UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Departamento de Estadística e Investigación Operativa

TESIS DOCTORAL

Modelos de Clasificación y Multidimensional Scaling y su tratamiento computacional

Rodrigo Macías Páez

Granada, España

Editor: Editorial de la Universidad de Granada
Autor: Rodrigo Macías Páez
D.L.: GR. 1731-2009
ISBN: 978-84-692-1315-5

Modelos de Clasificación y Multidimensional Scaling y su tratamiento computacional

Memoria presentada para optar al grado de Doctor en Estadística e Investigación Operativa, por Rodrigo Macías Páez.

Vº Bº
Director de Tesis:

Prof. Dr. José Fernando Vera Vera

Departamento de Estadística e Investigación Operativa

Facultad de Ciencias

Universidad de Granada
España

A Rodriguín, Luz Melania y Cony

A mamá

Agradecimientos

Quiero agradecer especialmente al Dr. José Fernando Vera por transmitirme su entusiasmo y dedicación por la investigación. Sus consejos y su guía durante todo este tiempo fueron fundamentales para llevar a buen término este trabajo. A Willem por sus valiosos comentarios y sugerencias en el desarrollo de este trabajo.

A Erika, Christian y Suyim por su entrañable amistad y por esas grandiosas veladas. Al Departamento de Estadística e Investigación Operativa de la Universidad de Granada por las facilidades otorgadas durante este tiempo. A Jacqueline Vreriks y Matthijs Warrens de la Universidad de Leiden por sus atenciones y apoyo durante mi estancia.

A CONACYT y a la Fundación Carolina por permitirme la oportunidad de realizar este proyecto.

Índice general

Capítulo 1

Introduction

Multidimensional scaling (MDS) is a statistical technique that has its origins at the end of XIX century, on the first models developed by psychology. Such models tried to study a relation between the physical intensity of the stimuli and the subjective sensation that these exert on the subjects. As input data MDS uses proximities between pairs of objects and the primary target is to represent these proximities by means distances between points in a low dimensional space.

Cluster analysis (CA) is a generic term used to define a wide range of multivariate methods, whose objective is to reveal or discover groups of homogenous observations. Cluster techniques have been used in diverse disciplines, e.g. psychiatry, where it has been used to redefine categories of existing diagnostic, archaeology, where it has been used to investigate the relation between several types of devices, and market research, producing consumer groups with diverse purchase pattern, amoung others.

Although MDS and CA have different objectives, joint application of CA and MDS methods facilitates the description of dataset in diverse situations. Originally, this application was achieved independent way as it is pronounced in Heiser and Groenen (1997). Nevertheless, interaction of both techniques could help understanding structure dataset better that if these was used of independent way, in this sense Bock (1986, 1987 and 1997) established the first models that consider an interaction of both methodologies. In particular, the simultaneous application of the techniques established in Heiser and Groenen (1997), by partitioning a set of N objects in classes or clusters according to dissimilarities among them and by constructing simultaneously a configuration of K points in a low dimensional space. Thus, points representing the

centers of K clusters, contribute significantly to improve the interpretation of large set of objects, by means a significantly smaller set points, each one representing a subgroup of objects with similar characteristics established only by dissimilarities matrix.

In this chapter, a general description of both techniques is presented as introduction and then a models revision of CA and MDS combined developed until now is provided. Next, both existing problems for its application and situations nonconsidered in its development that could allow a better performance, are discussed. Finally as central objective of present research work, the procedures proposed according to problematic exposed is briefly established, giving rise to presentation of the chapters.

1.1. Multidimensional Scaling

In order to introduce the MDS concept, we suppose that one is interested in the study of N cities with respect to its geographic position. In order to realise this we have a square matrix $(N \times N)$ consisting of distances between each pair of cities and it is tried to reconstruct a map from this matrix. This problem, in arbitrary dimension, has exact solution based on the works of Schoenberg (1935) and Young & Houselholder (1938), which are summarized in the following result and whose demonstration can be seen, for example, in Mardia (1980).

Theorem 1.1.1 (CLASSICAL MDS) *Let $D_{(N\times N)}$ be a distance matrix interpoints n points in a configuración space of K dimension and define $B_{(N\times N)}$ by B=HAH, being $H_{(N\times N)}$ given for $H = I - N^{-1}\mathbf{1}\mathbf{1}^t$ and $A_{(N\times N)} = (a_{rs})$ where $a_{rs} = -\frac{1}{2}d_{rs}^2$. Then, D is Euclidean if and only if B is p.s.d. In particular, the following results hold:*

1. *If $D_{(N\times N)}$ is the matrix of Euclidean interpoint distances for a configuration $Z_{(N\times K)} = (z_1, ..., z_N)'$, then $B = (HZ)(HZ)^t$, that is,*

$$b_{rs} = (z_r - \overline{z})^t(z_s - \overline{z}), \quad \forall r, s = 1, ..., N,$$

   so $B \geq 0$. Note that $\boldsymbol{B}$ can be interpreted as the centred inner product matrix for the configuration $\boldsymbol{Z}$.

2. *Conversely, if B is p.s.d and rank K then a configuration corresponding to $\boldsymbol{B}$ can be constructed as follows. Let $\lambda_1 > \cdots > \lambda_K$ denote the*

positive eigenvalues of $\boldsymbol{B}$ *with corresponding eigenvectors* $X_{(N\times K)} = (x_{(1)}, ..., x_{(K)})$ *normalized by*

$$x'_{(i)}x_{(i)} = \lambda_i, \ \forall i = 1, ..., K$$

.

Then the points P_r *in* $\mathbb{R}^K$ *with coordinates* $x_r = (x_{r1}, ..., x_{rK})'$ *(so* x_r *is the rth row of* X*) have interpoints distances given by* D*. Further, this configuration has centre of gravity* $\overline{x} = 0$ *and* B *represents the inner product matrix for this configuration.*

The previous result shows that there exist a unique solution for Euclidean distances in a space of dimension $K = rank(B)$, except for isometry, where K at most will be $(N-1)$, since $\mathbf{1}$ is a eigenvector of B associated to eigenvalue 0. The statistical problem that solves MDS arises when we want to obtain a representation in a dimension space smaller than K and/or available information between each pair of elements for represent is not given in terms of Euclidean distances but in terms of pseudo-distances or generally of *proximities* with respect to some criterion by means *dissimilarity* or *similarity* coefficients. In the formulated example, such situation could correspond to a case in which we wished to reconstruct map (two dimension) using distances measured by highway or proximity coefficients between cities based on some economic criterion, social criterion, etc ....

1.1.1. MDS Data

As described above, MDS is a method that describes relations between objects based on data that represent observed dissimilarities. Nevertheless, instead of dissimilarities, frequently we observe similarities between objects, for example, correlations. After transforming similarities into dissimilarities, MDS can be easily applied. There are several ways to transform the similarities into dissimilarities. For example, we can take the complement of a similarity or apply any monotone decreasing function that generates nonnegative values (dissimilarities cannot be negative). Later, when we formalize MDS, we will see that it is not necessary to transform similarities into dissimilarities. In order to indicate similarity or dissimilarity data, we will use the generic term proximity.

Proximities can be directly or indirectly obtained from observations on the objects. Distance between cities is an example of proximities obtained

directly. Proximities can also be derived from data gathered in a $(N \times q)$ matrix $X*$, that represents measurements realised on N objects of q variables. There exist different ways of calculating dissimilarities or similarities between variables or objects. If our interest is in representing variables, we can calculate the correlation matrix as a similarity measurement amoung variables, applying MDS for the variables representation in a low dimensional space. On contrary, if we want to represent objects, distances between objects can be computed in a q dimensional space. In this case, we can use MDS to represent objects in a low dimensional space. A detailed description of several (dis)similarity measures can be found in Gower and Legendre (1986).

MDS can be applied on data regarding objects, individuals, subjects or stimuli. These four terms are usually used indistinctively, although generally, *objects* and *stimuli* refer to elements over which judgment is emitted and *subjects* and *individuals* to emitters of these judgments.

Denoting by δ_{ij} the dissimilarity between a pair of objects (i, j), we consider the situation where the set of objects consists of ten bottles of red wine corresponding to different warehouses. Dissimilarity δ_{ij} between each pair of bottles (i, j) can correspond to a integer score between 0 and 10, interpreted like an emitted judgment by an expert wine taster with respect to $i-th$ and $j-th$ bottles. The judgment comes from a comparison between a glass from bottle i and one from bottle j, classifying its differences in a scale where 0 means that wines are identical, and 10 means that they are completely different. If instead of one, we suppose that there are several tasters, then data will be represented by δ_{ijr} where i, j refer to wines and r to $r-th$ taster.

In MDS, data usually are described by *number of modes* and *number of ways* . *Mode* is used for each set of objects underlying data for MDS analysis, and *Way* for each underlying index in dissimilarities measurement between objects. Following with the previous situation, if only 1 taster is considered, that is, δ_{ij}, data are one-mode two-way, whereas if several tasters are considered, δ_{ijr}, data are two-mode three-way. An example where two-mode two-way data are considered in certain MDS models, is called Unfolding models.

1.1.2. MDS models

Let N denote the number of objects and let δ_{ij} denote the dissimilarity between objects i and j (by simplicity we will refer to one-mode two-way data). Let X be a $N \times p$ matrix, where p represents dimensionality of the

solution that will be specified in advance by the user. Thus, row i of X will represent the coordinates for object i, where $d_{ij}(X)$ will denote generally the distance Euclidean between row i and row j of X, defined as

$$d_{ij}(X) = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right)^{1/2}. \tag{1.1}$$

The MDS goal is to find a matrix X such that $d_{ij}(X)$ approximates the corresponding δ_{ij}. This goal usually is formulated in a context of least squares, through raw STRESS function $\sigma^2(X)$,

$$\sigma^2(X) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} w_{ij}(\delta_{ij} - d_{ij}(X))^2, \tag{1.2}$$

defined by Kruskal (Kruskal, 1964a, 1964b), being the first in proposing a formal fitting measurement in MDS. The term w_{ij} denotes a weight factor defined for the user, which must be nonnegative, and usually takes value of $w_{ij} = 0$ for missing dissimilarities.

Up to now, it has been assumed that dissimilarities are known, although frequently this is not the case. Consider for example the situation in which objects have been ordained according to their rank. Values of the dissimilarities between the objects are not known, but their positions are known. In this case, numerical values are assigned to proximities such that these values exhibit same the order of the data. Usually, these numerical values are denominated as disparities or pseudo-distances, and are denoted by $\widehat{d}$. The MDS goal is to obtain disparities and a configuration simultaneously, such that the coordinates represent to disparities as well as possible. This objective can be expressed by minimizing

$$\sigma^2(\widehat{d}, X) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} w_{ij}(\widehat{d}_{ij} - d_{ij}(X))^2, \tag{1.3}$$

over $\widehat{\mathbf{d}}$ y X, where $\widehat{\mathbf{d}}$ is a vector that contains disparities $\widehat{d}_{ij}$ for all pairs (i, j), $i < j$. This process of finding disparities is called optimal scaling and was also introduced by Kruskal (1964a, 1964b). In order to avoid the trivial solution $X = 0$ and $\widehat{d}_{ij} = 0$ for all i, j, a length condition on disparities is imposed such that the sum of squared $\widehat{d}_{ij}$ is equal to a fixed constant. For example, $\sum_{i=2}^{N} \sum_{j=1}^{i-1} w_{ij} \widehat{d}_{ij}^2 = N(N-1)/2$ (De Leeuw, 1977).

Transformations or disparities, $\widehat{d}_{ij}$, are used frequently in MDS. The different forms define the different types of MDS, which are usually classified in metric and nonmetric models.

Metric MDS models

The metric MDS models consider transformations which relate dissimilarities to distances considering, besides order, intrinsic dissimilarities value. Metric MDS usually is associated to ratio transformations, that is, $\widehat{d}_{ij} = b\delta_{ij}$ where b is a scalar and to interval transformations, $\widehat{d}_{ij} = a + b\delta_{ij}$, which is a generalization of a ratio model, where a and b are unknown values. Thus, the goal of metric MDS models is to estimate simultaneously the $\widehat{d}_{ij}$ values and the configuration of points X, such that $d_{ij}(x) \approx \widehat{d}_{ij}$. Within these models, the following three stand out:

1. *Classical MDS*. This model treats dissimilarities directly as Euclidean distances by means of relation $d_{ij}(X) = \delta_{ij}$, using spectral decomposition of the doubly centered disimilarities matrix to obtain a configuration matrix. The result 1.1.1 summarizes development of this model. Here $a = 0$ and $b = 1$.

2. *Least squares MDS*. This type of models obtain configuration X by fitting distances $\{d_{ij}\}$ to the disparities $\widehat{d}_{ij}$, under a least squares scheme. The formulation established in 1.3 defines these models

3. *Maximum Likelihood MDS*. This model supposes an underlying probability distribution, for example, the Ramsay model (1977). Using normality or lognormality assumptions for the dissimilarity data. It is possible to fit dissimilarities to distances by estimating the configuration X and a dispersion parameter under a maximum Likelihood approach. This model allows the possibility to test the model parameters.

Some of the more important metric models are the Torgerson model (1958) and the INDSCAL model (Carroll and Chang, 1970).

Nonmetric MDS models

In nonmetric MDS models, disparities $\widehat{d}_{ij}$ are only restricted to preserve the proximities'order. In this case, $\widehat{d}_{ij}$ can be arbitrary and only obeys to

monotony restriction: $\delta_{ij} \leq \delta_{RS} \quad \Rightarrow \quad \widehat{d}_{ij} \leq \widehat{d}_{RS}$. The most important models of this type are:

1. *Least squares MDS* comprising the Kruskal model (1964a, 1964b) and the ALSCAL model (Takane, 1977), although this last one can also be considered metric.

2. *Maximum Likelihood MDS.* The Takane model (1981) under normality assumptions on errors allows maximum Likelihood estimation of parameters.

Preference data model. Unfolding

A particular case of MDS models, related to two-mode two-way preference data, are Unfolding models, originally developed by Coombs (1964) for the analysis of preference choice data. This approach assumes that there exists a relation of proximity between the elements of two different sets, such that each element i, $i = 1, \ldots, R$ in the first set denoted by V, called the individuals set, gives a preference rating s_{ij} over each of the elements j, $j = 1, \ldots, N$ in the second set denoted by O, called the objects set. The distance model for preference data gives a representation of individuals and objects in the same Euclidean space of dimension M. It is assumed that the distance from the point representing an individual i (called the ideal point) to the point representing an object j, is inversely related to the corresponding preference score s_{ij}. Thus, a large value of s_{ij} indicating strong preference will be associated with a small distance between the points representing the individual i and the object j. Conversely, a weak preference will be related to a large distance.

Technically, unfolding can be considered a special case of Multidimensional Scaling (MDS), where the within-sets proximities are missing (Heiser, 1981). The general formulation of the unfolding model includes the estimation of a monotone transformation, $\widehat{d}_{ij} = f(s_{ij})$, of the preference data, compared with the distances in the representation, which in the metric model is usually restricted to be linear, in order to deal with interval or ratio scale data. According to the nature of data s_{ij}, we can consider a transformation for each individual, in which case the model is denominated row conditional. In the analogous case for individuals, the model is called unconditional.

In order to formalize the general model, let us denote by $\boldsymbol{S} = (s_{ij})$ the $R \times N$, preference matrix, and consider the $R \times M$ matrix $\boldsymbol{A}$ and the $N \times M$

matrix $\boldsymbol{B}$, whose row vectors $\boldsymbol{a}_i$, $i = 1 \ldots, R$, and $\boldsymbol{b}_j$, $j = 1, \ldots, N$, are the coordinates of the R individuals and of the N objects respectively in dimension M. Let d_{ij} Euclidean distance between the $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ vectors, defined by $d_{ij} = d(\boldsymbol{a}_i, \boldsymbol{b}_j) = [(\boldsymbol{a}_i - \boldsymbol{b}_j)'(\boldsymbol{a}_i - \boldsymbol{b}_j)]^{1/2}$. Then, the objective of unfolding models is to find simultaneously the disparities value $\widehat{d}_{ij}$ and configurations $\boldsymbol{A}$ and $\boldsymbol{B}$, such that for each individual i and object j, $\widehat{d}_{ij}$ values are as close as possible to the distance d_{ij}. From a least squares approach, this objective can be expressed minimizing,

$$\sigma^2(\widehat{\mathbf{d}}, \boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{R} \sum_{j=1}^{N} (\widehat{d}_{ij} - d_{ij})^2, \tag{1.4}$$

where $\widehat{\mathbf{d}}$ is a vector of length RN that contains disparities $\widehat{d}_{ij}$ for $i = 1, , R$, $j = 1, , N$.

Degeneracy, is perhaps the largest single problem in unfolding, and specially so for the nonmetric situation. Several analytical procedures have been proposed to avoid the problem (see Busing et al. (2005), Borg and Groenen (2005) or Van Deun et al. (2005), for further details).

Optimization in MDS: SMACOF algorithm

The minimizing of least squares loss functions (1.3) or (1.4) is a complex problem that cannot be solved in exact form. Therefore, the software packages of MDS use numerical algorithms with an iterative way to find simultaneously $\widehat{d}$ and X such that these functions reach a minimum. One very efficient algorithm to realise this is **SMACOF**proposed by De Leeuw and Heiser (1977, 1980) and De Leeuw (1977, 1988), which is based on the principle of iterative majorization.

The central idea of the majorization method is to replace iteratively the original function to minimizing $\varphi(x)$ by an auxiliary function $\hat{\varphi}(x, y)$, where y in $\hat{\varphi}(x, y)$ is some fixed value, which must to meet the following requirements:

1. $\hat{\varphi}(x, y)$ should be more simple to minimize than $\varphi(x)$.

2. $\varphi(x)$ must always be smaller than or at most equal to the auxiliary function $\hat{\varphi}(x, y)$, i.e.,

$$\varphi(x) \leq \hat{\varphi}(x, y) \qquad \forall x, y$$

3. $\hat{\varphi}(x, y)$ should touch to $\varphi(x)$ at the so-called supporting point y, i.e.,

$$\varphi(y) = \hat{\varphi}(y, y).$$

If $\hat{\varphi}(x, y)$ meet these three conditions, is called *mayorización function* of $\varphi(x)$.

In order to understand the principle of minimizing a function by majorization, consider the following. Let the minimum of $\hat{\varphi}(x, y)$ over x be attained at x^* for x, y, x^* in the corresponding dominion. The last two conditions of the majorizing function imply the chain of inequalities:

$$\varphi(x^*) \leq \hat{\varphi}(x^*, y) \leq \hat{\varphi}(y, y) = \varphi(y) \qquad \forall x, y. \tag{1.5}$$

The iterative majorization algorithm is summarized as follow:

1. $y_0 \longrightarrow y$, where y_0 is a starting value.
2. Find x^+ for which $\hat{\varphi}(x^+, y) = \text{mín}_x\, \hat{\varphi}(x, y)$.
3. If $\varphi(y) - \varphi(x^+) < \epsilon$ then stop (ϵ is a small, positive constant).
4. $x^+ \longrightarrow y$ and go to 2.

By (1.5), the majorization algorithm yields a nonincreasing sequence of function values, which is an attractive aspect of iterative majorization.

According to the previous, STRESS function 1.3 (or equivalently 1.4) is majorized by function

$$\sigma_r^2(\widehat{d}, X, Y) = \eta_\delta^2 + tr\,(X'VX) - 2tr\,(X'B(Y)Y)\,, \tag{1.6}$$

which is a quadratic function in X. Setting the gradient respect of X to 0, we obtain its minimum:

$$X = V^+ B(Y)Y, \tag{1.7}$$

where V^+ is the Moore-Penrose inverse of $V = \{v_{ij}\}$, with $v_{ij} = -w_{ij}\ if\ i \neq j$ *and* $v_{ii} = \sum_{j=1, j\neq i}^{n} w_{ij}$. The majorization algorithm guarantees a series of nonincreasing STRESS values, and when the algorithm stops, the stationary condition $X = V^+ B(X)X$ holds. De Leeuw and Heiser (1980) call (1.7) the Guttman transform, in recognition of Guttman (1968). To see Borg and Groenen (2005)for further details over algorithm .

1.2. Cluster Analysis

Cluster Analysis (CA) has as objective to group elements in homogenous groups based on the similarity among them. Generally objects are grouped, but this analysis can also be applied to group variables. CA studies three kinds of problems:

Partition of data. We have data that we suspected are homogeneous and it is desired to divide these in a number of groups fixed beforehand, so that:

1. Each element belongs to one, and only one, of the groups.

2. All elements are classified.

3. Each group is homogenous within.

Construction of hierarchies. We wish to structure the elements of a set of hierarchical form by its similarity. A hierarchical classification implies that the data are ordained in levels, so that superior levels contain to inferior levels. This type of classification is frequently used in Biology, when classifying animals, plants, etc. Strictly, these methods do not define groups, but the structure of association in the chain that can exist between the elements. Nevertheless, the constructed hierarchy allows to obtain a partition of data in groups.

Variables Classification. In problems with many variables it is interesting to make a initial exploratory study to divide variables into groups. This study can orient us to raise formal models to reduce dimensionality. Variables can be classified in groups or be structured in a hierarchy.

Partition methods can be applied to any matrix of data, for example, a dissimilarity matrix (squared and symmetrical), whereas hierarchical algorithms use distances matrix or dissimilarities among elements. In order to group variables we start from a matrix of relations amoung variables; for continuous variables correlation matrix are usually considered, and for discrete variables the matrix is usually constructed from the chi-square distance.

We will concentrate only on partition methods, in which objects are grouped in a number specified clusters by minimizing or maximizing some numerical criterion. The underlying fundamental idea in these methods consists of associating an index $c(N,T)$, to each partition of N objects in T groups such that it measures some aspect of sufficiency of this particular partition.

Tabla 1.1: fitness measures of the t-th group that contains n_t objects, derived from dissimilarity matrix Δ, where elements $\delta_{ql,kv}$ measure the dissimilarity between object l in group q and object v in group k.

Measures	Index$(r \in 1, 2)$
Lack of homogeneity	$h_1(t) = \sum_{l=1}^{n_t} \sum_{v=1, v \neq l}^{n_t} (\delta_{tl,tv})^r$
Lack of homogeneity	$h_2(t) = \max_{\substack{l=1,\ldots,n_t \\ v=1,\ldots,n_t \\ v \neq l}} [(\delta_{tl,tv})^r]$
Lack of homogeneity	$h_3(t) = \min_{v=1,\ldots,n_t} \left[\sum_{l=1}^{n_t} (\delta_{tl,tv})^r \right]$
Separation	$i_1(t) = \sum_{l=1}^{n_t} \sum_{k \neq t} \sum_{v=1}^{n_k} (\delta_{tl,kv})^r$
Separation	$i_2(t) = \min_{\substack{l=1,\ldots,n_t \\ k \neq t \\ v=1,\ldots,n_k}} [(\delta_{tl,kv})^r]$

Depending on the nature of the used index, large values indicate a desired solution of group, whereas in other cases small index values reflect a suitable solution. Once an index is associated with each partition, these can be compared by means of some established criterion, choosing in this way the optimal partition. Several criteria to group have been suggested, some which are based on dissimilarities between objects whereas others are applied directly to the data matrix.

1.2.1. Techniques that group by means of optimization

Group criteria derived from a dissimilarity matrix

The concepts of *homogeneity and separation* can be used in order to develop suitable index. An appropiate partition of objects must produce groups such that objects within a group have a cohesive structure and such that the groups are well separated. In this approach it is particularly useful to define group criteria based on one-mode dissimilarity matrix $\Delta = (\delta_{ij})$. Table 1.1 summarizes some of common group criteria that, based on δ_{ij}, minimize lack of homogeneity or maximize separation of groups; other measures can be found in Rubin (1967) and Hansen and Jaumard (1997).

