
4	.2751	.1648	.1642	−699.3	165.3	8530
5	.2037	.1234	.1234	−554.6	170.8	7996
6	.1580	.0984	.0984	−593.2	172.6	7612
7	.1092	.0772	.0772	−414.2	177.9	7191
8	.0809	.0652	.0652	−267.2	179.0	6955
9	.0672	.0584	.0584	−252.8	181.0	6867
10	.0614	.0527	.0527	−72.4	187.2	6795*
11	.0658	.0511	.0511	−53.2	186.5	6910
12	.0715	.0500	.0500	−14.4	186.7	7043
13	.0784	.0497	.0497	−11.3	187.2	7216
14	.0855	.0495	.0495	−9.4	185.9	7391

*Minimum MDSIC.

5.2 Airline Distances Between Cities

Hartigan (1975, p. 192) provided airline distances between 30 principal cities of the world; these are shown in Figure 1. Cities are located on the surface of the earth, a three-dimensional sphere, and airplanes travel on the surface of the earth. Thus, airline distances are not exactly Euclidean distances and we may expect the dimension of $\mathbf{x}_i$ to be between 2 and 3.

The BMDS was applied to the data. In this example, initial values from CMDS and ALSCAL yielded almost the same BMDS results. The BMDS results from CMDS initial values are shown in Table 2.

BMDS yielded much smaller SSR than did CMDS and moderately smaller SSR than did ALSCAL in all cases. The estimated SSR from BMDS dropped very quickly until dimension 3 and then increased slightly at dimension 4. MDSIC selected dimension 3. We observed that the last coordinates of $\mathbf{x}_i$ in dimension 4 are almost equal to 0, indicating strong evidence for dimension 3.

Table 2. Analysis of City Data

dim	CMDS STRESS	ALSCAL STRESS	BMDS STRESS	LRT	Penalty	MDSIC
1	.6782	.4007	.3617	−704.2	95.7	5336
2	.4682	.1795	.1604	−548.5	91.4	4727
3	.3811	.0903	.0851	4.7	108.0	4270*
4	.4006	.0902	.0856	−2.4	88.9	4383
5	.4139	.0902	.0854	4.1	143.9	4469

*Minimum MDSIC.

Figure 2 is a plot of the observed airline distances versus the estimated Euclidean distances. A perfect fit would yield a 45-degree line, as shown in Figure 2. The estimated Euclidean distances from BMDS are represented as red dots, those from CMDS as blue dots, and those from ALSCAL as green dots. One can see that BMDS provided points very close to the 45-degree line, except for some points corresponding to large distances. The fit gets worse as the distance gets larger,

									NY	22	36	48	43	26	51	24	73	100	68
Azores	AZ									PY	54	33	59	33	31	37	93	88	84
Bagdad	39	BD									PS	57	7	56	72	50	57	105	61
Berlin	22	20	BN									RO	57	66	18	69	113	84	115
Bombay	59	20	39	BY									RE	63	74	57	57	101	61
Buenos Aires	54	81	74	93	BS									SF	59	7	61	74	52
Cairo	33	8	18	27	73	CO									SO	64	117	71	107
Capetown	57	49	60	51	43	45	CN									SE	57	77	48
Chicago	32	64	44	81	56	61	85	CH									SI	49	11
Guam	89	63	71	48	104	71	88	74	GM									SY	48
Honolulu	73	84	73	80	76	88	115	43	38	HU									TO
Istanbul	29	10	11	30	76	8	52	55	69	81	IL								
Juneau	46	61	46	69	77	63	103	23	51	28	55	JU							
London	16	25	6	45	69	22	60	40	75	72	16	44	LN						
Manila	83	49	61	32	111	57	75	81	16	53	57	59	67	MA					
Melbourne	120	81	99	61	72	87	64	97	35	55	91	81	105	39	ME				
Mexico City	45	81	61	97	46	77	85	17	75	38	71	32	56	88	84	MY			
Montreal	24	58	37	75	56	54	79	8	77	49	48	26	33	82	104	23	ML		
Moscow	32	16	10	31	84	18	63	50	61	70	11	46	16	51	90	67	44	MW	
New Orleans	36	72	51	89	49	68	83	8	77	42	62	29	46	87	93	9	14	58	NS
New York	25	60	40	78	53	56	78	7	80	50	50	29	35	85	104	21	3	47	12
Panama City	38	78	59	97	33	71	70	23	90	53	68	45	53	103	90	15	25	67	16
Paris	16	24	5	44	69	20	58	41	76	75	14	47	2	67	104	57	34	16	48
Rio De Janeiro	43	69	62	83	12	61	38	53	116	83	64	76	57	113	82	48	51	72	48
Rome	21	18	7	38	69	13	52	48	76	80	9	53	9	65	99	64	41	15	55
San Francisco	50	75	57	84	64	75	103	19	58	24	67	15	54	70	79	19	25	59	19
Santiago	57	88	78	100	7	80	49	53	98	69	81	73	72	109	70	41	54	88	45
Seattle	46	68	51	77	69	68	102	17	57	27	61	9	48	67	82	23	23	52	21
Shanghai	72	44	51	31	122	52	81	70	19	49	49	49	57	12	50	80	70	42	77
Sydney	121	83	100	63	73	90	69	92	33	51	93	77	106	39	4	81	100	90	89
Tokyo	73	52	56	42	114	60	92	63	16	39	56	40	60	19	51	70	65	47	69

Figure 1. Airline Distances Between Cities (100 miles, .62 mi = 1 km) (Hartigan 1975).

