DECLARATION FOR THE PhD THESIS

The undersigned Cugnata Federica, PhD Registration Number 1287663

Thesis title: Bayesian Three-Way Multidimensional Scaling

PhD in Statistics

Cycle XXIII

Candidate's tutor Prof. Raffaella Piccarreta

Year of discussion 2012

DECLARES

Under her responsibility:

1. that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove true, all benefits included in this declaration are automatically forfeited from the beginning;

2. that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted;

3. that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text;

4. that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to Società NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information:
   - thesis: Bayesian Three-Way Multidimensional Scaling;
   - by: Cugnata Federica;
   - discussed at Università Commerciale Luigi Bocconi - Milano in 2012;

- the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Università Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;

5. that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6. that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7. that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo.

January 31, 2012

Federica Cugnata

Bayesian Three-Way Multidimensional Scaling

A Thesis Submitted for the Degree of Doctor of Philosophy

by

Federica Cugnata

Department of Decision Sciences

Bocconi University

Contents

Introduction

Multidimensional Scaling (MDS) refers broadly to a set of spatial models utilized to discover hidden structures in various types of multidimensional data, including dissimilarity data. A dissimilarity indicates how dissimilar two objects are. A small score indicates that the objects are similar, a high score that they are dissimilar. Based on dissimilarities between objects MDS produces a spatial representation, consisting of a geometric configuration of points, as on a map. Each point in the configuration corresponds to one of the objects. This configuration reflects the "hidden structure" in the data since the smaller the dissimilarity between the two objects the closer they should be on the spatial map.

Typical application areas for MDS are, among others, social and behavioral sciences and marketing. MDS techniques are often used to understand how people perceive and evaluate certain signals and information. Political scientists use MDS techniques to understand why political candidates are perceived by voters as being similar or dissimilar. In marketing researchers use MDS techniques to shed light on the way consumers evaluate brands and to assess the relationship between product attributes. They are interested in understanding how several brands are perceived in relationship to each other by consumers. They use MDS solution to position new products appropriately, to modify an existing product image to emphasize brand differentiation.

If only one dissimilarity matrix is considered, MDS is referred to as two-way MDS (the ways correspond to the arrangement of the matrix, rows$\times$columns). There are many

two-way MDS techniques in the literature. Oh and Raftery (2001) have developed a Bayesian approach to two-way multidimensional scaling using Markov chain Monte Carlo. This approach provided good estimates of the object configuration in the cases studied, as well as a Bayesian criterion for choosing the object dimension.

If more than one dissimilarity matrix are considered the technique is referred to as three-way MDS. In many situations we have information on dissimilarities between n objects for S subjects but the S matrices may also represent different occasions on which the dissimilarities are measured, different conditions under which they are taken, or dissimilarities with respect to different characteristics. For the sake of exposition, in the following we will assume that the S matrices refer to subjects rather than any other objects. The simplest way of dealing with several dissimilarity matrices is to perform a separate MDS for each subject and then attempting a comparison. A second way is to perform a single MDS for the average subject. In order to perform an aggregate MDS analysis, the S single dissimilarity matrices can be combined by computing the average value for each cell and a single MDS is performed for the average subject. In this way we obtain a compromise configuration but we do not investigate individual differences. Another strategy between these two extremes is the Weighted Multidimensional Scaling (WMDS). Weighted MDS, or Individual difference scaling (INDSCAL), was first introduced by Carrol and Chang (1970). It postulates a common object configuration that applies to all individuals, but that dimensions in the common configuration are differentially weighted by different individuals to give rise to differences in dissimilarity judgments. Its basic assumption is that each individual responds to all the dimensions of the objects but may utilize the dimensions in varying degree. The weights are interpreted as the importance, relevance, or salience of each dimension to each subject. Other methods for fitting the WBMDS model are ALSCAL (Takane et al. 1977), which optimizes the fit of squared distances to the dissimilarities and the SMACOF algorithm (De Leeuw 1977), which optimizes the fit of distances to dissimilarities by a convergent algorithm.

We propose an extension of the Bayesian approach to MDS in Oh and Raftery (2001) in order to analyze several matrices of dissimilarities. We develop four models with several levels of aggregation. In the first and simplest model, a Replicated Bayesian Multidimensional scaling, we assume that individual differences are attributable only to differences in measurement errors in observations. In this way a single scaling solution captures the dissimilarity data of all matrices.

To investigate the individual differences we develop a Bayesian Weighted Multidimensional Scaling in which we assume that several dissimilarity matrices differ from each other in systematically ways. We compute S MDS solutions, but they are required to be related to each other. More specifically, we assume that all subjects use the same dimensions when evaluating the objects, but that they can apply individual weights to these dimensions.

A problem in Weighted Multidimensional Scaling is that the weights are rarely interpreted for individual subjects, and the improvement in goodness-of-fit measures seldom seems to justify so many additional parameters. Therefore, we propose two models in which individual differences are attributable to the difference in importance that groups of individuals attach to the various common scale. In the first model, Group Bayesian Multidimensional Scaling, we partition the subjects according to one or more categorical variables and the aim is to investigate if the variables considered are useful to explain the dissimilarities matrices. For example, we can analyze if there are differences in perception among specific subject's group as for instance male or female or single and married.

In the last model, Cluster Bayesian Multidimensional Scaling, we assume that each subject belongs to one and only one clusters or subpopulation but it isn't known in advance to which cluster a particular subject belongs. In order to identify such groups, we propose a model that combine a fuzzy form of cluster analysis with Weighted Multidimensional Scaling. This model is based on a finite mixture of distributions, in which each mixture component is taken to correspond to a different cluster or subpopulation.

An important issue in all MDS methods 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. For this reason we develop in each model a Bayesian selection criterion for dimension determination and in the last model we provide a Bayesian criterion to select the dimension of the configuration and the number of clusters. Since the Cluster Bayesian Multidimensional scaling is a general model that includes Replicated Bayesian Multidimensional Scaling and Weighted Bayesian Multidimensional Scaling as special cases, in this way we may compare these three models and study the heterogeneity among subjects.

The thesis is organized as follows. Two way MDS and three way MDS are briefly reviewed in Chapter 1. Replicated Bayesian MDS is described in Chapter 2, Weighted Bayesian MDS is described in Chapter 3, Groups Bayesian MDS is described in Chapter 4 and Cluster Bayesia MDS is described in Chapter 5. For each model we illustrate the Bayesian framework including prior specifications, likelihood function, joint posterior distribution, and an estimation procedure using a MCMC algorithm. Detailed descriptions of the full conditional distributions are also provided. For each model we also describe a Bayesian selection criterion and to better emphasize the contributions of the proposed methods we present the results of simulation studies. Finally, in Chapter 6 our procedures are applied to real datasets.

Chapter 1

Multidimensional scaling

Multidimensional scaling (MDS) is a set of methods for discovering hidden structures in multidimensional data. Based on dissimilarities between objects MDS produces a spatial representation of the objects. The derived spatial representation consists of a geometric configuration of points on a map, each point corresponding to one of the objects. The smaller the dissimilarity among the objects the closer the objects will be on the map. By "objects", we mean stimuli, political candidates, biological entities, variables, or whatever other objects the user seeks to position in a spatial representation.

The multidimensional scaling can be "metric" or "nonmetric". Metric scaling assumes that the dissimilarity data are measured at the interval or ratio levels of measurement and the spatial representation attempts to preserve the distance among the objects. Nonmetric scaling assumes that the data are at the ordinal level of measurement and the spatial representation only preserves the rank order among the dissimilarities. We will focus on metric multidimensional scaling in which one or more full symmetric dissimilarity matrices are considered.

1.1 Two-way Multidimensional Scaling

To formalize MDS, we need some notation. Let n be the number of different objects and let the observed dissimilarity for objects i and j be given by d_{ij}. $\mathbf{D} = (d_{ij})$ denote the $n \times n$ matrix of dissimilarities between the n objects. We assume that the dissimilarity matrix $\mathbf{D}$ (or a series of them in Three-way Multidimensional Scaling) contains entries that are symmetric about the main diagonal; that is $d_{ij} = d_{ji}$ for $i, j = 1, ..., n$. When the raw data satisfy this requirement, it is only necessary to read in a lower half matrix (or series of them in Three-way Multidimensional Scaling) d_{ij} $(i = 1, ..., n; j = 1, ..., i-1)$, since the upper triangular matrix would only contain redundant information.

The main assumption in MDS is the existence of an underlying multidimensional space that describes the objects displayed in the space and that the dissimilarity between objects is related to the distance of the corresponding points in the multidimensional space. The $n \times p$ matrix of objects coordinates $\mathbf{X}$ is a set of n points in p dimensional space and it denotes the underlying object configuration. Row i from $\mathbf{X}$ gives the coordinates for object i, namely $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denotes an unobserved vector representing the values of the attributes possessed by object i. We define δ_{ij} as the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$:

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

The aim of multidimensional scaling is to find a configuration such that the estimated distances $\hat{\delta}_{ij}$ "match", as well as possible, the dissimilarities d_{ij}. It is the different notions of "matching" that give rise to the different techniques of multidimensional scaling. There are many two-way MDS techniques in the literature. A commonly used MDS method was developed by Torgerson (1952, 1958). Maximum likelihood MDS methods have been developed by Ramsay (1982) and Takane (1982) among others.

Least squares MDS methods have been proposed by De Leeuw and Heiser (1982) and Groenen, Mathar, and Heiser (1995) among others. Oh and Raftery (2001) proposed a Bayesian MDS (BMDS) method using Markov chain Monte Carlo.

1.1.1 Classical Multidimensional Scaling

Classical scaling originated in the 1930s when Young and Householder (1938) showed how starting with a matrix of distances between points in a Euclidean space, coordinates for the points can be found such that distances are preserved. Torgerson (1952) brought the subject to popularity using the technique for scaling.
The basic idea of classical scaling is to assume that the observed dissimilarities are Euclidean distances and then find coordinates that explain them.
Formally, classical scaling assumes that the observed dissimilarity measure, d_{ij}, is the distance between objects i and j in a p-dimensional Euclidean space, i.e.,

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

where x_{ik} is the k-th element of $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$. 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

$$a_{ij} = -\tfrac{1}{2}(\delta_{ij}^2 - \delta_{i\cdot}^2 - \delta_{\cdot j}^2 + \delta_{\cdot\cdot}^2),$$

where

$$\delta_{i\cdot}^2 = \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\delta_{ij}^2, \quad \delta_{\cdot j}^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\delta_{ij}^2, \quad \delta_{\cdot\cdot}^2 = \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n}\sum_{j=1}\delta_{ij}^2.$$

It was shown by Togerson (1952, 1958) that

$$a_{ij} = \sum_{k=1}^{p} x_{ik} \cdot x_{jk}, \quad \text{for all } i, j, \quad \text{i.e.,} \quad A = \mathbf{X}\mathbf{X}' \tag{1.1}$$

where $\mathbf{X}$ is the $n \times p$ matrix of objects coordinates. The coordinates of $\mathbf{X}$ can recovered from the spectral decomposition of the matrix A in (1.1).

Because Euclidean distance is invariant under translation, rotation, and refection about the origin, there is no unique set of coordinate values that give rise to these distances. In other words, we cannot uniquely determine either the location or the orientation of the configuration. The location problem is usually overcome by placing the mean vector of the configuration at the origin. The orientation problem means that any configuration derived can be subjected to an arbitrary orthogonal transformation. Such transformations can often be used to facilitate the interpretation of solutions.

Several methods have been proposed to choose the dimensionality of the MDS solution. However, no definite strategy is present for choosing the effective dimension of $\mathbf{x}_i$, the number of object attributes that contribute significantly to the dissimilarities. A common way of assessing dimension is to look at the eigenvalues of the scalar product matrix A. The usual strategy is to plot the ordered eigenvalues (or some function of them) against dimension and then identify a dimension at which the eigenvalues become "stable" (i.e., do not change perceptively). At that dimension, we may observe an "elbow" that shows where stability occurs. Also a measure of fit, called stress, is commonly used to determine the dimensionality. Several definitions of stress have been proposed; in classical multidimensional scaling the mostly commonly used one, is

$$STRESS = \sqrt{\frac{\sum_{i>j}(d_{ij} - \hat{\delta}_{ij})^2}{\sum_{i>j} d_{ij}^2}} \tag{1.2}$$

where $\hat{\delta}_{ij}$ is the Euclidean distance obtained from the estimated object configuration. One approach to determine the dimensionality is to compute MDS solutions for a range of dimensions, plot the $STRESS$ against the dimension and identify a dimension at which the $STRESS$ become "stable".

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.

One way to incorporate measurement error into MDS is to adopt a more explicit modeling framework, such as a Bayesian viewpoint (Oh and Raftery, 2001).

1.1.2 Bayesian Multidimensional Scaling

Oh and Raftery (2001) have developed a Bayesian approach to multidimensional scaling in which they assume the dissimilarity, $d_{ij} > 0$, between objects i and j is observed with Gaussian error:

$$d_{ij} = \delta_{ij} + \epsilon_{ij}$$

where δ_{ij} is the true dissimilarity and $\epsilon \sim N(0, \sigma^2)$, $i, j = 1, 2, ..., n$. Thus, given δ_{ij}, the observed dissimilarity d_{ij}, which is assumed to be functionally related to the unknown $\mathbf{x}_i$, follows the truncated Gaussian distribution,

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

The likelihood function of $(\mathbf{X} = \{\mathbf{x}_i\}, \sigma^2)$, given $\mathbf{D} = (d_{ij})$, is given by

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

where $\Phi(\cdot)$ is the standard Gaussian cdf, and $m = n(n-1)/2$ is the number of non-diagonal elements in the lower triangle of the dissimilarity matrix. The second term in the exponent of the likelihood function is the modification to the normalizing constant due to the truncation. Next, they assume that the $\mathbf{x}_i$ are iid with a common multivariate Gaussian prior density, with mean 0 and a diagonal covariance matrix Λ, i.e., $\mathbf{x}_i \sim N_p(0, \Lambda)$. They assume that the error variance σ^2 has the (conjugate) prior $\sigma^2 \sim IG(a, b)$ where $IG(a, b)$ is the inverse Gamma distribution with mode $b/(a+1)$. Similarly, the prior for λ_k is taken to be $\lambda_k \sim IG(\alpha, \beta_k)$ independently for

each $j = 1, 2, ..., p$. Finally, the prior densities of $\mathbf{X}$, Λ, and σ^2 are assumed to be independent. The joint posterior density of $(\mathbf{X}, \sigma^2, \Lambda)$, given the dissimilarity matrix $\mathbf{D} = (d_{ij})$, is

$$\begin{aligned}\pi(\mathbf{X}, \sigma^2, \Lambda|\mathbf{D}) \quad \propto \quad & (\sigma^2)^{-\left(\frac{m}{2}+a+1\right)} \prod_{k=1}^{p} \lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \\ & \exp\left[-\frac{1}{2\sigma^2} \sum_{i>j} (d_{ij} - \delta_{ij})^2 - \sum_{i>j} \log \Phi\left(\frac{\delta_{ij}}{\sigma}\right) \right. \\ & \left. - \frac{1}{2} \sum_{i=1}^{n} \mathbf{x}_i' \Lambda^{-1} \mathbf{x}_i - \frac{b}{\sigma^2} - \sum_{k=1}^{p} \frac{\beta_k}{\lambda_k} \right], \end{aligned} \tag{1.3}$$

The posterior distribution (1.3) is a complicated function of the unknown quantities $(\mathbf{X}, \sigma^2, \Lambda)$. The numerical integration necessary to compute Bayes estimates of these quantities is best accomplished using Markov chain Monte Carlo (MCMC) methods. Oh and Raftery used a random-walk, Metropolis-Hastings algorithm. Initial values for $\mathbf{X}$ and the other unknown parameters, σ^2 and Λ, of the posterior distributions are taken from a classical scaling solution. They also proposed an easily computed Bayesian criterion called MDSIC for choosing object dimension.

1.2 Three-way Multidimensional Scaling

In three-way multidimensional scaling the observed dissimilarities data are obtained from a sample of S subjects and they are denoted by $d_{ij,s}$, where $s = 1, ..., S$ indexes subjects, and $i, j = 1, ..., n$ indexes pairs of objects. There were two basic approaches in the early work in this area. The first was to perform a two-way MDS for the average subject, the second to compare results subject by subject. A strategy between these two extremes to obtain a common object configuration that applies to all subjects and to analyze variation among subjects is the Weighted Multidimensional Scaling which assumes that all the subjects use the same dimensions when evaluating the objects but that dimensions can be differentially weighted by different subjects.

1.2.1 Weighted Multidimensional Scaling

Weighted Multidimensional Scaling assumes that for each individual the object configuration can differ from $\mathbf{X}$ and hence $\mathbf{X}^{(s)}$ denotes the object configuration for subject s. The Euclidean distance between object i and j, for the s-th subject is

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

The matrix $\mathbf{X}^{(s)}$ is assumed to be related to $\mathbf{X}$. The assumption is made that all subjects use the same dimensions when evaluating the objects, but that they might apply individual weights to these dimensions, $w_{k,s}$ $(s = 1, ..., S;\ k = 1, ..., p)$. Then $x_{ik}^{(s)} = \sqrt{w_{k,s}}x_{ik}$ where the weights $w_{k,s}$ are required to be nonnegative $w_{k,s} \geq 0$ and the weighted Euclidean distance between object i and j, for subject s is

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

The output of the analysis consist of two matrices, an $n \times p$ matrix of coordinates ($\mathbf{X}$) of the n objects on p dimensions (Group Space) and an $S \times p$ matrix of weights ($\mathbf{W}$) of S subjects on the p dimensions (Subject space). Points in the group space represent the objects. The subjects are represented as points in the subject space. The coordinates of each individual are the weights required to give the weighted Euclidean distances between the points in the objects space, the values that best represent the corresponding dissimilarities for that individual. The weights are interpreted as the importance, relevance, or salience of each dimension to each subject. A large weight means that the dimension is important to the subject, a small weight means the dimension is unimportant.

A geometric interpretation of the role of weights in the Weighted Multidimensional Scaling model, using artificial data is given in Figure 1.1. The two-dimensional group

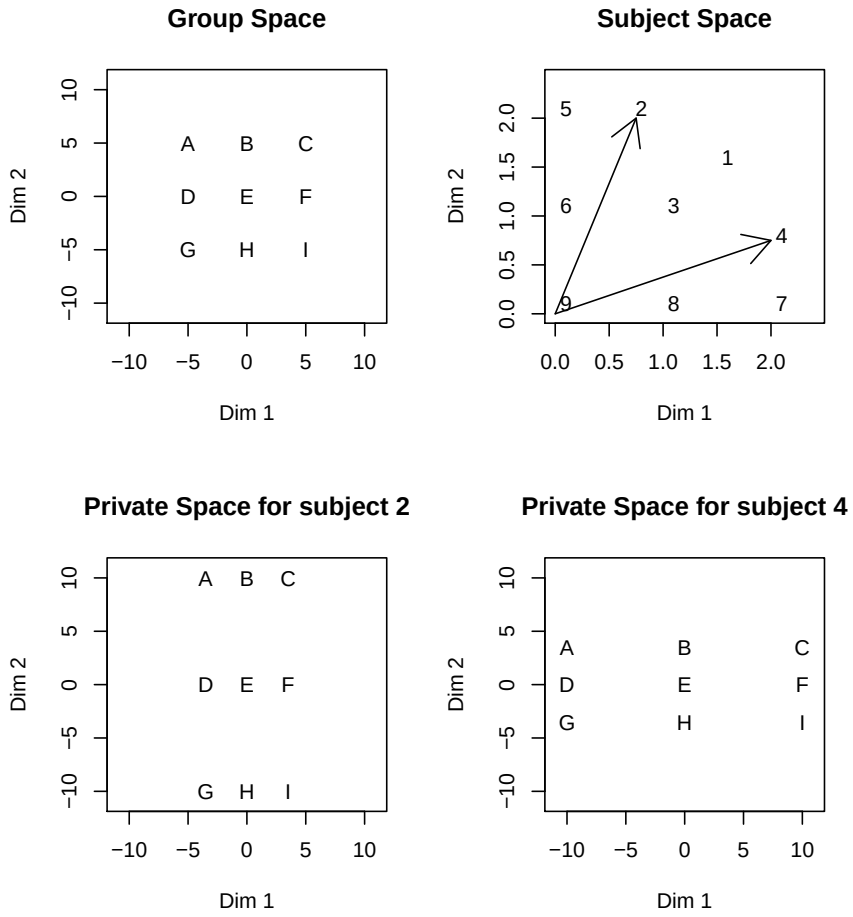


Figure 1.1: Hypothetical example illustrating the Weighted Multidimensional Scaling model. Weights from the Subject Space are applied to dimensions of the "group" object space to produce "private" spaces for subjects 2 and 4.

object space in the upper left panel of this figure shows nine objects, labeled A through I, in a lattice configuration, based on the **X** matrix of object coordinates. In the upper right panel, the **W** matrix of weights has been realized as a set of coordinates of nine hypothetical subjects, whose two-dimensional positions are indicated by the Arabic numerals.