The index $h_1(t)$ for example, quantifies the lack of homogeneity or heterogeneity among elements of the t-th group, by means of summing over all squared dissimilarities between two objects of the t-th group; when $r = 1$ and the dissimilarities are metrics, $h_2(t)$ can be interpreted as the diameter of the cluster. In a similar way to the homogeneity criteria, measurement $i_1(t)$ for example, quantifies the separation, $i(t)$, of the t-th group, by means of summing over all squared dissimilarities between an object in the group and an object outside the group.

After the index that measures the lack of homogeneity or separation of the groups is chosen, the optimal criteria can be defined as follow,

$$c_1(N,T) = \sum_{t=1}^{T} h(t) \tag{1.8}$$

$$c_2(N,T) = \max_{t=1,\ldots,T}[h(t)] \tag{1.9}$$

$$c_3(N,T) = \min_{t=1,\ldots,T}[h(t)]. \tag{1.10}$$

(In a similar way, some compendium of functions can be used as separation index.) The first optimal criterion reflects the *lack of homogeneity* average, whereas the second and the third measure are the lack of homogeneity of the worse and the best group, respectively. When using a lack of homogeneity criterion, the solution to cluster problem is found by minimizing criterion $c(N,T)$; when using a separation index, the objective is to maximize $c(N,T)$. Nevertheless it can be observed, that the optimal criterion, $\sum_{t=1}^{T} h_1(t)$, has the serious disadvantage that the number of dissimilarities that contribute to it depends on the size of group n_t ($\sum_{t=1}^{T} n_t = N$); Thus, the summing could simply be great because the particular partition in T groups implies that many dissimilarities are involved. The previous suggests a modification as follows

$$c_1^*(N,T) = \sum_{t=1}^{T} \frac{h_1(t)}{n_t}. \tag{1.11}$$

This group criterion is more appropriate for the index $h_1(m)$. Besides the previous criterions, others optimal criteria could be defined by means of a combination of homogeneity and separation measures.

Group criteria derived from a data matrix.

The usual group criteria from of a $n \times p$ matrix X of continuos data, consider the $p \times p$ matrix of total dispersion, T, defined as $T = \sum_{t=1}^{T} \sum_{l=1}^{n_t} (x_{tl} - \overline{x})(x_{tl} - \overline{x})'$, where x_{tl} is the p-dimensional vector of observations of the l-th object in the group t and $\overline{x}$ is p-dimensional vector of means for each variable. This matrix of total dispersion can be decomposed into a dispersion matrix *within-groups* W, and in a dispersion matrix *between-groups* B, that is, $T = W + B$, which has the explicit form

$$T = \sum_{t=1}^{T} \sum_{l=1}^{n_t} (x_{tl} - \overline{x}_t)(x_{tl} - \overline{x}_t)' + \sum_{t=1}^{T} n_t (\overline{x}_t - \overline{x})(\overline{x}_t - \overline{x})', \tag{1.12}$$

where $\overline{x}_t$ is the p-dimensional vector of means within group t.

In the univariate case, a natural criterion for grouping could be the election of the partition corresponding to the minimum value of the dispersion matrix within-groups W (homogeneity criterion), or equivalent to the maximum value of the sum of squared between groups, B (separation criterion). Nevertheless for the multivariate case ($p > 1$), the derivation of a criterion group from equation 1.12 is not clear. For this case several alternatives will be discussed. An obvious extension of the minimizing criterion of W in the univariate case to multivariate case, is to minimize $trace(W)$ as a homogeneity criterion, or equivalent to maximize $trace(B)$ (separation criterion). It can be shown that this criterion is equivalent to minimizing the sum of squared Euclidean distances between the objects and mean of its group, that is

$$\sum_{t=1}^{T} \sum_{l=1}^{n_t} (x_{tl} - \overline{x}_t)'(x_{tl} - \overline{x}_t) = \sum_{t=1}^{T} \sum_{i=1}^{n_t} d_{tl,t}^2$$

where $d_{tl,t}^2$ is the squared Euclidean distance between object l in group t and mean of the t-th group. Therefore, minimizing $trace(W)$ is equivalent to minimize of the lack of homogeneity criterion (1.11) for Euclidean distances and $r = 2$ in the definition of $h_1(m)$ in the table 1.1. The trace criterion was proposed by Ward (1963).Other criterions imply the minimizing of $det(W)$ or equivalently the maximization of $det(T)/det(W)$, or the maximization of $trace(BW^{-1})$ (to see Everitt, 2001 for more details).

Optimization algorithms

After deciding the appropiate group criterion, it is necessary to find an algorithm that optimize the criterion to obtain a partition in T groups. In theory the criterion value could be calculated for each possible partition and we may choose a partition that has an optimal value for the criterion. Nevertheless in practice, the task is not easy, because the number of different partitions of N objects in T groups is very large when N is moderately large. This problem leads to the development of designed algorithms to look for the optimal value of a group criterion, changing the partitions and considering the new value only if it provides an improvement. Generally, these algorithms follow the steps:

- Find some initial partition of N objects in T groups.
- Calculate the change in group criterion which is produced by moving each object from its own group to another group.
- Realise the change that leads to the greatest improvement in the group criterion value.
- Repeat two previous steps until no movement of a single object produces an improvement in the group criterion.

k-means algorithms

Some of the first algorithms based on the previous steps are the algorithms denominated *k-means*. These algorithms consist of an iterative update of a partition by reassigning simultaneously each object to the group whose mean it is nearest to, recalculating then the means of the groups. It is possible to show that under some regularity conditions, the *k-means* algorithms are equivalent to minimizing $trace(W)$. These algorithms are usually applied to a data matrix X. In fact, the majority of the statistical software packages (including SPSS) implement it this way. However, the algorithms can also be applied directly to a dissimilarity matrix.

In order to describe the operating of these algorithms, we will consider a data matrix that represents a sample of N objects with p variables. The objective is to divide this sample in a number of groups, say T, fixed beforehand. The steps to follow are:

1. Select T points as centers of the initials groups. This can be realised by:

   a) assigning objects to the groups randomly and taking the centers of groups formed.

   b) taking as centers the T points most distant to each other.

   c) constructing initial groups with *a priori* information and calculating its centers, or selecting *a priori* the centers .

2. Compute Euclidean distance of each element to centers of T groups, and assign each element to the group whose center is nearest to it. The allocation is realised sequentially and when introducing a new element in a group, the coordinates of the new group center are recalculated.

3. Verify if reassigning some of the elements the value of $trace(W)$ is reduced according to chosen optimal criterion for example $trace(W)$.

4. If it is possible to reduce $trace(W)$ by moving an element returns to step 2; if it is not possible to reduce $trace(W)$ the process is finished.

Although all *k-means* algorithms basically follow the four steps summarized above, they differ in the details of their implementations. For an extensive description of the implementations and the variants see, for example, Friedman and Rubin (1967), MacQueen (1967), Ball and Hall (1967), Hartigan and Wong (1979) and Ismail and Kamel (1989). In the algorithms of simulated annealing (Kirkpatrick *et al.*,1983) the reassignment of an object that produces a reduction in the quality of the partitions is not eliminated; instead, a small probability is assigned to avoid that the algorithm remains catched in a local minimum . Other algorithms to find optimal group criteria are discussed in Hansen *et. al.*,(1994), Hansen and Jaumard (1997) and Gordon (1999).

Choosing the appropiate number of clusters

In many applications of cluster analysis, the investigator will have that *to estimate* the number of groups in data. A great variety of methods has been suggested for this purpose. Nevertheless many of these are relatively informal and are based essentially on graphical criteria. Milligan and Cooper (1985), analyzed 30 methods and provided a detailed comparative study of

the efficiency of these for the determination of the number of groups. The method with good results was introduced by Calinski and Harabasz (1974) for continuous data. Calinski and Harabasz suggest to take the T value, which corresponds to the maximum value of $C(T)$, where $C(T)$ is given by

$$C(T) = \frac{trace(B)(N-T)}{trace(W)(T-1)}.$$

Another method with good results in the study of Milligan and Cooper was the "F"test proposal by Beale (1969a). Denoting by S_T^2 the sum of squared of the deviations from the clusters centroids in the sample, a partition of N objects in T_2 clusters is significantly better than a partition in T_1 clusters $(T_2 > T_1)$ if statistic

$$F(T_1, T_2) = \frac{S_{T_1}^2 - S_{T_2}^2 / S_{T_2}^2}{[(N-T_1)/(N-T_2)](T_2/T_1)^{2p} - 1}$$

exceeds the critical value of a distribution F with $p(T_2 - T_1)$ and $p(N - T_2)$ degrees of freedom. Another procedure to determine the number of groups consists in minimizing $det(W)$, which was proposed by Marriott (1971). Generally, chosen groups number depends on the cluster method used.

The methods described before suppose that the variables are measured on a continuous scale. For categorical data, there exist several procedures also described in Milligan and Cooper (1985) . Another diagnostic that is useful to determine the number of groups and that also operates based on a dissimilarity matrix, is the *silhouette* method suggested by Kaufman and Rousseeuw (1990).

1.2.2. Finite mixtures of densities as cluster analysis models

The methods of finding groups discussed above, are based on distances models, whereas determination of the number of clusters is based essentially on informal criteria, because of the descriptive nature of the methods. Nevertheless there exist other approaches to realise groups of a data set. Consider for example, the nonparametric estimation of densities to identify clusters. In a completely probabilistic context a particular class of parametric functions of density can be used to describe data that display a potential behavior in different sub-groups. The *finite mixtures of densities* have been

used in a variety of disciplines. Using these as models for clusters analysis, the group problem becomes a parameter estimation problem of supposed mixture models. These models are described next.

Finite mixtures of densities

A finite mixture of densities is a family of probability density functions of the form

$$f(x;\pi,\theta) = \sum_{j=1}^{T} \pi_j g_j(x;\theta_j), \tag{1.13}$$

where $\mathbf{x}$ is a p dimensional random variable, $\pi' = [\pi_1, \pi_2, , \pi_{T-1}]$ and $\theta' = [\theta_1,\ theta a_2, , \theta_T]$, where π_j are the mixture proportions and g_j, $j = 1, 2, , T$, are the densities that compose the mixture. Each density g_j is parametrized by θ_j. The mixture proportions are nonnegative values such that $\sum_j^T\ pi_j = 1$. The number of components that form the mixture is T.

The finite mixtures models provide suitable models for cluster analysis if we assume that each group of observations in a data set suspected to contain clusters, comes from a population with a different probability distribution. Each group can belong to the same family, but the groups may differ in their value for the parameter of the distribution. For example, densities components can be multivariate normal with different means vectors and possibly different covariances matrices. In other cases, the mixtures can include sums of different densities component. The first applications of mixtures of distributions was realised by Karl Pearson (1894) over two components with univariate normal densities, using for first time the method of moments for estimation of the model parameters.

Maximum likelihood estimation

Given a sample of observations $x_1, x_2, , x_N$, from a mixture of densities as in (1.13), the loglikelihood function, L, can be expressed as

$$L(\Phi) = \sum_{i=1}^{N} log\ f(x_i;\pi,\theta). \tag{1.14}$$

The parameter estimates in the densities could be obtained generally as a solution of the likelihood equations $\frac{\partial L(\Phi)}{\partial \Phi} = 0$, where $\Phi' = [\pi', \theta']$.

After estimating the distribution parameters of the mixture model the observations can be associated to a particular cluster based on the maximum value of the probability a posteriori estimated, given by

$$Pr(clusterj|x_i) = \frac{\widehat{\pi}_j g_j(x_i, \widehat{\theta}_j)}{f(x_j; \widehat{\pi}, \widehat{\theta})}, \quad j = 1, 2, ..., T. \tag{1.15}$$

Consider for example, a multivariate normal density mixture, with mean μ_j and covariances matrix, Σ_j, the maximum likelihood estimates of the parameters satisfies the following equations:

$$\widehat{\pi}_j = \frac{1}{N} \sum_{i=1}^{N} Pr(j|x_i) \tag{1.16}$$

$$\widehat{\mu}_j = \frac{1}{N\widehat{\pi}_j} \sum_{i=1}^{N} x_i \; Pr(j|x_i) \tag{1.17}$$

$$\widehat{\Sigma}_j = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_i)(x_i - \mu_i)' \; Pr(j|x_i) \tag{1.18}$$

where $Pr(j|x_i)$ are probabilities *a posteriori* estimated given by equation (1.15).

Hasselblad (1966; 1969), Day (1969) and Wolfe (1970), suggested an iterative scheme to solve likelihood equations. First we use initial estimations of the *a posteriori* probabilities from an initial estimation of the mixture parameters, and we use these in the right hand part of the equations(1.16)-(1.18). The parameter estimates can be used to obtain new estimation of *a posteriori* probabilities and the process is repeated until a suitable criterion of convergence is satisfied. This procedure is a particular application of the EM algorithm described by Dempster *et.al.*(1977) in the context of likelihood estimation for the problem of noncomplete data, a detailed description of the EM algorithm can be found in Krishnan and McLachlan (1997).

Scott and Symons (1971) and Banfield and Raftery (1993) also considered the use of mixtures models for cluster analysis, with the same probability model described above. However, they introduced the procedure of maximum likelihood *classification*, which involves choosing θ and γ to maximize the likelihood function given by:

$$l(\theta, \gamma) = \prod_{i=1}^{N} g_{\gamma_i}(x_i; \theta_{\gamma_i}) \tag{1.19}$$

where $\gamma' = [\gamma_1, , \gamma_N$ and $\gamma_i = j$ if x_i comes from the j-th subpopulation; the γ_i classify the subpopulation of each observation, $x_1, , x_N$.

The difference between the likelihood classification procedure and the formulation of mixtures is that, in last one it is assumed that x_i comes from a mixture of distributions, whose parameters are estimated and then the members of the clusters are determined by the maximum values of the probabilities *a posteriori* estimated. However, in the likelihood classification approach, it is assumed that x_i comes from a simple distribution $g_{\gamma_i}(x_i, \theta_{\gamma_i})$ which is determined by the unknown classification parameter γ_i. Then the integration of clusters is determined directly by finding the classification parameters that maximize the classification likelihood. The approach of classification by likelihood has an important role in the diverse models proposed in this work. Everitt, Landau and Leese (2001) describe models of mixtures for a variety of densities, and for categorical data (latent class model) or for a combination of continuous and categorical data.

Tests for number of components

The great advantage of the probabilistic approach in cluster analysis, is that it allows the possibility for testing several hypotheses on the model. In the present context, the fundamental hypothesis involves the determination of the number of mixture components. However, to determine the number of components T in a mixture is a very difficult problem, that not has been solved completely. A natural candidate to determine the hypothesis $T = T_0$ Vs $T = T_1$ $(T_1 > T_0)$, is the likelihood ratio statistic λ, given by

$$\lambda = \frac{Likelihood\ with\ T = T_0}{Likelihood\ with\ T = T_1}. \tag{1.20}$$

Unfortunately, this does not lead to a suitable significance test, due to the fact that in the mixtures models the regularity conditions are not satisfied so that $-2log\lambda$ has its usual null asymptotic distribution, that is, a chi squared distribution with degrees of freedom equal to the difference in the parameters number in the two hypotheses (to see McLachlan and Basford, 1988 for more details). The problem has been considered by numerous authors, for example

Wolf (1971), Hernandez-Avilia (1979), Everitt (1981), McLachlan (1987) and Thode *et al.* (1989).

For the simpler case that the decision to be made concerns solely the rejection or acceptance of the null hypothesis at a specified significance level α, Aitkin, Anderson and Hindle (1981) noted how analogous to the Monte Carlo test procedure of Hope (1968), the bootstrap replications can be used to provide a test of approximate size α. The test which rejects H_0 if $-2log\lambda$ for the original data is greater than the jth smallest of its K bootstrap replications, has size $\alpha = 1 - j/(K+1)$ approximately. The details of the implementation of this procedure can be found in McLachlan and Basford (1988). Other criteria considered for the election of the number of components are the information criterion of Akaike (AIC) and the Bayesian Information Criterion or BIC (Schwarz, 1978). Nevertheless both criteria depend essentially on the same necesary regularity conditions, so that $-2log\lambda$ has the mentioned asymptotic distribution under the null hypothesis. These criteria can therefore be used as guidelines and not as a automatic rule. Tibshirani *et al.* (2001) have proposed additional criteria.

1.3. Cluster-MDS models

Traditionally a close relation has existed among MDS and CA techniques (Kruskal, 1977). One of the first works on the matter was made by Bock (1986, 1987, 1997), he used a linear combination in an alternating method, optimizing between the MDS criteria and the classification methods. On the other hand, Heiser and Groenen (1997) indicate that the combined application of these techniques would provide a better understanding of the data if these were used separately. The general practice to reduce the number of dimensions by MDS and to superpose the results of a classification method is not advisable, proposing the representation, not of the original data, but the centers of the classes, in a space of low dimension. Therefore, arises the necessity to raise a unique model that represents all the objects and their associations respectively in a space of low dimension, using the original data.

In any case, it is necessary to remember that both methods are of different nature, so that MDS is allowed to obtain a reduction of the problems dimensionality in a function of the most relevant characteristics, which is used for the representation of the stimuli, whereas cluster analysis is used to classify the stimuli in homogenous groups in the space from which these

come.

As is shown in Heiser and Groenen (1997), some classification procedures based on partitions, for example k-means, when applied to rectangular matrices formed by vectors in $\mathbb{R}^v$, offer an acceptable representation if v or k take value of two or three. Otherwise, or when the procedure is applied directly to dissimilarities data, the number of dimensions for its possible representation is excessively high, being habitual from a point of view practical to procede by reducing the dimension for its representation and to interpret the groups previously obtained by means of classification techniques over the obtained configuration.

The MDS solution is optimal in the sense of representing the stimuli in reduced dimension, but not in obtaining conglomerates, whereas CA describes optimal conglomerates over the stimuli and therefore, on the original space from where the dissimilarity data measurement come, in that the great dimensionality causes that the representation of the stimuli lacks statistical interest. Although in the context of hierarchical classification, Kruskal (1977) shows that the use of the CA on the dissimilarity data has been used by the investigators to facilitate the interpretation of the aspects that have common the corresponding grouped points, that form to act can lead to deceptive conclusions with respect to the points that can seem related to each other, since the MDS solution is influenced by large disimilaridades but not by small or even by medium (Graef and Spence, 1979). An alternative to the independent analysis by means of both methods could consist of a methodology that allows to realise MDS and Classification simultaneously.

By means the transformation of Young-Householder for scalar products, Bock (1986) proposed a procedure so that, known previously a partition in k classes based directly in the dissimilarity data, allows to obtain the simultaneous representation of the k representatives of the classes joint to the n objects, defining the dissimilarity between two classes or from an object to a class, as the average of disimilaridades between the elements that compose them or between elements of the class and the object. Heiser (1993) and Heiser and Groenen (1997) in a least squares context propose a alternative to the previous approach and that which denominated *cluster differences scaling* (CDS). CDS simultaneously finds a classification by means of procedure equivalent to *k-means* and allows representation of the centers of the estimated classes in a space of low dimension.

In a probabilistic context, several models have been proposed in the literature to formulate latent class models for multidimensional scaling of paired

comparison data (Formann, 1989; BÄockenholt and BÄockenholt, 1990; De Soete, 1990; De Soete and Winsberg, 1993a), pick any/n data (BÄockenholt and BÄockenholt, 1990, 1991; De Soete and DeSarbo, 1991), multinomial choice data (Chintagunta, 1994), single stimulus preference data (DeSarbo, Howard and Jedidi, 1991; De Soete and Heiser, 1993; De Soete and Winsberg, 1993b), and three-way two mode data (Winsberg and De Soete, 1993), among others (see also DeSarbo et al., 1994, Wedel and DeSarbo, 1996 and Andrews and Manrai, 1999, who also provide a good review of the literature). For two-way one-mode data, Oh and Raftery (2007) recently have proposed a Bayesian approximation based on a mixture of multivariate normal distributions for the latent class positions, assuming an inverse Gamma, a Dirichlet, a Normal and an inverse Wishart distribution as priori distributions for the parameters

1.4. Some problems and extensions of the Cluster-MDS models

There exist diverse problematic situations, or well, that have not been considered in the development of the models of cluster-MDS and whose solution and consideration provides a greater efficiency in the description of the structure of the data. Schematically, some of them are the following:

1. *Applications of Cluster-MDS models for Non-stationary spatial covariance structure estimation.* In the analysis of spatiotemporal processes underlying environmental studies, the estimation of the non-stationary spatial covariance structure is a well known issue in which multidimensional scaling provides an important methodological approach (Sampson and Guttorp, 1992). It is also well known that approximating dispersion by a non-metric MDS procedure offers, in general, low precision when accurate differences in spatial dispersion are needed for interpolation purposes, specially if a low dimensional configuration is employed besides a high number of stations in oversampled domains. Therefore the application of cluster MDS models in this context is very appropiate, nevertheless according to the nature of the implied process, the group must be realised by considering as fundamental criterion the geographic location of the objects and later the characteristics under study measured for each object.

2. *Partition in latent class for two-way one-mode dissimilarities data.*

   In the probabilistic models of cluster-MDS developed until now, a partition of the two way one-mode dissimilarity matrix Δ in latent blocks has not been considered. Nevertheless, from a theoretical and practical point of view the development of this type of models is advisable to handle dissimilarity measures directly.

3. *A dual latent class unfolding model for two-way two-mode preference rating data.* In many situations, the analysis of stimuli preferences for a individuals set (two-way two-mode data) by means unfolding models, could be little efficient when the number of individuals is very great for its representation in a common space. Due to this situation, the categorization of the individuals set while the categories are represented in a low dimensional space may be an advisable procedure to facilitate their understanding. In addition to considering groups of individuals of a similar preference pattern, homogeneous groups of stimuli could also be considered, such that within each group there are clustered stimuli perceived to have similar attributes.

4. *Estimation of unrestricted variances in Finite Mixtures and MDS Models.* Even when a constant variance σ^2 is in accordance with most of the least squares MDS models formulations for two way one mode dissimilarities data or two way two mode preference data, the consideration of a variance depending on the latent block class does not significantly increase the number of parameters to be estimated, while it may lead to a model with fewer mixture components being selected. In addition, this assumption may be used to find the most objective possible the clusters structure in data.

5. *Consideration of asymmetric distributions in the components of the finite mixtures models.* The assumption of normality for two way one mode dissimilarities in the components of the finite mixture as group model, is a common fact. Nevertheless, although the assumption of normality in the dissimilarities means that minimizing the equivalent least squares loss function provides maximum likelihood estimators, regarding the non-negativity and the usual increase in spread with the location of δ_{ij}, the lognormal distribution can be an appropriate stochastic model for the residual variability (see Ramsay 1982). Furthermore, the

lognormal model is not as influenced by large dissimilarities as the normal model, and the near-zero dissimilarities are as important as the large ones for the solution.

6. *Indetermination problem in the partition in latent class.* The EM algorithm (Dempster, Laird and Rubin, 1977) traditionally can give an easy solution to computational parameter estimation problem in a finite mixtures model under a classification approach (see McLachlan y Krishnan, 1997). Nevertheless, in the situation where the two-way one-mode dissimilarity matrix is partitioned in blocks of the form Δ_{kl}, a dissimilarity δ_{ij} could have Δ_{kl} as the posteriori most probable associated latent class, while $\delta_{ij'}$ must be associated to the Δ_{tr} block by their largest estimated posterior probability $\pi_{ij',tr}$; this would be introducing an indetermination of what latent class the o_i objects is belonging to. This indetermination extends to the two way two mode preference data. Therefore it is necessary to establish an estimation procedure guaranteeing that the partition in blocks achieved from the dissimilarity or preference matrix can be associated to an objects classification in the case of two way one mode data and by the objects and individuals classifications for the case of two-way two-mode data.

7. *Determination of the suitable representation dimensionality and of the number of cluster.* The exploratory approaches and in particular the least squares method have been used as estimation procedures in MDS models. In addition, to improve the MDS solution interpretation and/or to obtain an adequate fit of the model when the number of objects is large for their representation, cluster-MDS methods have been developed demonstrating its utility as much in the classic framework as in the least squares framework. Nevertheless, due to its descriptive nature, these models do not provide an objective methodology to clarify the two fundamental problems underlying in these models: the suitable dimensionality of the clusters representation and the suitable number of clusters. This problem extends to the least squares unfolding models, in determining the number of clusters for individuals and objects and in the dimensionality of the joint representation of these clusters.