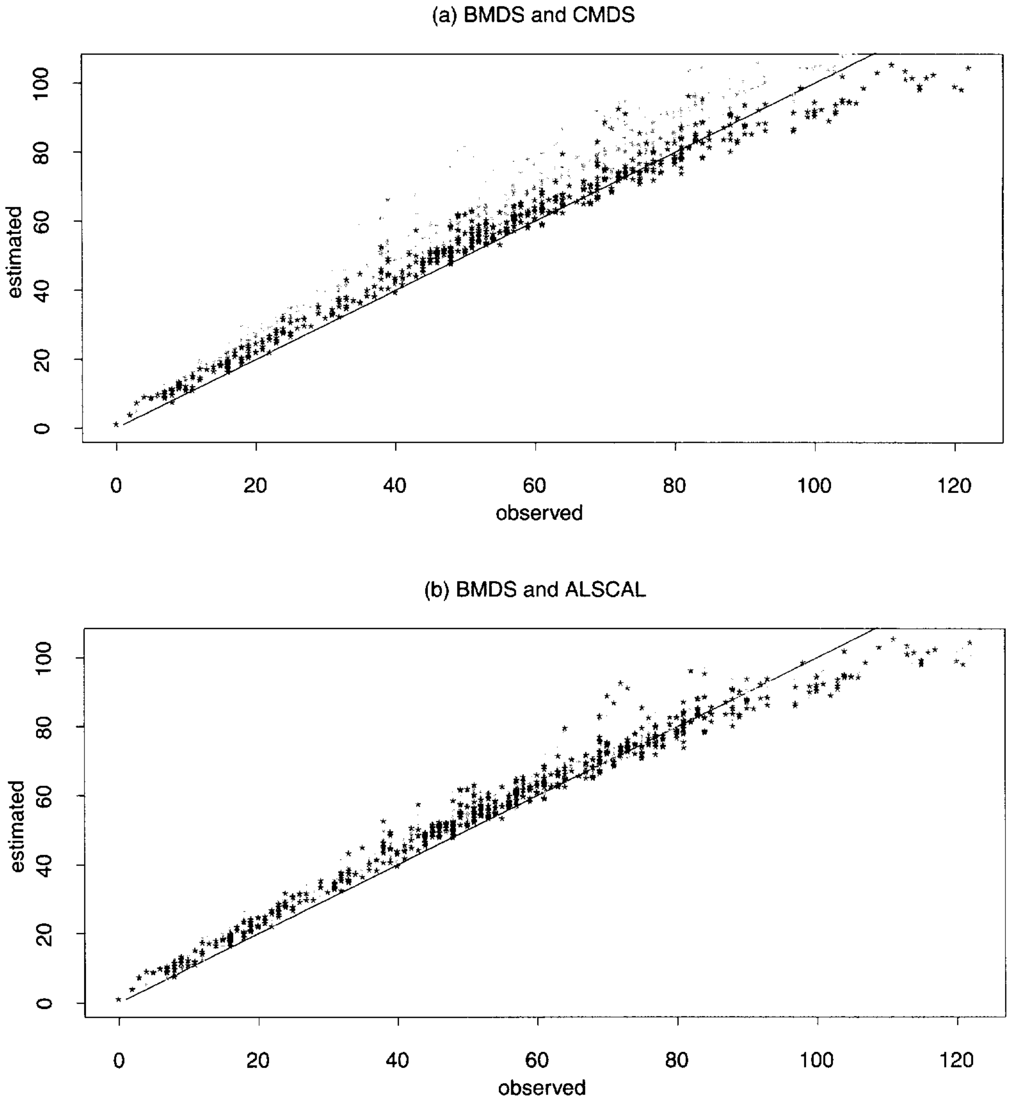

Figure 2. Observed and Estimated Distances for the Airline Distance Data (in Units of 100 mi). Red dots represent the estimated distances from BMDS, blue dots in (a) from CMDS, and blue dots in (b) from ALSCAL.

because when cities are farther apart, there is a greater discrepancy between airline distance and three-dimensional Euclidean distance.

Figure 3 shows plots for the locations of cities, obtained from BMDS with dimension $p = 3$ and rotated manually to fit the true locations of the cities approximately. One can observe that the cities are located approximately on the surface of a sphere with the radius of the earth and that the locations of the cities are well recovered.

5.3 Careers of Lloyds Bank Employees, 1905–1950

Sociologists are interested in characterizing and describing careers, to answer questions such as, What are the typical career patterns in a given period in a particular society? Have they been changing over time? Have people become more mobile occupationally?