A "Private Space" for each subject can be constructed, by applying (square roots of) the subject weights to object dimensions. The weights can be regarded graphically as stretching factors applied to the dimensions of the common object configuration. A "Private Space" is simply the group space with its dimensions stretched or contracted by the weights which that subject has assigned to them. In Figure 1.1, for example, the weights fitted for subject 3 are equal along the two dimensions. This subject's "private" object space would look much like the group object space, except for an overall rescaling factor applied uniformly to both dimensions, and thus leaving their relative importance or salience unchanged. The private object spaces for subjects 2 and 4 are shown in the left and right lower corners, respectively, of Figure 1.1. Subject 2, with greater weight for Dimension 2 than for 1, has a private space that is compressed along the axis of Dimension 1, or, equivalently, stretched along Dimension 2. This situation is reversed for subject 4, whose estimated weight is higher on the first dimension than on the second.

1.2.1.1 INDSCAL (INdividual Differences SCALing)

The first WMDS method, called INDSCAL (INdividual Differences SCALing), was developed by Carrol and Chang in 1970. The model assumes that the subject's data dissimilarities are a linear function of the distances of the solution

$$d_{ij,s} = L(\delta_{ij,s})$$

where

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

As with classical multidimensional scaling, dissimilarities $d_{ij,s}$ are converted to distances $\delta_{ij,s}$ and then $w_{k,s}$, x_{ik} are found by least squares.

The first step in INDSCAL method is to convert each subject's dissimilarities into estimates of euclidean distances by calculating an additive constant. Under "metric" assumptions we may assume, without loss of generality, that

$$\delta_{ij,s} \cong d_{ij,s} + c_s.$$

Dropping the (s) subscripts for the moment, the smallest constant c guaranteeing satisfaction of the triangle inequality ($\delta_{il} \leq \delta_{ij} + \delta_{jl}$) for all triples (i, j, l) can easily be shown to be:

$$c_{min} = max_{i,j,l}(\delta_{il} - \delta_{ij} - \delta_{jl}).$$

This is the "additive constant" scheme, for converting "comparative distances" (i.e., interval scale distance estimates) into absolute distances (i.e., ratio scale distance estimates).

Then, these distances are converted into scalar products in a three-way manner. As shown in Torgerson (1958), given a set of Euclidean distances, these may be converted into scalar products of vectors about an origin placed (arbitrarily) at the centroid of all the points by the equation:

$$a_{ij,s} = -\tfrac{1}{2}(\delta^2_{ij,s} - \delta^2_{i\cdot,s} - \delta^2_{\cdot j,s} + \delta^2_{\cdot\cdot,s}),$$

where

$$\delta^2_{i\cdot,s} = \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \delta^2_{ij,s}, \quad \delta^2_{\cdot j,s} = \tfrac{1}{n}\sum_{i=1}^{n} \delta^2_{ij,s}, \quad \delta^2_{\cdot\cdot,s} = \tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1} \delta^2_{ij,s}.$$

These scalar products can easily be shown to be of the form:

$$a_{ij,s} = \sum_{k=1}^{p} w_{k,s} x_{ik} \cdot x_{jk}, \quad \text{for all } i, j, s. \tag{1.4}$$

This is the three-way scalar products formulation of the INDSCAL model. A least-squares procedure is iteratively used to arrive at the optimal set of objects coordinates $\mathbf{X}$ and subject space $\mathbf{W}$. It consists of finding a preliminary estimate for the two objects coordinates (x_{ik} and x_{jk}), fixing them, and then estimating (by least squares) the subject weights $w_{k,s}$. Then the x_{ik} are estimated, with the $w_{k,s}$ and x_{jk} fixed, and so on. When a satisfactory approximation to the data is obtained, the process terminates, the two estimates of the objects coordinates are set equal, a final estimate of subject weights is made and the weights are then appropriately normalised before being output.

INDSCAL produces a unique orientation of the axes of the group space, in the sense that any rotation will destroy the optimality of the solution and will change the values of the subject weights. Because of this, it is not legitimate to rotate the axes of a Group Space to a more "meaningful" orientation, as is commonly done in the basic multidimensional scaling model.

1.2.1.2 ALSCAL (Alternating Least Squares SCAling)

Another method for fitting the Weighted Multidimensional Scaling is ALSCAL (alternating least squares scaling). It was developed by Takane, Young and De Leeuw (1977). ALSCAL can accommodate a wide variety of different models and data. It can be used either for two-way MDS or for fitting three-way MDS to data that may be defined on nominal, ordinal, interval or ratio scale, may be symmetric or asymmetric, and may have missing elements. ALSCAL uses an alternating least-squares algorithm to perform multidimensional scaling. We describe various technical aspects of ALSCAL in fitting the INDSCAL model for the case in which the data are assumed

symmetric, ratio scale, and without missing data or replications.

ALSCAL's optimization problem for this case can be generally stated as attempting to estimate $\mathbf{X}$, $\mathbf{W}$ and the "disparity" array Δ^* to minimize a measure called $S-STRESS$, given as:

$$S-STRESS=\sum_{s=1}^{S}\sum_{i>j}(\delta_{ij,s}^{*2}-\delta_{ij,s}^{2})^2 \tag{1.5}$$

where

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

and

$$\delta_{ij,s}^{*2}=f[d_{ij,s}^{2}]. \tag{1.6}$$

The function f in expression (1.6) is specified according to the restrictions that occur in the particular model and type of data.

$S-STRESS$ in (1.5) is minimized using an alternating least squares algorithm. Each iteration of the algorithm has two phases: an optimal scaling phase and a model estimation phase. The optimal scaling phase finds the least squares disparities Δ^* for fixed $\mathbf{X}$ and $\mathbf{W}$, which is followed by the model estimation phase which calculates new coordinates $\mathbf{X}$ and weights $\mathbf{W}$ for fixed Δ^*.

Note that $S-STRESS$ uses squared distances and dissimilarities 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).

1.2.1.3 SMACOF (Scaling by MAjorizing a Complicated Function)

Another method for fitting the Weighted Multidimensional Scaling is SMACOF (Scaling by MAjorizing a Complicated Function). This method based on the majorization algorithm was initially proposed by De Leeuw (1977). SMACOF is as flexible as

ALSCAL in the types of models it can fit and data it can analyze.

The SMACOF algorithm is based on the iterative majorization method for reducing a function of stress using the Guttman transform. The idea of the majorization method is to minimize a complicated function $f(x)$ by using a more simple function $g(x, y)$ that can be easily minimized. The function g has to satisfy the inequality $f(x) \leq g(x, y)$, for a given y such that $f(y) \leq g(y, y)$. Thus, for graphs of f and g, the function g is always above the function f, and g touches f at the point $x = y$. This leads to an interactive algorithm for minimizing f. First, an initial value x_0 is given, and a function $g(x_1, x_0)$ is minimized to get x_1. Next is minimized $g(x, x_l)$ to get the next value (x). The process is repeated until convergence.

The SMACOF algorithm satisfies the requirements for minimizing a function using majorization, as described above. We describe the three-way SMACOF for the case in which the data are assumed symmetric, ratio scale, and without missing data. In this case the main purpose is looking for a object configuration $\mathbf{X}$ and a subject space $\mathbf{W}$ that minimize the raw stress σ that measures the squared differences between weighted euclidean distances and observed dissimilarities:

$$\sigma = \sum_{s=1}^{S} \sum_{i>j} (d_{ij,s} - \delta_{ij,s})^2 \tag{1.7}$$

where

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

Thus, using the majorizing method for finding minimum stress σ simply has the Guttman transform as its updating equation. The algorithm gives rise to a non-decreasing sequence of stress values, which converge linearly. Its major feature is that it guarantees lower stress values in each iteration. However, it shares the problem of not necessarily finding the global minimum, but can get stuck at a local minimum.

1.2.1.4 Maximum likelihood MDS methods

Maximum likelihood MDS methods have been developed for handling measurement errors. Ramsay (1977) proposed a maximum likelihood methodology for the MDS analysis of either two or three-way dissimilarities in his MULTISCALE approach. In the three-way case Ramsey assumes that the $Sn(n-1)/2$ observations are independently and identically distributed with a probability density function given by $f(d_{ij,s}|\delta_{ij,s}, \sigma^2)$, where $\delta_{ij,s}$ represents the weighted Euclidean distance between two objects i and j for subject s, and σ is the standard error of a given input data value $d_{ij,s}$. Given this assumption of independence, the general form of the log likelihood function can be written as

$$\log l = \sum_{s=1}^{S} \sum_{i>j} \log f(d_{ij,s}|\delta_{ij,s}, \sigma^2). \tag{1.8}$$

Ramsay promotes the use of the lognormal distribution for the density function f assumed in (1.8). The objective of this maximum likelihood formulation is to maximize (1.8) with respect to $\mathbf{X}$, $\mathbf{W}$ and the dispersion parameter σ^2.

Winsberg and De Soete (1993) have observed that the subject weights included in Weighted Multidimensional Scaling model are rarely interpreted for individual subjects, even though these are useful for removing the rotational invariance. They have proposed to classify these weights into a reduced set of groups in order to obtain a more parsimonious model still preserving the rotational invariance property. The model has been called CLASCAL. It assumes that each of the S subjects belongs to one and only one of a small number T ($T \ll S$) of latent classes or subpopulations. The probability that a subject belongs to latent class t is p_t ($1 \leq t \leq T$). For those individuals in latent class t, their dissimilarities $d_{ij,s}$ are assumed to follow a common multivariate normal distribution. The coordinates of the points in the common object space and the weights in the subject space are then found by maximum likelihood.

Chapter 2

Replicated Bayesian Multidimensional Scaling

In this chapter, we generalized the Bayesian Multidimensional Scaling introduced by Oh and Raftery in order to analyze several matrices of dissimilarities. In this first and simplest model, a Replicated Bayesian Multidimensional Scaling (RBMDS), the individual differences are assumed to be due only to measurement errors. If no systematic individual differences exist, a single common Euclidean distance model may be fitted to all of them simultaneously. A single scaling solution captures the dissimilarity data of all matrices through separate relationships for each matrix.

We briefly describe the Bayesian framework including prior specifications, likelihood function, joint posterior distribution, and an estimation procedure using a Markov Chain Monte Carlo method. We also provide a Bayesian criterion to select the dimension of the configuration. Finally, some simulated examples are given.

2.1 Model and Prior

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 $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object i. We define δ_{ij} as the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$:

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

The observed dissimilarities data are obtained from a sample of S subjects and they are denoted by $d_{ij,s}$, where $s = 1, ..., S$ indexes subjects, and $i, j = 1, ..., n$ indexes pairs of objects. Let $m = n(n-1)/2$ the distinct number of objects pairs the m-component column vector $\mathbf{d}_s = (d_{12,s}, d_{13,s}, ..., d_{n(n-1),s})'$ contain the m dissimilarity values for subject s ($s = 1, ..., S$). The total data set will be indicated by the $S \times m$ matrix $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_S)'$.

Following Oh and Raftery (2001), we assume that the observed dissimilarity measure, $d_{ij,s}$ is equal to the true distance, δ_{ij}, plus a Gaussian error. In addition, since dissimilarity measures are typically given as positive values we restrict the observed dissimilarity to be always positive. Thus, given the Euclidean distance δ_{ij}, the observed dissimilarity measure $d_{ij,s}$ is assumed to follow the truncated normal distribution

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

and conditionally to δ_{ij}, the $d_{ij,s}$'s are independent.

From this, the likelihood function of the unknown parameters $\mathbf{X} = \{\mathbf{x}_i\}$, called the object configuration, and σ^2 is:

$$l(\mathbf{X}, \sigma^2) \propto (\sigma^2)^{-\frac{Sm}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s} - \delta_{ij})^2 - S\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]$$

where $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$, $m = n(n-1)/2$, 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., $\mathbf{x}_i \sim N_p(0, \Lambda)$, independently for $i = 1, 2, ...n$. For the prior of the error variance σ^2, we use a conjugate prior, that is the inverse Gamma distribution $\sigma^2 \sim IG(a, b)$ with mode $b/(a+1)$. Similarly, for a hyperprior for the elements of $\Lambda = Diag(\lambda_1, ..., \lambda_p)$, given dimension p, we assume a conjugate prior $\lambda_k \sim IG(\alpha, \beta_k)$, independently for each $k = 1, 2, ..., p$.

$$f(\mathbf{x}_i|\Lambda) = (2\pi)^{-p/2} |\Lambda|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i\right]$$
$$f(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1} \exp\left[-\frac{b}{\sigma^2}\right]$$
$$f(\lambda_k) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\lambda_k)^{-\alpha-1} \exp\left[-\frac{\beta}{\lambda_k}\right]$$

We assume prior independence among $\mathbf{X}$, Λ and σ^2.

2.2 Posterior Inference (MCMC)

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

$$\begin{aligned}\pi(\mathbf{X}, \sigma^2, \Lambda|\mathbf{D}) \quad \propto \quad & (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)} \prod_{k=1}^{p} \lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \\ & \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s} - \delta_{ij})^2 - S\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right) \right. \\ & \left. -\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i - \frac{b}{\sigma^2} - \sum_{k=1}^{p}\frac{\beta_k}{\lambda_k}\right]. \qquad (2.1)\end{aligned}$$

The full conditional distributions of the parameters involved in the model given all the other parameters and the data are:

$$\pi\left(\sigma^2|\cdot\right) \propto (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)} \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij})^2 - \frac{b}{\sigma^2} - S\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right],$$

$$\pi\left(\mathbf{x}_i|\cdot\right) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{j\neq i,j=1}^{n}(d_{ij,s}-\delta_{ij})^2 - \frac{1}{2}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i - S\sum_{j\neq i,j=1}^{n}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right].$$

Finally $\pi\left(\lambda_k|\cdot\right)$ is $IG\left(\alpha+n/2, \beta_k+\sum_{i=1}^{n}x_{ik}^2/2\right)$

$$\pi\left(\lambda_k|\cdot\right) \propto \lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{x_{ik}^2}{\lambda_k} - \frac{\beta_k}{\lambda_k}\right].$$

Iterative simulations of the unknown parameters from their full conditional distributions for a sufficiently long time yield samples of the parameters from the joint posterior distribution, and posterior inference can be done by using the samples. Simulation of samples from the full conditional distributions of the parameters λ_k, $k = 1, ..., p$ 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 to generate samples of $\mathbf{x}_i$ and σ^2. Oh and Raftery (2001) suggested a random walk Metropolis-Hastings algorithm for generating samples of $\mathbf{x}_i$ and σ^2 when $S = 1$. Thus, we can easily modify the algorithm of Oh and Raftery (2001) for generating samples of $\mathbf{x}_i$ and σ^2. Specifically, we use a normal proposal density in the random walk Metropolis-Hastings algorithm for generation of $\mathbf{x}_i$. To choose the variance of the normal proposal density we remember that $\pi\left(\mathbf{x}_i|\cdot\right) \propto \exp\left[-\frac{1}{2}\left(Q_1+Q_2\right) - S\sum_{j\neq i,j=1}^{n}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]$ where $Q_1 = \frac{1}{\sigma^2}\sum_{s=1}^{S}\sum_{j\neq i,j=1}^{n}(d_{ij,s}-\delta_{ij})^2$ and $Q_2 = \mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i$. Because $(d_{ij,s}-\delta_{ij})^2/\sigma^2$ is a quadratic function of $\mathbf{x}_i$ with leading coefficient equal to $1/\sigma^2$, and Q_1 has $S(n-1)$ of this kind whereas Q_2 has only one quadratic term with coefficient Λ, Q_1 would dom-

inate the full conditional posterior density function of $\mathbf{x}_i$ unless there is strong prior information. Thus we can consider Q_1 only, approximate the full conditional variance of $\mathbf{x}_i$ by $\sigma^2/S(n-1)$ and choose the variance of the normal proposal density to be proportional to $\sigma^2/S(n-1)$.

As concerns σ^2 we found that $\pi\left(\sigma^2|\cdot\right)$ can be well approximated by an inverse Gamma distribution $IG\left(Sm/2+a, SSR/2+b\right)$ where $SSR=\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij})^2$, which can, in turn, be approximated by a normal distribution, since the number of observed dissimilarities, $Sm=Sn(n-1)/2$, is generally large. Hence, the Metropolis-Hastings algorithm is applied with a normal proposal density with variance proportional to that of the inverse Gamma above: $\left(SSR/2+b\right)^2/\left[\left(Sm/2+a-1\right)^2\left(Sm/2+a-2\right)\right]$.

To initialize the algorithm it is necessary to specify suitable starting values. For initialization of $\mathbf{X}$ we can use the classical scaling solution, $\mathbf{X}^{(0)}$, for the average subject. The resulting 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. For σ^2, as in Oh and Raftery (2001), we can use the adjusted $\mathbf{X}^{(0)}$ to calculate $SSR^{(0)}$, and set $\sigma^{2(0)}=SSR^{(0)}/Sm$. For Λ, we can use diagonal elements of the covariance matrix of the adjusted $\mathbf{X}^{(0)}$.

Turning now attention to the parameters characterizing the prior distributions, we follow Oh and Raftery (2001), and assume that the prior distribution of σ^2 is $IG(a,b)$ with $a=5$, and $b=(a-1)\sigma_0^2$ so that the prior mean of σ^2 matches with SSR/Sm. For the hyperprior of λ_k we choose $\alpha=1/2$ and $\beta_k=\parallel \mathbf{x}_k^{(0)} \parallel^2/2n$ so that the prior mean of λ_k matches with the k-th diagonal element of the sample covariance matrix of $\mathbf{X}^{(0)}$.

2.3 The problem of indeterminacy

Euclidean distance is invariant under translation, rotation, and reflection of objects. Thus as emphasized by Oh and Raftery (2007),"the dissimilarity data provide infor-

mation 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". To get around this problem of weak identification, they propose a Procrustean similarity transformation (Sibson 1979; Borg and Groenen 2005) in order to obtain a stable estimate of the final configuration. Let $\mathbf{X}^*$ be a fixed 'reference' matrix of coordinates, the initial solution obtained from a preliminary run. In order to fix the orientation and location of the points, the sample of $\mathbf{X}$ at each iteration of MCMC is transformed 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. The steps to compute the considered transformation are:

- 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, $\mathbf{C} = \mathbf{P}\Phi\mathbf{Q}'$, where $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices and Φ is a diagonal matrix.

- Step 3: Let $\mathbf{T} = \mathbf{Q}\mathbf{P}'$ be the optimal rotation matrix.

- Step 4: Let $\mathbf{t} = 1/n(\mathbf{X}^* - \mathbf{X}\mathbf{T})'\mathbf{1}$ be the optimal translation matrix.

- Step 5: Transform $\mathbf{X}$ by $\mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{1}\mathbf{t}'$.

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.

2.4 Choice of dimension

One important decision that has to be made in MDS concerns the dimensionality of the solution space. The dimensionality refers to the number of coordinates needed to locate a point in the spatial representation of objects. Posterior inference as described in the previous section presumed that the dimension, p, of the object configuration, is given. In most cases, the dimension of the objects (the number of significant attributes) is unknown.