1.5. Chapter overview

The previous situations motivated the development of several models that give form to this work of investigation. The presentation order of these models is as follow.

In chapter 2 we propose a modification to the model *clusters differences scaling* of Heiser and Groenen (1997), consisting of including geographical spatial constraints, by which not the original stations but the cluster centres can be represented, while the stations and clusters retain their spatial relationships. A decomposition of the sum of squared dissimilarities into contributions from several sources of variation can be employed for an exploratory diagnosis of the model. Real data are analyzed and differences between several cluster-MDS strategies are discussed.

In chapter 3 we propose a cluster-MDS probabilistic model for two-way one-mode continuous rating dissimilarity data. The model aims at partitioning the objects into classes and simultaneously representing the cluster centers in a low-dimensional space. Under the normal distribution assumption, a latent class model is developed in terms of the set of dissimilarities in a maximum likelihood framework, assuming a unrestricted variances model, that is, considering different variances in the latent classes. In each iteration, the probability that a dissimilarity belongs to each of the blocks conforming to a partition of the original dissimilarity matrix, and the rest of the parameters, are estimated in a Simulated Annealing based algorithm. A model selection strategy is used to test the number of latent classes and the dimensionality of the problem. Both simulated and classical dissimilarity data are analyzed to illustrate the model.

In chapter 4 we formulate latent class cluster-MDS model that, under the normal or the lognormal distribution assumption for the mixture components, enables us to partition the sample stations into classes and simultaneously to represent the cluster centers in a low dimensional space, while the stations and clusters retain their spatial relationships. This model extends to models proposed in the chapters 2 and 3 for one-mode two-way continuous rating dissimilarity data, including geographical spatial constraints. Real and artificial data sets are also analyzed to prove their performance.

In chapter 5 a dual latent class model is proposed for a matrix of preference ratings data, which will partition the individuals and the objects into classes, and simultaneously represent the cluster centers in a low dimensional space, while individuals and objects retain their preference relationship.

Both the categories achieved and the unfolding configuration are estimated to be simultaneously optimal, by means of a conditional maximum likelihood estimation procedure, in a simulated annealing framework that enables us to take a statistical decision about the parameters of the model. The adjusted BIC statistic is employed to test the number of mixture components, and the dimensionality of the representation. Real and artificial data sets are analyzed to illustrate the performance model.

In chapter 6 an alternating least squares conditional procedure is proposed to partitioning the individuals and the objects into clusters, while the cluster centres are represented by unfolding. An extended minimum distance optimization procedure is employed that makes the algorithm efficient both in solution quality as in CPU time, that make it also advisable for large data sets. An extension of the jump method to determine the number of individual and objects clusters is proposed, and the adjusted BIC statistics is employed to determine the dimension- ality of the unfolding representation. Because of the dependence of the minimum distance optimization method of the initial solution, an enhanced Simulated An- nealing algorithm in the least squares framework is also proposed to deal with the local optimum problem, that also is efficient in terms of CPU time for small and medium sized data sets. Real and artificial data sets are analyzed to illustrate the model's performance.

Finally, in chapter 7 a series of general conclusions are presented from the results obtained of this work and recommendations on future investigations with the aim of extending the application perspective of the cluster-MDS models.

Capítulo 2

Un modelo de Cluster-MDS con restricciones de contigüidad espacial y su aplicación en procesos espacio-temporales

2.1. Introducción

En el análisis de procesos espacio-temporales subyacentes en estudios ambientales, generalmente la suposición de estacionaridad en estructuras de covarianza espacial no es una hipótesis adecuada, como lo demostraron Sampson y Guttorp (1992) mediante ejemplos convincentes. En este contexto, el Escalamiento Multidimensional recientemente ha tenido un papel fundamental en la estimación de la estructura de covarianza espacial no estacionaria.

Basados en una version ponderada del modelo de MDS no métrico (Kruskal, 1964a, 1964b), Sampson y Guttorp (1992) propusieron un enfoque no paramétrico para la estimación global de la estructura de covarianza espacial de una función aleatoria, $Z(x,t)$, observada repetidamente en el tiempo t_i, $i = 1, 2, ..., T$ en un número finito (y pequeño debido a las restricccciones de dimensionalidad de la configuración MDS) de estaciones muestrales, x_i, $i = 1, 2, ..., N$ en el plano, suponiendo estacionaridad temporal pero sin suponer estacionaridad espacial. Usando la notación del variograma, $D^2(x_i, x_j) = Var(Z_{it} - Z_{jt})$, donde Z_{it} y Z_{jt} representan las observaciones centradas (por la media) en las estaciones x_i y x_j, y denominando a $D^2(x_i, x_j)$

como la función de *dispersión espacial* debido a que el término de variograma convencionalmente sugiere la suposición de estacionaridad, la usaron como una métrica natural para la estructura de covarianza espacial, la cual después modelaron como una suave función general de las coordenadas geográficas de las estaciones (x_i, x_j).

Suponiendo una apropiada escala temporal para prevenir dependencia de la variabilidad espacial en el tiempo, o a la temporalidad, u a otros problemas relacionados (como observaron Suckling y Hay, 1978), el modelo se estableció considerando una matriz muestral Z, $N \times T$, centrada por columnas. Entonces $S = (1/T)ZH_TZ'$, donde $H_T = I - (1/T)\mathbf{1}\mathbf{1}'$ y $\mathbf{1}$ es un vector columna de longitud T, representa la matriz de covarianza muestral $(N \times N)$ entre las filas centradas (estaciones) de Z, cuyos elementos se denotan por s_{ij}, para $i, j = 1, ..., N$. Entonces $Var(Z_{it} - Z_{jt}) = s_{ii} + s_{jj} - 2s_{ij}$, representa una medida de disimilaridad para la dispersión espacial. Es fácil probar que si δ_{ij}, es la distancia Euclídea entre la i-ésima y j-ésima fila de Z, entonces

$$\frac{1}{T}\delta_{ij}^2 = Var(Z_{it} - Z_{jt}) = s_{ii} + s_{jj} - 2s_{ij},$$
$$\forall i, j = 1, ..., N, \quad \forall t = 1, ..., T.$$

De las disimilaridades espaciales δ_{ij} así definidas, el método de Sampson y Guttorp (1992) construye mediante MDS no métrico una configuración de las estaciones muestrales en un espacio de dos dimensiones, cuyas distancias entre puntos representan (en una versión suavizada) la dispersión espacial muestral. El procedimiento continúa buscando ahora una función apropiada que ajuste el gráfico de dispersion de las disimilaridades versus las distancias MDS, para obtener así un modelo condicional definido no positivo para las disimilaridades espaciales. Finalmente se propone una interpolación mediante *thin-plane spline* para relacionar las coordenadas geográficas con la configuración MDS de las estaciones.

La consideración del MDS en esta metodología es una estrategia recomendable para resolver el problema de la extrapolación en la situación no estacionaria, siendo adoptada recientemente, por ejemplo, por Arbia y Lafratta (2002), quienes usaron MDS para proponer un diseño muestral espacial anisotrópico. Una estrategia similar ha sido seguida por Løland y Høst (2003), ellos propusieron una aproximación métrica mediante MDS clásico para un modelo de covarianza espacial en un dominio costero complejo.

Es conocido el hecho de que una representación exacta MDS de N objetos se obtiene a lo máximo en $N - 1$ dimensiones, siendo la restricción para

obtener una aproximación en una dimensionalidad baja, en general, menos condicional para un procedimiento no métrico que para un procedimiento métrico (debido a su naturaleza menos restrictiva), con el costo instrínseco de que una representación no métrica usualmente es menos aproximada a la realidad. En cualquier caso, cuando el número de sitios muestreados incrementa en el procedimiento descrito arriba, resultará más difícil obtener una configuración en dos dimensiones que describa los datos de manera precisa, por lo tanto el mapeo de dispersión obtenido de acuerdo al procedimiento, podría ser un procedimiento no muy recomendable para la representación de las dispersiones espaciales, ni para los propósitos de predicción.

Por otro lado, si la muestra de sitios está distribuida irregularmente sobre el dominio, algunas vecindades locales pueden ser sobremuestreadas mientras que otras pueden ser submuestreadas, en cuyo caso, según lo precisado por Kovitz y Christakos (2004), se requieren estimadores modificados de estructuras de correlación. Pero aún cuando los sitios muestreados se distribuyan sobre alguna red o mallado regular, podrían aparecer clusters espaciales debido a la dependencia espacial heterogénea, entonces la interpretación después de aplicar procedimientos estandar de estimación para propósitos de extrapolación, podría resultar engañosa.

La aplicación combinada del análisis cluster con el escalamiento multidimensional métrico de mínimos cuadrados en las situaciones descritas anteriormente, podría conducir a una dispersión adecuada basada en la representación de las estaciones muestrales. Aún cuando la imposición de una estructura de cluster sobre la representación MDS podría parecer una metodo-logía alternativa para cualquiera de las situaciones previamente descritas , debe notarse que el espacio reducido de MDS es óptimo para los puntos representados, pero no para los cluster sobrepuestos, mientras que la estructura de cluster es óptima únicamente en el espacio original no reducido. Por lo tanto, integrar ambos procedimientos para modelar la dispersión entre clusters directamente en una dimensionalidad reducida es el enfoque más apropiado (Heiser y Groenen, 1997).

En este trabajo proponemos una modificación al procedimiento Cluster Differences Scaling (CDS) desarrollado por Heiser y Groenen (1997), mediante el cual no serán los puntos originales sino los centros de los clusters los que pueden ser representados en el plano, mientras que las estaciones y clusters mantienen sus relaciones espaciales. Esta metodología puede ser aplicada cuando el número de estaciones muestrales es demasiado grande para una representación MDS en el plano o cuando exista un esquema de clusters

subyacente en el plano geográfico y/o en el espacio de dispersiones. En todos los casos, la representación de los centros de los clusters en lugar de las estaciones ofrece una solución de MDS métrico de mínimos cuadrados para el problema del sobremuestreo.

Por lo tanto, la aplicación del procedimiento de estimación no paramétrica consistiría en dos pasos. Primero, se calcula una configuración 2D para las estaciones muestrales mediante el procedimiento CDS con restricciones espaciales, de esta forma transformamos el problema en uno en el cual la estructura de covarianza (en términos de dispersion espacial) es estacionaria e isotrópica. Entonces, la estimación no paramétrica de la estructura de covarianza puede realizarse siguiendo, por ejemplo, el enfoque de interpolación de Sampson y Guttorp (1992).

2.2. Modelo CDS con restricciones de contigüidad espacial

Sea $O = \{o_1, \ldots, o_N\}$ un conjunto de N estaciones (en general, N elementos de cualquier naturaleza), y $\Delta = (\delta_{ij})$, $i, j = 1 \ldots N$, una matriz simétrica de disimilaridades entre ellos. Se asume que, basado en Δ, cada estación será asignada a uno y solamente uno de K clusters, denotando por E a una matriz indicadora de orden $N \times K$, cuyos elementos e_{ik} son igual a uno si la estación o_i, pertenece al cluster k, o cero en otro caso. Así, si denotamos por $J_k = \{o_i | e_{ik} = 1\}$, para $k = 1, \ldots, K$, la hipótesis de que los clusters forman una partición se expresa como $J_k \bigcap J_l = \emptyset$, para $k \neq l$, y $\bigcup_k J_k = O$.

El problema del *cluster difference scaling* se puede expresar como la obtención de una configuración, X, de K puntos, x_i, $i = 1, \ldots K$, en un espacio euclídeo métrico de baja dimensión $M \leq N - 1$, usualmente $M = 2$ en este contexto, el cual es óptimo en el sentido de que el vector de distancias asociado, d, en $\mathbb{R}^{K(K-1)/2}$, se aproxima lo mejor posible a las correspondientes disimilaridades entre clusters. Este modelo se basa sobre la suposición de que cuando la asignación conduzca a la estación $i \in J_k$ y a la estación $j \in J_l$, sus disimilaridades serán representadas en el modelo como la distancia Euclídea $\mathbf{d}_{kl}$ entre los puntos clusters x_k y x_l, que será constante para los otros pares de estaciones en la cual la primera es elegida del cluster k y la segunda del cluster l. Para una partición, se supondrá que las disimilaridades variarán aleatoriamente dentro de un cluster, mientras que la distancia

correspondiente es constante dentro del mismo cluster, mientras que entre clusters, diferencias en distancia reflejarán la tendencia de las disimilaridades correspondientes a variar sistematicamente. Por lo tanto, el objetivo aquí es minimizar la función de pérdida (llamada STRESS), lo cual nos permite asignar una ponderación w_{ij} a cada par de estaciones (por ejemplo, podríamos tomar ponderaciones proporcionales a alguna función de distancia en el plano geográfico como en Sampson y Guttorp, 1992), la cual es definida como:

$$Stress = \sum_{k \leq l} \sum_{i \in J_k} \sum_{j \in J_l} w_{ij} (\delta_{ij} - \mathbf{d}_{kl})^2. \tag{2.1}$$

La obtención de cluster con restricciones es necesaria cuando deseamos que las estaciones y los clusters mantengan sus relaciones espaciales. Estas restricciones espaciales son generalmente bidimensionales (ocasionalmente, en tres o cuatro dimensiones si se incluye el tiempo). Un método comúnmente usado, especialmente para restricciones espaciales, implica el concepto de contigüidad entre estaciones y clusters, donde se define una matriz C $(N \times K)$, con elementos que son $c_{ik} = 1$, si la estación o_i es contigua al cluster k, y 0, de otro modo (Everitt *et al.* 2001). Una vez definida una apropiada matriz de contigüidad, los métodos de partición estandar o jerárquicos podrían ser modificados convenientemente para su aplicación. Diversas restricciones de contiguidad han sido ampliamente utilizadas en literatura, muchas de ellas para métodos de cluster jerárquicos tales como Maravalle *et al.* (1997), Ferligoj y Batagelj (1982) o Murtagh (1995). En el presente trabajo, proponemos una modificación al procedimiento de Heiser y Groenen (1997) basada en la inclusión de restricciones de contigüidad geográfica en la estructura de cluster. Además, tal modificación evitará el problema de clases vacías inherente al procedimiento sin restricciones. Incluso cuando el problema inherente de mínimos locales persiste, se pueden incorporar sin dificultad restricciones geográficas espaciales a la versión fuzzy del procedimiento CDS.

Después de la *Etapa Inicial*, en la cual se obtiene una partición inicial de las estaciones mediante el algoritmo de *k-means* y una configuración inicial clásica a partir de las disimilaridades de Sokal-Michener (Sokal y Michener, 1958) entre los clusters iniciales, el procedimiento de estimación alternante *k-mean/MDS* se puede resumir en dos fases (los detalles pueden ser consultados en Heiser, 1993 y Heiser y Groenen 1997): una Fase de MDS, en la cual la configuración de los centros de los clusters se obtiene a partir de las

disimilaridades de Sokal-Michener entre los clusters, usando SMACOF (De Leeuw y Heiser, 1980), y una *Fase de Asignación*, en la cual las estaciones son clasificadas a partir de las distancias MDS entre los centros de los clusters, minimizando el criterio equivalente al *k-means* clásico:

$$\kappa^2(E) = \sum_i \sum_k e_{ik} \|\mathbf{a}_i - \mathbf{b}_k^{(i)}\|^2, \tag{2.2}$$

where $\|\mathbf{a}_i - \mathbf{b}_k^{(i)}\|^2$ denota la distancia Euclídea cuadrada entre la $i-sima$ fila de la matriz $(N \times q)$ $A = \{a_{ir}\}$ y la $k-sima$ fila de la matriz $(K \times q)$ $B^{(i)} = \{b_{kr}^{(i)}\}$, con $q = N-1$, y con los elementos de $\mathbf{A}$ and $\mathbf{B}^{(i)}$, $i = 1, \ldots, N$, siendo especificados como $a_{ir} = \delta_{ir}^*$ y $b_{kr}^{(i)} = d_{kr}^*$, donde $\delta_{ir}^* = \delta_{is}$ y $d_{kr}^* = \sum_l e_{sl} d_{kl}(X)$, para $r = 1, \ldots, N-1$, con $s = r$ si $r < i$, y $s = r+1$ si $r \geq i$.

Si G es la matriz $N \times M$ de coordenadas geográficas para N estaciones en dimensión M (usualmente, $M = 2$), y denotando por D_G a la matriz $N \times N$ con elementos $d_{Gij} = d_G(o_i, o_j)$, i.e. las distancias Euclídeas entre la i-ésima fila y la$j$-ésima fila de G, entonces, una estación $o_i \in J_l$ se dice que es *contigua al cluster* J_k, $k \neq l$, si y solo si existe una estación $o_j \in J_k$ tal que $d_{Gij} = \text{mín}_{h \neq i}\{d_G(o_i, o_h), o_h \notin J_l\}$.

Para definir C, la matriz de restricciones de contigüidad geográfica asociada a G, denotaremos por $U = (u_1, \ldots, u_N)'$ al vector cuyos elementos están definidos por $u_i = l$, si $o_i \in J_l$, para $i = 1, \ldots, N$, y $l = 1, \ldots, K$, i.e. el vector que contiene los índices del cluster al cual pertenece cada estación en la partición actual, y consideramos el vector $d_G^- = (d_{G1}^-, \ldots, d_{GN}^-)'$ cuyos elementos son $d_{Gi}^- = \text{mín}_j\{d_{Gij}, o_j \notin J_{u_i}\}$, i.e. el vector columna que contiene la distancia mínima de la i-ésima fila en D_G a todas las estaciones fuera del cluster J_{u_i}.

Asociada a cada estación o_i, $i = 1, \ldots, N$ consideraremos el conjunto $L_i = \{u_j, j \neq i \mid \exists o_j \in J_{u_j}, d_{Gij} = d_{Gi}^-\}$, i.e., el conjunto de clusters diferentes a u_i en la misma distancia mínima d_{Gi}^- desde o_i. Sobre la base de L_i, para $i = 1, \ldots, N$, podemos construir la matriz $N \times K$, F, cuyos elementos están dados por $f_{il} = d_{Gi}^-$, si $l \in L_i$ y $d_{Gi}^- = \min\limits_{o_j \in J_{u_i}} \{d_{Gj}^- \mid l \in L_j\}$, y $f_{il} = 0$, en otro caso.

Con respecto ahora a cada estación o_i y su cluster correspondiente , J_{u_i}, podemos definir el conjunto $O_i = \{o_j \in J_{u_i} \mid f_{jl} \neq 0, l = 1, \ldots, K\}$ de todas las estaciones en el mismo cluster J_{u_i}, que podrían ser movidas a cualquier otro cluster. A partir de O_i, para $i = 1, \ldots, N$, podemos definir el conjunto asociado de *restricción de amplitud*, Φ_i, como

$$\Phi_i = \left\{ o_j \in O_i \mid \sum_{\substack{o_h, o_y \in J_{u_i} \\ h,y \neq j}} d_{Ghy} = \min_{o_w \in O_i} \sum_{\substack{o_h, o_y \in J_{u_i} \\ h,y \neq w}} d_{Ghy} \right\},$$

que consiste de las estaciones en el cluster J_{u_i} que pueden ser transferidas a otro cluster, sin dividir el cluster original.

Por último, los elementos de la matriz de contigüidad C, $N \times K$, se pueden definir como

$$c_{il} = \begin{cases} 1, & \text{si } \ l = u_i, \ \text{y } \ |J_{u_i}| \neq 1 \\ 1, & \text{si } \ o_i \in \Phi_i, \ \text{y } \ \nexists o_j \in J_{u_i}, j < i, \ \text{con } c_{jl} = 1, \forall \ l = 1, \ldots, K \\ 0, & \text{en otro caso} \end{cases}$$

donde $|J_{u_i}|$ representa la cardinalidad del cluster J_{u_i}. La primer condición posibilita a un punto a permanecer en su cluster actual, y la segunda condición está compuesta por dos criterios: primero, una estación puede ser movida únicamente si está situada en la frontera de su propio cluster y este movimiento no dejará subconjuntos disjuntos de estaciones en su cluster actual, y segundo, únicamente una estación puede ser movida al mismo tiempo.

2.2.1. Procedimiento alternante de estimación *k*-means

El ciclo de la fase de asignación ajusta E si una reasignación produce un decremento de $\kappa^2(E)$, y termina con el cálculo de las nuevas disimilaridades de Sokal-Michener entre cluster. Heiser y Groenen (1997) mostraron que un simple algoritmo convergente es posible si procedemos fila por fila encontrando

$$\min_{e_i} \sum_k e_{ik} \| a_i - b_k^{(i)} \|^2, \tag{2.3}$$

siendo $e_i = (e_{i1}, \ldots, e_{iK})'$, la i-ésima fila de la matriz E y manteniendo fija la asignación de las otras estaciones $j \neq i$. Este procedimiento se lleva a cabo sobre todos los K vectores binarios e_i. El valor mínimo se obtiene en alguna fila k de $B^{(i)}$ tal que si $k = u(i)$, el algoritmo se mueve a la siguiente estación, de otro modo la estación i es reasignada y la i-ésima fila de E es ajustada primero; entonces la siguiente estación es considerada hasta que la posición de todas las estaciones ha sido examinada, concluyendo la fase de asignación.

El esquema anterior de reasignación sin restricciones, llamado *método de la distancia mínima*, implica únicamente un ciclo de longitud N y puede generar clases vacías (Späth, 1985). Por lo tanto, que tal vez podría ser mas apropiado considerar una extensión con un cambio del método de reasignación basado en el algoritmo de Annealing Simulado con un esquema de enfriamiento similar al de Murillo *et al.* (2005) o *Vera et al.* (2007). No obstante, la inclusión de restricciones de contiguidad espacial implica un ciclo de longitud usualmente menor que N, debido a que únicamente serán reasignadas en (2.3) aquellas estaciones contiguas a cada estación o_i, además en todas las pruebas ejecutadas en el presente trabajo el problema de clases vacías no se presentó. De esta forma, es necesaria una nueva matriz $B_c^{(i)}$, $r^{(i)} \times q$, con $r^{(i)} \leq N$, el número de elementos no cero en la i-ésima fila c_i de la matriz C.

Denotando por $W^{(i)} = diag(c_i)$, i.e. el vector c_i expresado como una matriz diagonal $K \times K$, las filas de la matriz $W^{(i)}B^{(i)} = \widehat{B}^{(i)}$ tienen ceros excepto para los clusters candidatos en la fase de asignación. Buscamos entonces una matriz $R^{(i)}$ de rango $r^{(i)}$ tal que

$$R^{(i)}\widehat{B}^{(i)} = B_c^{(i)}. \tag{2.4}$$

Aplicando el *teorema de la factorización del rango* (Dhrymes, 1978), existe una matriz V_1, $K \times r^{(i)}$, y una matriz V_2, $r^{(i)} \times q$, ambas de rango $r^{(i)}$, tal que $\widehat{B}^{(i)} = V_1V_2$. Tomando $V_2 = B_c^{(i)}$, la relación

$$\widehat{B}^{(i)} = V_1B_c^{(i)} \tag{2.5}$$

se satisface si $R^{(i)}V_1 = I$. Pero definiendo $R^{(i)}$ como la matriz cuyas filas están dadas por los vectores $r^{(i)}$ $\{e_k|c_{ik} = 1\}$, se sigue que $R^{(i)}R^{(i)'} = I$, y tomando $V_1 = R^{(i)'}$, la ecuación (2.4) se cumple.