One approach to doing this views a career as a sequence of occupations held, for example, in successive years and then seeks to measure the similarity or dissimilarity between careers and, hence, to find groups of similar careers. Abbott and Hrycak (1990) proposed measuring the dissimilarity between the careers of two individuals by counting the minimum number of insertions, deletions, and replacements that would be necessary to transform one career into another. Costs are associated with each kind of change, and the dissimilarity between the two careers is then measured as the total cost of transforming one career into another. This approach, known as optimal alignment, is borrowed from molecular microbiology, where it is applied to the comparison of DNA and protein sequences (Sankoff and Kruskal 1983; Boguski 1992).

Here we reanalyze some data considered by Stovel, Savage, and Bearman (1996), consisting of the careers of 80 randomly selected employees of Lloyds Bank in England, whose careers started between 1905 and 1909. This is part of a much larger study aimed at discovering how career patterns in large organizations evolved over the course of the twentieth century. The more immediate goal here is to discover what the typical career sequences are, for data reduction and exploratory purposes, and also as a basis for further analysis. For each employee, information about his work position is available for

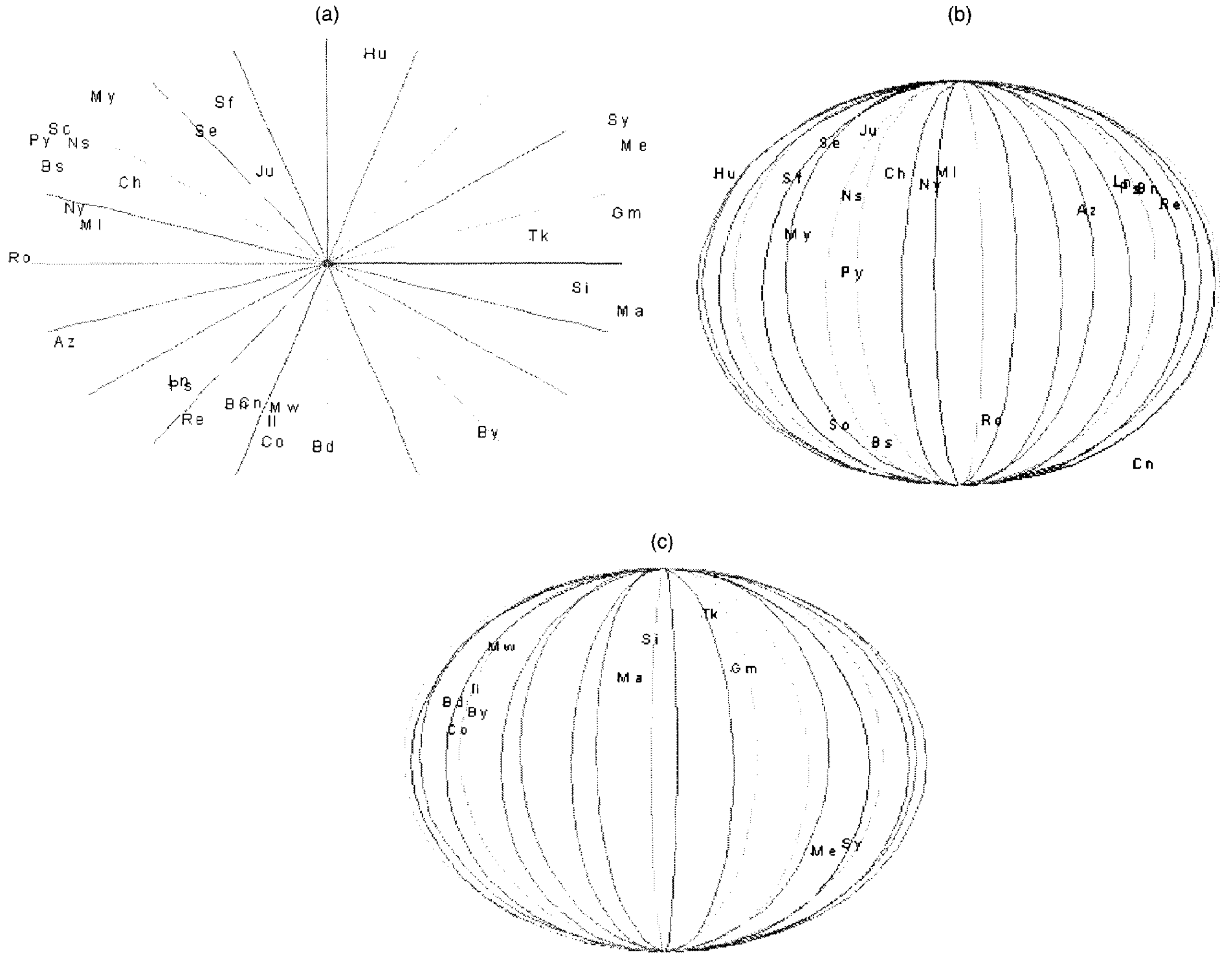

Figure 3. Estimated Locations of the Cities From BMDS. (a) View from the North Pole. (b) Hemisphere I. (c) Hemisphere II.

each year he was at Lloyds. The information consists of the nature of the position (four categories, from clerk to senior manager) and the kind of place they were in (six categories, from small rural place to large city).