Following Oh and Raftery (2001) we propose a Bayesian criterion to select the dimension of the configuration. We cast choice of dimension as a model selection problem. Bayesian approach to choosing between models M_p and M_{p+1} on the basis of data $\mathbf{D}$ is to compare models on the basis of the posterior probability of the model given the data. Using Bayes'rule, this is proportional to the prior probability for the model, multiplied by the likelihood of the data given the model. Thus the choice between model p and model $p+1$ can be made on the basis of the ratio of posterior model probabilities.

$$\frac{\pi(M_{p+1}|\mathbf{D})}{\pi(M_p|\mathbf{D})} = \frac{\pi(\mathbf{D}|M_{p+1})}{\pi(\mathbf{D}|M_p)} \times \frac{\pi(M_{p+1})}{\pi(M_p)} \tag{2.2}$$

where

$$\pi(\mathbf{D}|M_p) = \int \pi(\mathbf{D}|\xi_p, M_p)\pi(\xi_p|M_p)d\xi_p$$

is the integrated likelihood of model M_p, ξ_p is the vector of parameters in model M_p and $\pi(\xi_p|M_p)$ is its prior density. A large value of the ratio (2.2) gives support for M_{p+1} over M_p. However, in MDS we seek a best configuration rather than a best model, so the choice between dimension p and dimension $p+1$ is based on the ratio of the marginal posteriors of the configuration of dimension p and that of dimension $p+1$.

$$\frac{\pi(\mathbf{X}^{(p+1)}|\mathbf{D}, M_{p+1})}{\pi(\mathbf{X}^{(p)}|\mathbf{D}, M_p)}$$

where

$$\pi(\mathbf{X}^{(p)}|\mathbf{D}, M_p) = \int \pi(X^{(p)}|\Lambda, \sigma^2, \mathbf{D}, M_p)\pi(\Lambda, \sigma^2|\mathbf{D}, M_p)d\Lambda d\sigma^2$$

Hence we cast choice of dimension as a model selection problem in which the choice of the dimension p corresponds to the choice of the configuration $\mathbf{X}_p$ obtained by applying the algorithm described in the previous section with dimension p. The aim is then to choose the best configuration across p, which is considered as an unknown random variable, taking values between 1 and p_{max} with the same prior probability. Including p in the model, the posterior distribution is given by:

$$\begin{aligned}
\pi(\mathbf{X}, \sigma^2, \Lambda, p|\mathbf{D}) &\propto l(\mathbf{X}, \sigma^2, p|D)\pi(\mathbf{X}|\Lambda, p)\pi(\sigma^2)\pi(\Lambda|p) \\
&= (2\pi)^{-\frac{Sm}{2}}(\sigma^2)^{-\left(\frac{Sm}{2}\right)} \exp\left[-\frac{1}{2\sigma^2}SSR - S\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right] \\
&\quad \times(2\pi)^{-\frac{np}{2}}\prod_{k=1}^{p}\lambda_k^{-\frac{n}{2}} \exp\left[-\sum_{k=1}^{p}\frac{\sum_{i=1}^{n}x_{ik}^2}{2\lambda_k}\right] \\
&\quad \times\Gamma(a)^{-1}b^a\left(\sigma^2\right)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right] \\
&\quad \times\Gamma(\alpha)^{-p}\prod_{k=1}^{p}\beta_k^{\alpha}\lambda_k^{-(\alpha+1)} \exp\left[-\frac{\beta_k}{\lambda_k}\right] \\
&= A(p)\cdot h\left(\sigma^2, \mathbf{X}\right)\cdot g\left(\Lambda, \mathbf{X}, p\right), \qquad (2.3)
\end{aligned}$$

where

$$SSR = \sum_{s=1}^{S}\sum_{i>j}(d_{ij,s} - \delta_{ij})^2,$$

$$A(p) = (2\pi)^{-(Sm+np)/2}\Gamma(a)^{-1}b^a\Gamma(\alpha)^{-p}\prod_{k=1}^{p}\beta_k^{\alpha}\lambda_k^{-(\alpha+1)},$$

$$h\left(\sigma^2, \mathbf{X}\right) = (\sigma^2)^{-(Sm/2+a+1)} \exp\left[-\frac{1}{\sigma^2}\left(\frac{SSR}{2} + b\right) - S\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right], \qquad (2.4)$$

$$g\left(\Lambda, \mathbf{X}, p\right) = \prod_{k=1}^{p} \lambda_k^{-(n/2+\alpha+1)} \exp\left[-\frac{1}{\lambda_k}\left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)\right]. \tag{2.5}$$

Note that $\pi\left(\mathbf{X}, p|\mathbf{D}\right) \propto f\left(D|\mathbf{X}, p\right)\pi\left(\mathbf{X}, p\right)$. Thus, one needs only to compute the marginal likelihood and the marginal prior of $\mathbf{X}$ for each p. The integral of $g\left(\Lambda, \mathbf{X}, p\right)$, given in (2.5) with respect to Λ is

$$\int g\left(\Lambda, \mathbf{X}, p\right) d\Lambda = \Gamma^p\left(\frac{n}{2} + \alpha\right) \prod_{k=1}^{p} \left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)} \tag{2.6}$$

The integral of the function $h\left(\sigma^2, \mathbf{X}\right)$ given in (2.4) with respect to σ^2 is is not straightforward. However, in most cases, $Sm = Sn(n-1)/2$ is very large and the likelihood of σ^2 dominates the prior and hence $h\left(\sigma^2, \mathbf{X}\right)$ is approximately proportional to the likelihood

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

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

$$\int h\left(\sigma^2, \mathbf{X}\right) d\sigma^2 \approx \int l\left(\sigma^2, \mathbf{X}\right) d\sigma^2 \approx \left(2\pi\right)^{1/2} H^{-1/2} l\left(\mathbf{X}, \hat{\sigma}^2\right) \tag{2.7}$$

where H is the minus Hessian of the log-likelihood and $\hat{\sigma}^2$ is the MLE of σ^2.
The probability $\Phi\left(\frac{\delta_{ij}}{\sigma}\right)$ is unlikely to have much effect on the model comparison and can safely be ignored. Thus, following Oh and Raftery (2001) we use the approximation

$$l\left(\sigma^2, \mathbf{X}\right) \approx l^*\left(\sigma^2, \mathbf{X}\right) \equiv \left(\sigma^2\right)^{-(Sm/2)} \exp\left[-\frac{SSR}{2\sigma^2}\right] \tag{2.8}$$

Replacing l by l^* and H by the minus Hessian H^* of l^* in (2.7) and letting $\hat{\sigma}^2 =$

SSR/Sm, which maximizes l^*, gives the formula

$$\int h\left(\sigma^2, \mathbf{X}\right) d\sigma^2 \approx (2\pi)^{\frac{1}{2}} \left(\frac{Sm}{2}\right)^{-\frac{1}{2}} \left(\frac{SSR}{Sm}\right)^{-\frac{Sm}{2}+1} \exp\left[-\frac{Sm}{2}\right] \tag{2.9}$$

The above results lead to:

$$\begin{aligned} \pi\left(\mathbf{X}, p|\mathbf{D}\right) &= c \cdot A(p) \cdot \int h\left(\sigma^2, \mathbf{X}\right) d\sigma^2 \cdot \int g\left(\Lambda, \mathbf{X}, p\right) d\Lambda \\ &= c \cdot A^*(p) \cdot \left(\frac{SSR}{Sm}\right)^{-\frac{Sm}{2}+1} \cdot \prod_{k=1}^{p} \left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)}, \end{aligned} \tag{2.10}$$

where

$$A^*(p) = A(p) \cdot (2\pi)^{\frac{1}{2}} \cdot \left(\Gamma(\frac{n}{2} + \alpha)\right)^p \cdot \left(\frac{Sm}{2}\right)^{-\frac{1}{2}} \cdot \exp\left[-\frac{Sm}{2}\right]$$

As pointed out by Oh and Raftery (2001, 2007) there is a difficulty in directly comparing $(\mathbf{X}^{(p)}, p)$ and $(\mathbf{X}^{(p+1)}, p+1)$ because the marginal posterior $\pi(\mathbf{X}, p|\mathbf{D})$ is dependent on the scale of $\mathbf{X}$. The term $\prod_{k=1}^{p} \left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)}$ may change with the dimension p, even without improvement in the fit. This shrinking effect would provide larger $\pi(\mathbf{X}, p|\mathbf{D})$ for larger p even when the likelihoods are the same and hence would favor larger p.

To avoid this scale dependency, Oh and Raftery (2001) suggested that two configurations with different dimension should then be compared in the same dimension. More precisely let $\mathbf{X}^{*(p+1)} = [\mathbf{X}^{(p)} : 0]$, $(p+1)$-dimensional configuration 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. If p is the correct dimension, the optimal solution in $(p+1)$-dimensional space would be (close to) $\mathbf{X}^{*(p+1)}$. Hence, the comparison between the two dimensions p and $p+1$ is conducted by the ratio of the

marginal posteriors of $\mathbf{X}^{*(p+1)}$ and $\mathbf{X}^{(p+1)}$:

$$\begin{aligned} R_p &\equiv \frac{\pi\left(\mathbf{X}^{(p+1)}, p+1|\mathbf{D}\right)}{\pi\left(\mathbf{X}^{*(p+1)}, p+1|\mathbf{D}\right)} \\ &= \left(\frac{SSR_{p+1}}{SSR^*_{p+1}}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p+1} \frac{\frac{\sum_{i=1}^n x_{ik}^2}{2}+\beta_k}{\frac{\sum_{i=1}^n x_{ik}^{*2}}{2}+\beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \\ &= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{\frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{2}+\beta_k}{\frac{\sum_{i=1}^n (x_{ik}^{(p)})^2}{2}+\beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \left(\frac{\frac{\sum_{i=1}^n (x_{ip+1}^{(p+1)})^2}{2}+\beta_{p+1}}{\beta_{p+1}}\right)^{-\left(\frac{n}{2}+\alpha\right)} \end{aligned}$$

Clearly, the ratio R_p depends on the choice of the hyperparameters α and β_k of Λ. Since we have assumed that in $(p+1)$-dimensional space $\alpha = \frac{1}{2}$ and $\beta_k = \frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{2n}$, we obtain:

$$R_p = \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{r_k^{(p+1)}(n+1)}{n+r_k^{(p+1)}}\right)^{-\frac{n+1}{2}} (n+1)^{-\frac{n+1}{2}}$$

where $r_k^{(p+1)} = \frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{\sum_{i=1}^n (x_{ik}^{(p)})^2}$.
Taking minus twice the logarithm of the ratio we define the criterion to choose between p and $p+1$:

$$\begin{aligned} LR_p &\equiv -2\log R_P \\ &= (Sm-2)\log\left(\frac{SSR_{p+1}}{SSR_p}\right) \qquad (2.11) \\ &\quad + \left\{(n+1)\sum_{k=1}^{p}\left[\log\frac{r_k^{(p+1)}(n+1)}{n+r_k^{(p+1)}}\right] + (n+1)\log(n+1)\right\} \qquad (2.12) \end{aligned}$$

Term (2.11) corresponds roughly to the log-likelihood ratio and would be negative because higher dimension results in a smaller SSR. The others terms, in (2.12), play the role of penalty on the increase of dimension by 1 and would be positive if $\prod_{k=1}^{p}(n/r_k^{(p+1)}+1) < (n+1)^{p+1}$. In the case when there is no significant change in $\mathbf{X}$ between p and $(p+1)$-dimensional spaces, the penalty term is approximately

$(n+1)\log(n+1)$. If LR_p is positive we would prefer dimension p to $p+1$. Hence, the dimension can be chosen as the first p where LR_p shows a positive value.
Alternatively, as in Oh and Raftery (2001, 2007) we propose a selection criterion, which we call MSR, as follows.

$$MSR_1 = (Sm-2)\log SSR_1,$$

$$MSR_p = MSR_{p-1} - 2\log R_{p-1}$$

The dimension p that achieves the minimum of MSR is favored.

2.5 A Simulation Study

In this section, we present the results from a simulation to evaluate the performance of the model.
In the design of the simulation study, for $n = 30$ objects, $S = 100$ (subjects) dissimilarity matrices were generated from Euclidean distances with Gaussian errors. We generated 30 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 distaces δ_{ij}. Given these δ_{ij}'s, we generated the observed dissimilarities $d_{ij,s}$ from a normal distribution with mean δ_{ij} and standard deviation 0.3, truncated at 0. In this way we generated dissimilarities in which individual differences are attributable only to differences in measurement errors in observations and there are not systematic individual differences.
We applied classical multidimensional scaling on the average dissimilarity matrix (CMDS) and Replicated Bayesian Multidimensional Scaling (RBMDS) for various values of the dimension p. We ran the MCMC for 10000 iterations for various values of the dimension p. The first 3000 iterations were used as burn-in, and the last 7000 iterations were used for generating model estimates. Convergence was checked by starting

the chain from multiple starting points and by inspecting time series plots.

The performance of the method has been evaluated by using the R squared correlation (RSQ) and $STRESS$ index. The RSQ value is the squared correlation coefficient between the distance obtained from the estimated object configuration $\hat{\delta}_{ij}$ and the observed dissimilarities $d_{ij,s}$, and it is the variance accounted for in the solution. $STRESS$ index is used to evaluate how well (or poorly) a particular configuration reproduces the observed dissimilarities. It can be calculated as follows:

$$STRESS = \sqrt{\frac{\sum_{s=1}^{S} \sum_{i>j} \left(d_{ij,s} - \hat{\delta}_{ij} \right)^2}{\sum_{s=1}^{S} \sum_{i>j} d_{ij,s}^2}}$$

where $\hat{\delta}_{ij}$ is the distance obtained from the estimated object configuration. Unlike the goodness-of-fit measure, RSQ, which increases as the fit to the data improves, $STRESS$ is a badness-of-fit measure which decreases as the fit to data improves.

Since we used simulated data we can calculate also a measure of fit between the Euclidean distance obtained from the estimated object configuration and the Euclidean distance obtained from the true object configuration.

$$ST = \sqrt{\frac{\sum_{i>j} \left(\delta_{ij} - \hat{\delta}_{ij} \right)^2}{\sum_{i>j} \delta_{ij}^2}}$$

where δ_{ij} is the true distance and $\hat{\delta}_{ij}$ is the Euclidean distance obtained from the estimated object configuration.

In Table 2.1 are summarized RSQ, $STRESS$ and ST from CMDS and RBMDS.

Table 2.1: Analysis of Simulation Data

	CMDS			RBMDS			
dim (p)	RSQ	STRESS	ST	RSQ	STRESS	ST	MSR
1	0.4441	0.6253	0.6234	0.4792	0.4614	0.4576	524430.8
2	0.6611	0.4326	0.4286	0.6708	0.2721	0.2642	479364.2
3	0.7564	0.3220	0.3159	0.7640	0.1908	0.1788	448556.8
4	0.8404	0.2265	0.2168	0.8588	0.1336	0.1158	417344.1
5	0.8847	0.1708	0.1575	0.8989	0.1048	0.0801	396164.2
6	0.9079	0.1307	0.1126	0.9230	0.0883	0.0568	381846.2
7	0.9310	0.1006	0.0763	0.9375	0.0781	0.0405	370904.0
8	0.9407	0.0808	0.0471	0.9450	0.0724	0.0286	365138.9
9	0.9481	0.0717	0.0297	0.9494	0.0695	0.0196	361520.3
10	0.9508	0.0686	0.0221	0.9513	0.0680	0.0135	360060.3*
11	0.9508	0.0682	0.0225	0.9514	0.0681	0.0151	360166.6
12	0.9508	0.0684	0.0240	0.9515	0.0681	0.0159	360280.9

*Minimum MSR

Results show that the RBMDS performs much better than do CMDS. The improvement in performance of RBMDS is more pronounced when the dimension is too small than when the dimension is close to the true dimension 10. It is interesting because dimension $p = 2$ is often chosen for visualization purposes. In reality, the number of dimensions that can be perceived and interpreted is limited. We computed MSR for various values of p to select the dimension of the configuration. We may observe from Table 2.1 that MSR assumes a minimum at the correct dimension, 10.

Chapter 3

Weighted Bayesian Multidimensional Scaling

In Chapter 2 we have showed a model that can analyze several matrices of dissimilarity, but it treats the differences among the matrices as due to random errors only. In many situations, this assumption of no systematic individual differences is unrealistic. In such case, each dissimilarity matrix may be analyzed separately, yielding as many object configurations as there are dissimilarity matrices. A natural question is how they are related. In most cases, there are both common and unique aspects in dissimilarity judgments obtained from different subjects. If we want investigate relationships among objects as well as individual differences in perception of those objects (capture both commonality and individual differences in dissimilarity judgments) we have to choose a strategy to obtain a common object configuration that applies to all subjects and to analyze heterogeneity among subjects.

In this chapter we propose a model that captures both commonality and individual differences in dissimilarity judgments. As INDSCAL model we interpret difference between the subjects as arising from differences in the weights ascribed to the dimensions of common configuration. More specifically, we postulate a common object configura-

tion that applies to all subjects, but the dimensions in the common configuration are differentially weighted by different subjects in order to better fit model distances to the set of observed dissimilarities.

We assume that for each subject the object configuration can differ from the common object configuration $\mathbf{X}$ and hence $\mathbf{X}^{(s)}$ denotes the coordinate matrix for individual s. The Euclidean distance between object i and j, for the s-th individual is

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

The matrix $\mathbf{X}^{(s)}$ is assumed to be related to $\mathbf{X}$. The assumption is made that all subjects use the same dimensions when evaluating the objects, but that they may utilize the dimensions in varying degree. Each subject weights the dimensions separately and, maybe, differently. Each subject might apply individual weights to these common dimensions, $w_{k,s}$ $(s = 1, ..., S;\ k = 1, ..., p)$. Then $x_{ik}^{(s)} = \sqrt{w_{k,s}}x_{ik}$ where the weights $w_{k,s}$ are required to be nonnegative $w_{k,s} \geq 0$.

The weighted Euclidean distance between object i and j, for subject s is

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

The output of the analysis consist of the common configuration given by a $n \times p$ matrix $\mathbf{X}$ and a $S \times p$ matrix of weights $\mathbf{W}$ for S subjects on p dimensions. The weights are interpreted as the importance of each dimension to each subject. A large weight means that the dimension is important to the subject, a small weight means the dimension is unimportant. If the dissimilarity matrices correspond to experimental conditions, say, rather than to people, the interpretation is that the weights reflect the importance of each dimension in the various experimental conditions.

In this way we produce a unique orientation of the axes of the common configuration,

in the sense that any rotation will destroy the optimality of the solution and will change the values of the subject weights. Because of this, it is not legitimate to rotate the axes of a common configuration to a more 'meaningful' orientation, as is commonly done in two-way multidimensional scaling model.

In this chapter we develop a Weighted Bayesian Multidimensional scaling (WBMDS) in which observed dissimilarities are modeled as equal to Weighted Euclidean distances plus a Gaussian error. We use Markov Chain Monte Carlo to obtain a Bayesian solution and we provide a Bayesian dimension selection criterion based on the WBMDS output.

3.1 Model and Prior

As in the previous chapter, the observed dissimilarities data are obtained from a sample of S subjects and they are denoted by $d_{ij,s}$, where $s = 1, ..., S$ indexes subjects, and $i, j = 1, ..., n$ indexes pairs of objects. The total data set will be indicated by the $S \times m$ matrix $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_S)'$ where the vector $\mathbf{d}_s = (d_{12,s}, d_{13,s}, ..., d_{n(n-1),s})'$ contain the $m = n(n-1)/2$ distinct dissimilarity values for subject s $(s = 1, ..., S)$.

Let $\delta_{ij,s}$ indicate the dissimilarity measure between objects i and j for subject s, which is assumed to be functionally related to p unobserved attributes of the objects and p unobserved attributes of the subject. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object i, and $\mathbf{w}_s = (w_{s1}, ..., w_{sp})$ denote an unobserved vector representing the weight associated with the p dimensions for the subject s. Unlike previous model,we model the true dissimilarity measure $\delta_{ij,s}$ as the Weighted Euclidean distance between the i-th and the j-th objects for the s-th subject: $\delta_{ij,s} = \sqrt{\sum_{k=1}^{p} w_{k,s}(x_{ik} - x_{jk})^2}$.

We assume that the observed dissimilarity measure, $d_{ij,s}$ is equal to the true distance, $\delta_{ij,s}$, plus a Gaussian error. In addition, since dissimilarity measures are typically given as positive values we restrict the observed dissimilarity to be always positive. Thus, given the Weighted Euclidean distance $\delta_{ij,s}$, the observed dissimilarity measure $d_{ij,s}$ is

assumed to follow the truncated normal distribution

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

and conditionally to $\delta_{ij,s}$, the $d_{ij,s}$'s are independent.
From this, the likelihood function of the unknown parameters $\mathbf{X} = \{\mathbf{x}_i\}$, $\mathbf{W} = \{\mathbf{w}_s\}$, and σ^2 is

$$l(\mathbf{X}, \mathbf{W}, \sigma^2) \propto (\sigma^2)^{-\frac{Sm}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s} - \delta_{ij,s})^2 - \sum_{s=1}^{S}\sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)\right]$$

where $m = n(n-1)/2$ and $\Phi(\cdot)$ is the standard normal cdf.
As in the previous chapter we assume that the $\mathbf{x}_i$ are iid with a common multivariate Gaussian prior density, with mean 0 and a diagonal covariance matrix Λ, i.e., $\mathbf{x}_i \sim N_p(0, \Lambda)$. We assume that the error variance σ^2 has the (conjugate) prior $\sigma^2 \sim IG(a, b)$ where $IG(a, b)$ is the inverse Gamma distribution with mode $b/(a+1)$. Since the weights $w_{k,s}$ have to be nonnegative $w_{k,s} \geq 0$ for the prior distribution of $w_{k,s}$, we use a negative exponential distribution, $w_{k,s} \sim Exp(\theta_k)$ independently for $k = 1, ..., p$ and $s = 1, ..., S$. For a hyperprior for the elements of $\Lambda = Diag(\lambda_1, ..., \lambda_p)$, following a standard approach, given dimension p, we assume a conjugate prior $\lambda_k \sim IG(\alpha, \beta_k)$, independently for each $k = 1, 2, ..., p$.

$$f(\mathbf{x}_i|\Lambda) = (2\pi)^{-p/2}|\Lambda|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i\right],$$
$$f(w_{k,s}) = \theta_k \exp\left[-\theta_k w_{k,s}\right],$$
$$f(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\left[-\frac{b}{\sigma^2}\right],$$
$$f(\lambda_k) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\lambda_k)^{-\alpha-1}\exp\left[-\frac{\beta}{\lambda_k}\right].$$

Finally, the prior densities are assumed to be independent.