2.3. Aplicación

La estructura especial del modelo CDS nos permite localizar las estaciones con respecto a la configuración de los centros de los clusters mediante una función matricial simple de las coordenadas de los centros de los clusters, denominada configuración implícita de las estaciones (los detalles se pueden encontrar en Heiser y Groenen, 1997). Considerando la matriz $Y = EX$, donde E es la clasificación óptima y X la configuración final de los centros de los clusters , definimos la matriz $P = \{p_{ij}\}$ de orden $N \times N$, con elementos

fuera de la diagonal $p_{ij} = -w_{ij}$, y elementos de la diagonal $p_{ii} = \sum_{j \neq i} w_{ij}$, y la matriz $Q(Y)$, con elementos fuera de la diagonal $q_{ij}(Y) = -w_{ij}\delta_{ij}/d_{ij}(Y)$, si $d_{ij}(Y) > 0$, y cero en otro caso, mientras que los valores de la diagonal de $Q(Y)$ son tal que sus filas y columnas suman a cero. Entonces, la configuración implícita de las estaciones se obtiene mediante la siguiente expresión:

$$I(Y) = P^{+}Q(Y)Y, \tag{2.6}$$

donde P^{+} es la inversa Moore-Penrose de P.

Denotamos por N_k al número de estaciones en el cluster J_k, $k = 1, \ldots, K$, y por $\widetilde{\delta}_{kl}$ a las disimilaridades Sokal-Michener (Sokal y Michener, 1958), definidas por

$$\widetilde{\delta}_{kl} = \sum_{i \in J_k} \sum_{j \in J_l} \frac{w_{ij}\delta_{ij}}{\widetilde{w}_{kl}}, \quad \forall k, l = 1, \ldots, K,$$

donde

$$\widetilde{w}_{kl} = \sum_{i \in J_k} \sum_{j \in J_l} w_{ij}.$$

El enfoque de mínimos cuadrados permite una descomposición de la suma de cuadrados de las disimilaridades en contribuciones de varias fuentes de variación, lo cual es el análogo multidimensional de una tabla de Análisis de Varianza para una clasificación a una vía. La tabla 2.1 muestra tal descomposición y los grados de libertad apropiados para cada fuente, teniendo en cuenta que el término grados de libertad se utiliza aquí en un sentido descriptivo, como el número de términos lógicamente independientes en los cuales se basa un estadístico, es decir, sin referirse a ninguna distribución específica (los detalles pueden encontrarse en Heiser y Groenen, 1997). El criterio DAF (***D**ispersion-**A**ccounted **F**or*) Entre-Cluster es la suma de las distancias cuadradas entre los centros de los clusters ajustados, que combinada con el *Stress total* (2.1) representan la suma de las disimilaridades al cuadrado. Mas aún, el *Stress total* es descompuesto en cuatro componentes: la *suma de cuadrados del error Entre Cluster*(SSQ), que mide la variabilidad en las disimilaridades no explicadas por el agrupamiento ni por el modelo espacial MDS; la *falta de ajuste* del modelo espacial MDS, que mide la desviación ponderada entre las distancias en la configuración de los clusters y las disimilaridades de Sokal-Michener; la *suma de cuadrados del error Dentro de Cluster*, mide la suma de la variabilidad residual de las disimilaridades en

Tabla 2.1: Analisis de Dispersión para el modelo CDS.

Fuente	SSQ	GL
Between	$\sum_{k\leq l} \widetilde{w}_{kl}\widetilde{\delta}^2_{kl}$	$1/2K(K+1)$
Entre-clusters D.A.F.	$\sum_{k<l} \widetilde{w}_{kl} d^2_{kl}(\widetilde{X})$	$KM - M(M+1)/2$
Falta de homogeneidad	$\sum_{k} \widetilde{w}_{kk}\widetilde{\delta}^2_{kk}$	K
Falta de ajuste espacial	$\sum_{k<l} \widetilde{w}_{kl}(\widetilde{\delta}_{kl} - d_{kl}(\widetilde{X}))^2$	$(K-1)(K/2-M) + M(M-1)/2$
Error	$\sum_{k\leq l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \widetilde{\delta}_{kl})^2$	$[N(N-1) - K(K+1)]/2$
Entre-clusters	$\sum_{k<l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \widetilde{\delta}_{kl})^2$	$\sum_{k<l}(N_k N_l - 1)$
Dentro-clusters	$\sum_{k}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \widetilde{\delta}_{kk})^2$	$\sum_k (N_k(N_k-1)/2 - 1)$
Total	$\sum_{i<j} w_{ij}\delta^2_{ij}$	$N(N-1)/2$

cada bloque k con respecto a sus medias, y finalmente, la *falta de homogeneidad*, que mide la suma de la compactación relativa de cada cluster k. De esta forma, un valor alto del DAF y un bajo valor de la falta de ajuste espacial para la configuración de los cluster son señales de un análisis satisfactorio. La suficiencia de la suposicion espacial para un numero dado de clusters (es decir, los clusters están bien localizados por sus centros en un espacio metrico de baja dimensión) se puede obtener comparando la falta de ajuste espacial MDS con el error Entre-Cluster SSQ, y la suficiencia del numero de clusters se obtiene comparando el error Entre-Cluster SSQ con el error Dentro-Cluster SSQ. El componente Entre-Cluster debería ser aproximadamente $(K-1)/2$ veces más grande que el componente Dentro-Cluster. Finalmente, un valor grande del componente Between comparado con el componente Error será una señal para sentir confianza de que las variaciones en las disimilaridades de Sokal-Michener son sistemáticas.

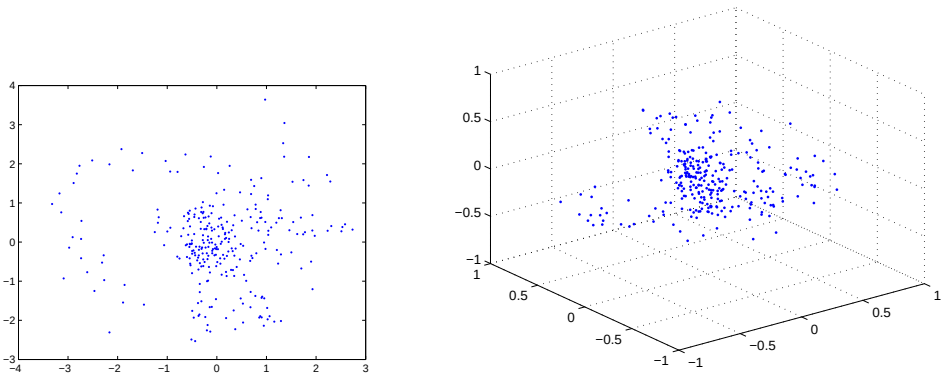

Figura 2.1: Configuración de MDS no métrico para las 289 estaciones con los datos sobre la velocidad del viento en dos y tres dimensiones.

2.4. Análisis de una red regular ajustada

La metodología fue aplicada a los datos de Cressie y Huang (1999) que representan la velocidad del viento de componente este-oeste (en m/s) sobre una región en el oeste tropical del oceano pacífico. Los datos fueron recolectados de una muestra de 289 sitios distribuidos en una red regular rectangular de 17×17, esparcidos 120 km, en cada sitio se midió la velocidad del viento repetidamente 480 veces. De la matriz de disimilaridades obtenida usando el variograma, y calculado este a partir de la matriz de datos centrados, se obtuvo una configuración mediante MDS no métrico de las 289 estaciones en dos y tres dimensiones, como se muestra en la figura 2.1. El valor del STRESS normalizado para la configuración en dos dimensiones fue de 0.2079, en tres dimensiones el STRESS fue de 0.1402 y en 10 dimensiones el valor es de 0.0261, lo cual indica que aun para soluciones no métricas, serán necesarias más de dos dimensiones para obtener una adecuada representación para un número de estaciones tan grande.

Para ilustrar el método propuesto en el caso de un dominio regular grande, se consideraron cinco clusters para su representación simultanea, siguiendo el procedimiento de Heiser y Groenen (1997). La falta de ajuste espacial (1129.5) fue más grande que el Error SSQ Entre Cluster (671.7) para dos dimensiones, lo cual sugiere como en la situación del MDS no métrico, que quizás son necesarias más dimensiones. Sin embargo, cuando se consideraron

Tabla 2.2: Análisis de dispersión para cinco clusters en dos dimensiones para el modelo CDS-R (con restricciones), de los datos de la velocidad del viento.

CDS-R (5 clusters) de los datos de la velocidad del viento.

Fuente	SSQ	%	DF	MS
Between	*39203.7*	*94.2*	*15*	*2613.579*
Entre-Cluster D.A.F.	30639.3	73.6	7	4377.043
Falta de Homogeneidad	7140.5	17.2	5	1428.107
Falta de ajuste espacial	1423.9	3.4	3	474.616
Error	*2412.3*	*5.8*	*41601*	*0.057*
Entre-Clusters	1888.9	4.5	33305	0.056
Dentro-Clusters	523.4	1.3	8296	0.063
Total	*41616.0*	*100.0*	*41616*	

restricciones espaciales, la falta de ajuste espacial (1423.9) es 1.32 veces más pequeña que el Error SSQ Entre Cluster (1888.9), como se puede apreciar en la tabla 2.2; Esta conclusión no es contradictoria con la suposición espacial en dos dimensiones en la situación previa, debido a la baja variabilidad entre las disimilaridades obtenidas del variograma. El alto DAF (73.6 %) y la baja falta de ajuste espacial para la configuración de los cluster en dos dimensiones (3.4 %) indican un análisis satisfactorio. El cociente Entre-Cluster/Dentro-Cluster es $1888.9/523.4 = 3.6$, lo cual es ligeramente más alto que el valor esperado ($4/2 = 2$), y la falta de homogeneidad (17.2 %) no es demasiado baja, lo cual sugiere que quizás podría ser más apropiado considerar un número de clusters más grande.

La figura 2.2 muestra la estructura de partición sobre el plano geográfico usando un *convex hull* para identificar estaciones que pertenecen al mismo cluster (panel izquierdo). Para ilustrar las diferencias entre la configuración de los centros de los clusters con una obtenida después de aplicar el procedimiento de k-means sobre la configuración MDS no métrica y con la obtenida sobre las coordenadas geograficas de las estaciones, se realizó una representación simultanea de la solución CDS-R y la transformación procrustes para los centros de los clusters obtenidos con los mencionados procedimientos alternativos, la cual es mostrada en el panel derecho de la figura 2.2. Como podría esperarse, diferentes estructuras son evidenciadas con las diferentes configuraciones.

Cuando siete clusters son considerados para su representación simultanea en dos dimensiones, con y sin la consideración de restricciones espaciales,

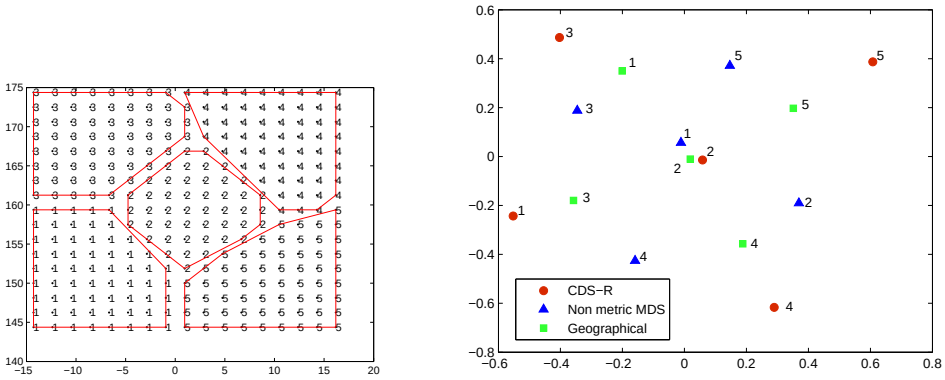

Figura 2.2: Estructura de partición para cinco clusters con restricciones mostradas sobre el plano geográfico original (panel izquierdo) y representación simultanea después de procrustes, de los centros de los clusters obtenidos por CDS-R, con los centros geográficos y los obtenidos con MDS no métrico y k-means en dos dimensiones (panel derecho) para los datos de la velocidad del viento.

la falta de ajuste espacial (1505.6 para el modelo CDS y 2263.6 para el modelo CDS-R) fue más grande que el Error SSQ Entre-Clusters (721.4 para el modelo CDS y 1992.3 para el modelo CDS-R), en ambas situaciones, lo cual sugiere que deben tomarse en cuenta más dimensiones. Una posible razón de estos valores para el error se podría encontrar en la baja variabilidad entre las disimilaridades, porque en general el error SSQ Entre-Cluster no es tan alto como el valor de la falta de ajuste espacial, para estos datos.

La tabla 2.3 muestra la descomposición de la dispersión para el modelo CDS sobre los datos de la velocidad del viento. El alto DAF (87.6 %) y la baja falta de ajuste espacial para la configuración de los clusters en tres dimensiones (1.5 %) son señal de que este análisis es satisfactorio. La falta de ajuste espacial (654.8) es ahora más pequeña que el error SSQ Entre-Cluster (746.8), y dada la baja variabilidad en las disimilaridades, parece justificable cambiar de 2 a tres dimensiones. La razón entre el Error SSQ Entre-Cluster (746.8) y el Error SSQ Dentro-Cluster (199.4) es 3.7, que es escasamente más grande que el valor de 3.0 que debería esperarse. En términos de cuadrados medios (MS), el valor Entre componentes (1452.493) es muy grande comparado al valor del error (0.022) lo cual indica una variación sistemática de las

Tabla 2.3: Análisis de Dispersión del modelo de Cluster-MDS métrico para siete clusters en tres dimensiones para el modelo CDS, sobre los datos de la velocidad del viento.

Fuente	SSQ	%	DF	MS
Entre	*40669.8*	*97.7*	*28*	*1452.493*
Entre-Clusters D.A.F	36453.1	87.6	15	2430.204
Falta de homogeneidad	3561.9	8.6	7	508.852
Falta de ajuste espacial	654.8	1.5	6	109.132
Error	*946.2*	*2.3*	*41588*	*0.022*
Entre-Clusters	746.8	1.8	35262	0.021
Dentro-clusters	199.4	0.5	6326	0.031
Total	*41616.0*	*100.0*	*41616*	

disimilaridades de Sokal-Michener.

La figura 2.3 muestra la estructura de clusters en el plano geográfico original (panel izquierdo) y la representación MDS de los clusters y la configuración transformada por procustes de los clusters obtenida después de aplicar el procedimiento de $k-means$ sobre la configuración MDS no métrica en tres dimensiones de las 289 estaciones (panel derecho). Como se puede observar de la gráfica, diferentes conclusiones se podrían obtener para propósitos de interpolación, dependiendo de la configuración de referencia que se considere.

La descomposición de la variabilidad para describir la bondad de ajuste del modelo cuando se consideran restricciones espaciales se muestra en la tabla 2.4. De nuevo, el DAF (80 %) es alto y la falta de ajuste espacial (2.1 %) baja, lo cual significa que el análisis es aceptable. El error SSQ Entre-Cluster (1992.3) es aproximadamente 2.3 veces más grande que la falta de ajuste espacial (876.0), lo cual es indicativo de una suposición espacial apropiada. La razón Entre/Dentro de los dos componentes del error es 4.8, la cual es ligeramente más alta que en la situación sin restricciones. Esto se debe a que el error se incrementa por la introducción de restricciones espaciales y, como en la situación previa, debido a que el componente Between del cuadrado medio (1400.477) es muy grande comparado al correspondiente componente del error (0.057), por lo que se asume una variación sistemática de las disimilaridades.

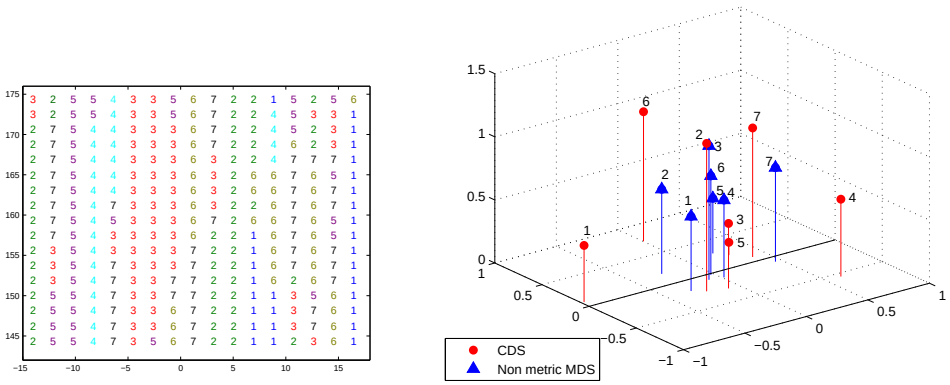

Figura 2.3: Estructura de partición para los 7 cluster mostrada sobre el plano geografico original (panel izquierdo) y la representación de Cluster-MDS en tres dimensiones (panel derecho) para los datos de la velocidad del viento.

Tabla 2.4: Análisis de Dispersión para siete clusters en tres dimensions para el modelo CDS-R sobre los datos de la velocidad del viento.

CDS-R (7 clusters) sobre los datos de la velocidad del viento

Fuente	SSQ	%	DF	MS
Entre	*39213.4*	*94.2*	*28*	*1400.477*
Entre-Cluster D.A.F.	33289.5	80.0	15	2219.301
Falta de Homog	5047.9	12.1	7	721.130
Falta de ajuste espacial	876.0	2.1	6	145.987
Error	*2402.6*	*5.8*	*41588*	*0.057*
Entre-cluster	1992.3	4.8	35687	0.055
Dentro-cluster	410.3	1.0	5901	0.069
Total	*41616.0*	*100.0*	*41616*	

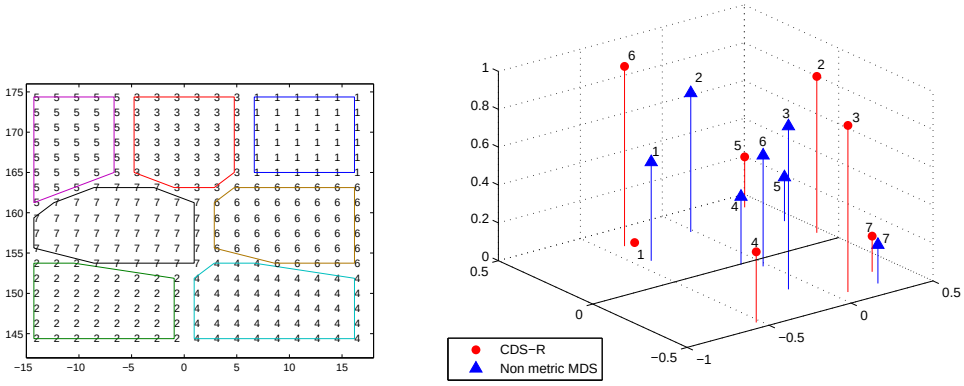

Figura 2.4: Estructura de partición para siete clusters con restricciones mostradas sobre el plano geográfico original (panel izquierdo) y la representación clusters-MDS en tres dimensiones (panel derecho) para los datos de la velocidad del viento.

Incluyendo restricciones de contiguidad espacial, la figura 2.4 muestra la estructura de partición sobre la representación geográfica en el plano mediante un convex hull, debido a que dos estaciones que pertenecen al mismo cluster deben ser ahora adyacentes. De nuevo, la configuración resultante de los centros de los cluster es representada con la transformación procrustes de la obtenida aplicando k-means sobre la solución MDS no métrica, presentándose una diferencia entre los dos esquemas de cluster y MDS utilizados.

Para el modelo de CDS no condicionado (Heiser y Groenen (1997)), no se pudo obtener una clasificación adecuada con más de siete clusters debido a que se encontraron una o más clases vacías en todas las ejecuciones. En el análisis de los datos de la velocidad del viento bajo restricciones espaciales, se obtuvieron resultados aceptables para 3 dimensiones con diecisiete clusters, como se muestra en la tabla 2.5. El DAF (85.5 %) es alto y la falta de ajuste espacial (4.7) es bajo, como se esperaba. El cociente del error Entre/Dentro es 10.57 ligeramente más alto que el valor esperado de 8 y como podría esperarse debido a que el número de clusters se incrementa, la falta de ajuste espacial (1956.5) se aproxima al error SSQ Entre-Cluster (2029.5), indicando que quizás el número de clusters no es demasiado bajo para tres dimensiones. La figura 2.5 muestra la estructura de partición sobre el plano geográfico y la configuración CDS-R en tres dimensiones.

Tabla 2.5: Análisis de Dispersión para diecisiete clusters en tres dimensiones para el modelo CDS-R sobre los datos de la velocidad del viento.

CDS-R (17 clusters) sobre los datos de la velocidad del viento

Fuente	SSQ	%	DF	MS
Entre	*39394.5*	*94.7*	*153*	*257.480*
Entre-Cluster D.A.F.	35560.6	85.5	45	790.236
Falta de Homo	1877.4	4.5	17	110.437
Falta de ajuste espacial	1956.5	4.7	91	21.499
Error	*2221.5*	*5.3*	*41463*	*.053*
Entre-cluster	2029.5	4.8	39124	.051
Dentro-cluster	192.0	0.5	2339	.082
Total	*41616.0*	*100.0*	*41616*	

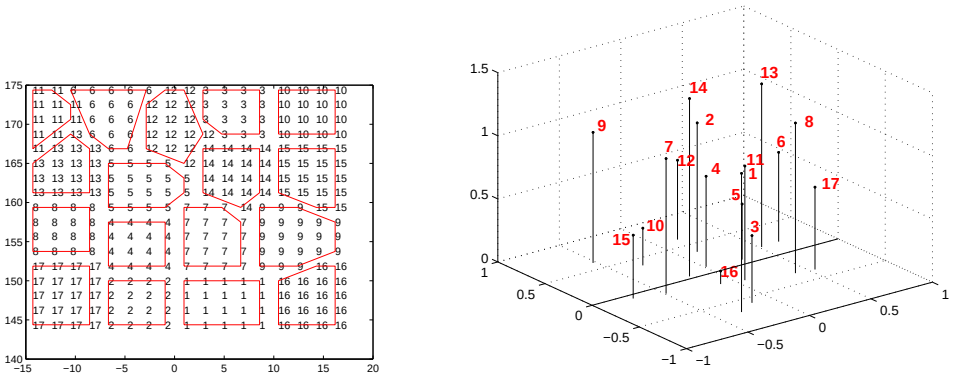

Figura 2.5: Estructura de partición para diecisiete clusters con restricciones mostradas sobre el plano geográfico original (panel izquierdo) y la representación clusters-MDS en tres dimensiones (panel derecho) para los datos de la velocidad del viento.

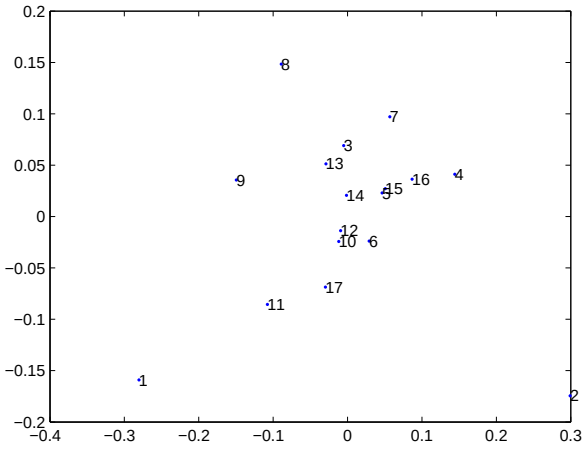

Figura 2.6: Configuración MDS no métrico para 17 estaciones sobre los datos del dioxido de sulfuro.

2.5. Análisis en un dominio distribuido irregulamente

La metodología fue aplicada a un segundo conjunto de datos que presentan una agrupación natural y que representan las concentraciones de dioxido de sulfuro recolectadas diariamente de 17 estaciones de monitoreo distribuidas en un mallado rectangular (Arbia y Lafratta, 1997). La figura 2.6 muestra la configuración de MDS no métrico en dos dimensiones con un valor del STRESS normalizado de 0.0967.

Las disimilaridades obtenidas del variograma, fueron analizadas considerando cinco clusters y escalando los centros de los clusters en dos dimensiones. Aplicando el modelo CDS se obtuvo la estructura de cluster presentada en su plano geográfico original, como se muestra en la figura 2.7 (panel izquierdo) y, debido a que el número de estaciones no es muy alto, la representación MDS de los centros de los clusters y la configuración implícita de las estaciones en dos dimensiones son representadas simultáneamente(panel derecho). De nuevo, varias estaciones pertenecientes al mismo cluster están separadas geográficamente.