From the sequence data, an 80 × 80 matrix of dissimilarity measures was obtained, by using the method of Abbott and Hrycak (1990); for details, see Stovel et al. Clearly, the dissimilarities are not Euclidean distances, and they may not satisfy certain geometric properties that hold for Euclidean distances, such as the triangle inequality. Our approach is to model the dissimilarities as before, with the idea that the non-Euclidean nature of the dissimilarities can be modeled at least approximately as part of the error. As we show, this supposition turns out to be reasonable in practice.

We applied BMDS to the dissimilarity data with initial values from CMDS. (Initial values from ALSCAL gave almost the same results.) Table 3 presents the results of the analysis along with STRESS from CMDS and ALSCAL. Again, BMDS performed much better than did CMDS and moderately better than did ALSCAL, especially when the dimension was too small. The improvement in performance of BMDS is more pronounced in this example than in the two previous examples. This suggests that BMDS is more robust than CMDS or ALSCAL to variations in the alleged dimension and to violations of the Euclidean model assumption.

Dimension 8 is chosen as optimal because MDSIC attains its minimum at 8. Thus, the estimated configuration **X** when $p = 8$ can be used as a final estimate of **X**. Figure 4 shows the fitted and observed dissimilarities for the three MDS techniques. The BMDS fitted dissimilarities fit the observed ones very well, considerably better than the CMDS ones and

Table 3. Analysis of the Lloyd Bank Data

dim	*CMDS STRESS*	*ALSCAL STRESS*	*BMDS STRESS*	*LRT*	*Penalty*	*MDSIC*
1	.5357	.4648	.3545	−4228.1	325.8	26924
2	.3390	.2228	.1815	−3380.3	315.7	23022
3	.2190	.1297	.1063	−2924.9	310.9	19957
4	.1280	.0853	.0669	−1540.1	317.5	17343
5	.0891	.0623	.0524	−941.2	330.5	16120
6	.0725	.0530	.0452	−600.9	326.7	15510
7	.0619	.0472	.0411	−392.7	330.3	15236
8	.0558	.0414	.0386	−221.8	330.5	15173*
9	.0547	.0384	.0372	27.8	367.6	15282
10	.0556	.0374	.0374	7.7	396.1	15677
11	.0600	.0370	.0375	−17.3	310.3	16081
12	.0637	.0363	.0374	−21.2	444.5	16374

*Minimum MDSIC.