3.2 Posterior Inference (MCMC)

From the likelihood and the prior, the posterior density function of the unknown parameters is

$$\begin{aligned}\pi(\mathbf{X},\mathbf{W},\sigma^2,\Lambda|\mathbf{D}) \;\propto\; & (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)}\prod_{k=1}^{p}\lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \\ & \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2-\sum_{s=1}^{S}\sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)\right. \\ & \left.-\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i-\frac{b}{\sigma^2}-\sum_{k=1}^{p}\frac{\beta_k}{\lambda_k}-\sum_{s=1}^{S}\sum_{k=1}^{p}\theta_k w_{k,s}\right].\end{aligned}$$

The full conditional distributions of the parameters involved in the model given all the other parameters and the data are:

$$\pi\left(\sigma^2|\cdot\right) \;\propto\; (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)}\exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2-\frac{b}{\sigma^2}-\sum_{s=1}^{S}\sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)\right],$$

$$\pi\left(\mathbf{x}_i|\cdot\right) \;\propto\; \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{j\neq i,j=1}^{n}(d_{ij,s}-\delta_{ij,s})^2-\frac{1}{2}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i-\sum_{s=1}^{S}\sum_{j\neq i,j=1}^{n}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)\right],$$

$$\pi\left(w_{k,s}|\cdot\right) \;\propto\; \exp\left[-\frac{1}{2\sigma^2}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2-\sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)-\theta_k w_{k,s}\right].$$

Finally $\pi\left(\lambda_k|\cdot\right)$ is $IG\left(\alpha+n/2,\beta_k+\sum_{i=1}^{n}x_{ik}^2/2\right)$

$$\pi\left(\lambda_k|\cdot\right)\propto\lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{x_{ik}^2}{\lambda_k}-\frac{\beta_k}{\lambda_k}\right].$$

We use an MCMC algorithm to simulate from the posterior distributions. For generation of Λ, σ^2, and $\mathbf{X}$ we can easily modify the algorithm of Replicated Bayesian Multidimensional Scaling. The full conditional posterior distributions of $w_{k,s}$ do not have closed forms, and so we apply a random walk Metropolis-Hastings algorithm to

generate samples of $w_{k,s}$. To choose the variance of the normal proposal density we remember that $\pi\left(w_{k,s}|\cdot\right) \propto \exp[-\frac{1}{2\sigma^2}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2 - \sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right) - \theta_k w_{k,s}]$. Because $(d_{ij,s}-\delta_{ij,s})^2/\sigma^2$ is a quadratic function of $w_{k,s}$ with leading coefficient equal to $1/\sigma^2$, $\frac{1}{\sigma^2}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2$ would dominate the full conditional posterior density function of $w_{k,s}$. Thus we can consider $\frac{1}{\sigma^2}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2$ only, approximate the full conditional variance of $w_{k,s}$ by σ^2/m and choose the variance of the normal proposal density to be proportional to σ^2/m.

To initialize the MCMC algorithm, we can run one WMDS methods such as INDSCAL or the SMACOF algorithm as a preliminary run to obtain starting values for $\mathbf{X}$ and $\mathbf{W}$. For σ^2, as in Oh and Raftery (2001), we can set $\sigma^{2(0)} = SSR^{(0)}/Sm$ where $SSR = \sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2$. For Λ, we can use diagonal elements of the covariance matrix of $\mathbf{X}^{(0)}$.

The prior parameters are chose as in the previous model. For the prior of $w_{k,s}$ we choose $\theta_k = S/\sum_{s=1}^{S} w_{k,s}^{(0)}$ so that the prior mean of $w_{k,s}$ matches with $\sum_{s=1}^{S} w_{k,s}^{(0)}/S$.

3.3 Choice of dimension

In this section we propose a Bayesian criterion with which to choose an appropriate dimension of the configuration. The choice of the dimension p corresponds to the choice of the configuration $(\mathbf{X}_p, \mathbf{W}_p)$ obtained by applying the algorithm described in the previous section with dimension p. We base our model selection criteria on $\pi\left(\mathbf{X}, \mathbf{W}, p|\mathbf{D}\right)$, the posterior density function of $\mathbf{X}$, $\mathbf{W}$ and p given data $\mathbf{D}$, evaluated at $\mathbf{X} = \mathbf{X}_p$ and $\mathbf{W} = \mathbf{W}_p$.

$$\pi(\mathbf{X}^{(p)}, \mathbf{W}^{(p)}|\mathbf{D}) = \int \pi(\mathbf{X}^{(p)}, \mathbf{W}^{(p)}|\Lambda, \sigma^2, \mathbf{D}, M_p)\pi(\Lambda, \sigma^2|D)d\Lambda d\sigma^2$$

Consider the dimension p as an unknown variable, and assume equal prior probability for all values among p_{min} and p_{max}.

Then the posterior is given by:

$$\begin{aligned}
\pi(\mathbf{X},\mathbf{W},\sigma^2,\Lambda,p|\mathbf{D}) &\propto l(\mathbf{X},\mathbf{W},\sigma^2,p|\mathbf{D})\pi(\mathbf{X}|\Lambda,p)\pi(\mathbf{W}|p)\pi(\sigma^2)\pi(\Lambda|p) \\
&= (2\pi)^{-\frac{Sm}{2}}(\sigma^2)^{-\left(\frac{Sm}{2}\right)}\exp\left[-\frac{1}{2\sigma^2}SSR-\sum_{s=1}^{S}\sum_{i>j}\log\Phi\left(\frac{\delta_{ij,s}}{\sigma}\right)\right] \\
&\quad\times(2\pi)^{-\frac{np}{2}}\prod_{k=1}^{p}\lambda_k^{-\frac{n}{2}}\exp\left[-\sum_{k=1}^{p}\frac{\sum_{i=1}^{n}x_{ik}^2}{2\lambda_k}\right] \\
&\quad\times\prod_{s=1}^{S}\prod_{k=1}^{p}\theta_k\exp\left[-\theta_k w_{k,s}\right] \\
&\quad\times\Gamma(a)^{-1}b^a\left(\sigma^2\right)^{-(a+1)}exp\left[-\frac{b}{\sigma^2}\right] \\
&\quad\times\Gamma(\alpha)^{-p}\prod_{k=1}^{p}\beta_k^{\alpha}\lambda_k^{-(\alpha+1)}exp\left[-\frac{\beta_k}{\lambda_k}\right] \\
&= A(p)\cdot h\left(\sigma^2,\mathbf{X},\mathbf{W}\right)\cdot g\left(\Lambda,\mathbf{X},\mathbf{W},p\right), \qquad (3.1)
\end{aligned}$$

where

$$SSR=\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s}-\delta_{ij,s})^2$$

$$A(p)=(2\pi)^{-(Sm+np)/2}\Gamma(a)^{-1}b^a\Gamma(\alpha)^{-p}\prod_{k=1}^{p}\beta_k^{\alpha}\lambda_k^{-(\alpha+1)}$$

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

$$g\left(\Lambda,\mathbf{X},p\right)=\prod_{k=1}^{p}\lambda_k^{-(n/2+\alpha+1)}\exp\left[-\frac{1}{\lambda_k}\left(\frac{\sum_{i=1}^{n}x_{ik}^2}{2}+\beta_k\right)\right]\times\prod_{s=1}^{S}\prod_{k=1}^{p}\theta_k\exp\left[-\theta_k w_{k,s}\right]$$

Note that $\pi\left(\mathbf{X},\mathbf{W},p|\mathbf{D}\right)\propto f\left(\mathbf{D}|\mathbf{X},\mathbf{W},p\right)\pi\left(\mathbf{X},\mathbf{W},p\right)$. Thus, one needs only to compute the marginal likelihood and the marginal prior of $\mathbf{X}$ and $\mathbf{W}$ for each p. The

marginalized likelihood term increases as p increases and it is approximately equal to:

$$\int h\left(\sigma^2, \mathbf{X}\right) d\sigma^2 \approx (2\pi)^{\frac{1}{2}} \left(\frac{Sm}{2}\right)^{-\frac{1}{2}} \left(\frac{SSR}{Sm}\right)^{-\frac{Sm}{2}+1} \exp\left[-\frac{Sm}{2}\right].$$

The marginal prior term decreases as p increases, so this term penalizes more complex models, and it is approximately equal to:

$$\int g\left(\Lambda, \mathbf{X}, \mathbf{W}, p\right) d\Lambda = \Gamma^p\left(\frac{n}{2}+\alpha\right) \prod_{k=1}^{p} \left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)} \prod_{s=1}^{S}\prod_{k=1}^{p} \left(\theta_k \exp\left[-\theta_k w_{k,s}\right]\right)$$

From the above results,

$$\begin{aligned}
\pi\left(\mathbf{X}, \mathbf{W}, p|\mathbf{D}\right) &= c \cdot A(p) \cdot \int h\left(\sigma^2, \mathbf{X}, \mathbf{W}\right) d\sigma^2 \cdot \int g\left(\Lambda, \mathbf{X}, \mathbf{W}, p\right) d\Lambda \\
&= c \cdot A^*(p) \cdot \left(\frac{SSR}{Sm}\right)^{-\frac{Sm}{2}+1} \\
&\quad \times \prod_{k=1}^{p} \left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)} \prod_{s=1}^{S}\prod_{k=1}^{p} \left(\theta_k \exp\left[-\theta_k w_{k,s}\right]\right),
\end{aligned}$$

where

$$A^*(p) = A(p) \cdot (2\pi)^{\frac{1}{2}} \cdot \left(\Gamma\left(\frac{n}{2}+\alpha\right)\right)^p \cdot \left(\frac{Sm}{2}\right)^{-\frac{1}{2}} \cdot \exp\left[-\frac{Sm}{2}\right]$$

There is a difficulty in directly comparing $(\mathbf{X}^{(p)}, \mathbf{W}^{(p)}, p)$ and $(\mathbf{X}^{(p+1)}, \mathbf{W}^{(p+1)}, p+1)$. The marginal posterior $\pi(\mathbf{X}, \mathbf{W}|\mathbf{D})$ can be larger in a higher dimension, although there is no change in the distance and the fit. Then to avoid shrinking effect, we compare configurations in the same dimensional space. Let $\mathbf{X}^{*(p+1)} = [\mathbf{X}^{(p)} : 0]$ and $\mathbf{W}^{*(p+1)} = [\mathbf{W}^{(p)} : 0]$ in $(p+1)$-dimensional space, which have the first p coordinates and weights equal to $\mathbf{X}^{(p)}$ and $\mathbf{W}^{(p)}$, and the last coordinates and weights all equal to 0. Then $(\mathbf{X}^{*(p+1)}, \mathbf{W}^{*(p+1)})$ provide the same Weight Euclidean distance and the same fit as $(\mathbf{X}^{(p)}, \mathbf{W}^{(p)})$ and may be considered as an embedding of $(\mathbf{X}^{(p)}, \mathbf{W}^{(p)})$in $(p+1)$-dimensional space.

We base the choice between dimension p and dimension $p+1$ on the ratio of the

marginal posteriors of $(\mathbf{X}^{*(p+1)}, \mathbf{W}^{*(p+1)})$ and $(\mathbf{X}^{(p+1)}, \mathbf{W}^{(p+1)})$

$$
\begin{aligned}
R_p &\equiv \frac{\pi\left(\mathbf{X}^{(p+1)}, \mathbf{W}^{(p+1)}, p+1|\mathbf{D}\right)}{\pi\left(\mathbf{X}^{*(p+1)}, \mathbf{W}^{*(p+1)}, p+1|\mathbf{D}\right)} \\
&= \left(\frac{SSR_{p+1}}{SSR^*_{p+1}}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p+1} \frac{\frac{\sum_{i=1}^n x_{ik}^2}{2}+\beta_k}{\frac{\sum_{i=1}^n x_{ik}^{*2}}{2}+\beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \left(\prod_{k=1}^{p+1}\prod_{s=1}^{S} \frac{\exp[-\theta_k w_{k,s}]}{\exp[-\theta_k w^*_{k,s}]}\right) \\
&= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{\frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{2}+\beta_k}{\frac{\sum_{i=1}^n (x_{ik}^{(p)})^2}{2}+\beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \left(\frac{\frac{\sum_{i=1}^n (x_{ip+1}^{(p+1)})^2}{2}+\beta_{p+1}}{\beta_{p+1}}\right)^{-\left(\frac{n}{2}+\alpha\right)} \\
&\quad \times \prod_{k=1}^{p} \left(\frac{\exp[-\theta_k \sum_{s=1}^S w_{k,s}^{(p+1)}]}{\exp[-\theta_k \sum_{s=1}^S w_{k,s}^{(p)}]}\right) \exp\left[-\theta_{p+1} \sum_{s=1}^{S} w_{p+1,s}^{(p+1)}\right]
\end{aligned}
$$

If $\alpha = \frac{1}{2}$, $\beta_k = \frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{2n}$ and $\theta_k = S/\sum_{s=1}^S w_{k,s}^{(p+1)}$ we obtain:

$$
\begin{aligned}
R_p &= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{r_k^{(p+1)}(n+1)}{n+r_k^{(p+1)}}\right)^{-\frac{n+1}{2}} \\
&\quad \times (n+1)^{-\frac{n+1}{2}} \prod_{k=1}^{p} \left(\exp\left[-S + S z_k^{(p+1)}\right]\right) \exp\left[-S\right]
\end{aligned}
$$

where

$$
r_k^{(p+1)} = \frac{\sum_{i=1}^n (x_{ik}^{(p+1)})^2}{\sum_{i=1}^n (x_{ik}^{(p)})^2},
$$

$$
z_k^{(p+1)} = \frac{\sum_{s=1}^S w_{k,s}^{(p)}}{\sum_{s=1}^S w_{k,s}^{(p+1)}}.
$$

Taking minus twice the logarithm of the ratio we define the criterion to choose between p and $p+1$:

$$\begin{aligned} LR_p &\equiv -2\log R_P \\ &= (Sm-2)\log\left(\frac{SSR_{p+1}}{SSR_p}\right) && (3.2) \\ &\quad +\left\{(n+1)\sum_{k=1}^{p}\left[\log\frac{r_k^{(p+1)}(n+1)}{n+r_k^{(p+1)}}\right]+(n+1)\log(n+1)\right\} && (3.3) \\ &\quad -2S\sum_{k=1}^{p} z_k^{(p+1)}+2S(p+1) && (3.4) \end{aligned}$$

A positive value of LR_p would lead to prefer dimension p to $(p+1)$. The (3.2) term correspond to the log-likelihood ratio and would be negative because higher dimension results in a smaller SSR. The terms (3.3) and (3.4) play the role of penalty on the increase of the dimension from p to $p+1$. The term (3.3) would be positive if $\prod_{k=1}^{p}(n/r_k^{(p+1)}+1) < (n+1)^{p+1}$. When there is no significant change in $\mathbf{X}$ between p and $(p+1)$-dimensional spaces, this penalty term is approximately $(n+1)\log(n+1)$. The term (3.4) would be positive if $\sum_{k=1}^{p} z_k^{(p+1)} < p+1$. When there is no significant change in $\mathbf{W}$ between p and $(p+1)$-dimensional spaces, this penalty term is approximately $2S$.

Alternatively, it is possible to define MSW as:

$$MSW_1 = (Sm-2)\log SSR_1,$$

$$MSW_p = MSW_{p-1} - 2\log R_{p-1}$$

then the optimal dimension is the one that achieves the minimum of MSW_p.

3.4 A Simulation Study

In this section, we present the results from a simulation to evaluate the performance of proposed method. The purpose is to recover the underlying classification structure of subjects together with the common object configuration.

Two datasets with S=60 subjects and n=8 objects in p=4 dimensions each were generated. In the first case, we generated dissimilarities in which there isn't systematic individual differences. We generated 8 random samples of $\mathbf{x}_i$ from a 4-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 dissimilarities $d_{ij,s}$ from a normal distribution with mean δ_{ij} and standard deviation 0.3, truncated at 0. Thus, the data consist of 60 8×8 symmetric matrices of dissimilarities computed from Euclidean distances with Gaussian errors.

In the second case we postulate a common object configuration that applies to all subjects, but the dimensions in the common configuration are differentially weighted by different subjects. We generated a random configuration in 4 dimensions for 8 objects and the weight vectors for 4 subjects' groups. 60 subjects were assigned to one of the 4 groups. Once the coordinates of parameters and the weights were randomly generated we used the Weighted Euclidean distances as dissimilarities $\delta_{ij,s}$. We generated the observed distances $d_{ij,s}$ from a normal distribution with mean $\delta_{ij,s}$ and standard deviation 0.3, truncated at 0. Thus, the data consist of 60 8×8 symmetric matrices of dissimilarities computed from Weighted Euclidean distances with Gaussian errors.

CMDS, Replicated Bayesian MDS and Weighted Bayesian MDS were applied to each generated data set for various values of the dimension p. In all the examples, for each Bayesian models we ran the MCMC for 10000 iterations. The first 3000 iterations were used as burn-in, and the last 7000 iterations were used for generating model estimates. Convergence was checked by starting the chain from multiple starting points and by inspecting time series plots.

As in the previous chapter the performance of the methods has been evaluated by using the R squared correlation (RSQ), the STRESS index

$$STRESS = \sqrt{\frac{\sum_{s=1}^{S}\sum_{i>j}\left(d_{ij,s} - \hat{\delta}_{ij,s}\right)^2}{\sum_{s=1}^{S}\sum_{i>j} d_{ij,s}^2}}$$

and a measure of fit between the Euclidean distance obtained from the estimated object configuration and the Euclidean distance obtained from the true object configuration

$$ST = \sqrt{\frac{\sum_{s=1}^{S}\sum_{i>j}\left(\delta_{ij,s} - \hat{\delta}_{ij,s}\right)^2}{\sum_{s=1}^{S}\sum_{i>j} \delta_{ij,s}^2}}$$

Table 3.1 displays the results of the simulation experiment for the first data sets. The table shows that the two Bayesian Multidimensional Scaling procedures are characterized by better fit as compared to the standard ones, and are more or less characterized by the same fit. This is expected since in our simulations all individuals have a common object configuration, that is all subjects consider the dimensions in the common configuration to have equal importance. The improvement in performance of RBMDS and WBMDS is more pronounced when the dimension is too small than when the dimension is close to the true dimension 4. For each Bayesian models we computed the Bayesian selection criteria for various values of p and in all cases the Bayesian selection criteria correctly identified the true dimension p=4.

Figure 3.1 displays the subject space obtained from Weighted Bayesian MDS with four dimensions. Examination of this figure reveals that the subjects are quite homogeneous in their perceptions, because the weights show few substantive differences on either dimension and no distinct “clusters” of subjects emerge.

Table 3.2 displays the results of the simulation experiment for the second data set. In order to compare WBMDS with an other Weighted Multidimensional Scalings model we also applied SMACOF algorithm. Results in Table 3.2 show that the WBMDS

performs much better than do CMDS, RBMDS and SMACOF.

Table 3.1: Results of the simulation experiment 1

Classical MDS (CMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.5055	0.8784	0.9211	0.9419	0.9420
STRESS	0.5171	0.2011	0.1345	0.1083	0.1083
ST	0.4145	0.1262	0.0728	0.0590	0.0654

Replicated Bayesian MDS (RBMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.5426	0.8910	0.9290	0.9420	0.9420
STRESS	0.4127	0.1598	0.1210	0.1079	0.1079
ST	0.4067	0.1184	0.0532	0.0532	0.0532
MSR	12776.06	9606.219	8765.663	8337.433*	8355.076

*Minimum MSR

Weighted Bayesian MDS (WBMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.6013	0.8920	0.9305	0.9460	0.9464
STRESS	0.3853	0.1604	0.1200	0.1044	0.1041
ST	0.3745	0.0964	0.0481	0.0442	0.0401
MSW	12546.34	9536.985	8481.381	8045.292*	8094.993

*Minimum MSW

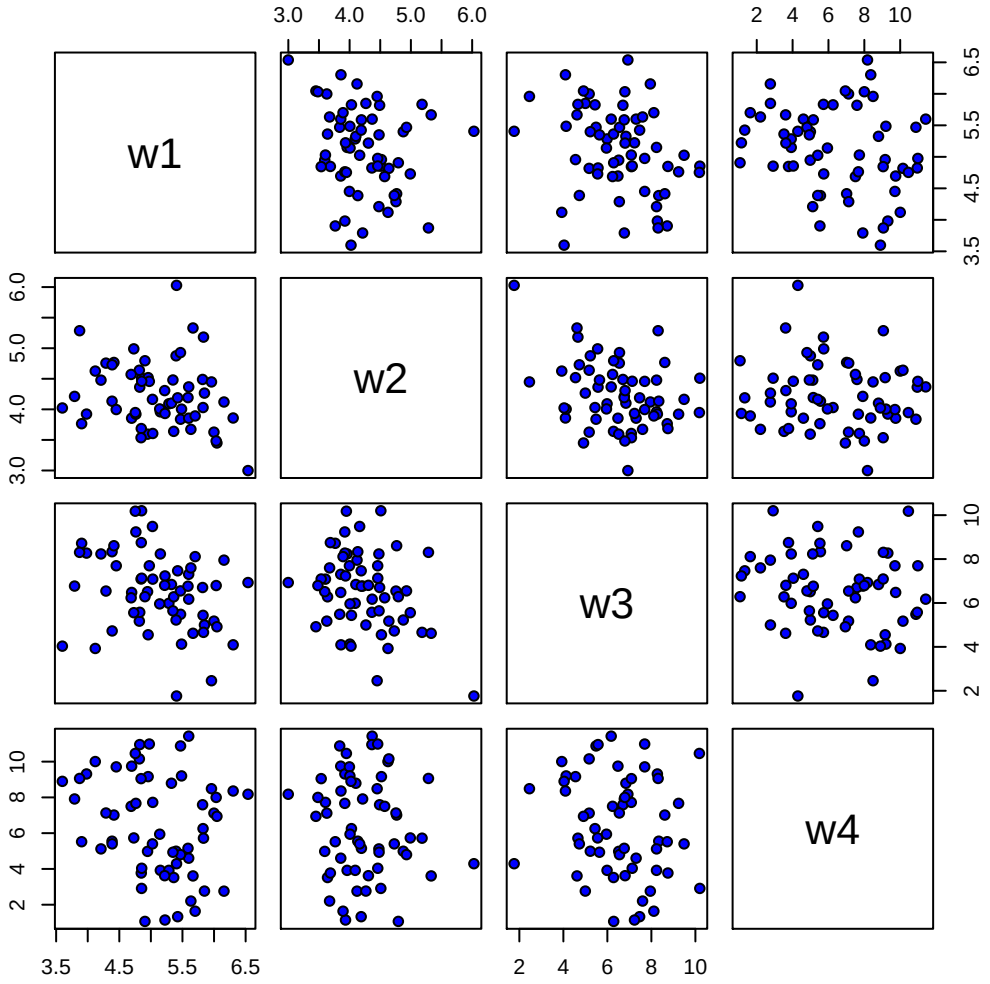


Figure 3.1: Estimated Subject Configuration of the simulation experiment 1

Table 3.2: Results of the simulation experiment 2

Classical MDS (CMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.6626	0.7674	0.8466	0.8512	0.8512
STRESS	0.4153	0.2895	0.2047	0.2004	0.2005
ST	0.4034	0.2801	0.1901	0.1826	0.1818