La tabla 2.6 muestra la descomposición de la dispersión para el modelo

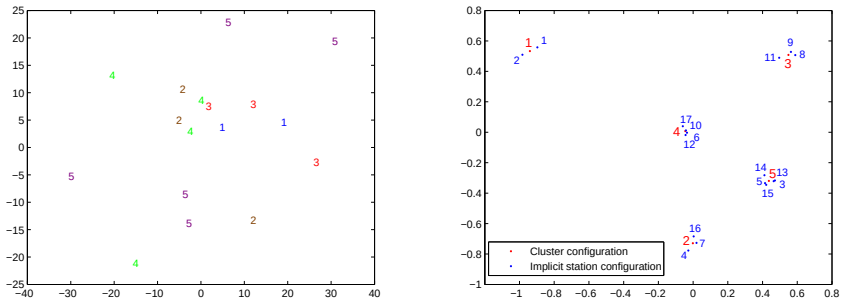

Figura 2.7: Estructura de partición para 5 clusters mostrada en diferentes colores sobre el plano geográfico original (panel izquierdo) y la representación de Cluster-MDS en dos dimensiones (panel derecho) para los datos de dioxido de sulfuro.

CDS. La falta de ajuste espacial (1.8) es un factor 1.6 veces mas pequeño que el error SSQ entre clusters (2.9) indicando que dos dimensiones pueden ser usadas para explicar las relaciones entre las estaciones. De nuevo, el elevado porcentaje de D.A.F (85.3) y el bajo porcentaje de la falta de ajuste espacial (1.3) es un indicio de un análisis satisfactorio. El cociente del cuadrado medio del Error Entre Clusters (0.027)/ Dentro de clusters (0.013) es 2.1, similar a $(K-1)/2 = 2$, y por lo tanto no existe evidencia de que el número de clusters elegido sea inadecuado.

Imponiendo restricciones de contiguidad espacial sobre la composición de los clusters, la figura 2.8 muestra la estructura de partición dada sobre el mapa geográfico original, la cual coincide con una clasificación geográfica natural y además con la representación de los centros de los clusters obtenida con el modelo CDS-R en dos dimensiones. Como puede apreciarse de la disposición geográfica de las estaciones, éstas no fueron clasificadas en clusters geográficamente disjuntos y la representación estadística muestra diferencias con respecto a la configuración obtenida sin restricciones.

La tabla 2.7 muestra la descomposición de la dispersión para la partición en cinco clusters bajo restricciones espaciales. La falta de ajuste espacial (1.4) es aproximadamente 5.3 veces mas pequeña que el error SSQ Entre-Cluster (7.4). Se encontró un valor relativamente alto del DAF (69.8 %), con respecto a la falta de ajuste espacial (1.0 %), indicando un análisis aceptable, a pesar de que las restricciones penalizan la bondad de ajuste en el análisis.

Tabla 2.6: Análisis de Dispersión para cinco clusters en el modelo CDS sobre los datos de dioxido de sulfuro.

Fuente	SSQ	%	DF	MS
Entre	*132.9*	*97.7*	*15*	*8.859*
Entre-Clusters D.A.F	116.0	85.3	7	16.567
Falta de homogeneidad	15.1	11.1	5	3.015
Falta de ajuste espacial	1.8	1.3	3	0.615
Error	*3.1*	*2.3*	*121*	*0.025*
Entre-Clusters	2.9	2.1	103	.027
Dentro-Clusters	0.2	0.2	18	.013
Total	*136.0*	*100.00*	*136*	

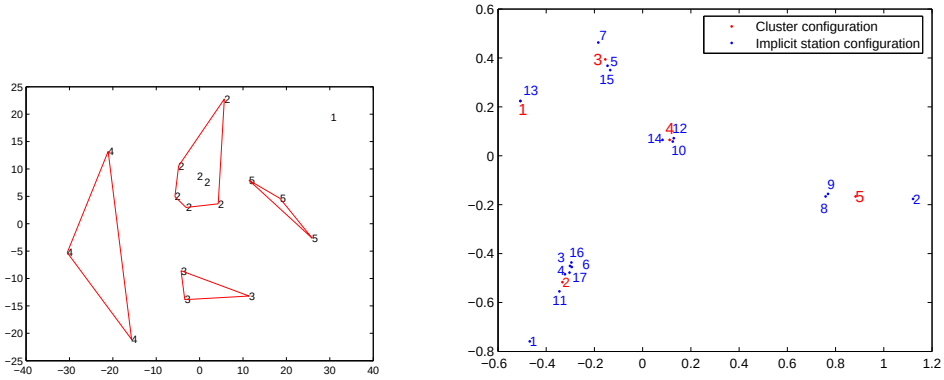

Figura 2.8: Estructura de partición para cinco clusters con restricciones mostradas sobre el plano geográfico original(panel izquierdo) y la representación clusters-MDS en dos dimensiones (panel derecho) para los datos de dioxido de sulfuro.

Tabla 2.7: Análisis de dispersión para cinco clusters en el modelo CDS-R para los datos del dióxido de sulfuro.

CDS-R (5 clusters) sobre los datos de dióxido de sulfuro

Fuente	SSQ	%	DF	MS
Entre	*125.1*	*92.0*	*15*	*8.33*
Entre-Cluster D.A.F.	94.9	69.8	7	13.55
Falta de homog	28.8	21.2	5	5.75
Falta de ajuste espacial	1.4	1.0	3	0.45
Error	*10.9*	*8.0*	*121*	*0.09*
Entre-cluster	7.4	5.4	96	0.07
Dentro-cluster	3.5	2.6	25	0.14
Total	*136.0*	*100.0*	*136*	

El cuadrado medio del error Entre-Cluster (0.07) y el cuadrado medio del error Dentro-Cluster(0.14) son muy pequeños y aproximadamente iguales, y la razón entre sus valores del stress bruto es $7.4/3.5 = 2.1$, similar al que podría esperarse (2.0). El componente del MS *Entre* (8.33) es grande comparado al componente *Error* (0.09), lo cual nos permite confiar en que la variación de las disimilaridades es sistemática.

2.6. Conclusiones

Sampson y Guttorp (1992) propusieron un algoritmo de MDS no métrico usando la formulación del variograma para la aproximación paramétrica a la estimación global de la estructura de covarianza espacial, lo cual constituye una estrategia recomendable para el propósito de extrapolación en la situación no estacionaria para conjuntos de datos de tamaño moderado, como se ha señalado en la literatura al respecto.

Sin embargo, cuando son necesarias diferencias precisas en las dispersiones espaciales, la aproximación de éstas mediante un procedimiento de MDS no métrico ofrece en general poca exactitud (debido a su naturaleza menos restrictiva), especialmente si se utiliza una configuración en una dimensión baja y considerando un gran número de estaciones. Cuando las localizaciones de la muestra están distribuidas irregularmente sobre el dominio, se requerirán estimadores para la estructuras de correlación más confiables para tratar con conjuntos de datos agrupados, como señalaron Kovitz y Christakos (2004). Pero incluso, en mallados regulares ajustados, se podrían presentar cluster es-

paciales debido a información rebundante. Una estrategia recomendable para tratar con tales problemas, es realizar un análisis de cluster en combinación con MDS Métrico, y por lo tanto un algoritmo en el que ambos métodos puedan ser utilizados de manera simultánea, puede ser altamente útil. El agrupamiento con restricciones es necesario cuando deseamos mantener la relación espacial entre estaciones y clusters; estas restricciones espaciales son generalmente bidimensionales, y se basan en la contigüidad geográfica de las estaciones, y ocasionalmente en tres o cuatro dimensiones si se incluye el tiempo.

En este trabajo proponemos una modificación al procedimiento de Cluster-MDS de mínimo cuadrados o modelo *CDS* desarrollado por Heiser y Groenen (1997), mediante el cual no los puntos originales sino los centros de los cluster pueden ser representados en un espacio de baja dimensión, mientras que las relaciones espaciales entre las estaciones y los clusters se mantienen. Esta metodología se puede aplicar cuando el tamaño de la muestra de estaciones es muy grande para ser representada en un espacio de baja dimensión o bien cuando un esquema de clusters es subyacente en el plano geográfico y /o en el espacio de las dispersiones. En todo caso, la representación de los centros de los clusters en lugar de las mismas estaciones, ofrece una solución MDS métrica de mínimos cuadrados para el problema del sobremuestreo. Además, el procedimiento nos permite establecer una representación conjunta de las estaciones y los centros de los clusters, por medio de la configuración implícita de las estaciones.

El enfoque de mínimos cuadrados permite una descomposición de la suma de cuadrados de las disimilaridades en contribuciones de varias fuentes de variación que constituye el análogo multidimensional de una tabla de Análisis de Varianza para una clasificación a una vía. Esta descomposición se puede emplear para un diagnóstico exploratorio del modelo que puede ayudar a decidir sobre la suficiencia del número de clusters y las dimensiones para la representación de los centros de los clusters.

Para ilustrar el desempeño del modelo, fueron analizados dos conjuntos de datos, el primero representa a un gran número de estaciones sobre un mallado regular ajustado y el segundo a un pequeño número de estaciones en una muestra de localizaciones irregularmente distribuidas sobre el dominio. En ambas situaciones, aún cuando en general el error se incrementa por la introducción de restricciones espaciales, el modelo CDS-R ofrece resultados satisfactorios para todos los datos analizados. Para analizar la eficiencia de la metodología propuesta, se empleó un procedimiento de k-means para de-

terminar las coordenadas de los centroides de los clusters directamente de la representación MDS no métrico de las estaciones en ambas situaciones, lo cual representa la estimación no métrica de la dispersion espacial muestral. Los centroides obtenidos de esta manera fueron comparados gráficamente mediante un análisis de procrustes con la configuración de cluster-MDS, con y sin la consideración de restricciones. Se encontraron diferencias notables con este enfoque no paramétrico a la dispersión espacial, que deben tomarse en cuenta. El conocido problema de mínimos locales en el modelo CDS está todavía presente bajo la consideración de restricciones espaciales. Aquí, las soluciones del procedimiento de agrupamiento k means sobre las coordenadas geográficas de las estaciones fueron empleadas como configuraciones iniciales para los datos analizados. Sin embargo, la versión *fuzzy* del modelo CDS puede ser extendida sin dificultad para incluir las restricciones espaciales geográficas propuestas.

Cuando los centros de los clusters son representados adecuadamente en un espacio de dimensionalidad baja, se puede proponer un *thin plane spline* o cualquier otro procedimiento de interpolación de manera que la relación entre las dispersiones espaciales es reflejada de forma precisa por el espacio MDS implantado, y así los resultados de interpolación son aceptables.

Para concluir, enfatizaremos que las metodologías de modelización de interacción espacio-tiempo, de estimación y de diseño de muestras implicando aspectos cruciales de dimensionalidad (ver, por ejemplo, Angulo *et al.*,2000, 2005; Ruiz-Medina *et al.*, 2003) constituyen un contexto importante donde la implementación apropiada de la metodología CDS propuesta ofrece significativas aplicaciones potenciales, que se están considerando actualmente para continuar la investigación.

Capítulo 3

Un modelo de clases latentes MDS para disimilaridades unimodales a dos vías

3.1. Introducción

El objetivo principal del escalamiento Multidimensional (MDS) es la representación de un conjunto $O = \{o_1, \ldots, o_N\}$, de N objetos, en un espacio generalmente euclídeo de baja dimensionalidad M, mediante una matriz de configuración $(N \times M)$ $Y = (y_{ia})$, $i = 1, \ldots, N$, $a = 1, \ldots, M$, de manera que se preserve la información de las proximidades entre los objetos. Sea $\Delta = (\delta_{ij})$ la matriz simétrica $N \times N$ de disimilaridades entre los objetos, por ejemplo evaluaciones sobre una escala continua (ver Ramsay, 1973, para una discusión sobre escala discreta), y $D = (d_{ij})$ la matriz de distancia Euclídea entre los puntos de la configuración

$$d_{ij} = \left(\sum_{a=1}^{M} (y_{ia} - y_{ja})^2 \right)^{1/2}.$$

Aunque el enfoque exploratorio y en particular el método de mínimos cuadrados ha sido el procedimiento de estimación más empleado en MDS, también se han desarrollado enfoques confirmatorios. Si el error resultante en la aproximación de distancias d_{ij} a disimilaridades δ_{ij}, se considera de naturaleza aleatoria en lugar de determinística como en el método de mínimos cuadrados, se puede formular un modelo probabilístico para MDS asumiendo

una distribución de probabilidad para las disimilaridades (ver, por ejemplo, Ramsay, 1982). En esta situación, se puede utilizar el método de máxima verosimilitud, no únicamente para la estimación de parámetros, sino también para tomar decisiones sobre las características del modelo

Para realzar la interpretación de la solución MDS y /u obtener un ajuste adecuado del modelo cuando el número de objetos es grande para su representación, se han desarrollado métodos de cluster-MDS demostrando su utililidad tanto en el marco clásico como en el de mínimos cuadrados, (Bock 1986, 1987; Heiser, 1993; Heiser y Groenen 1997; Vera, Macías y Angulo, 2008). En un contexto probabilístico, las formulaciones de mezclas de distribuciones se han considerado como un modelo de clases latentes para datos continuos, asumiendo que dentro de cada grupo homogéneo o clase latente, los datos son independientes y normalmente distribuidos (ver De Soete, 1992). También se han propuesto varios modelos de clases latentes para escalamiento multidimensional de comparación de pares de datos (Formann, 1989; Böckenholt y Böckenholt, 1990; De Soete, 1990; De Soete y Winsberg, 1993a), pick any/n data (Böckenholt y Böckenholt, 1990, 1991; De Soete y DeSarbo, 1991), elección de datos multinomiales (Chintagunta, 1994), datos de preferencia de estímulos (DeSarbo, Howard y Jedidi, 1991; De Soete y Heiser, 1993; De Soete y Winsberg, 1993b), y datos bimodales a tres vías (Winsberg y De Soete, 1993), entre otros (ver tambien DeSarbo *et al.*, 1994, Wedel y DeSarbo, 1996 y Andrews y Manrai, 1999, quienes también proporcionan una buena revisión de la literatura). Para datos unimodales a dos vías, Oh y Raftery (2007) propusieron recientemente una aproximación Bayesiana basada en una mezcla de distribuciones normales multivariadas para las posiciones de las clases latentes, asumiendo una Gamma inversa, una Dirichlet, una normal y una Wishart inversa como distribuciones a priori para los parámetros.

En este trabajo, proponemos un modelo de clases latentes para datos continuos que representan disimilaridades unimodales a dos vías cuyo objetivo es particionar los objetos en clases y estimar simultáneamente una representación espacial de dimensionalidad baja de K puntos que representan a los clusters. Para cada clasificación de prueba de los objetos, se encuentra la correspondiente partición en bloques de la matriz original de disimilaridades. Entonces, asumiendo que las disimilaridades siguen una distribución normal (ver Ramsay, 1982), se estiman condicionalmente una configuración de los centros de los clusters en un espacio de dimensionalidad baja y las dispersiones de los clusters, mediante un procedimiento de estimación de An-

nealing Simulado (**S**imulated **A**nnealing) denominado LACSSCAL (LAtent Class Simulated annealing SCAling). A diferencia del procedimiento usual de estimación basado en el algoritmo EM, la propuesta heurística asigna primero los objetos a K clusters, y entonces se evalua la logverosimilitud en la partición derivada de las disimilaridades. Por lo tanto, el algoritmo asegura que no únicamente al final, sino cada vez mediante SA, se preserva la relación en el espacio de objetos entre los objetos y las clases latentes. De esta forma, al final del procedimiento SA, los objetos son asignados a la clase a la cual son mas probables de pertenecer conjuntamente con la configuración óptima de los centros de clusters. Luego se utiliza una estrategia de selección del modelo para probar el número de clases latentes así como la dimensionalidad del problema.

3.2. Modelo

Siguiendo la notación usual en la formulación de modelos de clases latentes sea $\mathcal{P}(O)$ una partición en el espacio de los objetos O, en un número pequeño $K(K \ll N)$ de clases latentes, es decir, cada objeto pertenece a uno y solo uno de los K subconjuntos O_k, $k = 1 \ldots, K$, con n_k elementos donde $n_1 + \cdots + n_K = N$, sin conocer de antemano a que clase latente pertenece un objeto particular o_i. Sin pérdida de generalidad, se asumirá que las matrices de disimilaridades y de distancias se particionan en bloques permutando los índices de fila y columna de acuerdo a la secuencia en los conjuntos de índices de clases latentes $O_1, \ldots, O_K$. Por lo tanto, tal partición $\mathcal{P}(O)$, induce una partición correspondiente $\mathcal{P}(\Delta)$ en la matriz de disimilaridades Δ, en K^2 bloques de dimensión $n_k \times n_l$, $k, l = 1, \ldots, K$ que serán denotados por $\Delta_{kl} = (\delta_{ij})$, con $o_i \in O_k$ y $o_j \in O_l$. De esta forma, Δ_{kl} es la matriz de todas las disimilaridades entre los objetos de la clase latente O_k y de la clase latente O_l.

Como en otras formulaciones clásicas o de mínimos cuadrados de cluster-MDS, el modelo propuesto en este trabajo asume que no serán los objetos en sí los que serán representados sino los centros de los clusters mediante una matriz de configuración $X = (x_{ta})$ $(K \times M)$, donde $k = 1, \ldots, K$ y $a = 1, \ldots M$, $(M \leq K - 1)$. Así, se asume que las disimilaridades varían aleatoriamente dentro de un bloque, mientras que la distancia correspondiente es constante dentro del mismo bloque. Entonces, estamos asumiendo que cuando la asignación de objetos en clusters conduce a $o_i \in O_k$ y $o_j \in O_l$,

sus disimilaridades δ_{ij} serán representadas en el modelo como la distancia Euclídea entre las coordenadas de los centros de los clusters x_k y x_l en $\mathbb{R}^M$,

$$d_{kl} = d(x_k, x_l) = \left(\sum_{a=1}^{M} (x_{ka} - x_{la})^2 \right)^{1/2}. \tag{3.1}$$

La partición óptima de las disimilaridades $\mathcal{P}(\Delta)$, está completamente determinada por la partición óptima $\mathcal{P}(O)$ de los objetos originales, pero el inverso en general no es verdad en cuanto a que la conexión en el espacio de los objetos entre los objetos y las clases latentes podría perderse, a menos que consideremos restricciones en la forma del bloque en la construcción de $\mathcal{P}(\Delta)$ (esto se hará explícito cuando se discuta la estimación condicional de máxima verosimilitud). Por lo tanto, se desarrolla un modelo de clases latentes en el conjunto de disimilaridades preservando la relación entre los objetos o_i y los puntos de los clusters latentes x_k. En cada iteración, la probabilidad de que una disimilaridad pertenezca a cada uno de los bloques que conforman una partición $\mathcal{P}(\Delta)$ de la matriz original de disimilaridades, se estima condicionalmente de la correspondiente partición provisional $\mathcal{P}(O)$ del espacio de objetos, en un algoritmo basado en Annealing Simulado que asegura el vínculo directo entre objetos y clases latentes.

Para formular el modelo de clases latentes en términos del conjunto de disimilaridades continuas δ_{ij}, $i < j$, $i, j = 1, \ldots, N$, se asumirá que existen $K \times (K+1)/2$ bloques o grupos homogéneos de disimilaridades en la matriz triangular inferior Δ, donde los bloques diagonales también son considerados. Se asume que cada disimilaridad δ_{ij} pertenece exactamente a un bloque latente Δ_{kl}, $k \leq l$, pero no se conoce de antemano a cual bloque latente pertenece. De esta forma, la probabilidad incondicional de que una disimilaridad δ_{ij} pertenezca a un bloque latente Δ_{kl}, mientras se preservan las restricciones sobre la forma del bloque, es el tamaño relativo del bloque latente que será denotada por λ_{kl}, con $0 \leq \lambda_{kl} \leq 1$, y

$$\sum_{k \leq l} \lambda_{kl} = 1. \tag{3.2}$$

Se asumirá que las δ_{ij} pertenecientes a un bloque particular Δ_{kl} son observaciones de variables aleatorias independientes que siguen una distribución normal de media μ_{kl}, y varianza σ^2_{kl},

$$\delta_{ij} \sim \mathcal{N}(\mu_{kl}, \sigma^2_{kl}), \text{ para } \delta_{ij} \in \Delta_{kl}, \tag{3.3}$$

donde las medias de los bloques estarán relacionadas geométricamente a los centros de los clusters asumiendo $\mu_{kl} = d_{kl}$. Aún cuando la hipótesis de una varianza proporcional a la media (como argumentó Ramsay, 1982) puede considerarse como una situación particular que reduce el número de parámetros a estimar, en este trabajo se ha considerado únicamente la situación más general sin restricciones. La suposición de independencia entre pares de estímulos dentro de cada clase latente es congruente con la suposición de independencia local del análisis clásico de clases latentes (ver De Soete, 1992), y la suposición de independencia local no implica independencia global entre los pares de estímulos. La condición de independencia local es más débil que la suposición usual de independencia global de los modelos probabilísticos clásicos de MDS (Ramsay 1977, 1982) o de la mayoría de los modelos MDS de mínimos cuadrados. Como argumentó Ramsay (1997, p. 65), también con datos de escala de evaluación, las dependencias no son generalmente bastante serias para preocuparse (ver por ejemplo De Soete y Heiser, 1993 o Winsberg y De Soete, 1992 para más detalles sobre este punto).

El modelo propuesto de clases latentes para disimilaridades continuas unimodales a dos vías tiene $3 \times K \times (K+1)/2$ parámetros si las medias de las clases μ_{kl} no están condicionadas geométricamente. Tomando en cuenta la condición (3.2), los grados de libertad del modelo sin condiciones son $(3 \times K \times (K+1)/2) - 1$. Cuando se consideran relaciones geométricas de las medias de los clusters con los parámetros de la configuración X, es decir, ajustando $\mu_{kl} = d_{kl}$, y tomando en cuenta la invarianza rotacional y traslacional del modelo MDS Euclídeo unimodal a dos vías, los grados de libertad del modelo de clases latentes propuesto se transforman en

$$K \times ((K+1) + M) - \frac{M \times (M+1)}{2} - 1, \tag{3.4}$$

lo cual permite determinar un límite superior en el número de dimensiones M en el modelo, tal que los grados de libertad para el modelo condicionado sean más pequeños que para la situación incondicional.

En todo caso, el número de parámetros a estimar en el modelo propuesto será generalmente más pequeño comparado con el número de parámetros en el modelo de Ramsay, mientras tengamos un número suficientemente pequeño de clases latentes en relación a objetos en el modelo.