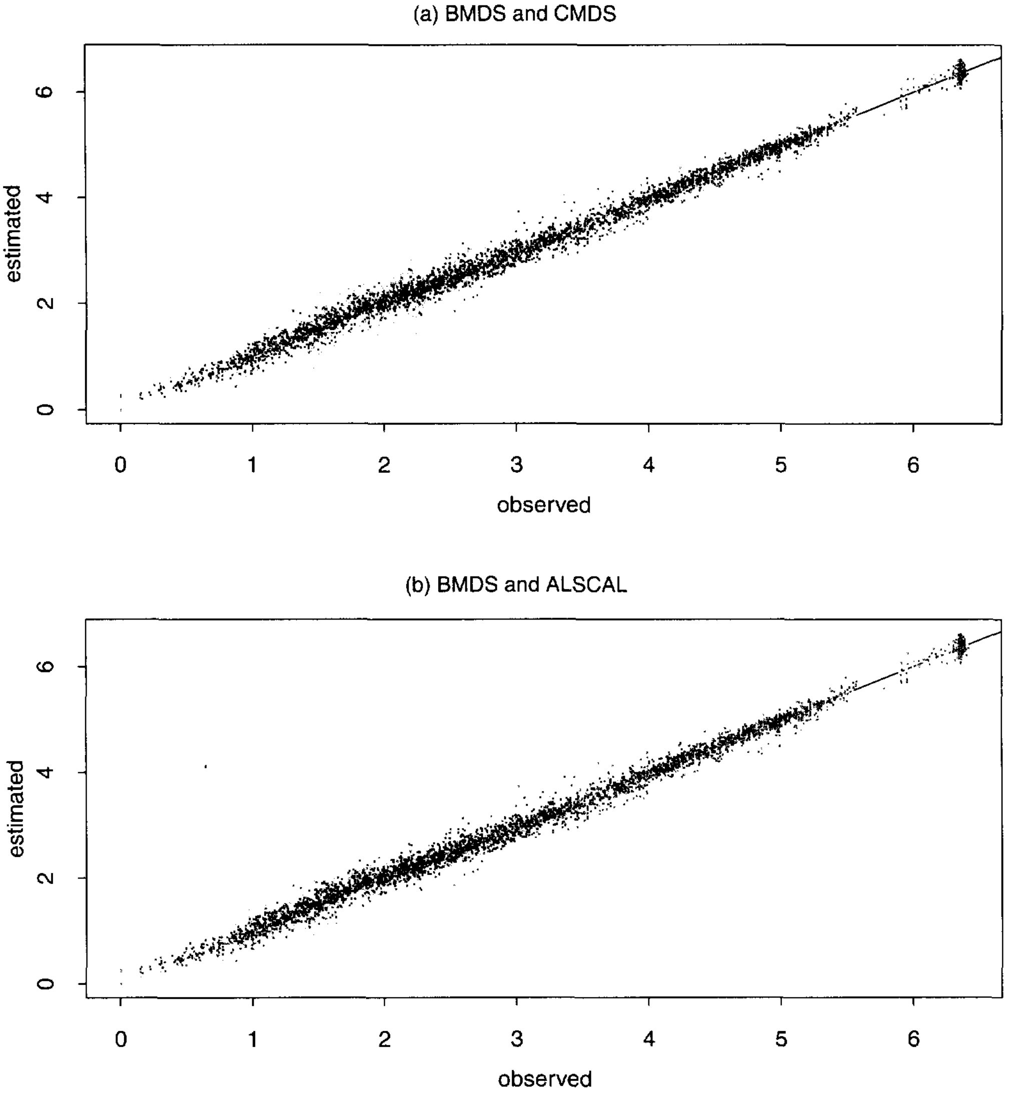

Figure 4. Fitted and Observed Dissimilarities for the Lloyd Bank Data. Red dots represent the estimated distances from BMDS, blue dots in (a) from CMDS, and blue dots in (b) from ALSCAL.

moderately better than the ALSCAL ones; the sum of squared residuals for BMDS is 48% of that for CMDS and 87% of that for ALSCAL.

Figure 5 gives pairwise scatterplots of the first four dimensions of the BMDS estimates of **X**. The fourth dimension clearly separates two outliers. On closer inspection of the data, it turned out that these were individuals who had very short careers at Lloyds. They spent only a few years there, whereas all the other employees were at Lloyds for at least 10 years.

The sociologists' interest in these data is primarily to characterize the typical career patterns at Lloyds in this period. To try to answer this question, we applied model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 1998) to the BMDS estimate of **X**, after removing the two clear outliers. This models the data as a mixture of multivariate normals, allowing for possible geometrically motivated constraints on the covariance matrices of the different groups. The number of groups and the clustering model are chosen by using approximate Bayes factors, approximated via BIC.

Model-based clustering clearly identified three groups. These are shown in Figure 6, which displays the first two components of the BMDS solution. The three groups selected make clear substantive sense: Group 1 consists of 16 employees who had shorter careers (22 years or fewer) and spent all or almost all of their career at the lowest clerk rank. Group 2 consists of 30 employees with long careers (40 years or more), almost all of whom ended their careers at the lowest clerk level. Group 3 consists of 32 employees, most of whom were promoted and ended their careers as managers.

Another interest in these data would be in whether the career pattern of individual i is more similar to that of individual j than to that of individual k, or whether the difference in the career patterns between individuals i and j is significantly different from that between individuals i and k. This can be answered by considering the posterior distribution of $\delta_{ij} - \delta_{ik}$. Because BMDS generates posterior samples of δ's after convergence, one can easily perform the test by using the MCMC samples of δ's. To illustrate this, we picked some individuals i, j, k and compared δ_{ij} with δ_{ik}. Empirical 2.5%, 5%, 50%,