Scaling by MAjorizing a Complicated Function (SMACOF)

	p=1	p=2	p=3	p=4	p=5
RSQ		0.8535	0.9193	0.9373	0.9391
STRESS		0.2124	0.1465	0.1399	0.1426
ST		0.2006	0.1461	0.1391	0.1398

Replicated Bayesian MDS (RBMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.7208	0.7992	0.8476	0.8506	0.8512
STRESS	0.3223	0.2383	0.2027	0.2004	0.2005
ST	0.3137	0.1839	0.1227	0.1419	0.1261
MSR	13391.55	12390.9	11855.62	11853.36*	11874.2

*Minimum MSR

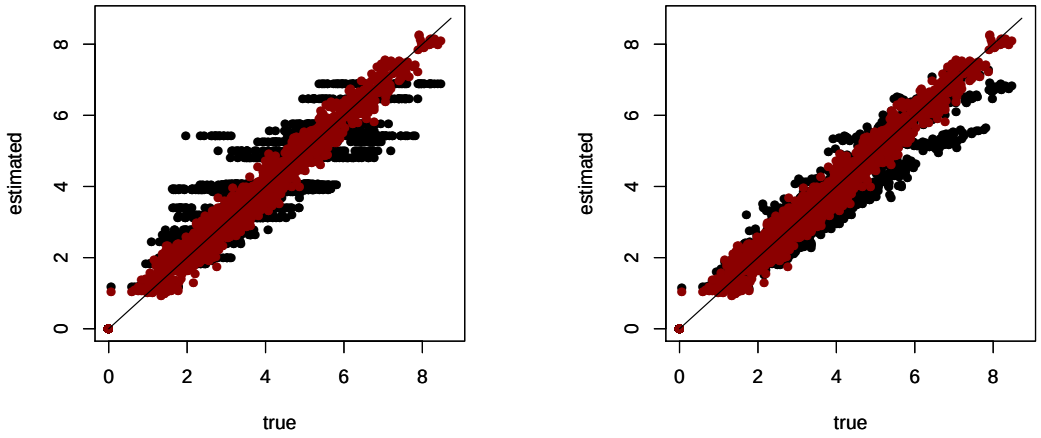
Weighted Bayesian MDS (WBMDS)

	p=1	p=2	p=3	p=4	p=5
RSQ	0.7285	0.8692	0.9223	0.9486	0.9592
STRESS	0.3207	0.1683	0.0795	0.0716	0.0714
ST	0.3020	0.1540	0.0708	0.0692	0.0690
MSW	11278.06	9197.607	6772.744	6598.226*	6699.826

*Minimum MSW

For each Bayesian models we computed the Bayesian selection criteria for various values of p and in all cases the Bayesian selection criteria correctly identified the true dimension p=4.

Figure 3.2 is a plot of the true dissimilarities versus the estimated distances. It is a scatterplot of true dissimilarities against output distances for every pair of items scaled and for every subjects. The X-axis corresponds to the true dissimilarities and the Y-axis corresponds to the MDS distances. A perfect fit would yield a 45-degree line. The estimated Euclidean distances from WBMDS are represented as red dots. Black dots in (a) represent the estimated distances from CMDS and in (b) from SMACOF. One can see that WBMDS provides points very close to the 45-degree line.



(a) CMDS and WBMDS (b) SMACOF and WBMDS

Figure 3.2: True and Estimated Distances. Red dots represent the estimated distances from WBMDS, black dots in (a) from CMDS, and black dots in (b) from SMACOF

Examination of the weights in Figure 3.3 reveals that in this case the subjects are not homogeneous in their perceptions and we can clearly detect the four subjects' groups.

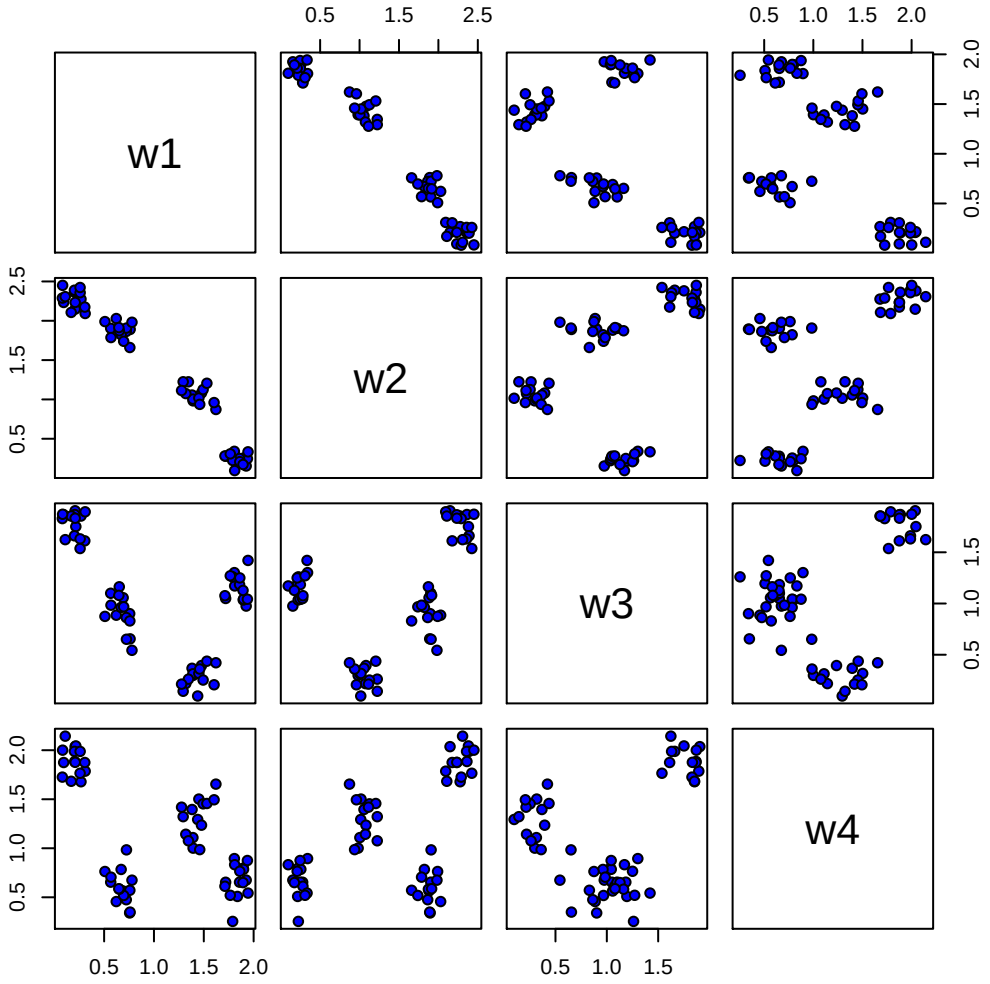


Figure 3.3: Estimated Subject Configuration of the simulation experiment 2

Chapter 4

Groups Bayesian Multidimensional Scaling

In Chapter 3 we have presented a Weighted Bayesian Multidimensional Scaling. The output of this model consist of a common object configuration and a subject configuration that contains the weight for each individual on each dimension of the object configuration. The weights are rarely interpreted for individual subjects, and the improvement in goodness-of-fit measures seldom seems to justify so many additional parameters. When we have to analyze data sets where the number of subjects S is very large the resulting subject space tends to be cluttered often rendering interpretation impossible. Given the traditional role of market segmentation in marketing strategy, marketers are rarely interested in the particular responses of consumers at the individual level but they are more concerned with identifying and targeting market segments - homogeneous groups of consumers who share some designated set of characteristics (e.g. demographics, psychographics, consumption patterns, etc.) - that are relevant to the purchase of the brand under study (DeSarbo, Manrai, and Manrai, 1994).
In this chapter, we consider differences among preselected groups of subjects $t = 1, ...T$. The subjects are classified into groups according to the categories of one or more qual-

itative variales. Each group of subjects is assumed to have his own configuration space of objects, which is the common configuration with its dimensions stretched or shrinked in proportion to the relevance that group gives to dimensions. The aim is to investigate if the qualitative variables considered bias the difference importance subjects attach to the various common scale, namely, if the partition of subjects is useful to explain the dissimilarities matrices. In essence, what we obtain is a Weighted Multidimensional solution where the Subject Space does not represent individual subjects but groups of subjects. Subjects can be grouped together based on a wide variety of factors. For example, in marketing research individuals can be grouped together based on their gender and we can investigate if there are differences in perception among brands between male and female.
We develop a Group Bayesian Multidimensional scaling using Markov Chain Monte Carlo to obtain a Bayesian solution and on the basis of the GBMDS estimate of object and subject configuration over a range of dimensions, we propose a Bayesian criterion with which to choose an appropriate dimension.

4.1 Model and Prior

The observed dissimilarities data are obtained from a sample of S subjects and they are denoted by $d_{ij,s}$, where $s = 1, ..., S$ indexes subjects, and $i, j = 1, ..., n$ indexes pairs of objects. The total data set will be indicated by the $S \times m$ matrix $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_S)'$ where the vector $\mathbf{d}_s = (d_{12,s}, d_{13,s}, ..., d_{n(n-1),s})'$ contain the $m = n(n-1)/2$ dissimilarity values for subject s $(s = 1, ..., S)$.
Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object i, and $\mathbf{w}_t = (w_{t1}, ..., w_{tp})$ denote an unobserved vector representing the weights associated with the p dimension for the subjects belong to group t. We define $\delta_{ij,t}$ as the Weighted Euclidean distance between the i-th and the

j-th objects for subjects belong to group t,

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

We assume that, given the Weighted Euclidean distance $\delta_{ij,t}$, the observed dissimilarity measure $d_{ij,s}$ follows the truncated normal distribution

$$d_{ij,s} \sim N(\delta_{ij,t|s\in t}, \sigma^2) I(d_{ij,s} > 0), \quad i \neq j,\ i,j = 1,...,n \quad s = 1,...,S \quad t = 1,...,T \quad T \leq S$$

From this, the likelihood function of the unknown parameters $\mathbf{X} = \{\mathbf{x}_i\}$, $\mathbf{W} = \{\mathbf{w}_t\}$ and σ^2 is

$$l(\mathbf{X}, \mathbf{W}, \sigma^2) \propto (\sigma^2)^{-\frac{Sm}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i>j}(d_{ij,s} - \delta_{ij,t|s\in t})^2 - \sum_{t=1}^{T} S_t \sum_{i>j} \log\Phi\left(\frac{\delta_{ij,t}}{\sigma}\right)\right]$$

where $m = n(n-1)/2$, S_t is the number of subjects belong to group t, and $\Phi(\cdot)$ is the standard normal cdf.

We use the following priors for the model parameters:

$$f(\mathbf{x}_i|\Lambda) = (2\pi)^{-p/2} |\Lambda|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i\right]$$

$$f(w_{k,t}) = \theta_k \exp\left[-\theta_k w_{k,t}\right]$$

$$f(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1} \exp\left[-\frac{b}{\sigma^2}\right]$$

$$f(\lambda_k) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\lambda_k)^{-\alpha-1} \exp\left[-\frac{\beta}{\lambda_k}\right]$$

The prior densities are assumed to be independent.

4.2 Posterior Inference (MCMC)

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

$$\begin{aligned}\pi(\mathbf{X}, \mathbf{W}, \sigma^2, \Lambda|\mathbf{D}) \quad \propto \quad & (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)} \prod_{k=1}^{p} \lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \\ & \exp\left[-\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{i>j} (d_{ij,s} - \delta_{ij,t|s\in t})^2 - \sum_{t=1}^{T} S_t \sum_{i>j} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma}\right) \right. \\ & \left. - \frac{1}{2} \sum_{i=1}^{n} \mathbf{x}_i' \Lambda^{-1} \mathbf{x}_i - \frac{b}{\sigma^2} - \sum_{k=1}^{p} \frac{\beta_k}{\lambda_k} - \sum_{t=1}^{T} \sum_{k=1}^{p} \theta_k w_{k,t} \right],\end{aligned}$$

where $\mathbf{D}$ is the matrix of observed dissimilarities.
The full conditional distributions of the parameters involved in the model given all the other parameters and the data are:

$$\pi\left(\sigma^2|\cdot\right) \propto (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)} \exp\left[-\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{i>j} (d_{ij,s} - \delta_{ij,t|s\in t})^2 - \frac{b}{\sigma^2} - \sum_{t=1}^{T} S_t \sum_{i>j} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma}\right) \right],$$

$$\pi\left(\mathbf{x}_i|\cdot\right) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{j\neq i, j=1}^{n} (d_{ij,s} - \delta_{ij,t|s\in t})^2 - \frac{1}{2} \mathbf{x}_i' \Lambda^{-1} \mathbf{x}_i - \sum_{t=1}^{T} S_t \sum_{j\neq i, j=1}^{n} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma}\right) \right],$$

$$\pi\left(w_{k,t}|\cdot\right) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{s\in t} \sum_{i>j} (d_{ij,s} - \delta_{ij,t})^2 - S_t \sum_{i>j} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma}\right) - \theta_k w_{k,t} \right].$$

Finally $\pi\left(\lambda_k|\cdot\right)$ is $IG\left(\alpha + n/2, \beta_k + s_k = \sum_{i=1}^{n} x_{ik}^2/2\right)$

$$\pi\left(\lambda_k|\cdot\right) \propto \lambda_k^{-\left(\frac{n}{2}+\alpha+1\right)} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \frac{x_{ik}^2}{\lambda_k} - \frac{\beta_k}{\lambda_k} \right].$$

We use an MCMC algorithm to simulate from the posterior distributions. For generation of Λ, σ^2, and $\mathbf{X}$ we can easy modify the algorithm of Weighted Bayesian Multidimensional Scaling. For generation of $w_{k,t}$ the proposal density used in the Metropolis Hastings algorithm is a normal proposal density with variance proportional to $\sigma^2/S_t m$.

To initialize the MCMC algorithm, we can use the solution obtained by applying one WMDS methods such as INDSCAL or the SMACOF algorithm for the average subjects for each group.

The prior parameters are chose as in the previous models. For the prior of $w_{k,t}$ we choose $\theta_k = T/\sum_{t=1}^{T} w_{k,t}^{(0)}$ so that the prior mean of $w_{k,t}$ matches with $\sum_{t=1}^{T} w_{k,t}^{(0)}/T$.

4.3 Choice of dimension

Posterior inference as described in the previous section presumed that the dimensionality of the solution space, p is given. As in the previous models we propose a Bayesian criterion with which to choose an appropriate dimension among p_{min} e p_{max}. Let $\mathbf{X}^{*(p+1)} = [\mathbf{X}^{(p)} : 0]$ and $\mathbf{W}^{*(p+1)} = [\mathbf{W}^{(p)} : 0]$ in $(p+1)$-dimensional space, which have the first p coordinates and weights equal to $\mathbf{X}^{(p)}$ and $\mathbf{W}^{(p)}$, and the last coordinates and weights all equal to 0. We base the choice between dimension p and dimension $p+1$ on the ratio of the marginal posteriors of $(\mathbf{X}^{*(p+1)}, \mathbf{W}^{*(p+1)})$ and $(\mathbf{X}^{(p+1)}, \mathbf{W}^{(p+1)})$

$$
\begin{aligned}
R_p &\equiv \frac{\pi\left(\mathbf{X}^{(p+1)}, \mathbf{W}^{(p+1)}, p+1|\mathbf{D}\right)}{\pi\left(\mathbf{X}^{*(p+1)}, \mathbf{W}^{*(p+1)}, p+1|\mathbf{D}\right)} \\
&= \left(\frac{SSR_{p+1}}{SSR^*_{p+1}}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p+1} \frac{\frac{\sum_{i=1}^{n} x_{ik}^2}{2} + \beta_k}{\frac{\sum_{i=1}^{n} x_{ik}^{*2}}{2} + \beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \left(\prod_{k=1}^{p+1}\prod_{t=1}^{T} \frac{\exp[-\theta_k w_{k,t}]}{\exp[-\theta_k w^*_{k,t}]}\right) \\
&= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{\frac{\sum_{i=1}^{n} (x_{ik}^{(p+1)})^2}{2} + \beta_k}{\frac{\sum_{i=1}^{n} (x_{ik}^{(p)})^2}{2} + \beta_k}\right)^{-\left(\frac{n}{2}+\alpha\right)} \left(\frac{\frac{\sum_{i=1}^{n} (x_{ip+1}^{(p+1)})^2}{2} + \beta_{p+1}}{\beta_{p+1}}\right)^{-\left(\frac{n}{2}+\alpha\right)} \\
&\quad \times \prod_{k=1}^{p} \left(\frac{\exp[-\theta_k \sum_{t=1}^{T} w_{k,t}^{(p+1)}]}{\exp[-\theta_k \sum_{t=1}^{T} w_{k,t}^{(p)}]}\right) \exp\left[-\theta_{p+1} \sum_{t=1}^{T} w_{p+1,t}^{(p+1)}\right]
\end{aligned}
$$

where

$$SSR = \sum_{s=1}^{S} \sum_{i>j} (d_{ij,s} - \delta_{ij,t|s\in t})^2$$

If $\alpha = \frac{1}{2}$, $\beta_k = \frac{\sum_{i=1}^{n}(x_{ik}^{(p+1)})^2}{2n}$ and $\theta_k = T/\sum_{t=1}^{T} w_{k,t}^{(p+1)}$ we obtain:

$$\begin{aligned} R_p &= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-\frac{Sm}{2}+1} \left(\prod_{k=1}^{p} \frac{r_k^{(p+1)}(n+1)}{n + r_k^{(p+1)}}\right)^{-\frac{n+1}{2}} \\ &\quad \times (n+1)^{-\frac{n+1}{2}} \prod_{k=1}^{p} \left(\exp\left[-T + T z_k^{(p+1)}\right]\right) \exp\left[-T\right] \end{aligned}$$

where $r_k^{(p+1)} = \frac{\sum_{i=1}^{n}(x_{ik}^{(p+1)})^2}{\sum_{i=1}^{n}(x_{ik}^{(p)})^2}$ and $z_k^{(p+1)} = \frac{\sum_{t=1}^{T} w_{k,t}^{(p)}}{\sum_{t=1}^{t} w_{k,t}^{(p+1)}}$.

Taking minus twice the logarithm of the ratio we define the criterion to choose between p and $p+1$:

$$\begin{aligned} LR_p &\equiv -2 \log R_P \\ &= (Sm-2) \log\left(\frac{SSR_{p+1}}{SSR_p}\right) && (4.1) \\ &\quad + \left\{(n+1)\sum_{k=1}^{p}\left[\log \frac{r_k^{(p+1)}(n+1)}{n + r_k^{(p+1)}}\right] + (n+1)\log(n+1)\right\} && (4.2) \\ &\quad -2T\sum_{k=1}^{p} z_k^{(p+1)} + 2T(p+1) && (4.3) \end{aligned}$$

A positive value of LR_p would lead to prefer dimension p to $(p+1)$. The (4.1) term correspond to the log-likelihood ratio and would be negative because higher dimension results in a smaller SSR. The terms (4.2) and (4.3) play the role of penalty on the increase of the dimension from p to $p+1$. The term (4.2) would be positive if $\prod_{k=1}^{p}(n/r_k^{(p+1)} + 1) < (n+1)^{p+1}$ and when there is no significant change in $\mathbf{X}$ between p and $(p+1)$-dimensional spaces, it is approximately $(n+1)\log(n+1)$. The term (4.3) would be positive if $\sum_{k=1}^{p} z_k^{(p+1)} < p+1$ and when there is no significant change in $\mathbf{W}$ between p and $(p+1)$-dimensional spaces, it is approximately $2T$.

Alternatively, it is possible to define MSG as:

$$MSG_1 = (Sm - 2) \log SSR_1,$$

$$MSG_p = MSG_{p-1} - 2 \log R_{p-1}$$

then the optimal dimension is the one that achieves the minimum of MSG_p.

4.4 A Simulation Study

In this section we consider the second data set introduced in Section 3.4. Groups Bayesian MDS was applied for $T = 4$ groups which correspond with the four cases we have generated.