3.3. Estimación de máxima verosimilitud

Bajo el modelo normal en (3.3), y relacionando las medias de las distribuciones a la configuración de los centros de los clusters X, la función de densidad de probabilidad (f.d.p.) de una disimilaridad δ_{ij} que pertenece a un bloque Δ_{kl}, puede escribirse como

$$f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma^2_{kl}) = \frac{1}{\sigma_{kl}(2\pi)^{1/2}} \exp\left[-\frac{(\delta_{ij} - \mu_{kl})^2}{2\sigma^2_{kl}}\right], \tag{3.5}$$

y debido a que no se conoce de antemano a cual bloque o clase latente pertenece una disimilaridad, la f.d.p. de la variable aleatoria δ_{ij} se convierte en una mezcla finita de densidades normales univariadas dadas por (3.5). La mezcla de distribuciones se puede expresar como

$$g(\delta_{ij} \mid X, \Sigma, \lambda) = \sum_{k \leq l} \lambda_{kl} f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma^2_{kl}), \tag{3.6}$$

donde $\Sigma = (\sigma^2_{kl})$ denota la matriz $K \times K$ de varianzas en los bloques, y λ el vector columna $(K \times (K+1)/2)$ de λ_{kl}, $k \leq l$, $k, l = 1, \ldots, K$. Entonces, la función de log-verosimilitud se puede escribir como

$$\log L(X, \Sigma, \lambda \mid \Delta) = \sum_{i<j} \log(g(\delta_{ij}|X, \Sigma, \lambda)), \tag{3.7}$$

cuyo valor máximo bajo las condiciones impuestas por (3.2), es obtenido por los estimadores de máxima verosimilitud de los parámetros dados por

$$\widehat{\lambda}_{kl} = \frac{\sum_{i<j} \pi_{ij,kl}}{N(N-1)/2} \tag{3.8}$$

$$\widehat{\mu}_{kl} = \frac{\sum_{i<j} \pi_{ij,kl}\delta_{ij}}{\sum_{i<j} \pi_{ij,kl}} \tag{3.9}$$

$$\widehat{\sigma}^2_{kl} = \frac{\sum_{i<j} \pi_{ij,kl}(\delta_{ij} - \widehat{\mu}_{kl})^2}{\sum_{i<j} \pi_{ij,kl}} \tag{3.10}$$

donde

$$\pi_{ij,kl} = \frac{\lambda_{kl} f_{kl}(\delta_{ij})}{\sum_{k \leq l} \lambda_{kl} f_{kl}(\delta_{ij})}. \tag{3.11}$$

Los coeficientes $\pi_{ij,kl}$ representan la probabilidad a posteriori de que un valor observado de la disimilaridad δ_{ij} pertenezca a la clase latente Δ_{kl}, es decir, que δ_{ij} provenga de una f.d.p. normal $f_{kl}(\delta_{ij})$. Para resolver las ecuaciones (3.8), (3.9) y (3.10), son necesarios los valores de las probabilidades $\pi_{ij,kl}$, pero para obtener estos mediante (3.11), son necesarios los valores estimados de los parámetros y entonces se aplica el teorema de Bayes.

El algoritmo EM (Dempster, Laird y Rubin, 1977) tradicionalmente proporciona una solución fácil a este problema computacional de estimación (ver McLachlan y Krishnan, 1997). Sin embargo en la presente situación no se puede asegurar que los estimadores resultantes de $\pi_{ij,kl}$ (y consecuentemente de λ_{kl}), están representando las probabilidades asociadas a una partición a dos vías de la matriz de disimilaridades Δ en bloques. Si no se consideran restricciones adicionales, la partición $\mathcal{P}(\Delta)$ obtenida mediante el algoritmo EM, no necesariamente debe ser una partición en bloques en la matriz original de disimilaridades Δ y consecuentemente, no puede estar asociada a una partición $\mathcal{P}(O)$ en el espacio de objetos, como se asume en la formulación del modelo de clases latentes propuesto. En un algoritmo EM genérico, una disimilaridad δ_{ij} podría tener a Δ_{kl} como el bloque latente asociado a la probabilidad a posteriori más grande, mientras que $\delta_{ij'}$ podría estar asociada al bloque latente Δ_{tr} debido a su probabilidad a posteriori estimada más grande $\pi_{ij',tr}$; esto podría introducir una indeterminación sobre la clase latente a la que pertenece realmente el objeto o_i. El mismo problema también está presente usando cualquiera de los procedimiento de optimización heurístico Monte Carlo en los cuales las soluciones de prueba pueden ser generadas directamente por la asignación aleatoria de los valores de los parámetros en el modelo de mezclas. Para resolver este problema, se ha propuesto un procedimiento de estimación condicional de máxima verosimilitud para estimar los parámetros en el modelo, que garantiza que la particion obtenida en las disimilaridades puede estar asociada a una clasificación de objetos.

Como en la formulación del algoritmo del EM, se introducen las siguientes variables indicadoras de cluster,

$$z_{ij,kl} = \begin{cases} 1, & \text{si } \delta_{ij} \in \Delta_{kl},\ i < j,\ k \leq l, \\ 0, & \text{otro caso.} \end{cases}$$

Definimos el vector columna $K(K+1)/2 \times 1$, $\mathbf{z}_{ij} = (z_{ij,11}, \ldots, z_{ij,KK})'$, y la matriz $N(N-1)/2 \times K(K+1)/2$, $\mathbf{Z}$, escrita por sus vectores fila como $(\mathbf{z}'_{12}, \ldots, \mathbf{z}'_{(N-1)N})'$. Se asumirá que las $\mathbf{z}_{ij}$ son variables independientes e identicamente distribuidas multinomialmente con probabilidades λ_{kl}, tal que

$$\sum_{k \leq l} z_{ij,kl} = 1, \quad \text{y} \quad \sum_{i<j} \sum_{k \leq l} z_{ij,kl} = N(N-1)/2.$$

Entonces, la f.d.p. de δ_{ij}, dada $\mathbf{z}_{ij}$, se puede escribir como,

$$g(\delta_{ij} \mid \mathbf{z}_{ij}, X, \Sigma, \lambda) = \prod_{k \leq l} f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma^2_{kl})^{z_{ij,kl}} \tag{3.12}$$

y la f.d.p. de $\mathbf{z}_{ij}$, dada λ, adopta la expresión,

$$p(\mathbf{z}_{ij} \mid \lambda) = \prod_{k \leq l} \lambda_{kl}^{z_{ij,kl}}. \tag{3.13}$$

Usando (3,12) y (3,13), la f.d.p completa de δ_{ij} y $\mathbf{z}_{ij}$ se puede escribir como

$$\begin{aligned} f(\delta_{ij}, \mathbf{z}_{ij} | X, \Sigma, \lambda) &= g(\delta_{ij} | \mathbf{z}_{ij}, X, \Sigma, \lambda) p(\mathbf{z}_{ij} | \lambda) \\ &= \prod_{k \leq l} (\lambda_{kl} f_{kl}(\delta_{ij} | x_k, x_l, \sigma^2_{kl}))^{z_{ij,kl}}, \end{aligned} \tag{3.14}$$

y la log verosimilitud de los datos completos Δ y $\mathbf{Z}$ puede expresarse como

$$\begin{aligned} \log L(X, \Sigma, \lambda \mid \Delta, \mathbf{Z}) = &\sum_{i<j} \sum_{k \leq l} z_{ij,kl} \log \lambda_{kl} \\ &+ \sum_{i<j} \sum_{k \leq l} z_{ij,kl} \log f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma^2_{kl}). \end{aligned} \tag{3.15}$$

En la práctica, las variables indicadoras $\mathbf{Z}$ no son observadas. Sin embargo, si fuese conocido a que clase pertenece cada objeto, las probabilidades a posteriori $\pi_{ij,kl}$ serán 1 si $\delta_{ij} \in \Delta_{kl}$ y cero en otro caso, y el problema de máxima verosimilitud se convierte en un procedimiento general de estimación de máxima verosimilitud. Conceptualmente, el problema puede ser resuelto

mediante un algoritmo en el cual, iniciando de una partición provisional en el espacio de objetos, el resto de los parámetros son estimados condicionalmente maximizando (3.15), dada Δ y los valores previamente conocidos de $\mathbf{Z}$. Este procedimiento de estimación condicional se convierte en parte de un proceso iterativo en el cual continuamente, la partición provisional cambia aleatoriamente, terminando cuando se alcanza algún criterio de convergencia. Por lo tanto, los métodos de optimización Monte Carlo, y en particular, un algoritmo basado en Annealing Simulado con soluciones de prueba basadas en la clasificación aleatoria de los objetos originales en K clusters, puede dar una solución al problema de estimación total. Específicamente, en nuestro método de estimación aseguramos que las variables indicadoras para las disimilaridades satisfacen

$$z_{ijkl} = e_{ik}e_{jl},$$

donde e_{ik} es una variable binaria que indica si el objeto o_i está o no en el cluster O_k, y similarmente e_{jl} indica si el objeto o_j está o no en el cluster O_l. Esta misma estructura fue empleada en Heiser y Groenen (1997, ver teorema 2). Tal optimización heurística toma en cuenta la naturaleza de la forma en bloque de la partición en la matriz de disimilaridades, preservando el vínculo directo entre objetos y objetos latentes en términos de las variables indicadoras e_{ik}.

Por lo tanto, en la s-ésima iteración del algoritmo propuesto, los valores de $\widehat{Z}^{(s)}$ son encontrados directamente mediante la partición en bloques $\mathcal{P}(\Delta)^{(s)}$, derivada de una partición obtenida $\mathcal{P}(O)^{(s)}$, y la estimación condicional de las probabilidades a posteriori está dada por

$$\widehat{\pi}_{ij,kl}^{(s)} = \widehat{z}_{ij,kl}^{(s)} = \begin{cases} 1, & \text{si } \ \delta_{ij} \in \Delta_{kl}^{(s)},\ i < j,\ k \leq l, \\ 0, & \text{otro caso.} \end{cases}, \tag{3.16}$$

de lo cual, cuando los valores no observados $\mathbf{Z}$ son sustituidos en (3.15) por $\widehat{Z}^{(s)}$, la función a maximizar se convierte en,

$$\begin{aligned} \mathcal{Q}(X, \Sigma, \lambda \mid \Delta, \widehat{Z}^{(s)}) = &\sum_{i<j} \sum_{k \leq l} \widehat{z}_{ij,kl}^{(s)} \log \lambda_{kl} \\ &+ \sum_{i<j} \sum_{k \leq l} \widehat{z}_{ij,kl}^{(s)} \log f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma_{kl}^2). \end{aligned} \tag{3.17}$$

Entonces, bajo la hipótesis de los valores conocidos $\mathbf{Z}$, (3.17) se maximiza con respecto a los parámetros X, Σ, and λ, con los valores de $\widehat{z}_{ij,kl}^{(s)}$ previa-

mente estimados. Se puede demostrar fácilmente que sustituyendo los valores de $\hat{\pi}_{ij,kl}^{(s)}$ en (3.8), la expresión para el estimador de λ_{kl} en la s-ésima iteración está dada por

$$\widehat{\lambda}_{kl}^{(s)} = \frac{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)}}{N(N-1)/2}. \tag{3.18}$$

Mediante la condición geométrica de nuestro modelo, $\mu_{kl} = d_{kl}$, la estimación de los centros de los cluster puede realizarse maximizando (3.17) con respecto a X, o equivalentemente minimizando

$$q(X) = \sum_{i<j} \sum_{k\leq l} \widehat{z}_{ij,kl}^{(s)} (\delta_{ij} - d(x_k, x_l))^2. \tag{3.19}$$

Debido a que $q(X)$ puede ser descompuesto ortogonalmente en un componente *dentro* de clases y un componente *entre* clases

$$q(X) = \sum_{i<j} \sum_{k\leq l} \widehat{z}_{ij,kl}^{(s)} (\delta_{ij} - \overline{\delta}_{kl})^2 + \sum_{k\leq l} \gamma_{kl} (\overline{\delta}_{kl} - d(x_k, x_l))^2, \tag{3.20}$$

donde

$$\overline{\delta}_{kl} = \frac{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)} \delta_{ij}}{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)}}, \quad \text{y} \quad \gamma_{kl} = \sum_{i<j} \widehat{z}_{ij,kl}^{(s)},$$

debería ser necesario minimizar únicamente el último término de (3.20), que denotamos por $\phi(X)$, pero usando la condición geométrica $\mu_{kl} = d(x_k, x_l)$, y considerando que $d_{ll} = 0$, para $k, l = 1, \ldots, K$, este término puede ser de nuevo descompuesto como

$$\begin{aligned} \phi(X) &= \sum_{k\leq l} \gamma_{kl} (\overline{\delta}_{kl} - d(x_k, x_l))^2 \\ &= \sum_{k<l} \gamma_{kl} (\overline{\delta}_{kl} - d(x_k, x_l))^2 + \sum_{l} \gamma_{ll} (\overline{\delta}_{ll})^2. \end{aligned} \tag{3.21}$$

Entonces, para la estimación de la configuración, el primer término de (3.21) puede minimizarse con respecto a X usando cualquier algoritmo estándar de

MDS. En particular, la presente implementación usa SMACOF (De Leeuw y Heiser, 1980), para obtener la estimación de la configuración $\hat{X}^{(s)}$, en la s-ésima iteración.
Finalmente, sustituyendo $\hat{\pi}_{ij,kl}^{(s)}$ de (3.16) y $\hat{\mu}_{kl}^{(s)} = d(x_k^{(s)}, x_l^{(s)})$ de la configuración solución en (3.10), la estimación de los elementos de Σ se obtiene mediante la expresión

$$\widehat{\sigma}_{kl}^{2^{(s)}} = \frac{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)} (\delta_{ij} - d(\widehat{x}_k^{(s)}, \widehat{x}_l^{(s)}))^2}{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)}}, \tag{3.22}$$

en la cual, para evitar soluciones degeneradas para los bloques diagonales de disimilaridades, se emplea la varianza muestral como $\widehat{\sigma}_{kk}^{2^{(s)}}$, $\forall k = 1, \ldots K$.

3.3.1. Algunas consideraciones en la estimación de parámetros

A diferencia del algoritmo EM, que inicia a partir de algunos valores que deben ser especificados, y donde diferentes estrategias de inicio y reglas de detención pueden conducir a estimaciones totalmente distintas en este contexto (ver Seidel, Mosler y Alker, 2000), el algoritmo de Annealing Simulado no depende de una solución inicial debido a su naturaleza aleatoria, a costa de la posibilidad de elegir deliberadamente una adecuada combinación de los valores de los parámetros. La estimación inicial para los datos no observados $\widehat{\mathbf{z}}_{ij,kl}^{(0)}$ se realiza simplemente a partir de una clasificación aleatoria de los N objectos en K grupos bajo la restricción computacional de $n_k > 2$, $k = 1, \ldots, K$, la cual se preservará a lo largo del algoritmo para evitar la presencia de columnas ceros en la matriz $\widehat{\mathbf{Z}}$ como consecuencia de clases que contengan un solo objeto y a la presencia de varianza cero de un bloque diagonal comprendido por únicamente una disimilaridad. Para estimar la configuración inicial en cada iteración, en SMACOF se empleó MDS clásico sobre las disimilaridades de Sokal-Michener (1958) como los valores iniciales de $\widehat{X}^{(0)}$ (ver Heiser y Groenen, 1997).

Asociada a cada clasificación provisional en el algoritmo propuesto, se debe calcular las diferencias entre dos valores consecutivos de la función de logverosimilitud condicional. Entonces, los cambios en la estimación de los parámetros entre dos iteraciones consecutivas se han simplificado eficiente-

mente para reducir el costo computacional. En la s-ésima iteración, si el objeto $o_v \in O_t$ es seleccionado para moverse al cluster O_r, todas las entradas de $\widehat{Z}^{(s-1)}$ son preservadas en $\widehat{Z}^{(s)}$, excepto aquellas para las cuales los índices v, t o r aparecen en su posición correspondiente. Entonces, cuando $\widehat{Z}^{(s)}$ es actualizada de $\widehat{Z}^{(s-1)}$, únicamente se deben cambiar $(N-1) \times (2K-1)$ entradas mientras que el resto se preservan.

Únicamente los bloques relacionados a los índices t o r necesitan actualizarse para calcular las probabilidades λ_{kl} en (3.18). Además, si no se consideran restricciones espaciales, el mismo *atajo* se puede aplicar para la estimación de las medias μ_{kl} y varianzas σ^2_{kl}, cuando los valores $\hat{\pi}_{ij,kl}$ son sustituidos en (3.9) y (3.10) respectivamente. Entonces, sustituyendo $d(x_k, x_l)$ por μ_{kl} en (3.15), y organizando los componentes de la logverosimilitud, $f_{kl}(\delta_{ij} \mid \hat{\mu}_{kl}, \hat{\sigma}^2_{kl})$, en una matriz $\mathcal{F}$ $(N(N-1)/2 \times K(K+1)/2)$, únicamente los términos relacionados a los mismos índices que fueron considerados en la $\widehat{Z}^{(s)}$ actualizada deberían ser recalculados entre dos iteraciones consecutivas

3.4. Algoritmo de estimación de máxima verosimilitud basado en Annealing Simulado

Annealing Simulado (SA) es una técnica estocástica para la optimización global usando el algoritmo de Metropolis *et al.* (1953) en analogía al proceso de enfriamiento en termodinámica. Fue introducido por Kirkpatrick, Gelatt y Vecchi (1983), y Černy (1985). Para optimizar una función sobre un dominio compacto, en la versión homogénea del algoritmo, SA genera una secuencia de cadenas de Markov disminuyendo un parámetro positivo $\mathcal{T}$, llamado temperatura de acuerdo con la analogía física. Annealing se basa en la analogía del principio termodinámico de cristalización, en el cual primero se calienta el material y posteriormente la temperatura se reduce lentamente hasta que el material ha sido enfriado de tal forma que el cristal resultante alcanza una configuración de enrejado cristalino lo más regular posible (estado de energía mínima). El esquema de enfriamiento es un aspecto fundamental del procedimiento; si el material es enfriado demasiado rápido, esto implica la existencia de impurezas (la energía mínima no será alcanzada) y así no se obtendrá el óptimo. Además, el procedimiento proporciona medios para escapar de óptimos locales aceptando puntos que pueden tener energía más alta que los previos con probabilidad no cero, la cual es llamada la regla de

aceptación de Metropolis. Al final, valores pequeños de $\mathcal{T}$ aseguran aceptar únicamente buenas soluciones. Winkler (1995) y Andrieu y Doucet (1998) dieron una demostración de la convergencia asintótica del algoritmo SA a un óptimo global. Sin embargo, en cualquier implementación, el algoritmo de annealing simulado es un procedimiento de aproximación, garantizado a detenerse por lo menos en un óptimo local. El algoritmo general SA es como sigue:

- *Paso 1.* Se elige un punto aleatorio de una vecindad del punto previamente seleccionado, y se evalua la energía del sistema, ε, en este punto
- *Step 2.* Si la energía del sistema, ε, decrece, es decir, si la diferencia en la energía del sistema de la evaluación previa, $\triangle\varepsilon$, es negativa, el nuevo punto es aceptado. Si la energía incrementa en $\triangle\varepsilon$, el nuevo punto puede ser aceptado con probabilidad de $\exp(-\triangle\varepsilon/\mathcal{T})$, donde $\mathcal{T}$ es la temperatura actual. Este paso se denomina regla de aceptación de Metropolis (Metropolis et al., 1953).
- *Step 3.* Repetir los pasos 1 y 2 después de actualizar las cantidades apropiadas, de acuerdo a la longitud de la cadena de Markov, después de lo cual, la temperatura se disminuye y el proceso se repite hasta que el sistema alcanza un equilibrio.

Varios autores en la literatura han propuesto diferentes SA heurísticos para tratar con el problema de escalamiento unidimensional (De Soete, Hubert y Arabie, 1988; Brusco, 2001) y el problema de escalamiento multidimensional (Brusco, 2001), sin conclusiones alentadoras sobre su utilidad. Murillo, Vera y Heiser (2005) en el contexto de escalamiento unidimensional, y Vera, Heiser y Murillo (2007) en el contexto de escalamiento multidimensional para cualquier metrica Minkowsky, fueron los primeros en demostrar que Annealing Simulado puede constituir un heurístico sistema autónomo eficiente. Además, el uso del SA se ha ilustrado extensamente en la literatura del análisis cluster, como en Klein y Dubes (1989), Selim y Asultan (1991), o Sun, Xie, Song, Wang y Yu (1994), y fue comparado con el algoritmo EM en modelos de descomposición de mezclas como en Ingrassia (1991) e Ingrassia (1992), entre otros.

La figura 3.1 muestra el pseudo-código del algoritmo de Annealing Simulado propuesto, en la cual se utiliza la siguiente notación:

LC: Longitud de truncamiento de la cadena de Markov en cada fase o nivel de temperatura, la cual se incrementa por un número IC cada m iteraciones.

IC: Incremento en LC para m iteraciones.

m: Numero de iteraciones en la cual LC permanece constante.

γ: Factor de enfriamiento, que controla las reducciones en temperatura.

$\mathcal{T}$: Temperatura del sistema en cada iteración.

$\mathcal{T}_f$: Temperatura final del sistema.

$It_{\text{máx}}$: Número máximo de iteraciones de enfriamiento

$R_{\text{máx}}$: Número máximo de iteraciones en la cual una solución se mantiene sin cambios.

Ma: Número de elementos de la muestra inicial que empeoran el valor de la función de pérdida (en el calculo de la temperatura inicial).

$Ma_{\text{máx}}$: Número máximo de translaciones de la muestra inicial para obtener Ma.

χ: Probabilidad de que peores soluciones en la muestra inicial serán aceptadas cuando el sistema inicia (en el cálculo de la temperatura inicial).

Asumiendo una partición $\mathcal{P}(O) = \{O_1, \ldots, O_K\}$ en K clases, se define el vector columna $\mathcal{C}_{\mathcal{P}} = (c_1, \ldots, c_n)'$, tal que $c_i = k$, si $o_i \in O_k$, $i = 1, \ldots, N$, $k = 1, \ldots, K$. El algoritmo primero calcula la partición inicial de bloques $\mathcal{P}(\Delta)^{(0)}$ a partir de una clasificación inicial aleatoria $\mathcal{P}(O)^{(0)}$ de los N objectos en K clusters. Entonces las probabilidades a posteriori $\hat{\pi}_{ij,kl}$ son directamente calculadas de $\mathbf{Z}$ y el resto de los parámetros son estimados condicionados a la partición inicial dada, siendo estimada la configuración de los centros de los clusters en un ciclo iterativo interno mediante el algoritmo SMACOF. Por último, se evalua la función de pérdida $\log L(X, \Sigma, \lambda \mid \Delta, \mathbf{Z})$. El valor inicial de $\mathcal{T}$ es calculado, de acuerdo al esquema dado en la figura 3.2. La inicialización de $\mathcal{T}$ es realizada objetivamente, usando muestreo aleatorio simple como en Murillo *et al.* (2005) o en *Vera et al.* (2007). Eventualmente, se realizan $Ma_{\text{máx}}$ asignaciones con el objetivo de promediar los posibles incrementos Ma de las soluciones que empeoran la función objetivo. De esta forma obtenemos un valor para la temperatura inicial tal que, en las primeras iteraciones del proceso, el $100\chi\,\%$ de las peores soluciones son generalmente aceptadas.

La fase principal en el algoritmo consiste de un ciclo interno de longitud LC en el cual se aplica la regla de aceptación de Metropolis a la nueva

Lectura de datos: Lee la matriz de datos $\Delta_{N\times N}$
Inicialización: Genera $\mathcal{P}(O)$ presentando $\mathcal{C}_{\mathcal{P}} = (c_1, \ldots, c_N)'$, con $c_i \in [1, K]$ aleatoriamente.
Estima λ, X (usando SMACOF) y Σ, dada $\mathbf{Z}$
$\log L \leftarrow \log L(X, \Sigma, \lambda \mid \Delta, \mathbf{Z})$
Calcule: $Tm \leftarrow$ Temperatura inicial (ver figura 3.2)
Inicialización de parámetros: $LC, IC, \gamma, Tm_f, R_{\text{máx}}, m$
$It_{\text{máx}} \leftarrow \ln(Tm_f/Tm)/\ln(\gamma)$
$Iter \leftarrow 1$
Repetir
Para 1 **a** LC **Hacer**
Genera $i \in [1, n]$, y $t \in [1, K]$, aleatoriamente. Asigna $c_i^{(s)} = t$
$\mathcal{C}_{\mathcal{P}}^{(s)} \leftarrow (c_1, \ldots, c_i^{(s)}, \ldots, c_n)'$
Estima $Z^{(s)}$, $\lambda^{(s)}$, $X^{(s)}$ y $\Sigma^{(s)}$
$\log L^{(s)} \leftarrow \log L(X^{(s)}, \Sigma^{(s)}, \lambda^{(s)} \mid \Delta, Z^{(s)})$
Calcula: $\triangle \log L = \log L^{(s)} - \log L$
Si $\triangle \log L > 0$ **Entonces**
$\mathcal{P}(O) \leftarrow \mathcal{P}(O)^{(s)}$
De lo contrario
Si $\left(\exp\left(\frac{dt}{T}\right) > Random(0,1)\right)$ **Entonces** $\mathcal{P}(O) \leftarrow \mathcal{P}(O)^{(s)}$
Fin(Condición)
Fin(Ciclo)
$Tm \leftarrow (Tm * \gamma)$
Si $MOD(Iter, m) = 0$ **Entonces** $LC \leftarrow (LC + IC)$
Actualización: $ContRep \leftarrow$ Contador de repetición
$Iter \leftarrow (Iter + 1)$
Hasta $(Iter = It_{\text{máx}})$ o $(ContRep = R_{\text{máx}})$

Figura 3.1: Pseudo-código del algoritmo LACSSCAL

TEMPERATURA INICIAL:

Inicialización: Ma, $Ma_{\text{máx}}$, χ
$Promedio \leftarrow 0$
$cont \leftarrow 0$
$Prueba \leftarrow 1$
Repetir
    Genera $\mathcal{C}_{\mathcal{P}}$, aleatoriamente
    Estima Z, λ, X y Σ
    Calcula $\triangle \log L$
    Si $\triangle \log L < 0$ **Entonces**
        $Promedio \leftarrow (Promedio + dt)$
        $cont \leftarrow (cont + 1)$
    Fin(Condición)
    $Prueba \leftarrow (Prueba + 1)$
Hasta $(Prueba = Ma_{\text{máx}})$ o $(cont = Ma)$
$Tm_0 \leftarrow \dfrac{Promedio/cont}{\ln(\chi)}$

Figura 3.2: Pseudo-código para calcular la temperatura inicial

partición elegida aletoriamente. Esta nueva partición se obtiene asignando un objeto elegido aleatoriamente $o_i \in O_l$ a un nuevo cluster elegido aleatoriamente O_k en su correspondiente vecindad. La temperatura del sistema es reducida mediante un sencillo procedimiento de enfriamiento propuesto por Kirkpatrick *et al.* (1983), que consiste en calcular, $\mathcal{T}^{(s+1)} = \gamma \times \mathcal{T}^{(s)}$ para cada iteración.