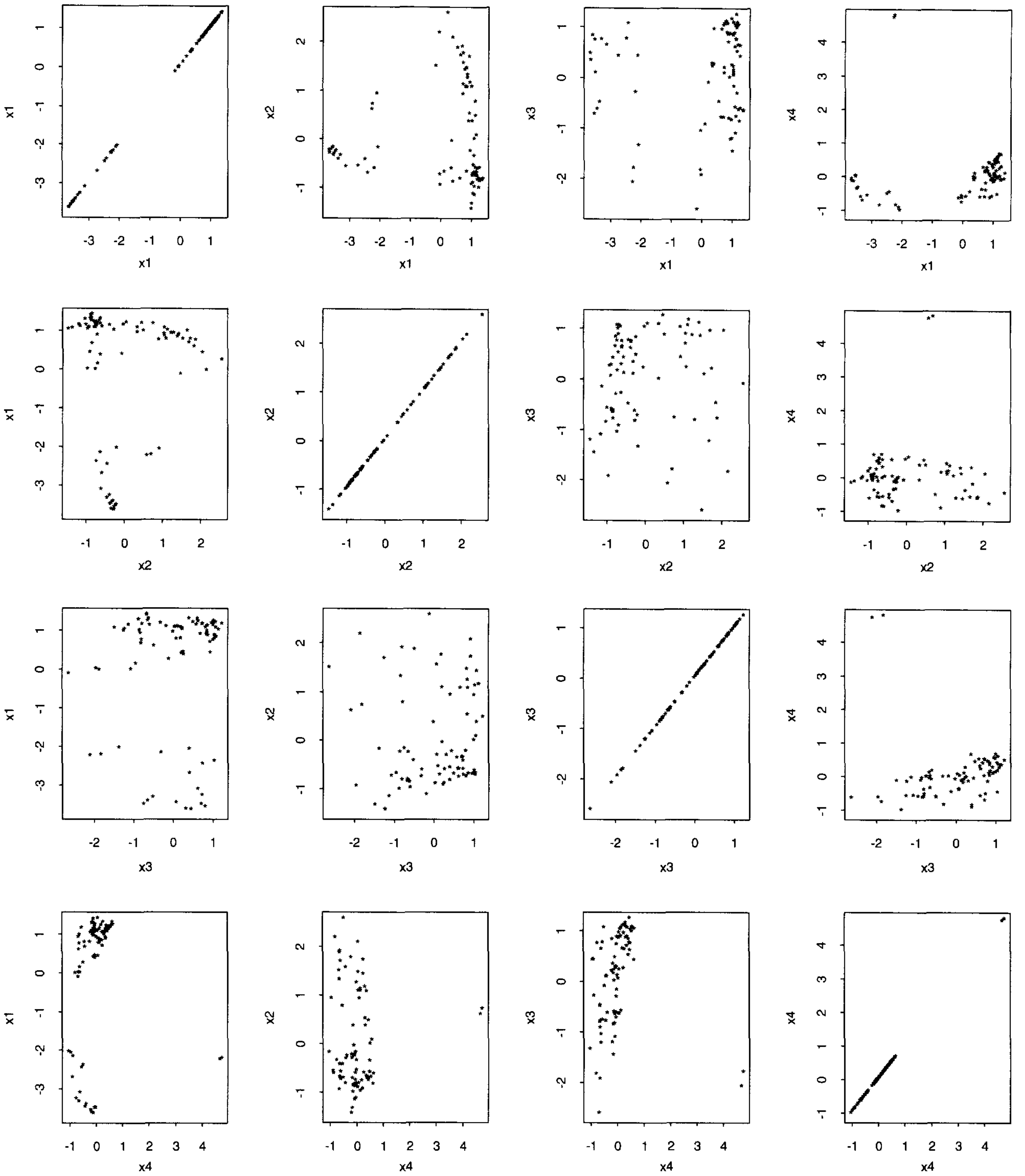

Figure 5. Pairwise Scatterplots of the Estimated Object Configuration From BMDS for the Lloyd Bank Data.

95%, and 97.5% posterior percentiles of $\delta_{ij} - \delta_{ik}$ and its standard error, obtained from the MCMC samples, are presented in Table 4 for some selected values of (i, j, k). One can see that the career pattern of individual 21 (72) is significantly closer to that of individual 38 (79) than to that of individual 75 (24). On the other hand, individual 29 is not significantly closer to either 79 or 80 in terms of career patterns. Note that individuals 21, 38, and 75 are in the first group, individuals 72 and 79 in the second group, and individuals 24, 29, and 80 in the third group.

6. DISCUSSION

In this article, we proposed a Bayesian approach to object configuration in multidimensional scaling and a simple Bayesian dimension choice criterion, MDSIC, based on the results from BMDS. Key advantages of the proposed Bayesian multidimensional scaling are as follows. First, it provided a significantly better fit than classical MDS and a moderately better fit than ALSCAL overall. This implies that BMDS explores the posterior distribution quite well compared to the other MDS methods. The improvement in performance of

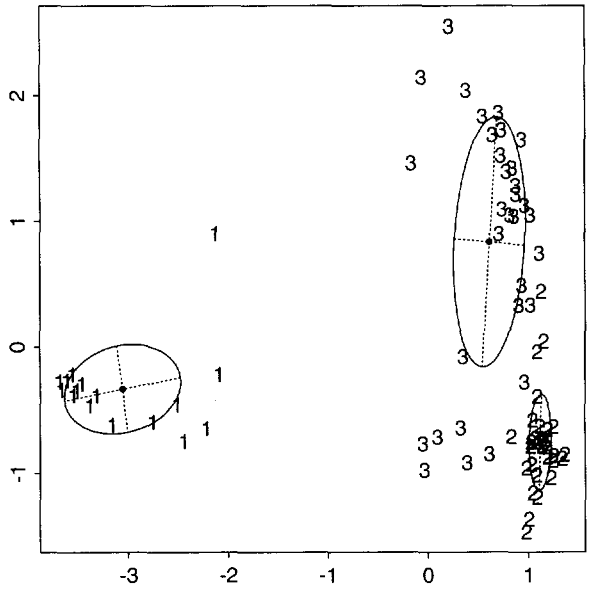

Figure 6. Lloyds Bank Data: First Two BMDS Dimensions, With the Three-Group Model-Based Clustering Classification Shown. The ellipses show the one-standard-deviation contours of the densities of each of the component multivariate normal distributions, and the dotted lines show their principal axes.

BMDS is more pronounced when the dissimilarities are different from Euclidean distances and the effective dimension is ambiguous. This sort of robustness is useful in practice, because in applications dissimilarities often are not Euclidean distances, and the concept of dimension may not even arise in their formulation. Another consideration is that one may often want to use two or three dimensions for visual display, although the true dimension may be much higher. Second, the proposed dimension selection criterion, MDSIC, is easy to compute and gives a direct indication of optimal dimensionality. In this article, we based MDSIC on the Bayesian object configuration. However, it may, in fact, be used independently of MDS techniques, and any good MDS solution can be used for MDSIC for choosing the dimension. In our three examples, MDSIC based on ALSCAL solutions gave the same dimension choices as those based on BMDS solutions. Finally, BMDS used MCMC to generate posterior samples of unknown parameters. Thus, unlike other MDS methods, it can provide estimation errors of the distances, as illustrated in the analysis of careers of Lloyds Bank employes.