Table 4.1: Analysis of Simulation Data

	WBMDS			GBMDS			
dim (p)	RSQ	STRESS	ST	RSQ	STRESS	ST	MSG
1	0.7285	0.3207	0.3020	0.7332	0.3158	0.3194	13311.56
2	0.8692	0.1683	0.1540	0.8614	0.1691	0.1583	11802.14
3	0.9223	0.0795	0.0708	0.9370	0.0798	0.0725	10797.69
4	0.9486	0.0716	0.0692	0.9414	0.0725	0.0686	10385.56*
5	0.9592	0.0714	0.0690	0.9412	0.0726	0.0694	10582.03

*Minimum MSG

Results in Table 4.1 show that true to form the WBMDS and GBMDS are more or less characterized by the same goodness of fit. This means that the partition of subjects is significative to explain the dissimilarity matrices. Indeed, when we have generated the data we have postulated a common object configuration, and differential weighting of the dimensions in the common configuration by each subjects' groups. Since the GBMDS is a model more parsimonious than WBMDS in this case would be better use

the GBMDS than WBMDS.

The output of the analysis consist of a common object configuration, and a matrix of weights of the 4 groups on the p dimensions. With these obtained results, we applied MSG described in Section 4.3 to select the dimension of the solution space. MSG assumes a minimum at the correct dimension $p = 4$.

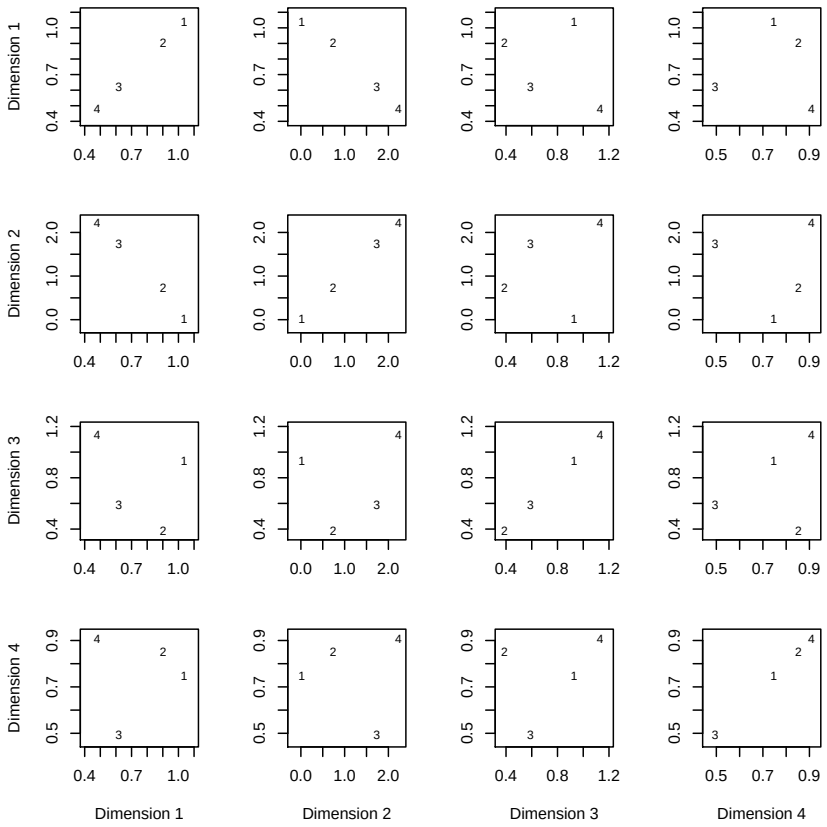


Figure 4.1: Pairwise Scatterplots of the Estimated Subject Configuration

Figure 4.1 gives pairwise scatterplots of weights when we consider four dimensions. We can see the importance of each dimension for each group of subjects.

A Private Space for each group can be constructed, by applying (square roots of) the weights to object dimensions. The estimated private spaces are presented in Figure 4.2.

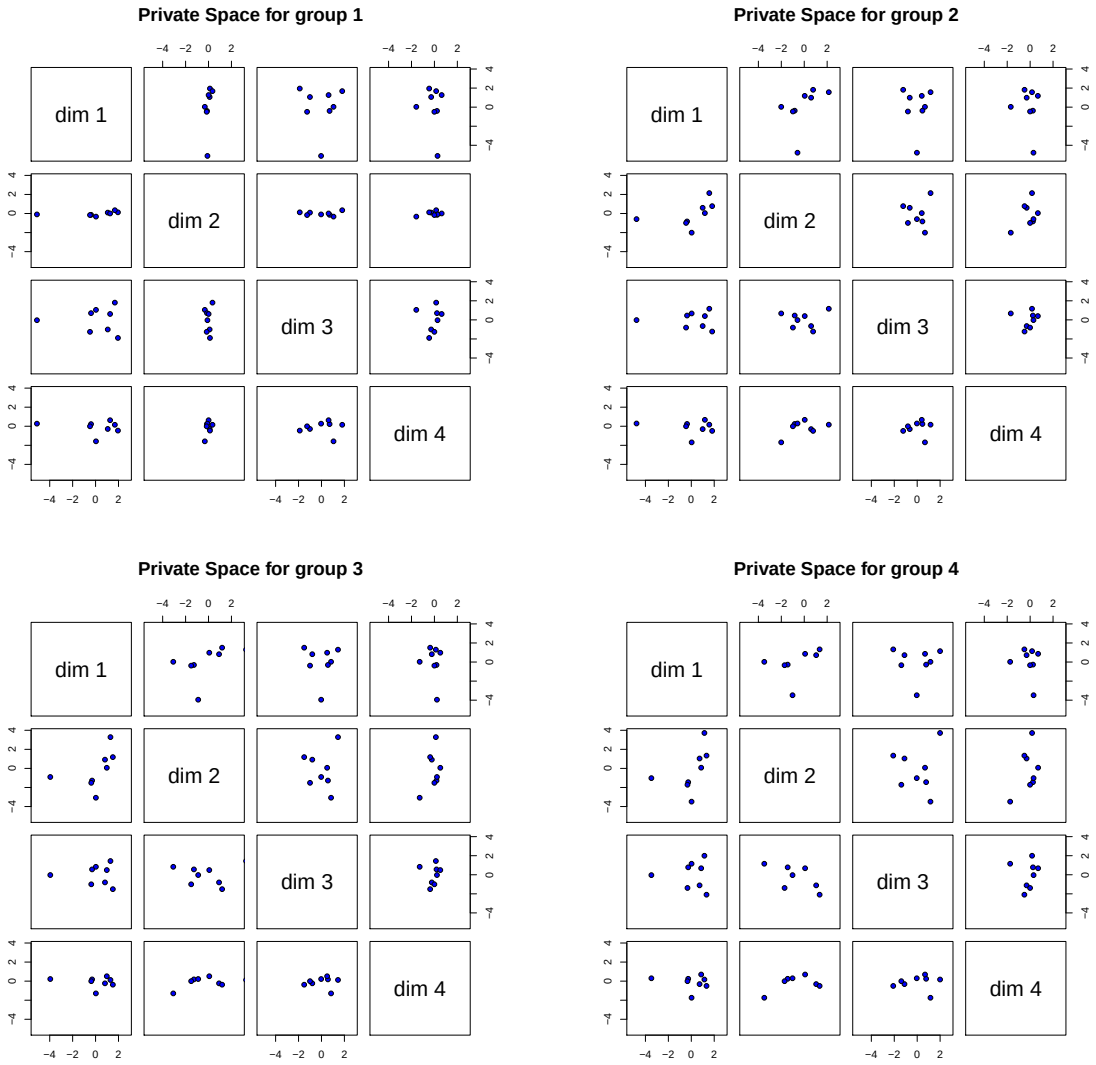


Figure 4.2: Scatterplots of the Estimated Private Spaces for simulation data.

Examination of this figure reveals the differences among the groups. We can note, for example, that group 1, with a smaller weight for Dimension 2 than other groups, has a private space that is compressed along the axis of Dimension 2.

Chapter 5

Cluster Bayesian Multidimensional Scaling

In Chapter 4 we have showed a model in which individual differences are attributable to the difference in importance that groups of individuals attach to the various common scale. We have assumed that each subject s belongs to one and only one group $t = 1, ..., T$ or subpopulation and it is known in advance to which group a particular subject belongs. In this chapter, we assume that each subject belongs to one and only one cluster but it isn't known in advance to which cluster a particular subject belongs. In order to identify such groups, we propose a model that combines a fuzzy form of cluster analysis with Weighted Bayesian Multidimensional Scaling. This model is based on a finite mixture of distributions, in which each mixture component is taken to correspond to a different cluster or subpopulation. The clustering is then done by assigning each observation to the cluster to which it is most likely to belong a posteriori, conditionally on the selected model and its estimated parameters. If the number of clusters, T, is equal to 1, Cluster Bayesian Multidimensional Scaling corresponds to Replicated Bayesian Multidimensional Scaling. If $T = S$, Cluster Bayesian Multidimensional Scaling corresponds to Weighted Bayesian Multidimensional Scaling. For

$1 < T < S$, Cluster Bayesian Multidimensional Scaling estimates the probability that each subject belongs to cluster T and, furthermore, computes a private space MDS solution for each cluster separately. In this model we obtain the cluster from respondents judgments about the dissimilarity of objects. It does not depend on researchers judgments.

We describe the Bayesian framework including prior specifications and an estimation procedure using a Markov Chain Monte Carlo method. Finally, we provide a Bayesian criterion to select the dimension of the configuration and the number of clusters. Since the Cluster Bayesian Multidimensional scaling is a comprehensive model that includes the Replicated BMDS and the Weighted BMDS as special cases, in this way we may compare these three models and study the heterogeneity among subjects.

5.1 Model and Prior

Let the m-component column vector $\boldsymbol{\delta}_t = (\delta_{12,t}, \delta_{13,t}, ..., \delta_{n(n-1),t})'$ contain the $m = n(n-1)/2$ dissimilarity values among n objects for a subject belonging to group t, which are assumed to be functionally related to p unobserved attributes of the objects and p unobserved attributes of the subject. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of p attributes possessed by object i, and $\mathbf{w}_t = (w_{t1}, ..., w_{tp})$ denote an unobserved vector representing the weights associated with the p dimensions for subjects belong to group t. We define $\delta_{ij,t}$ as the Weighted Euclidean distance between the i-th and the j-th objects for the t-th group:

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

The observed dissimilarities data are obtained from a sample of S subjects and they are denoted by $d_{ij,s}$, where $s = 1, ..., S$ indexes subjects, and $i, j = 1, ..., n$ indexes

pairs of objects. Let the m-component column vector $\mathbf{d}_s = (d_{12,s}, d_{13,s}, ..., d_{n(n-1),s})'$ contain the m dissimilarity values for subject s ($s = 1, ..., S$). The total data set will be indicated by the $S \times m$ matrix $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_S)'$. Conditionally on $s \in t$, the distribution of $\mathbf{d}_s$ is assumed to be a multivariate truncated normal

$$\mathbf{d}_{s\in t} \sim N_m(\boldsymbol{\delta}_t, \mathrm{I}\sigma^2) I(\mathbf{d}_s > 0), \;\; s = 1, ..., S, \; t = 1, ..., T, \; T \leq S$$

Because it is not known in advance to which class a subject belongs, the p.d.f. of the random variable $\mathbf{d}_s$ becomes a finite mixture of multivariate truncated normal densities. The mixture distribution can be expressed as

$$\mathbf{d}_s \sim \sum_{t=1}^{T} \tau_t N_m(\boldsymbol{\delta}_t, \mathrm{I}\sigma^2) I(\mathbf{d}_s > 0),$$

where τ_t are the mixture weights s.t. $\tau_t \geq 0 \; \forall t$ and $\sum_{t=1}^{T} \tau_t = 1$. The parameter τ_t is the unknown probability for a subject to belong to group t. We need to specify priors for $\mathbf{X}$, $\mathbf{W}$, σ^2 and $\tau_1, ..., \tau_T$. For the prior distribution of $\tau_1, ..., \tau_T$ we use a conjugate prior $\tau_1, ..., \tau_T \sim Dirichlet(\varrho_1, ..., \varrho_T)$. 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., $\mathbf{x}_i \sim N_p(0, \Lambda)$, independently for $i = 1, 2, ...n$. 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 the prior distribution of $w_{k,t}$, we use a negative exponential distribution, $w_{k,t} \sim Exp(\theta_k)$ independently for $k = 1, ..., p$ and $t = 1, ..., T$. For a hyperprior for the elements of $\Lambda = Diag(\lambda_1, ..., \lambda_p)$, given dimension p, we assume a conjugate prior $\lambda_k \sim IG(\alpha, \beta_k)$, independently for each $k = 1, 2, ..., p$. Finally, we assume prior independence among the prior densities.

5.2 Posterior Inference (MCMC)

In order to avoid the computational problems, we introduce latent variables (labels) $J_1, ..., J_S$ which indicate the group memberships of subjects. We define latent variables J_s such that $P(J_s = t) = \tau_t$ and $\mathbf{d}_s$ belongs to class t if $J_s = t$, so that

$$\mathbf{d}_s | J_s = t \sim N_m(\boldsymbol{\delta}_t, \mathrm{I}\sigma^2) I(\mathbf{d}_s > 0),$$

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

$$\pi(\mathbf{x}_i|\cdot) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{j\neq i, j=1}^{n} \left(d_{ij,s} - \left(\sum_{t=1}^{T} \delta_{ij,t} I(J_s = t) \right) \right)^2 - \frac{1}{2} \mathbf{x}_i' \Lambda^{-1} \mathbf{x}_i - \sum_{t=1}^{T} n(t) \sum_{j\neq i, j=1}^{n} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma} \right) \right],$$

$$\pi(w_{k,t}|\cdot) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(\sum_{i>j} (d_{ij,s} - \delta_{ij,t})^2 I(J_s = t) \right) - n(t) \sum_{i>j} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma} \right) - \theta_k w_{k,t} \right]$$

$$\pi\left(\sigma^2|\cdot\right) \propto (\sigma^2)^{-\left(\frac{Sm}{2}+a+1\right)} \exp\left[-\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{i>j} \left(d_{ij,s} - \left(\sum_{t=1}^{T} \delta_{ij,t} I(J_s = t) \right) \right)^2 - \frac{b}{\sigma^2} - \sum_{t=1}^{T} n(t) \sum_{i>j} \log \Phi\left(\frac{\delta_{ij,t}}{\sigma} \right) \right],$$

$$\pi(\tau_1, ...\tau_T|\cdot) \propto Dir\left(\varrho_1 + n(1), ..., \varrho_T + n(T)\right),$$

$$\pi(\lambda_k|\cdot) \propto IG\left(\alpha + \frac{n}{2}, \beta_k + \frac{\sum_{i=1}^{n} x_{ik}^2}{2} \right),$$

$$\pi\left(J_s = t|\cdot\right) \propto \phi(\mathbf{d}_s; \boldsymbol{\delta}_t, \mathrm{I}\sigma^2) / \sum_{t=1}^{T} \phi(\mathbf{d}_s; \boldsymbol{\delta}_t, \mathrm{I}\sigma^2)$$

where ϕ is the probability density function of the multivariate truncated normal distribution, $n(t) = \sum_{s=1}^{S} I(J_s = t)$ and I is the indicator function.

Simulation of samples from the full conditional distributions of the parameters τ_t and λ_k is straightforward since the full conditional posterior distributions all have convenient forms. However, the full conditional posterior distributions of $\mathbf{x}_i$, $w_{k,t}$ and σ^2 do not have closed forms, and so we apply the Metropolis-Hastings algorithm to generate samples of $\mathbf{x}_i$, $w_{k,t}$ and σ^2. Given the latent indicator variable J_s, we can easily modify the algorithm of Weighted Bayesian Multidimensional Scaling for generating $\mathbf{x}_i$, $w_{k,t}$ and σ^2 from their full conditional posterior distributions.

When Bayesian approach is applied to parameter estimation and clustering analysis for mixture models, the so-called label switch problem might occur. A mixture of density functions is invariant with respect to the permutation of the 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 unreasonable 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. By relabeling, we obtain stable samples of the unknown parameters from which posterior estimates can be computed. Initial values for $\mathbf{X}$, $\mathbf{W}$, σ^2 and Λ, of the posterior distributions can be taken from a WMDS solution and the subjects may be grouped randomly to start. Alternatively, we can group the subjects into clusters on the basis of the derived individual subject weights.

5.3 Choice of dimension and the number of clusters

Until now, we've worked with a given dimension p, of the object configuration, and a given number of clusters, T. These are typically unknown. We develop a way of choosing both the dimension p and the number of clusters T in one step. Since CBMDS is a general model which includes RBMDS and WBMDS as special cases, in this way we can compare these three models. We can investigate if subjects weight differently the dimensions when evaluating the objects and if subjects can be grouped according to these differences.

We base our model selection criteria on $\pi\left(\mathbf{X}, \mathbf{W}, p, T|\mathbf{D}\right)$, the posterior density function of $\mathbf{X}$, $\mathbf{W}$, p, T given data $\mathbf{D}$, evaluated at $\mathbf{X} = \mathbf{X}^{(p)}$ and $\mathbf{W} = \mathbf{W}^{(pT)}$, the best configuration given p and T.

Under equal prior probabilities for all p in $p_{min} \leq p \leq p_{max}$ and for all T in $T_{min} \leq T \leq T_{max}$,

$$\pi\left(\mathbf{X}^{(p)}, \mathbf{W}^{(pT)}, p, T|\mathbf{D}\right) = f(\mathbf{D}|\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)}, p, T)\pi(\mathbf{X}^{(p)}, \mathbf{W}^{(pT)}, p, T)$$

Thus, one needs only to compute the marginal likelihood and the marginal prior of $\mathbf{X}$ and $\mathbf{W}$ for each p and T. We have showed that the marginal prior is

$$\begin{aligned}\int \pi(\Lambda, \mathbf{X}^{(p)}, \mathbf{W}^{(pT)}, p, T)d\Lambda \;\; = \;\; & \prod_{t=1}^{T}\prod_{k=1}^{p}\left(\theta_k \exp\left[-\theta_k w_{k,t}\right]\right)\left(\Gamma\left(\frac{n}{2}+\alpha\right)\right)^p \\ & \times \prod_{k=1}^{p}\left(\frac{\sum_{i=1}^{n} x_{ik}^2}{2}+\beta_k\right)^{-\left(\frac{n}{2}+\alpha\right)}\end{aligned}$$

Now consider computation of the marginal likelihood. $f(\mathbf{D}|\mathbf{X}^{(p)}, \mathbf{W}^{(pT)})$ is not given in closed form and needs to be estimated.

We do this using the Laplace-Metropolis estimator (Raftery, 1996; Lewis and Raftery, 1997). It holds

$$f(\mathbf{D}|\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)}) = \int f(\mathbf{D}|\xi, \mathbf{X}^{(p)}, \mathbf{W}^{(pT)})\pi(\xi|\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)})d\xi$$

where $\xi = (\tau, \sigma^2)$. The Laplace method for integrals is based on a Taylor series expansion of the real-valued function. Thus, applying a Laplace approximation to the integral gives

$$f(\mathbf{D}|\mathbf{X}^{(p)}, \mathbf{W}^{(pT)}) \approx (2\pi)^{d/2}|\Psi|^{1/2}\pi(\mathbf{D}|\tilde{\xi})\pi(\tilde{\xi})$$

where d is the dimension of ξ, $\tilde{\xi}$ is the posterior mode of ξ and Ψ is minus the inverse Hessian of $h(\xi) = \log \pi(\mathbf{D}|\xi)\pi(\xi)$ evaluated at $\xi = \tilde{\xi}$. Thus the Laplace method requires the posterior mode $\tilde{\xi}$ and $|\Psi|$. The idea of the Laplace-Metropolis estimator is to avoid the limitations of the Laplace method by using posterior simulation to estimate the quantities it needs. The simplest way to estimate $\tilde{\xi}$ from posterior simulation output, and probably the most accurate, is to compute $h(\xi)$ for each draw from the posterior simulation output and take the value ξ for which it is largest. The other quantity needed for the Laplace Metropolis estimator is Ψ. This is asymptotically equal to the posterior variance matrix, as sample size tends to infinity, and so it would seem natural to approximate Ψ by the estimated posterior variance matrix from the posterior simulation output.