Al final del proceso, se deben comprobar dos criterios para que el algoritmo finalice. Uno es de naturaleza computacional, evaluando hasta obtener el número máximo de iteraciones, determinado por

$$It_{\text{máx}} = \frac{\ln(\mathcal{T}_f/\mathcal{T}_0)}{\ln(\gamma)},$$

tal que si el parámetro $\mathcal{T}_f$ es muy cercano a cero, el proceso se comporta eventualmente como una técnica de gradiente descendiente, garantizando que cualquier método SA finaliza en un óptimo local (Winkler, 1995). El otro criterio es el de la convergencia, cuando una solución permanece inalterada durante un número $R_{\text{máx}}$ de iteraciones, establecidas previamente por el investigador

3.5. Selección del modelo

Una de las preocupaciones principales cuando se aplican modelos probabilísticos de MDS es la pregunta sobre cuántas dimensiones M necesitan especificarse en el modelo. Cuando se comparan soluciones para dos dimensiones consecutivas, para un número fijo de clusters K, el valor más pequeño del criterio de información AIC (Akaike, 1977) y BIC (Schwarz, 1978) o la prueba de la razón verosimilitud, pueden ser útiles para resolver este problema (ver Ramsay, 1982). Sin embargo las condiciones de regularidad para una prueba de razón de verosimilitud no se sostienen cuando el número de clusters no es conocido previamente como es comentado abajo, y ninguno de estos procedimientos son apropiados para comparaciones en esta situación (ver McLachlan y Basford, 1988).

En la aplicación del modelo MDS de clases latentes, el problema de la determinación del número de cluster K se presenta naturalmente además del problema del número de dimensiones. Una manera obvia de abordar este problema sería utilizar el estadístico de la razón de verosimilitud $\mathcal{U} = -2\log(\hat{L}^{(K)}/\hat{L}^{(K+1)})$, donde $\hat{L}^{(K)}$ y $\hat{L}^{(K+1)}$, denotan el máximo de la función de verosimilitud para las K y $(K+1)$ clases respectivamente, para probar el valor mas pequeño de K compatible con los datos. Desafortunadamente, con modelos de mezclas las condiciones de regularidad no se cumplen para que $\mathcal{U}$ tenga su distribución usual chi-cuadrada, con grados de libertad igual a la diferencia entre el número de parámetros bajo las hipótesis nula y alternativa. Aunque este problema ha sido considerado por un gran número de autores, el problema no se ha resuelto completamente (ver McLachlan y Peel, 2001 para una descripción detallada).

Como en otras aplicaciones relacionadas de clases latentes (e.g. De Soete, 1990; De Soete y DeSarbo, 1991; De Soete y Winsberg, 1993a, 1993b; De Soete y Heiser, 1993), en este trabajo se utilizó un procedimiento de bootstrap para determinar la distribución muestral de $\mathcal{U}$ como sugirió Hope (1968) y aplicado por Aitkin *et al.* (1981) en el contexto de modelos de clases latentes. Se generan muestras bootstrap del modelo ajustado bajo la hipótesis nula de K clases sin condiciones geométricas. Esto es, los estimadores de máxima verosimilitud de los parámetros $\hat{\lambda}_{kl}$, $\hat{\sigma}^2_{kl}$ y $\hat{\mu}_{kl}$ derivados de los datos son sustituidos en el modelo bajo la hipótesis nula. Entonces, $B-1$ muestras aleatorias independientes de matrices simétricas Δ^* $(N\times N)$ son elegidas de la población sin restricciones. El valor de $\mathcal{U}$ es calculado para la muestra original, denotado por $\mathcal{U}_\Delta$ y cada muestra bootstrap entonces es ajustada a los modelos

de clases latentes para K y $K+1$ alternadamente. Los valores replicados de $\mathcal{U}$ obtenidos mediante el proceso de bootstrap proporciona una valoración de la verdadera distribución de $\mathcal{U}$ si B es por lo menos igual a 20 para un nivel de significancia de $\alpha = 0{,}05$ (Hope, 1968), pero preferimos valores más grandes de B para aumentar la potencia de la prueba. Si denotamos por j al número de valores replicados en la muestra de U por debajo del valor U_Δ, el j-ésimo estadístico de orden de las $B-1$ replicaciones bootstrap se puede usar para estimar el quantil de orden j/B. Así, la hipótesis nula de K clases será rechazada a un nivel de significancia α en favor de la hipótesis alternativa de $(K+1)$-clases si los valores de $\mathcal{U}_\Delta$ para los datos observados exceden $B(1-\alpha)$ de los valores para la muestra de Monte Carlo $\mathcal{U}$ obtenida.

3.6. Aplicación

El procedimiento propuesto fue implementado en Fortran, trabajando en un ordenador Pentium IV 3.00 GHz con 2 Gb de RAM bajo Microsoft Windows XP. Para el algoritmo LACCSCAL, se siguió la siguiente estrategia en la implementación de SA en el problema de MDS. Se eligió el mejor óptimo local en 20 réplicas independientes como la mejor solución, usando los valores de los parámetros χ=0.95, $Ma = 50K$ y $Ma_{max} = 500$, para la fase de la temperatura inicial, y los valores de γ =0.70, $\mathcal{T}_f = 10^{-7}$, $R_{\text{máx}} = 10$, $LC = (N+K)$, $m = 20$ y $IC = 25K$ para el resto de los parámetros. Para el procedimiento SMACOF en la etapa de estimación de la configuración, se seleccionó la solución clásica como la configuración inicial, empleándose el criterio de convergencia con un número máximo de 200 iteraciones, así como una diferencia en valores subsiguientes del STRESS menor que 10^{-6}. Finalmente, para probar el número de clases por el procedimiento Monte Carlo y para asegurar una estimación muy precisa del p-valor, se utilizó un valor de $IC = 100K$, y el tamaño de muestra fue fijado en $B = 1000$, con un nivel de significancia de α=0.05.

Esta metodología se aplicó primero a un conjunto de datos agrupados artificialmente de distancias Euclídeas, calculadas de las coordenadas de 75 puntos en el plano, clasificados en grupos de quince valores seleccionados de cinco distribuciones normales bivariadas diferentes, con los siguientes vectores de medias y matrices de covarianzas:

Tabla 3.1: Resultados Monte Carlo para probar el número de clases para $K = 1, 2, 3, 4, 5$ en el conjunto de datos generados de una normal bivariada.

Análisis sin restricciones sobre las medias de las clases				
No. de Clases (K)	gl del modelo	Log-Verosimilitud	Prueba de Significancia de Monte Carlo de K versus K+1 Clases	
			Razón de verosimilitud	Probabilidad
1	2	-2076.50	150.93	0.002
2	8	-2001.04	845.41	0.002
3	17	-1578.33	474.28	0.002
4	29	-1341.19	262.70	0.002
5	44	-1209.84	9.28	0.578

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix},\ \mu_3 = \begin{pmatrix} 4 \\ 3 \end{pmatrix},\ \mu_4 = \begin{pmatrix} 7 \\ 6 \end{pmatrix},\ \mu_5 = \begin{pmatrix} 4 \\ -5 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 4 & 1{,}7 \\ 1{,}7 & 1 \end{pmatrix},\ \Sigma_2 = \Sigma_4 = \begin{pmatrix} {,}25 & 0 \\ 0 & {,}25 \end{pmatrix},\ \Sigma_3 = \Sigma_5 = \begin{pmatrix} 4 & -1{,}7 \\ -1{,}7 & 1 \end{pmatrix},$$

Las primeras tres distribuciones fueron las mismas que propusieron Diday y Govaert (1977), mientras que las dos últimas fueron agregadas aquí. Siguiendo un procedimiento similar al descrito por Gordon (1999), de cada distribución, se generaron cincuenta puntos, luego se seleccionó una muestra de quince candidatos con distancia mínima a la media, elegidos aleatoriamente entre el quinto vecino más cercano (entre los puntos que todavía no habían sido seleccionados), en una region de confianza del 75 % para la media. Los puntos generados y la muestra seleccionada se pueden ver en el panel izquierdo de la figura 3.3.

La tabla 3.1 muestra la bondad de ajuste del procedimiento de Monte Carlo descrito previamente, para probar el número adecuado de clases latentes en el modelo sin considerar restricciones geométricas sobre los datos generados. Los p-valores correspondientes muestran que debe seleccionarse el modelo de cinco clases como el modelo más apropiado de acuerdo con la prueba de significancia Monte Carlo. Entonces, usando un modelo de cinco clases con restricciones espaciales, se calcularon los estadísticos AIC y BIC para

Tabla 3.2: Resultados de los criterio de información para el modelo de $K = 5$ clases cuando los centros de los clusters son restringidos a ser escalados en dos y en tres dimensiones para el conjunto de datos generados de una normal bivariada.

Análisis con 5-clases restringiendo la media de la clases					
No. de Clases (K)	No. dimensiones	gl. modelo	logVer	AIC	BIC
5	2	36	-1209.69	2491.38	2704.80
5	3	38	-1209.97	2495.95	2721.23

probar una dimensionalidad adecuada del modelo, considerando $M = 2, 3$. Los resultados se resumen en la tabla 3.2. Los valores más pequeños de los estadísticos AIC y BIC se encontraron para un modelo en dos dimensiones, y de esta forma la estructura natural de los datos parece recuperarse satisfactoriamente.

La configuración resultante de los centros de los clusters se muestra en el panel derecho de la figura 3.3, después de aplicar un procedimiento de procrustes con respecto a los centros de los clusters originales para propósitos comparativos. Como se aprecia en la tabla 3.3, se encontraron probabilidades iguales para bloques de disimilaridades que están compuestos de clusters distintos, así como para los bloques diagonales. Los valores más pequeños de μ_{kl} corresponden a los bloques diagonales, mientras que el resto de valores son congruentes con la distancia real entre clusters distintos (sin considerar la transformación procrustes de la configuración).

Para probar el desempeño del algoritmo propuesto en términos del tiempo CPU, se generaron adicionalmente dos conjuntos de datos simulados de tamaño $N = 50$ y $N = 100$, respectivamente, usando el procedimiento descrito anteriormente. Estos conjuntos de datos fueron analizados para probar el número de clusters en 20 réplicas, el tiempo CPU promedio por réplica sin imponer restricciones espaciales se muestra en la figura 3.4. Se puede apreciar que el tiempo crece cuando N y K aumentan, lo cual se debe al incremento en términos del costo computacional intrínseco al procedimiento SA, por lo tanto el procedimiento propuesto es recomendable hasta conjuntos de datos de tamaño mediano.

El modelo propuesto se aplicó tambien a las disimilaridades obtenidas de un conjunto de datos reales que corresponden a 159 peces de 7 especies que

Tabla 3.3: Resultados LACSSCAL bajo restricciones espaciales en dos dimensiones y $K = 5$, para el conjunto de datos generados de una normal bivariada. Para cada bloque diagonal, las disimilaridades de Sokal-Michener y la varianza muestral representan los valores de los parámtros μ_{kk} y σ^2_{kk}, respectivamente.

Δ_{kl}	λ_{kl}	μ_{kl}	σ^2_{kl}
Δ_{11}	0.0378	0.1464	0.0084
Δ_{12}	0.0811	1.0581	0.0068
Δ_{13}	0.0811	1.7140	0.0044
Δ_{14}	0.0811	1.2122	0.0025
Δ_{15}	0.0811	1.4322	0.0058
Δ_{22}	0.0378	0.1037	0.0032
Δ_{23}	0.0811	1.3512	0.0067
Δ_{24}	0.0811	0.7037	0.0085
Δ_{25}	0.0811	0.4311	0.0034
Δ_{33}	0.0378	0.0967	0.0018
Δ_{34}	0.0811	0.6541	0.0045
Δ_{35}	0.0811	1.1589	0.0052
Δ_{44}	0.0378	0.1072	0.0031
Δ_{45}	0.0811	0.6178	0.0095
Δ_{55}	0.0378	0.0993	0.0024

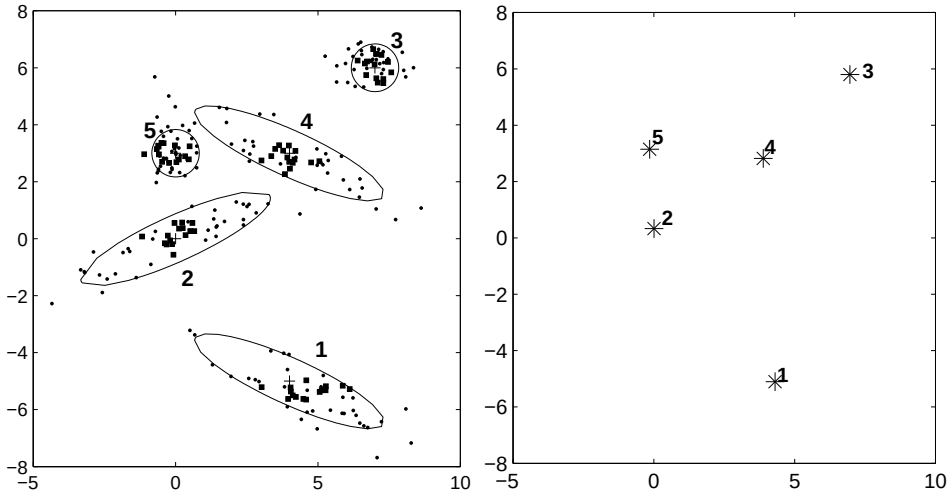

Figura 3.3: Conjunto de datos generados de una distribución normal bivariada. Los puntos expresados en cuadrados negros en cada uno de los cincuenta puntos generados, conforman la muestra aleatoria seleccionada en la región de confianza del 75 % para cada cluster (panel izquierdo) y la representación de clusters latentes-MDS obtenida en dos dimensiones (panel derecho).

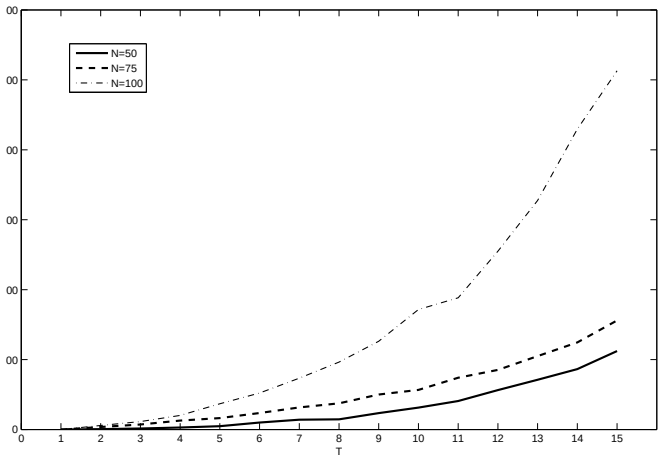

Figura 3.4: Tiempo CPU promedio por réplica (en segundos), para tres conjuntos de datos simulados de tamaño 50, 75 y 100 analizados en veinte réplicas, considerando los valores de K=1,...,15 para probar el número de clases.

fueron capturados y medidos del lago Laengelmavesi, cerca de Tampere en Finlandia (El conjunto de datos está disponible libremente en el sitio web de *the Journal of Statistical Education* http://www.amstat.org/publications/jse /datasets/fishcatch.txt). De acuerdo a lo indicado antes, debido a que el tiempo CPU crece para valores grandes de N, se analizó una muestra compuesta de 100 peces únicamente. Para el análisis se emplearon seis variables continuas: *peso del pez*, *longitud desde la nariz hasta el inicio de la cola*, *la longitud desde la nariz hasta el valle de la cola*, *longitud desde la nariz hasta el final de la cola* , *altura maxima* y *anchura*, de las cuales y para propósitos ilustrativos en este trabajo, se consideró a *peso de los peces* como la principal variable de clasificación. Entonces, las disimilaridades se calcularon como las distancias Euclídeas al cuadrado entre las filas de la matriz muestral de los datos *brutos*. La representación MDS de los 100 peces en dos dimensiones obtenida mediante SMACOF se muestra en la figura 3.5.

Siguiendo el procedimiento bootstrap descrito en la sección 3.5, primero se probó el número de clases mediante la distribución empírica generada para el estadístico U, sin considerar restricciones espaciales sobre el modelo estimado considerando K y $K+1$ clases, para los valores de $K = 1, 2, 3, ..., 15$.

Tabla 3.4: Resultados Monte Carlo para probar el número de clases para K=1,2,...,10 en el conjunto de datos de peces.

Análisis sin restricciones sobre las medias de las clases				
No. de Clases (K)	Gl Modelo	LogVer	Prueba de Significancia de Monte Carlo de K versus K+1 Clases	
			Razón de verosimilitud	Prob
1	2	-6229.82	7441.15	0.002
2	8	-2509.25	2312.86	0.002
3	17	-1352.82	2524.93	0.002
4	29	-90.35	365.52	0.002
5	44	92.41	480.27	0.002
6	62	332.55	434.93	0.020
7	83	550.01	261.44	0.030
8	107	680.73	245.34	0.038
9	134	803.40	203.12	0.044
10	164	904.96	77.72	0.668

La tabla 3.4 presenta los resultados de la bondad de ajuste obtenidos por el procedimiento Monte Carlo para valores de K hasta 10 (siendo limitado por razones de espacio), de lo cual se puede inferir que un modelo de $K = 10$ clases resulta adecuado para la estructura de disimilaridades.

Para el modelo de 10 clases restringido espacialmente, se calcularon los estadísticos AIC y BIC para probar la dimensionalidad adecuada del modelo en una, dos y tres dimensiones. Los valores más pequeños de los estadísticos AIC (14.38) y BIC (782.23) se encontraron en ambos casos para un modelo unidimensional, lo que es congruente con la mayor importancia de la variable *peso del pez* en el análisis, esto también puede apreciarse de la figura 3.5 donde se presenta mayor dispersión para la dimensión 1 que para la dimensión dos. La configuración obtenida en una dimensión se muestra en la figura 3.6. El cluster O_4 está compuesto por los peces o_{63}, o_{64} y o_{65} que pertenecen a la especie Pike, y que se caracterizan por tener un gran peso, como se aprecia en la figura 3.5. Del mismo modo, los clusters uno, cinco, seis y ocho, están conformados principalmente por peces de un tamaño pequeño en la muestra, lo que es congruente con su posición en las gráficas.

Para probar el desempeño del procedimiento de cluster MDS propuesto, fueron analizados los 10 centros de los clusters obtenidos sin usar restricciones espaciales usando SMACOF en un procedimiento de dos pasos, que

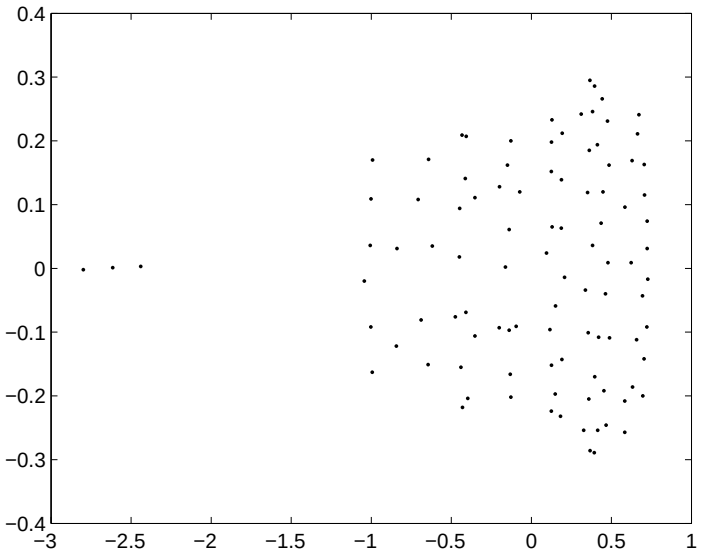

Figura 3.5: Configuración MDS en dos dimensiones para el conjunto de datos de peces.

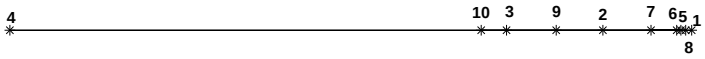

Figura 3.6: Representación clusters latentes-MDS en una dimension para el conjunto de datos de peces.

primero realiza clusters y entonces los centros de los clusters obtenidos son representados mediante MDS en una dimensión. Con la metodología propuesta de cluster MDS se encontró un valor menor del STRESS normalizado de 0.02608 mientra que con el procedimiento de dos pasos el STRESS menor fue de 0.03754, lo que demuestra que la metodología de cluster MDS simultáneo tiene un mejor funcionamiento, como se había indicado antes.

3.7. Conclusiones

En este trabajo se desarrolló un modelo de clases latentes para disimilaridades continuas unimodales a dos vías. Como en el modelo de Cluster Differences Scaling (Heiser y Groenen, 1997), el modelo propuesto nos permite encontrar de manera simultánea la clasificación más apropiada y la representación de los centros de los clusters en un espacio de dimensionalidad baja mediante MDS. Como en metodologías probabilísticas de MDS previas, una de las mayores ventajas del modelo propuesto es que permite la posibilidad de contrastar varias hipótesis en el modelo, donde la selección de número de clases es quizás la más importante. Además, los estadísticos AIC y BIC se pueden emplear para seleccionar la dimensionalidad apropiada para el modelo, siguiendo el principio de parsimonia según lo propuesto por Lee (2001).

La busqueda de la clasificación óptima de los objetos se formula como un problema de partición en bloques en la matriz de disimilaridades, asumiendo que las disimilaridades en un bloque son independientes y normalmente distribuidas, y estableciendo una mezcla de normales univariadas como la distribución de una disimilaridad sin clasificar previamente. La hipótesis de normalidad en las disimilaridades, considerada en este trabajo, es consistente con resultados anteriores en la literatura, pero podrían ser propuestas otras distribuciones de probabilidad para modelar las disimilaridades como la log-normal o la inversa Gaussiana. Sin embargo, en general no existe garantía a priori que asumiendo cualquier otra distribución se puedan obtener resultados más plausibles en el modelo, mientras que en todas las ejecuciones probadas, el modelo normal propuesto parece ser adecuado.