Compared with CMDS, a key feature of BMDS is that when the dimension increases, the coordinates for lower dimensions are changed, whereas in CMDS the coordinates for a lower dimension are always a subset of those for a higher one. The coordinates obtained from the lower dimensions are not necessarily an optimal choice when the dimension increases, and retaining them in higher dimensions may adversely affect the performance of CMDS. Compared with ALSCAL, BMDS gives an object configuration that minimizes STRESS, whereas ALSCAL gives one that minimizes S-STRESS—the sum of the squared difference in the squared dissimilarities and the squared distances. Squaring dissimilarities and distances causes S-STRESS to emphasize larger dissimilarities more than smaller ones (Borg and Groenen 1997).

Table 4. Posterior Percentiles and Standard Error of $\delta_{ij} - \delta_{ik}$

(i,j,k)	2.5%	5%	50%	95%	97.5%	SE
(21,38,75)	−1.609	−1.568	−1.331	−1.076	−1.033	.151
(72,24,79)	.037	.065	.222	.390	.419	.098
(29,79,80)	−.265	−.228	−.028	.168	.199	.121

A common reason for doing MDS is to cluster the objects. In our third example, we showed how model-based clustering can be used to do this, providing a formal basis for choosing the number of groups. The results were substantively reasonable and useful. Combining BMDS and model-based clustering thus provides a fully model-based approach to clustering dissimilarity data, including ways to choose the dimension of the data and the number of groups.

A more comprehensive approach to this problem would be to build a single model and carry out Bayesian inference for it. This could be done by using a prior distribution of **X** that is based on a mixture of multivariate normal distributions, rather than on a single one as here. Then MCMC could be used to estimate both object configuration and group membership simultaneously. This approach could provide a way to choose the dimension and the number of groups simultaneously, rather than sequentially, as we did in our example. This seems desirable because there may be a trade-off between dimension and number of groups. A maximum likelihood approach to the problem of clustering with multidimensional scaling of two-way dominance or profile data was proposed by DeSarbo, Howard, and Jedidi (1991), but this is somewhat different from the present context, where the data come in the form of dissimilarities.

We modeled dissimilarities as being equal to Euclidean distances plus error. This corresponds to metric scaling, and so our approach would perhaps best be called Bayesian *metric* multidimensional scaling. There is a great emphasis in the MDS literature on nonmetric scaling, however. In nonmetric scaling, dissimilarities are modeled as equal to a nonlinear function of distance plus error. This could be incorporated in the present framework by replacing (3) by

$$d_{ij} \sim N\big(g(\delta_{ij}), \sigma^2\big) I(d_{ij} > 0), \quad i \neq j, \quad i,j = 1, \ldots, n, \qquad (15)$$

where $g(\cdot)$ is a nonlinear but monotonic function. One could postulate a parametric model, or a family of parametric models, for g; one such family of models was proposed by Ramsey (1982). Then standard Bayesian inference via MCMC would again be possible, leading to Bayesian nonmetric multidimensional scaling.

We used the truncated normal distribution for the observed distance because it is approximately a conjugate form for the normal prior distribution of **X**, if the restriction is ignored, and this makes it easy to find a reasonable proposal density in the Metropolis–Hastings algorithm for generating **X**. The truncated normal distribution seems to work well in practice, as illustrated in the examples. Other distributions, such as the log-normal and noncentral chi-squared, may be used instead of the truncated normal distribution (Suppes and Zinnes 1963; Ramsay 1969; Ramsay 1977). The use of other distributions would require only a slight modification of the Metropolis–Hastings step for generating **X** in the proposed algorithm.

We assumed equal variances for objects but different variances for coordinates. When there are replicated measures on each pair of objects, one may further assume different variances for different objects to incorporate the object variation. The proposed BMDS can be easily modified to handle the case of both object and coordinate variations.

Apart from the present work, we do not know of any Bayesian analyses of multidimensional scaling for dissimilarity data. DeSarbo, Kim, and Fong (1999) proposed a Bayesian approach to multidimensional scaling when the data are in the form of binary choice data, but this is rather different from the present context, where the data take the form of dissimilarities.

A Fortran program implementing the proposed method is freely available through StatLib.