There is a difficulty in directly comparing $(\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)}, p, T)$ and $(\mathbf{X}^{(p+1\,T)}, \mathbf{W}^{(p+1\,T)}, p+1, T)$. The marginal posterior $\pi(\mathbf{X}, \mathbf{W}|\mathbf{D})$ can be larger 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 configurations in the same dimensional space. Let $\mathbf{X}^{*(p+1,\,T)} = [\mathbf{X}^{(p\,T)} : 0]$ and $\mathbf{W}^{*(p+1\,T)} = [\mathbf{W}^{(p\,T)} : 0]$ in $(p+1)$-dimensional space, which have the first p coordinates and weights equal to $\mathbf{X}^{(p\,T)}$ and $\mathbf{W}^{(p\,T)}$, and the last coordinates and weights all equal to 0. Then $(\mathbf{X}^{*(p+1\,T)}, \mathbf{W}^{*(p+1\,T)})$ provide the same Weight Euclidean distance and the same fit as $(\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)})$ and may be

considered as an embedding of $(\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)})$in $(p+1)$-dimensional space.
We base the choice of p and T on the ratio of the marginal posteriors and we define a selection criterion, which we call MSC, as follows.

$$
\begin{aligned}
MSC_{1T} &= -2\log f(\mathbf{D}|\mathbf{X}^{(1T)}, \mathbf{W}^{(1T)}) - 2\log \pi\big(\mathbf{X}^{(1T)}, \mathbf{W}^{(1T)}\big) \\
MSC_{pT} &= MSC_{1T} + \sum_{q=1}^{p-1} -2\log \frac{\pi\big(\mathbf{X}^{(q+1\,T)}, \mathbf{W}^{(q+1\,T)}|\mathbf{D}\big)}{\pi\big(\mathbf{X}^{*(q+1\,T)}, \mathbf{W}^{*(q+1\,T)}|\mathbf{D}\big)} \\
&= MSC_{1T} + \sum_{q=1}^{p-1} -2\log \Bigg[\frac{f(\mathbf{D}|\mathbf{X}^{(q+1\,T)}, \mathbf{W}^{(q+1\,T)})}{f(\mathbf{D}|\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})} \\
&\qquad \times \frac{\pi(\mathbf{X}^{(q+1\,T)}, \mathbf{W}^{(q+1\,T)})}{\pi(\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})} \frac{\pi(\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})}{\pi(\mathbf{X}^{*(q+1\,T)}, \mathbf{W}^{*(q+1\,T)})} \Bigg] \\
&= MSC_{1T} + \sum_{q=1}^{p-1} -2\log \frac{f(\mathbf{D}|\mathbf{X}^{(q+1\,T)}, \mathbf{W}^{(q+1\,T)})}{f(\mathbf{D}|\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})} \frac{\pi(\mathbf{X}^{(q+1\,T)}, \mathbf{W}^{(q+1\,T)})}{\pi(\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})} - 2\log A_q \\
&= -2\log f(\mathbf{D}|\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)}) - 2\log \pi\big(\mathbf{X}^{(pT)}, \mathbf{W}^{(pT)}\big) - 2\sum_{q=1}^{p-1} \log A_q
\end{aligned}
$$

where $A_q = \frac{\pi(\mathbf{X}^{(q\,T)}, \mathbf{W}^{(q\,T)})}{\pi(\mathbf{X}^{*(q+1\,T)}, \mathbf{W}^{*(q+1\,T)})}$.
$-2\log f(\mathbf{D}|\mathbf{X}^{(p)}, \mathbf{W}^{(pT)})$ can be considered as a measure of fit, $-2\log \pi\big(\mathbf{X}^{(p)}, \mathbf{W}^{(pT)}\big)$ plays the role of a penalty for complexity, and $-2\sum_{q=1}^{p-1} \log A_q$ is a cumulative correction factor for the shrinking effect. The values of p and T that yield the minimum of MSC are viewed as best.

5.4 A simulation study

In this section, we present the results from a simulation to evaluate the performance of the model. The two main purposes of the simulation are: to demonstrate the proposed model ability to uncover the true dissimilarity structure and to understand whether model selection criteria based on derived solutions point to the true structure.
We consider the two data sets introduced in Section 3.4. In the first dataset all subjects

evaluate in the same way the dimensions of the common object configuration while in the second dataset subjects in different groups attach different importance to the various dimensions.

We applied the Cluster Bayesian MDS for various values of p and T to each generated dataset. Table 5.1 presents values of STRESS and Table 5.2 presents values of the Bayesian selection criteria to choose the number of dimensions and number od clusters in the first case.

Table 5.1: CBMDS Stress values for Simulation 1

T	1	2	3	4	5	60
p=1	0.3799	0.4130	0.4137	0.4145	0.3796	0.3818
p=2	0.1598	0.1597	0.1597	0.1596	0.1596	0.1595
p=3	0.1210	0.1209	0.1206	0.1205	0.1198	0.1192
p=4	0.1079	0.1075	0.1071	0.1071	0.1070	0.1045
p=5	0.1079	0.1076	0.1070	0.1071	0.1063	0.1046

Table 5.2: Bayesian selection criteria for Simulation 1

T	1	2	3	4	5	60
p=1	4827.1207	5115.5893	5135.9477	5138.1112	4849.1241	5145.385
p=2	1904.4904	1914.4863	1920.8581	1939.5998	1954.3293	3000.83
p=3	967.4546	984.1919	986.8433	1005.4640	1035.4849	2627.916
p=4	594.5149*	610.7422	619.5565	653.6447	658.0260	2518.435
p=5	595.0351	615.7886	625.7075	641.0854	679.4327	3113.639

*Minimum Bayesian selection criterion

The Bayesian selection criteria chooses the correct dimension $p = 4$ and correctly select $T = 1$. So in this case as we expected the best model is the aggregate model.

Table 5.3 and Table 5.4 show the results obtained from the second dataset.

Table 5.3: CBMDS Stress values for Simulation 2

T	1	2	3	4	5	60
p=1	0.3222	0.3219	0.3189	0.3187	0.3199	0.3189
p=2	0.2354	0.1791	0.1696	0.1650	0.1651	0.1655
p=3	0.2026	0.1192	0.1063	0.0797	0.0796	0.0794
p=4	0.2003	0.1155	0.1023	0.0724	0.0728	0.0720
p=5	0.2004	0.1155	0.1021	0.0725	0.0727	0.0718

Table 5.4: Bayesian selection criteria for Simulation 2

T	1	2	3	4	5	60
p=1	5733.755	5733.023	5722.226	5696.6270	5736.8639	5965.601
p=2	4683.540	3812.208	3778.072	3657.1252	3925.8788	4675.068
p=3	4143.780	2453.116	1791.680	1241.1810	2144.3472	2890.701
p=4	4108.505	2347.539	1604.685	935.4417*	941.5958	2761.145
p=5	4110.127	2348.891	1997.857	970.6540	2056.2413	3337.408

*Minimum Bayesian selection criterion

The Bayesian criterion (MSC) correctly identifies the true dimension $p = 4$ and the correct number of clusters $T = 4$ and the posterior probabilities classify each subject correctly for each dimension. Figure 5.1 shows plots of MSC for various p. The choice of T seems to depend not on p but on the clustering structure of the data.

To evaluate 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 $\sigma^2 = 0.6,\ 0.9,\ 1.5,\ 3.0$. The optimal values of (p, T) for each σ^2 are always $(4, 4)$.

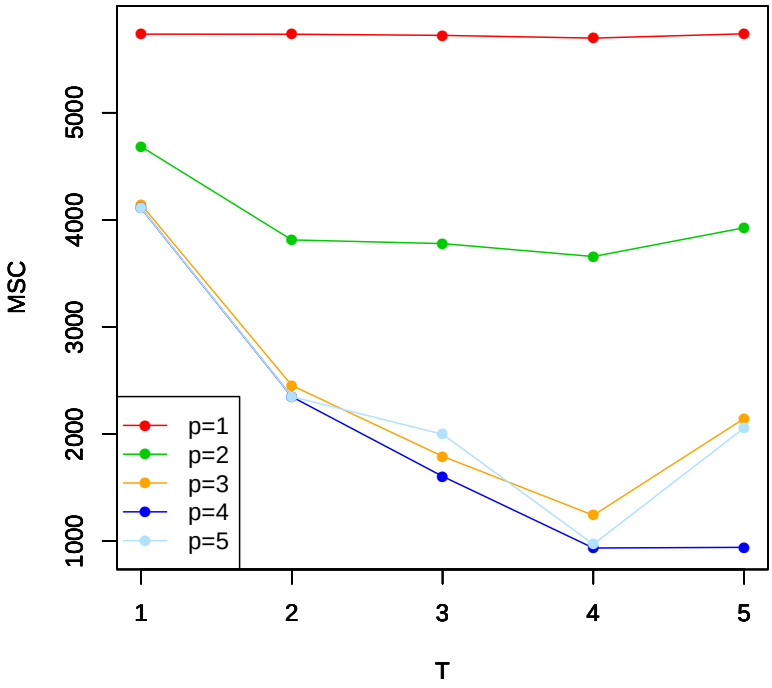


Figure 5.1: Plots of MSC in Simulation 2. MSC chooses the correct number of clusters in each dimension.

Chapter 6

Empirical application

In this chapter, two applications of our procedures are discussed. At first, data on consumer perceptions of various automobiles are studied. Secondly, patent applications data are investigated to study the relationship among technological fields.
We applied our algorithms, with 10000 MCMC iterations, the first 3000 of which were discarded for burn-in. The Bayesian selection criterion was used to select the best configuration.

6.1 The Ford Ka Case

We consider data on $n = 10$ small cars. The survey was conducted in 1995 by Ford France on a sample of $S = 188$ potential small car buyers.
Ford's decision to develop the Ka was response to the changes in the small car market. Ford had long been present in the market with the Fiesta. In 1992, however, Ford had to re-evaluate its strategy because of the dramatic success of the newly launched Renault Twingo. Within 12 months, the Twingo had gained market share of 8.9 and become the third best-selling model in the small car market in France. The Twingo changed the characteristics of the category, forcing other manufactures to offer more than basic transportation to remain competitive. The Twingo gave customers a very

distinct, original style, flashy colors, greater functionality and more interior space than cars of similar size. The Twingo was the pioneer that confirmed a demand for an urban, stylish car. Ford management felt that the Fiesta brand line did not have the appeal to compete with the Renault Twingo and decided to lunch a new small car, the Ford Ka.

In the market research to identify target customers for the Ka, personal interviews of potential customer were used to better understand the competitors in the small car market and customer perceptions of the Ka relative to its potential competition. Marketing researcher conducted a series of studies among potential small car buyers. In this occasion the names of two automobiles were presented each time to the subjects, who were asked to indicate the degree of dissimilarity between them on a nine point scale, with one being Very similar and nine being Very different. The ten automobiles are:

1. Opel Tigra
2. Ford Ka
3. Toyota Rav4
4. Renault Twingo
5. Opel corsa
6. Ford Fiesta
7. Volkswagen Polo
8. Peugeot 106
9. Nissan Micra
10. Fiat500

The 45 pairs of cars were presented to 188 subjects. Thus our dataset consists of $S = 188$ dissimilarity matrices between n=10 cars.

Marketing researchers also asked potential target consumers to choose their three most preferred and three least preferred cars from the list of 10 small cars. We call "Choosers" respondents who put the Ka among their top three choices and "Non-

choosers" respondents who put the Ka among their bottom three choices.
Our main purposes in this particular application are to understand the customer perceptions of the Ford Ka relative to its potential competition and to investigate the existence of pereptual heterogeneity among customers using the proposed methodology.
At first, we applied Classical MDS on the average occasion, Replicated Bayesian MDS and Weighted Bayesian MDS. The Table 6.1 shows that the Weighted Bayesian MDS is characterized by lower STRESS as compared to the others models.

Table 6.1: Stress values for the Ford Ka Data

	CMDS	RBMDS	WBMDS	GBMDS
dim (p)	STRESS	STRESS	STRESS	STRESS
1	0.3399	0.3091	0.3066	0.3076
2	0.2380	0.2333	0.2264	0.2286
3	0.2305	0.2294	0.2031	0.2090
4	0.2299	0.2284	0.1973	0.2035
5	0.2308	0.2285	0.2001	0.2058

For each Bayesian model we computed the Bayesian selection criteria for various values of p and in all cases the Bayesian selection criteria identified the dimension p=4. Figure 6.1 displays the common objects configuration obtained from Weighted Bayesian MDS with dimension p=4. In this figure, we can see the position of the Ford Ka respect with the others small car. In the first two dimensions we can see a group made of Ford Ka, Renault Twingo and Opel Tigra, in the third dimension the Ford Ka is isolated and in the last dimension is near all the others automobiles except Opel Tigra and Toyota Rav4.

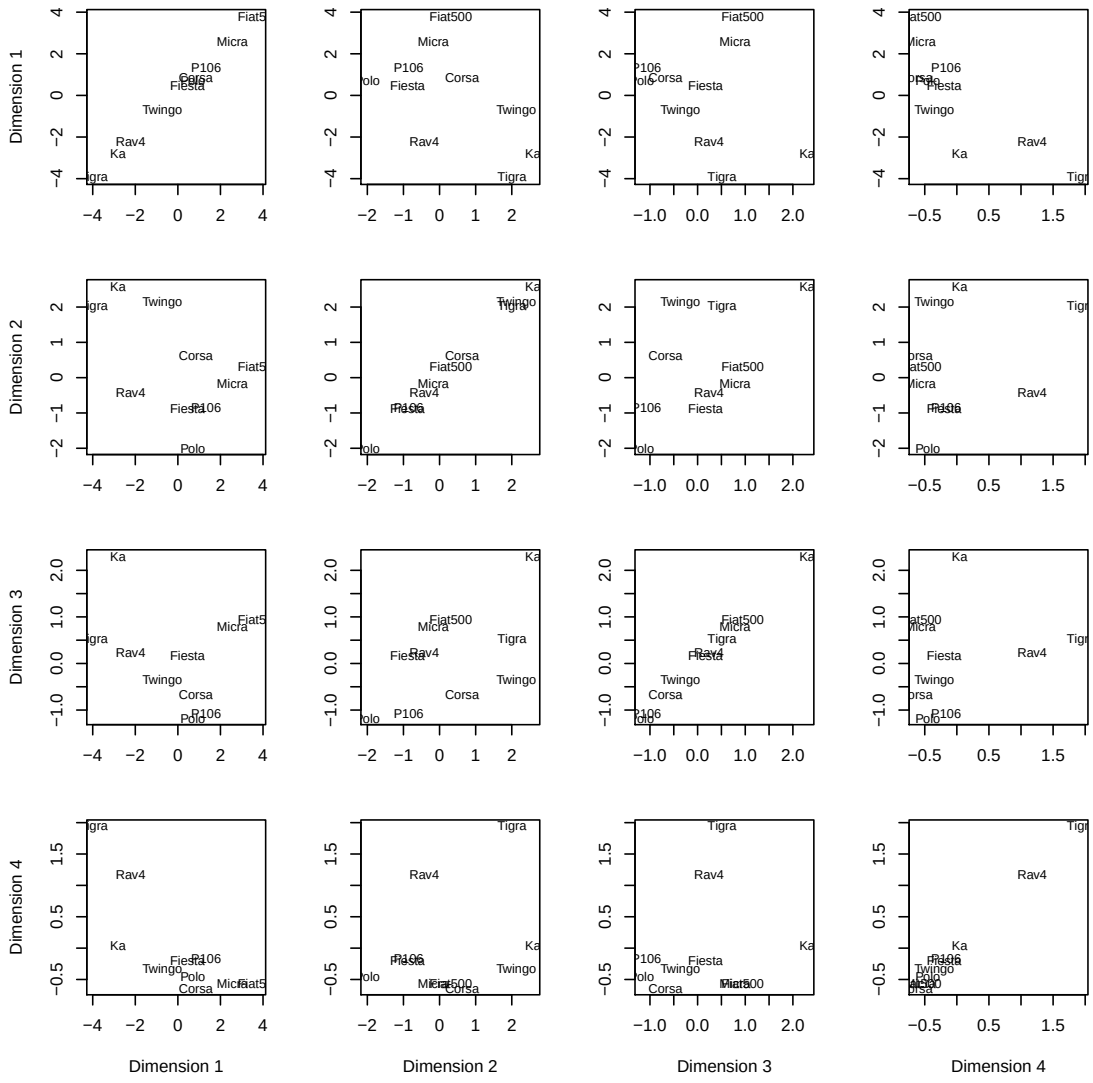


Figure 6.1: Estimated Object Configuration from WBMDS for the Ford Ka Data

Figure 6.2 displays the subject space obtained from Weighted Bayesian MDS with four dimensions. In this figure, we can see the estimated weights for each subject on each dimension of the common object configuration and different groups of subjects can be detected. Since we can clearly detect the two subjects groups, "Choosers" and "Non-choosers" we applied Group Bayesian MDS for $T = 2$ groups which correspond with "Choosers" and "Non-choosers". Results in Table 6.1 show that true to form GBMDS and WBMDS are more or less characterized by the same goodness of fit. The

bayesian criteria (MSG) selects a four dimensional representation.

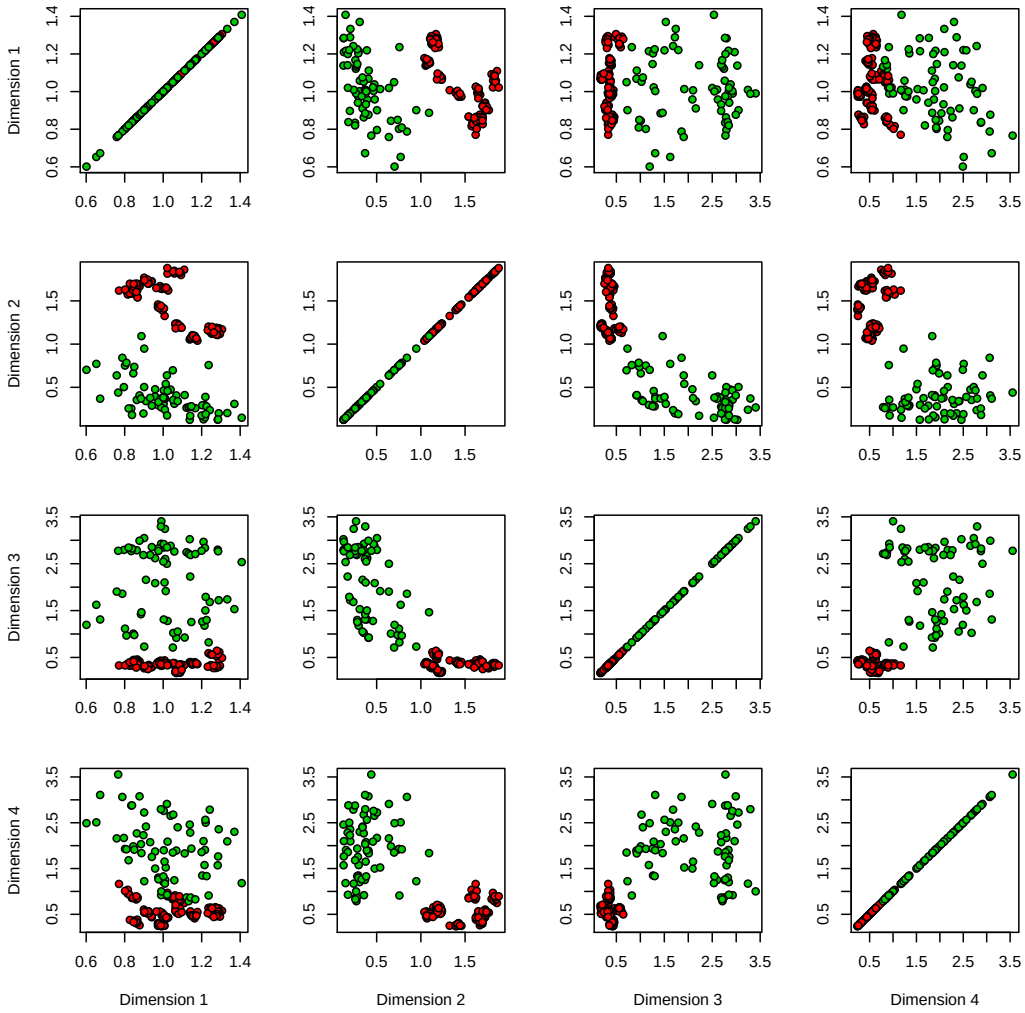
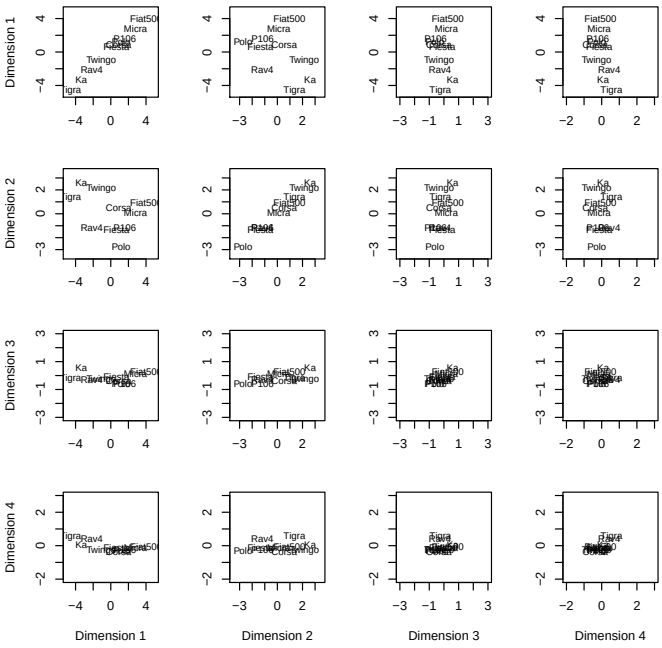
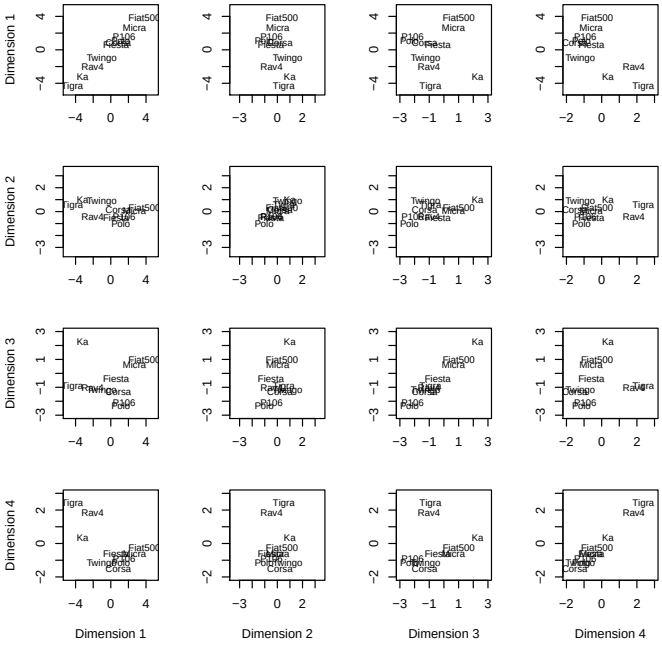


Figure 6.2: Estimated Subject Configuration from WBMDS for the Ford Ka Data. Red dots represent "Choosers", green dots "Non-choosers"

From Groups Bayesian MDS with $T = 2$ we obtained the common object configuration, the subject space and two private spaces that map the automobiles under consideration for "Choosers" and "Non-choosers". Private space is obtained for any group by applying the square root of his weights on dimensions to the associated dimensions in the Group objects space. As expected, there are differences in the perceptions between "Choosers" and "Non-choosers". The private spaces obtained for



(a) Choosers



(b) Non-choosers

Figure 6.3: (a) Private Space for "Choosers" (b) Private Space for "Non-Choosers"

"Choosers" and "Non-choosers" with the Groups Bayesian MDS model are displayed in Figure 6.3. Upon examining Figure 6.3 it is noted that the first dimension have the same importance for both groups. Indeed, the two groups are distinguished by the weights on the others three dimensions. The second dimension is important only for "Choosers". In this dimension the Ford Ka is near Renault Twingo and Opel Tigra. The Dimension 3 in which Ford Ka is separated from the rest of the automobiles and the Dimension 4 are important only for "Non-choosers". The subjects who put the Ka among their top three choices consider Ford Ka, Renault Twingo and Opel Tigra highly similar while the subjects who put the Ka among their bottom three choices consider Ford Ka is most closely associated with Fiat 500 and Micra. In order to provide a clear answer about the target customers a demographic description of "Choosers" and "Non-Choosers" should be necessary.