APPENDIX: JUSTIFICATION OF (10)

Integration of $h(\sigma^2,\mathbf{X})$, where

$$h(\sigma^2,\mathbf{X})=(\sigma^2)^{-(m/2+a+1)}\exp\left[-\frac{SSR/2+b}{\sigma^2}-\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]$$

is not straightforward. However, in most cases, $m=n(n-1)/2$ is very large and the likelihood of σ^2 dominates the prior and hence $h(\sigma^2,\mathbf{X})$ is approximately proportional to the likelihood

$$l(\sigma^2,\mathbf{X})\equiv(\sigma^2)^{-m/2}\exp\left[-\frac{SSR}{2\sigma^2}-\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]. \qquad (A.1)$$

In addition, because of the large m, the likelihood $l(\sigma^2,\mathbf{X})$ is well approximated by a normal density function. Thus, applying a Laplace approximation to the integral of $l(\sigma^2,\mathbf{X})$ gives

$$\int h(\sigma^2,\mathbf{X})d\sigma^2\approx\int l(\sigma^2,\mathbf{X})d\sigma^2\approx(2\pi)^{1/2}H^{-1/2}l(\mathbf{X},\hat{\sigma}^2), \qquad (A.2)$$

where H is the minus Hessian of the log-likelihood and $\hat{\sigma}^2$ is the MLE of σ^2.

We now argue that the probability $\Phi(\delta_{ij}/\sigma)$ is unlikely to have much effect on the model comparison and can safely be ignored. Suppose we are comparing dimension p with dimension $(p+1)$. We distinguish between two situations. Suppose first that the true dimension is $(p+1)$. Then, asymptotically, the term $(-SSR/2\sigma^2)$ will dominate the exponent on the right side of (A.1), dimension $(p+1)$ will be preferred, and the term $\sum_{i>j}\log\Phi(\delta_{ij}/\sigma)$ will be immaterial. Second, suppose instead that the true dimension is p. Then, the fitted δ_{ij} will be the same, asymptotically, for dimension p as for dimension $(p+1)$, and so the term $\sum_{i>j}\log\Phi(\delta_{ij}/\sigma)$ will be the same for both dimensions. Thus, it will cancel in the comparison and can again be ignored.

Thus, we ignore the term $\sum_{i>j}\log\Phi(\delta_{ij}/\sigma)$ and use the approximation

$$l(\sigma^2,\mathbf{X})\approx l^*(\sigma^2,\mathbf{X})\equiv(\sigma^2)^{-m/2}\exp\left[-\frac{SSR}{2\sigma^2}\right].$$

Replacing l by l^* and H by the minus Hessian H^* of l^* in (A.2) and letting $\hat{\sigma}^2=SSR/m$, which maximizes l^*, gives the formula (10).

[Received July 2000. Revised February 2001.]

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.

Boguski, M. S. (1992), "Computational Sequence Analysis Revisited," *Journal of Lipid Research*, 33, 957–974.

Borg, I., and Groenen, P. (1997), *Modern Multidimensional Scaling*, New York: Springer-Verlag.

Cox, D. R. (1982), "Comment," *Journal of the Royal Statistical Society*, Ser. A, 145, 308–309.

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

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

DeSarbo, W. S., Howard, D. J., and Jedidi, K. (1991), "MULTICLUS—A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis," *Psychometrika*, 56, 121–136.

DeSarbo, W. S., Kim, Y., and Fong, D. (1999), "A Bayesian Multidimensional Scaling Procedure for the Spatial Analysis of Revealed Choice Data," *Journal of Econometrics*, 89, 79–108.

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

Gelman A., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis Jumping Rules," *Bayesian Statistics*, 5, 599–608.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

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

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.

Kass, R. E. and Wasserman, L. A. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Kruskal, J. B. (1964), "Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis," *Psychometrika*, 29, 1–28.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

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

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.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research (with Discussion)," *Sociological Methodology*, 25, 111–193.

——— (1999), "Bayes Factors and BIC—Comment on 'A Critique of the Bayesian Information Criterion for Model Selection,' " *Sociological Methods and Research*, 27, 411–427.

——— (1969), "Some Statistical Considerations in Mutlidimensional Scaling," *Psychometrika*, 34, 167–182.

——— (1977), "Maximum Likelihood Estimation in Multidimensional Scaling," *Psychometrika*, 42, 241–266.

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

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*, 74–81.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–466.

Spence, I., and Graef, J. C. (1974), "The Determination of the Underlying Dimensionality of an Empirically Obtained Matrix of Proximities," *Multivariate Behavioral Research*, 9, 331–341.

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.

Suppes, P., and Zinnes, J. L. (1963), "Basic Measurement Theory," in *Handbook of Mathematical Psychology*, Vol. I, eds. R. D. Luce, R. R. Bush, and E. Galanter, New York: Wiley, pp. 2–76.

Takane, Y., Young, F. W., and de Leeuw, J. (1977), "Nonmetric Individual Difference Multidimensional Scaling: An Alternating Least Squares Method With Optimal Scaling Features," *Psychometrika*, 42, 7–67.

Takane, Y., and Carroll, J. D. (1981), "Nonmetric Maximum Likelihood Multidimensional Scaling from Directional Rankings of Similarities," *Psychometrika*, 46, 389–405.

——— (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.

Torgerson, W. S. (1952), "Multidimensional Scaling: I. Theory and Method," *Psychometrika*, 17, 401–419.

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, New York: Wiley.

Young, F. W. (1987), *Multidimensional Scaling—History, Theory, and Applications*, Hauer, R. M. (editor), Hillsdale Erlbaum Association, Hillsdale, NJ: Lawrence Erlbaum.