Lastly, we applied the Cluster Bayesian MDS for various values of dimension p and number of clusters T in order to see whether potential consumers in the small car market can be grouped into clusters and to choose the best model. The MSC selects a four dimensional configuration and it yields four as the appropriate number of clusters, which means that the subjects are not homogeneous in their perceptions. Figure 6.4 shows plots of MSC for various p. The choice of T seems to depend not on p but on the clustering structure of the data. The first cluster is identical to the group "Choosers" while the subjects belong to group "Non-choosers" are divided among the others 3 clusters.

6.2 Technological diversification

In this section we analyze the relations between fields of technology. We intend to examine whether firms extend their innovative activities across related technological fields. We intend to examine whether knowledge relatedness remains stable or relations between fields of technology change over time, and whether knowledge relatedness is

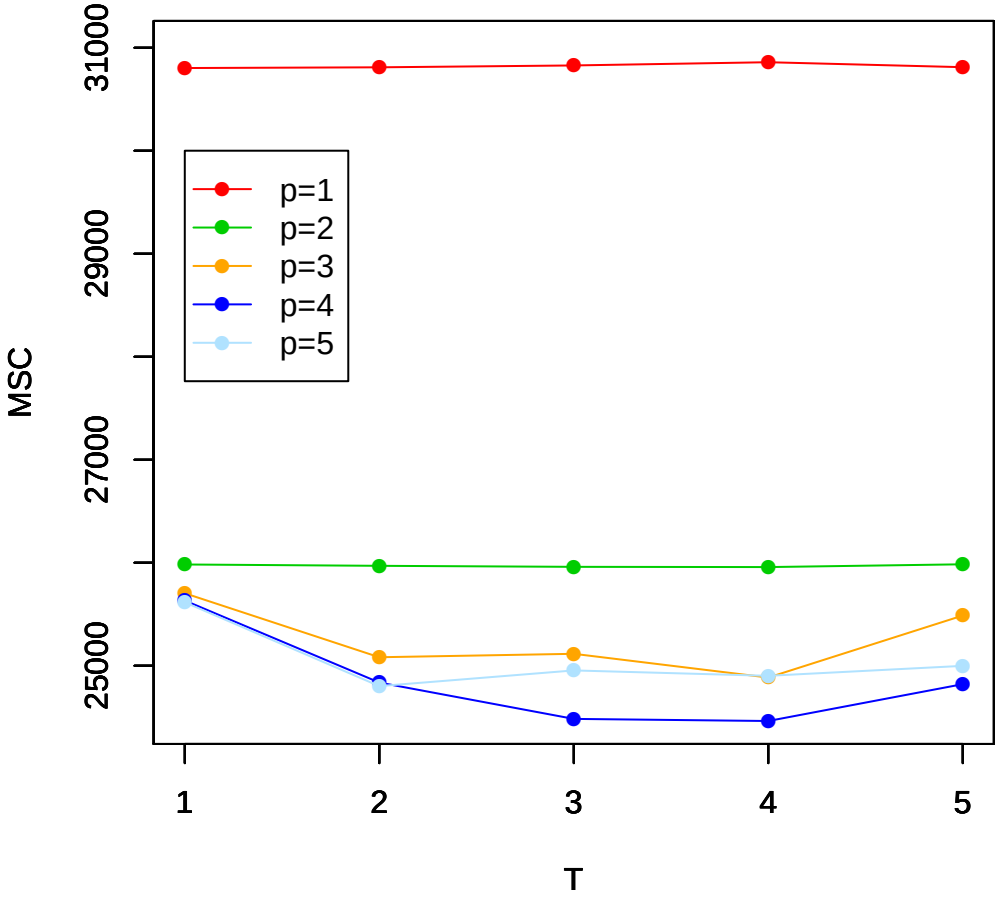


Figure 6.4: Plot of MSC. MSC chooses $p = 4$ and $T = 4$.

affected by differences among countries or whether it is relatively invariant across countries.

We used patent applications to measure the dissimilarity between different technological fields. We considered the PATSTAT database (i.e. EPO Worldwide PATent STATistical Database). It provides raw patent data coming from around 90 patent offices worldwide, including of course the most important and largest ones such as the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). The dataset includes a set of variables concerning each patent application as for example the country of residence of inventor for the technologies for which patent protection is sought, the year and exact date of application of the patent and the technological class to which the patent belongs.

The International Patent Classification (IPC) is a system in which the whole area of technology is divided in 30 different fields of technology. This classification has been elaborated jointly by Fraunhofer Gesellschaft-ISI (Karlsruhe), Institut National de la Proprit Industrielle (INPI, Paris) and Observatoire des Sciences and des Techniques (OST, Paris) and it is reported in the Table 6.2.

All patent documents are classified at least by one (main or primary) classification code of the International Patent Classification (IPC) during their examination by the patent office based on the technological nature of the underlying invention, but usually more classification codes (secondary or supplementary) are assigned to the documents. When different codes of the International Patent Classification (IPC) exist in a single patent document we call this a case of co-occurrence. We measured the dissimilarity between fields of technology analyzing the co-occurrence of classification codes assigned to individual patent documents. The assumption which is made is that the frequency by which two classification codes are jointly assigned to the same patent document measures the strength of the similarity, in terms of knowledge links and spillovers, between the technological fields which the codes refer to. That is, it represents an inverse measure of the technological distance between the two fields.

Table 6.2: Technology classification (30 fields) based on the IPC

OST30-code	OST30-name
1	Electrical engineering
2	Audiovisual technology
3	Telecommunications
4	Information technology
5	Semiconductors
6	Optics
7	Technologies for Control/Measures/Analysis
8	Medical engineering
9	Nuclear technology
10	Organic chemistry
11	Macromolecular chemistry
12	Basic chemistry
13	Surface technology
14	Materials; Metallurgy
15	Biotechnologies
16	Pharmaceuticals; Cosmetics
17	Agricultural and food products
18	Technical processes (chemical, physical, mechanical)
19	Handling; Printing
20	Materials processing, textile, glass, paper
21	Environmental technologies
22	Agricultural and food apparatuses
23	Machine tools
24	Engines; Pumps; Turbines
25	Thermal processes
26	Mechanical elements
27	Transport technology
28	Space technology; Weapons
29	Consumer goods
30	Civil engineering

The procedure we followed to build a measure of similarity was originally proposed by Engelsman and van Raan (1992).

Formally, let us indicate with M the universe being studied, e.g. all patent applications to the EPO in a certain period of time. Each of the M patent applications has been assigned by patent examiners to one or more classification codes. Let $F_{im} = 1$ if patent document m contains the classification code i ($i = 1, ..., 30$) and 0 otherwise. The number of patents assigned to technological field i is thus $N_i = \sum_m F_{im}$. We can, therefore, indicate the number of observed joint occurrences of classification codes i and j as: $C_{ij} = \sum_m F_{im}F_{jm}$. By applying the latter to all possible pairs of classification codes, we built a square (30×30) symmetrical matrix of co-occurrences ($\mathbf{C}$), whose generic cell C_{ij} represents the number of patent documents classified in both technological fields i and j, which represents the fundamental input to derive a measure of similarity between technological fields.

We used the the cosine index S_{ij}, which measures the angular separation between the vectors representing the co-occurrences of technological fields i and j with all the other fields:

$$S_{ij} = \frac{\sum_{k=1}^{30} C_{ik}C_{jk}}{\sqrt{\sum_{k=1}^{30} C_{ik}^2}\sqrt{\sum_{k=1}^{30} C_{jk}^2}}$$

The cosine index is a measure that reflects the similarity between two technological fields in terms of their mutual relationships with all the other fields. S_{ij} is the greater the more the two fields i and j co-occur with the same technological fields. It is equal to one for pairs of technological fields with identical distribution of co-occurrences with all the other technological fields, while it delivers a value close to zero for pairs of technological fields that do not overlap.

Applying the cosine index calculation to a 30×30 matrix of the joint occurrences, we generated a new matrix of the same size, whose elements are the various S_{ij} ($i = 1, ..., 30$; $j = 1, ..., 30$), which describe the level of similarity between technological fields. Dissimilarity measures were obtained from the similarities simply considering

the dissimilarity index $d_{ij} = 1 - S_{ij}$.

We extracted all the patents filed from 1978 to 2003 by inventors from five countries: United States, Germany, Japan, Great Britain and Italy. The three main inventors are Japan, United States and Germany. United States account for over 30% of the world's patents filed from 1978 to 2003. Japan and Germany account for about 19%. Great Britain accounts for about 5% and Italy accounts for about 3%.

Since it was intended to verify stability of knowledge relatedness over time and across countries we calculated 15 dissimilarity matrices based on patent applications filed during the periods 1978-1987, 1988-1997 and 1998-2007 by inventors from the five considered countries.

We applied the Cluster Bayesian MDS to these 15 dissimilarity matrices for an increasing number of dimensions (p) and an increasing number of clusters (T). Then we applied the Bayesian selection criterion (MSC) to select the best fitting model. This is one of the main advantages of our approach over the others approaches where no criterion is provided to select the model. Results of model selection criteria suggest that the best model for this data is the aggregate model with $p = 30$, which is the true dimension. Since MSC leads us to select $T = 1$ for each dimension we can conclude that there is not difference among countries and the relationships among technological fields remain stable over time. To prove these conclusions further in Figure 6.5 we present the results with $T = 15$ clusters in two dimensions for visualization purposes. This case corresponds to the Weighted Bayesian Multidimensional Scaling. The left panel of this figure displays the common object configuration which visualizes relations between fields of technology. The right panel displays the subject space which visualizes the weights and reveals that the sample is rather homogeneous.

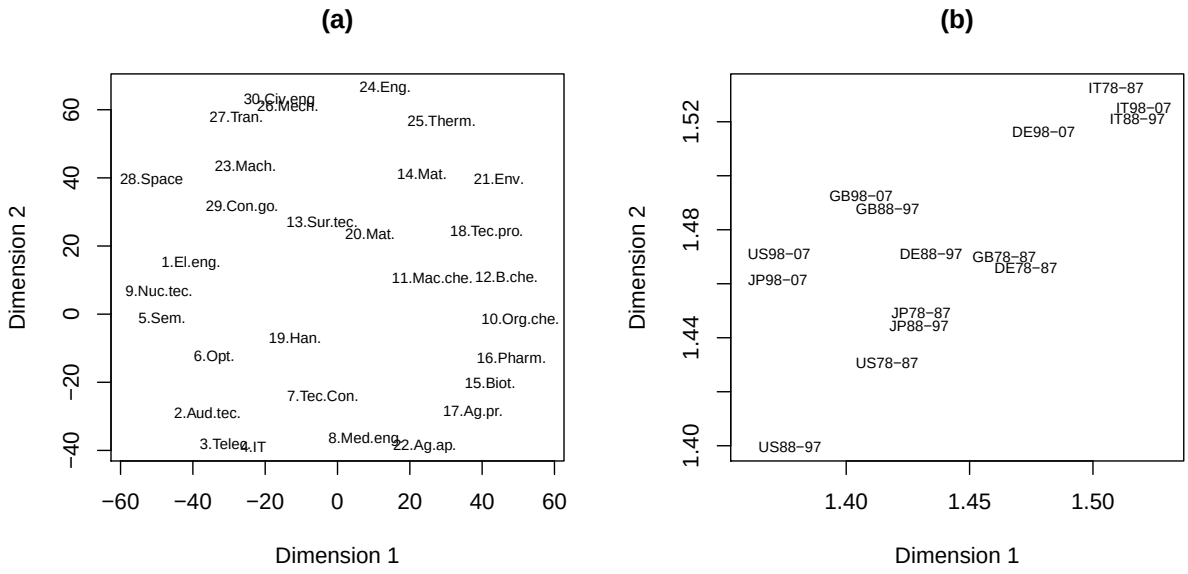


Figure 6.5: (a) Estimated Object Configuration and (b) Estimated Subject Configuration from CBMDS with $p = 2$ and $T = 15$.

Conclusion

The work in this thesis dealt with problems in three-way multidimensional scaling that is the situation where the data consist of dissimilarity measures between pairs of objects for more than one subject. We have proposed four models with different levels of aggregation. First, we have proposed a Replicated Bayesian MDS in which individual differences are attributable only to differences in measurement errors in observations, and a Weighted Bayesian MDS, in which difference between the subjects are attributable to the differences in the weights ascribed to the dimensions of a common configuration. These methods fit much better than Classical MDS and SMACOF. An important issue in MDS methods is dimensionality so we have proposed a Bayesian dimension choice criterion for each model.

The weights are rarely interpreted for individual subjects, and the improvement in goodness-of-fit measures seldom seems to justify so many additional parameters. Then, we have proposed two models in which individual differences are attributable to the difference in importance that groups of individuals attach to the various common scale. We have introduced a model with known groups, a Groups Bayesian MDS, and a model with unknown groups, a Cluster Bayesian MDS based on a finite mixture of distributions, in which each mixture component is taken to correspond to a different cluster. This last model is a comprehensive model that includes the Replicated Bayesian MDS and the Weighted Bayesian MDS. We have proposed a Bayesian criterion, MSC, for simultaneously selecting the object dimension and the number of clusters of subjects. In this way, we are able to compare the cluster based model (CBMDS), the aggregate

model (RBMDS), and the Weighted Bayesian MDS. The proposed Bayesian selection criteria performs well in our simulations and in real examples. The results demonstrate the proposed models ability to uncover the true dissimilarity structure and the Bayesian selection criteria ability to choose the true configuration.

References

Arabie, P., Carroll, J. D., and DeSarbo, W. S. (1987). *Three-Way Scaling and Clustering*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-065, Beverly Hills and London: Sage Publications.

Arce, C., and Garling, T. (1989). Multidimensional scaling, *Anuario de Psicologia*, **43**, 63-80.

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.

Bergmann Tiest, W. M., and Kappers, A. M. L. (2005). *Analysis of haptic perception of materials by multidimensional scaling and physical measurements of roughness and compressibility*, Preprint submitted to Acta Psychologica.

Bocci, L., and Vichi, M. (2011). The K-INDSCAL Model for Heterogeneous Three-Way Dissimilarity Data, *Psychometrika*, **76**, 691-714.

Borg, I., and Groenen, P. (2005). *Modern Multidimensional Scaling*, New York: Springer-Verlag.

Breschi, S., Lissoni F., and Malerba, F. (2003). Knowledge-relatedness in firm technological diversification, *Research Policy*, **32**, 69-87.

Carroll, J. D., and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, **35**, 283-319.

Carroll, J. D., and Green, P. E. (1997). Psychometric Methods in Marketing Research: Part II, Multidimensional Scaling, *Journal of Marketing Research*, **34**, 193-204.

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

Claxkson, D. B., and Gonzalez, R. (2001). Random effects diagonal metric multidimensional scaling models, *Psychometrika*, **66**, 25-43.

Cothier, G., Christen, M., and Soberman, D. (2001). Ford Ka (A) Breaking New Ground in the Small Car Market, *INSEAD*.

Cothier, G., Christen, M., and Soberman, D. (2003). Ford Ka: The market research problem (A), *INSEAD*.

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.

De Leeuw, J. (1977). Applications of Convex Analysis to Multidimensional Scaling, in JR Barra, F Brodeau, G Romier, B van Cutsem (eds.), *Recent Developments in Statistics*, 133-145. North Holland Publishing Company, Amsterdam.

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

De Leeuw, J. and Mair, P. (2009), Multidimensional Scaling Using SMACOF, *Journal of Statistical Software*, **31**, 1-30.

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

DeSarbo, S. W., Manrai A. K. and Manrai, L. A. (1994). Latent Class Multidimensional Scaling: A Review of Recent Developments in the Marketing and Psychometric Literature, *Advanced Methods of Marketing Research*, ed. R. P. Bagozzi, Oxford: Blackwell.

Engelsman, E. C., and van Raan, A. F. J., (1992). A patent-based cartography of technology. *Research Policy*, **23**, 1-26.

Fraley, C., and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation, *Journal of the American Statistical Association*, **97**, 611-631.

Green, P. E., and Rao, V. R. (1971). Multidimensional Scaling and Individual Differences, *Journal of Marketing Research*, **8**, 71-7.

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.

Groenen, P. J. F., and van de Velden, M. (2004). Multidimensional Scaling. *Econometric Institute Report*, EI 2004-15.

Izenman, A. (2008). *Multidimensional Scaling and Distance Geometry*, Modern Multivariate Statistical Techniques. Springer Texts in Statistics, Springer New York.

Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-011, Beverly Hills and London: Sage Publications.

Joo, S. H., and Kim, Y. (2010). Measuring relatedness between technological fields, *Scientometrics*, **83**, 435-454.

Lewis, S. M., and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, **92**, 648-655.

Mardia, K. V., Kent, J. T., and Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

Mead, A. (1992). Review of the Development of Multidimensional Scaling Methods, *Journal of the Royal Statistical Society*, Ser. D **41**, 27-39.

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. E. (2001). Bayesian Multidimensional Scaling and Choice of Dimension, *Journal of the American Statistical Association*, **96**, 1031-1044.

Oh, M.-S. and Raftery, A. E. (2007). Model-Based Clustering With Dissimilarities: A Bayesian Approach, *Journal of Computational and Graphical Statistics*, **16**, 559-585.

Park, J., DeSarbo, W. S., and Liechty, J. (2008). A Hierarchical Bayesian Multidimensional Scaling Methodology for Accommodating Both Structural and Preference Heterogeneity. *Psychometrika*, **73**, 451-472.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research, *Sociological Methodology*, **25**, 111-163.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, 163188. London: Chapman & Hall.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling, *Psychometrika*, **42**, 337-360.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society*, Ser. A **145**, 285-312.

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

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

Steyvers, M. (2002). Multidimensional Scaling, *Encyclopedia of Cognitive Science*. Nature Publishing Group, London, UK.

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

Takane, Y. (2007). Applications of multidimensional scaling in psychometrics. In Rao, C. R., and Sinharay, S. (Eds.), *Handbook of Statistics*, **26**, Pyschometrics, 359-400. Amsterdam: Elsevier.

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.

Tucker L. R., and Messick S., (1963). An indivilual differences model for multidimensional scaling, *Psychometrika*, **28**, 333-367.

Winsberg, S., and De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model. CLASCAL, *Psychometrika*, **58**, 315-330.

Young, G., and Householder, A. S. (1938). Discussion of a set of point in terms of their mutual distances. *Psychometrika*, **3**, 19-22.

Zhang, Z., and Takane, Y. (2010). Statistics: Multidimensional scaling. In E. Baker, B. McGaw, and P. P. Peterson (Eds.), *International Encyclopedia of Education* (3rd Edition), 304-311. Oxford: Elsevier.