La consideración de una varianza específica por clase en la estructura de partición de las disimilaridades generaliza el caso de la hipótesis de una distribución normal con desviación estandar proporcional a la media, sugerida por Ramsay (1982). De este modo, es posible condicionar los parámetros de

las varianzas de las clases para que sean comunes a todos los bloques en la estructura de clases latentes, para reducir el número de parámetros estimados en el modelo si el número de objetos N, es pequeño en relación al número de clases latentes K.

El algoritmo EM constituye en general una herramienta fundamental en el ajuste de modelos de mezclas por máxima verosimilitud, pero en su forma habitual no es apropiado para el modelo de clases latentes propuesto, debido a que no garantiza que la partición dada en las disimilaridades esté relacionada a la clasificación en el espacio de objetos. Como se mencionó en la sección 3.3.1, diferentes valores iniciales para el algoritmo EM conducen a diferentes estimaciones. La convergencia con el algoritmo EM es lenta y la situación se exacerbará si los valores iniciales de los parámetros son elegidos deficientemente. Además, en varias situaciones, la secuencia de estimaciones generada por el algoritmo EM podría diverger si los valores iniciales de los parámetros son elegidos muy cercanos a la frontera (ver McLachlan y Peel, 2001). Finalmente, otro problema importante con los modelos de mezclas es que la ecuación de verosimilitud tiene generalmente múltiples máximos locales y entonces, se recomienda un algoritmo heurístico de optimización que evite óptimos locales. El desarrollo de Annealing Simulado en este contexto, nos permite encontrar cada vez, una partición en bloques en la matriz de disimilaridades, examinando directamente la partición en el espacio de objetos. Esta también es una estrategia recomendable para tratar con el problema de óptimos locales.

El alto costo de calculos intrínseco a cualquier procedimiento combinatorio de optimización heurística comparado a otros procedimientos de optimización, constituye la principal desventaja de SA, que puede ser resulta parcialmente mediante el uso de fórmulas de reducción en los cálculos, para evitar que el tiempo CPU crezca cuando N y K aumentan. De esta forma, se pueden explorar implementaciones paralelas de SA en estudios futuros sobre este aspecto. Como en el procedimiento EM, el problema es especialmente relevante en el *bootstrapping* de la prueba estadística de la razón de verosimilitud para determinar el número de clases latentes en el modelo, y se podrían investigar otros criterios de información para la selección del modelo para reducir el tiempo CPU.

El procedimiento propuesto también se puede considerar para conjuntos de datos a tres vías bimodales, para determinar simultáneamente grupos de objetos y/o sujetos similares (ver por ejemplo, Brusco y Cradit, 2005), mientras que los centros de los clusters son representados es un espacio de baja

dimensión, este tema está siendo actualmente investigado por los autores.

Capítulo 4

Un modelo de clases latentes MDS con restricciones espaciales para la estimación de la covarianza espacial no estacionaria

4.1. Introducción

Consideremos una función aleatoria, $\mathcal{Z}(x,t)$, observada repetidamente en el tiempo t_i, $i = 1,2,...,T$ en un número finito de estaciones muestrales, x_i, $i = 1,2,...,N$, en el plano, suponiendo estacionaridad temporal pero no espacial. Usando el variograma muestral, como una métrica natural para la estructura de covarianza espacial $D^2(x_i,x_j) = Var(\mathcal{Z}_{it} - \mathcal{Z}_{jt})$, donde $\mathcal{Z}_{it}$ y $\mathcal{Z}_{jt}$ representan las observaciones en las estaciones x_i y x_j (centradas por la media) en el tiempo t, y basados en una version ponderada del método MDS no métrico (Kruskal, 1964a, 1964b), Sampson y Guttorp (1992) propusieron modelar la dispersión espacial como una función general *suave* de las coordenadas geográficas de las estaciones.

De la matriz muestral $N \times T$, $\mathcal{Z}$, podemos obtener la matriz $N \times N$ de las covarianzas muestrales s_{ij}, $i,j = 1,...,N$, entre las filas centradas (estaciones) de $\mathcal{Z}$, denotada por $S = (1/T)\mathcal{Z}H_T\mathcal{Z}'$, donde $H_T = I - (1/T)\mathbf{1}\mathbf{1}'$ y $\mathbf{1}$ es un vector columna de longitud T, y así $\delta_{ij}^2 = Var(Z_{it} - Z_{jt}) = s_{ii} + s_{jj} - 2s_{ij}$, de-

nota una medida de disimilaridad para la dispersión espacial entre estaciones. Basado en la raíz del variograma δ_{ij}, el método de Sampson y Guttorp (1992) usa las distancias entre puntos derivadas de la configuracion MDS no métrico de Kruskal en dos dimensiones, como una medida suavizada de la dispersión espacial de la muestra. Entonces se ajusta una apropiada función al gráfico de dispersión de las disimilaridades versus las distancias MDS para obtener un modelo condicional definido no positivo para las disimilaridades espaciales, relacionando luego las coordenadas geográficas con la configuración MDS de las estaciones mediante un procedimiento de interpolación *thin-plane spline* (ver también Arbia y Lafratta 2002, para otras aplicaciones).

Løland y Høst (2003), proposieron un enfoque diferente no estacionario para un modelo de covarianza espacial, que a diferencia del método de Sampson y Guttorp (1992), no se basa en observaciones repetidas de una red de monitoreo. Mediante MDS clásico, se obtiene una configuración Euclídea a partir de una distancia acuática aproximada previamente dada, usando interpolación lineal para situar las localizaciones de los datos en el espacio MDS.

Según lo precisado recientemente por Vera, Macías y Angulo (2008), cuando el número de localizaciones muestrales se incrementa, la configuración MDS en un espacio de dimensión baja podría no ser conveniente, incluso para un procedimiento de MDS no-métrico. Entonces es necesario reducir el número de localizaciones muestrales o aumentar el número de dimensiones de la representación para alcanzar una solución adecuada. Por otro lado, cuando las localizaciones muestrales están distribuida irregularmente sobre el dominio, algunas vecindades locales pueden ser sobremuestreadas y otras submuestreadas (Kovitz y Christakos, 2004); para datos recolectados en un mallado regular estrecho, en algunas situaciones podrían aparecer clusters espaciales debido a la dependencia heterogénea espacial.

Para tratar este problema, Vera, Macías y Angulo (2008) propusieron un modelo para la estimación de la estructura de covarianza espacial no estacionaria, basado en la inclusión de restricciones espaciales geográficas en el modelo CDS de Heiser y Groenen (1997). De esta forma, no serán las estaciones originales sino los centros de los clusters los que son representados, mientras que las estaciones y clusters mantienen sus relaciones espaciales. El método integra simultáneamente cluster y MDS desde una perspectiva exploratoria, proponiendo un procedimiento puramente descriptivo (i.e., sin hacer referencia a una distribución específica) para determinar la idoneidad de la suposición espacial y el número apropiado de clusters, basado en una

descomposición de la suma de cuadrados de las disimilaridades en contribuciones de varias fuentes de variación. Aunque los resultados obtenidos fueron muy satisfactorios en todos los conjuntos de datos analizados, quizás la falta de una naturaleza estocástica subyacente dificulta tomar decisiones sobre los parámetros estructurales del modelo.

Suponiendo una distribución normal para las disimilaridades, Vera, Macías y Heiser (2007) propusieron un modelo probabilístico de cluster-MDS unimodal a dos vías en el cual la clasificación óptima, la dispersión asociada de los clusters y la configuración de los centros de los clusters son estimadas por máxima verosimilitud. Se desarrolló un modelo de clases latentes en el espacio de las disimilaridades de tal forma que los parámetros son estimados condicionalmente a la clasificación de objetos, mediante un algoritmo basado en Annealing Simulado (SA). La hipótesis de normalidad en las disimilaridades significa que minimizando la función de pérdida de mínimos cuadrados equivalente se obtienen estimadores de máxima verosimilitud. Sin embargo, considerando la no negatividad y el incremento usual en propagación con la localización de δ_{ij}, la distribución lognormal puede ser un modelo estocástico apropiado para la variabilidad residual (ver Ramsay 1982). Esta situación puede ser de especial interés en los modelos en los cuales las disimilaridades no se obtienen directamente del variograma muestral, sino que son aproximadas directamente por MDS como en Løland y Høst (2003). Además, el modelo lognormal no está tan influenciado por disimilaridades grandes como el modelo normal, y las disimilaridades cercanas a cero son tan importantes como las grandes para la solución.

En este trabajo proponemos un modelo general de clases latentes con restricciones espaciales, que bajo la suposición de que las disimilaridades siguen una distribución normal o lognormal nos permite particionar las estaciones muestrales en clases y representar simultáneamente los centros de los clusters en un espacio de dimensionalidad baja, mientras que las estaciones y los clusters conservan sus relaciones espaciales. El modelo introduce una modificación del modelo de clases latentes de Vera, Macías y Heiser (2007) incluyendo restricciones de contigüidad espacial bajo una mezcla de distribuciones normales o lognormales; se propone una estrategia de selección de modelo para probar el número de clases latentes y la dimensionalidad del problema. Al igual que en el modelo de mínimos cuadrados, este modelo probabilístico métrico ofrece una solución para la estimación de una medida suavizada de dispersión espacial cuando el número de estaciones es alto como para obtener una representación MDS en dos dimensiones o cuando está

presente un esquema natural de agrupamiento, en el plano geográfico y /o en el espacio de dispersión. A diferencia del enfoque exploratorio previo, la naturaleza estocástica de este modelo proporciona la valiosa posibilidad de probar hipótesis sobre los parámetros estructurales del modelo.

Por lo tanto, la aplicación del procedimiento de estimación no paramétrico consistiría en tres pasos. Primero, bajo las restricciones de contigüidad espacial, se determina el número de clases suponiendo una distribución normal o lognormal para los componentes de la mezcla, pero sin considerar las restricciones geométricas. Segundo, se imponen las restricciones geométricas para el número dado de clases latentes y se determina la dimensionalidad apropiada del problema, estimándose los parámetros del modelo. Tercero, las distancias Euclídeas entre los centros de los clusters de la configuración se puede usar como un estimador isotrópico y estacionario de la dispersión espacial, y entonces se puede emplear el método de Sampson y Guttorp (1992), el método de interpolación de Løland y Høst (2003), o cualquier otro procedimiento similar para la estimación no paramétrica de la estructura de covarianza.

4.2. Un modelo general de mezclas de disimilaridades con restricciones de contigüidad espacial

Sea $O = \{o_1, \ldots, o_N\}$ un conjunto de N estaciones, y $\Delta = (\delta_{ij})$, $i, j = 1 \ldots N$, una matriz simétrica de disimilaridades entre ellos. Considerando $\mathcal{P}(O)$, una partición en el espacio de los objetos O en un numero pequeño $K (K \ll N)$ de clases latentes, es decir, cada estación pertenece a uno y solo uno de los K subconjuntos O_k, $k = 1 \ldots, K$, con n_k elementos y $n_1 + \cdots + n_K = N$, sin conocer de antemano a que clase latente pertenece un objeto particular o_i.

De $\mathcal{P}(O)$, podemos derivar una partición $\mathcal{P}(\Delta)$ de la matriz de disimilaridades Δ en K^2 bloques, $\Delta_{kl} = (\delta_{ij})$, de dimensión $n_k \times n_l$,$k, l = 1, ..., K$, con $o_i \in O_k$ y $o_j \in O_l$, $i, j = 1, ..., N$, donde Δ_{kl} es la matriz de todas las disimilaridades entre objetos en la clase latente O_k y en la clase latente O_l, respectivamente. Entonces, consideramos los $K \times (K+1)/2$ bloques Δ_{kl}, $k \leq l$, en la matriz triangular inferior Δ, si los bloques de la diagonal son tambien tomados en cuenta. Entonces, se asume que cada disimilaridad δ_{ij},$i < j$,

pertenece a exactamente un bloque latente Δ_{kl}, $k \leq l$, pero sin conocer de antemano a cual bloque latente pertenece una disimilaridad particular δ_{ij}. Denotamos por λ_{kl} a la probabilidad incondicional de que $\delta_{ij} \in \Delta_{kl}$, con

$$\sum_{k \leq l} \lambda_{kl} = 1. \tag{4.1}$$

En el modelo, los centros de los clusters son representados por una matriz de configuración $X = (x_{km})$, $K \times M$, en un espacio Euclídeo métrico de baja dimensión ($M \leq K$), usualmente $M = 2$. Entonces se asume que mientras la distancia es constante dentro de cada bloque, las disimilaridades correspondientes variarán aleatoriamente dentro del mismo bloque, siguiendo una distribución normal o lognormal, de parámetros μ_{kl} y σ^2_{kl}, que dependen de cada bloque.

La suposición de independencia entre pares de estaciones dentro de cada clase latente es congruente con la suposición de independencia local del análisis clásico de clases latentes, y la suposición de independecia local no implica independencia global entre los pares de estaciones. Así, las medias de los bloques de las disimilaridades originales o del logaritmo de las disimilaridades estarán relacionadas geométricamente a los centros de los clusters ajustando $\mu_{kl} = d_{kl}$ o $\mu_{kl} = logd_{kl}$, en el modelo de mezclas normales o lognormales, respectivamente, donde $d_{kl} = d(x_k, x_l)$ representa la distancia Euclídea entre los centros de los clusters.

Debido a que no se conoce de antemano el bloque o clase latente al que pertenece una disimilaridad, la f.d.p. de la variable aleatoria δ_{ij} se convierte en una mezcla finita de densidades normales univariadas $N(\mu_{kl}, \sigma^2_{kl})$ o en una mezcla finita de densidades lognormales univariadas $\Lambda(\mu_{kl}, \sigma^2_{kl})$, adoptando la expresión,

$$g(\delta_{ij} \mid X, \Sigma, \lambda) = \sum_{k \leq l} \lambda_{kl} f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma^2_{kl}), \tag{4.2}$$

donde f_{kl} denota la función de densidad de probabilidad correspondiente en la mezcla, $\Sigma = (\sigma^2_{kl})$ denota la matriz $K \times K$ de los parámetros de dispersión dentro de los bloques, y $\lambda = (\lambda_{kl})$ denota el vector columna $K \times (K+1)/2$ de proporciones de mezclas, $k \leq l$, $k, l = 1, \ldots, K$. Entonces, la función de log-verosimilitud se puede escribir como

$$\log L(X, \Sigma, \lambda \mid \Delta) = \sum_{i<j} \log(g(\delta_{ij}|X, \Sigma, \lambda)). \tag{4.3}$$

Denotando por $\widetilde{\delta} = \delta$, o $\widetilde{\delta} = log\delta$, en el modelo normal o lognormal, respectivamente, los estimadores de máxima verosimilitud de los parámetros, bajo la condición impuesta por (4.1), están dados por

$$\widehat{\lambda}_{kl} = \frac{\sum_{i<j} \pi_{ij,kl}}{N(N-1)/2} \tag{4.4}$$

$$\widehat{\mu}_{kl} = \frac{\sum_{i<j} \pi_{ij,kl} \widetilde{\delta}_{ij}}{\sum_{i<j} \pi_{ij,kl}} \tag{4.5}$$

$$\widehat{\sigma}^2_{kl} = \frac{\sum_{i<j} \pi_{ij,kl} (\widetilde{\delta}_{ij} - \widehat{\mu}_{kl})^2}{\sum_{i<j} \pi_{ij,kl}} \tag{4.6}$$

donde

$$\pi_{ij,kl} = \frac{\lambda_{kl} f_{kl}(\delta_{ij})}{\sum_{k\leq l} \lambda_{kl} f_{kl}(\delta_{ij})}. \tag{4.7}$$

Para resolver las ecuaciones (4.4), (4.5) y (4.6), son necesarios los valores de las probabilidades $\pi_{ij,kl}$, que representan la probabilidad a posteriori de que el valor observado de la disimilaridad δ_{ij} pertenezca a la clase latente Δ_{kl}, pero para obtener estos mediante (4.7), son necesarios los valores estimados de los parámetros. Definamos la matriz $N(N-1)/2 \times K(K+1)/2$, $\mathbf{Z} = (z_{ij,kl})$ de las variables indicadoras de cluster

$$z_{ij,kl} = \begin{cases} 1, & \text{si } \ \delta_{ij} \in \Delta_{kl},\ i < j,\ k \leq l, \\ 0, & \text{otro caso.} \end{cases}$$

tal que

$$\sum_{k\leq l} z_{ij,kl} = 1, \quad \text{y} \quad \sum_{i<j} \sum_{k\leq l} z_{ij,kl} = N(N-1)/2.$$

Puesto que desde un punto de vista práctico, los indicadores $\mathbf{Z}$ son variables no observadas, el algoritmo del modelo de clases latentes es un procedimiento iterativo de estimación condicional de máxima verosimilitud. Entonces, en la s-ésima iteración, los valores de $\widehat{Z}^{(s)}$ se encuentran a partir de una partición de prueba en el espacio de las estaciones, preservando la restricción de contigüidad espacial, y tomando $\widehat{\pi}_{ij,kl} = z_{ij,kl}$, los parámetros restantes son estimados maximizando la ecuación

$$\begin{aligned} logL_c(X, \Sigma, \lambda \mid \Delta, \widehat{Z}^{(s)}) = & \sum_{i<j} \sum_{k \leq l} \widehat{z}_{ij,kl}^{(s)} \log \lambda_{kl} \\ & + \sum_{i<j} \sum_{k \leq l} \widehat{z}_{ij,kl}^{(s)} \log f_{kl}(\delta_{ij} \mid x_k, x_l, \sigma_{kl}^2), \end{aligned} \tag{4.8}$$

mediante un procedimiento basado en Annealing Simulado. Denotando por $\widetilde{d}_{kl}(X) = d_{kl}(X)$, ó $\widetilde{d}_{kl}(X) = log\ d_{kl}(X)$, en el modelo de mezclas normales o lognormales, respectivamente, e imponiendo las restricciones geométricas $\mu_{kl} = \widetilde{d}_{kl}(X)$, la configuración de los centros de los clusters X es estimada maximizando 4.8, o equivalentemente minimizando

$$q(X) = \sum_{i<j} \sum_{k<l} \widehat{z}_{ij,kl}^{(s)} (\widetilde{\delta}_{ij} - \widetilde{d}_{kl})^2, \tag{4.9}$$

de lo cual, usando la descomposición ortogonal de $q(X)$ en un componente dentro de clases y un componente entre clases,

$$q(X) = \sum_{i<j} \sum_{k<l} \widehat{z}_{ij,kl}^{(s)} (\widetilde{\delta}_{ij} - \overline{\delta}_{kl})^2 + \sum_{k<l} \gamma_{kl} (\overline{\delta}_{kl} - \widetilde{d}(x_k, x_l))^2, \tag{4.10}$$

donde

$$\overline{\delta}_{kl} = \frac{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)} \widetilde{\delta}_{ij}}{\sum_{i<j} \widehat{z}_{ij,kl}^{(s)}}, \quad \text{y} \quad \gamma_{kl} = \sum_{i<j} \widehat{z}_{ij,kl}^{(s)},$$

la configuración de los centros de los clusters puede ser estimada minimizando el término final en 4.10, dado por

$$\phi(X) = \sum_{k<l} \gamma_{kl}(\overline{\delta}_{kl} - \widetilde{d}(x_k, x_l))^2, \tag{4.11}$$

usando el algoritmo SMACOF (de Leeuw y Heiser 1980). Finalmente, y para evitar soluciones degeneradas en los bloques diagonales, los parámetros de dispersión dentro de los bloques, σ^2_{kl}, son estimados como

$$\widehat{\sigma}^{2^s}_{kl} = \begin{cases} \dfrac{\sum\limits_{i<j} \widehat{z}^{(s)}_{ij,kl}(\widetilde{\delta}_{ij} - \widetilde{d}(\widehat{x}^{(s)}_k, \widehat{x}^{(s)}_l))^2}{\sum\limits_{i<j} \widehat{z}^{(s)}_{ij,kl}} & \text{si} \quad k < l \\ \dfrac{\sum\limits_{i<j} \widehat{z}^{(s)}_{ij,kk}(\widetilde{\delta}_{ij} - \overline{\delta}_{kk})^2}{\sum\limits_{i<j} \widehat{z}^{(s)}_{ij,kk}} & \text{si} \quad k = l \end{cases} \tag{4.12}$$

4.2.1. Un algoritmo de annealing simulado de clases latentes con restricciones de contigüidad espacial

Para el problema de estimación de parámetros, el algoritmo EM es probablemente el procedimiento más utilizado en el análisis de clases latentes. Sin embargo, en la presente situación, EM o algoritmos similares no pueden ser aplicados directamente a menos que se consideren restricciones en la estimación del parámetro λ_{kl}, asegurando que las probabilidades están asociadas a una partición en bloques en Δ. Así, Vera, Macías y Heiser (2007) propusieron un procedimiento iterativo de estimación condicional de máxima verosimilitud tal que cada vez, una clasificación de prueba, $\mathcal{P}(\Delta)$, y los valores relacionados de Z se derivan de la clasificación de prueba correspondiente $\mathcal{P}(O)$, de los cuales el resto de los parámetros son estimados, todo en un algoritmo basado en SA.

Annealing Simulado es un buscador meta-heurístico que proporciona medios para escapar de óptimos locales permitiendo movimientos de *subir colinas* en la búsqueda del óptimo global. Su nombre procede de una analogía con el proceso físico de calentamiento de sólidos, en el cual primero se calienta un sólido cristalino a una temperatura muy alta, $\mathcal{T}_0$, de tal forma que el sistema alcanza una energía inicial ε_0. Entonces se permite enfriar el sistema muy lentamente hasta que alcanza la configuración de enrejado cristalino

más regular posible, es decir, la energía de enrejado mínima ε_f. La evolución del sistema físico es simulado en optimización, en el cual se genera una secuencia de cadenas de Markov mientras la temperatura decrece, y un nuevo estado del sistema es aceptado mediante la regla de aceptación de Metrópolis (Metropolis et al. 1953). Se elige un punto aleatorio de la vecindad del punto previamente seleccionado, y se evalua la energía del sistema, ε, en este punto. Si ε decrece en $\triangle\varepsilon$, el nuevo punto es aceptado y la temperatura se disminuye, pero si ε aumenta, el nuevo punto es aceptado con una probabilidad de $\exp(-\triangle\varepsilon/T)$, donde $\mathcal{T}$ es la temperatura actual. La temperatura se disminuye y el proceso se repite hasta que el sistema alcanza un equilibrio, así se garantiza que el algoritmo se detiene por lo menos en un mínimo local.

Para incluir las condiciones de contigüidad espacial en el modelo, consideramos una generalización de la matriz de contigüidad definida en Vera, Macías y Angulo (2008) como sigue.

Considerando G la matriz $N \times M^*$ de coordenadas geográficas para N estaciones en dimensión M^* (usualmente, $M^* = 2$), y denotando por D_G a la matriz $N \times N$ con elementos $d_{Gij} = d_G(o_i, o_j)$, i.e. las distancias Euclídeas entre la i-ésima fila y la$j$-ésima fila de G. Para una partición $\mathcal{P}(O)$ en el espacio de estaciones, una estación $o_i \in O_l$ se dice que es *contigua al cluster* O_k, $k \neq l$, si y solo si existe una estación $o_j \in O_k$ que satisface $d_{Gij} = \text{mín}_{h \neq i}\{d_G(o_i, o_h), o_h \notin O_l\}$. Denotando por $U = (u_1, \ldots, u_N)'$ al vector cuyos elementos estan definidos por $u_i = l$, if $o_i \in O_l$, for $i = 1, \ldots, N$, y $l = 1, \ldots, K$, es decir, el vector que contiene los índices de los clusters al cual pertenece cada estación en la partición actual, y consideramos el vector $d_G^- = (d_{G1}^-, \ldots, d_{GN}^-)'$ cuyos elementos son $d_{Gi}^- = \text{mín}_j\{d_{Gij}, o_j \notin O_{u_i}\}$, i.e. el vector columna que contiene la distancia mínima de la i-ésima fila en D_G a todas las estaciones fuera del cluster O_{u_i}.

Con respecto a cada estación o_i, $i = 1, \ldots, N$ consideraremos el conjunto $L_i = \{u_j, j \neq i \mid \exists o_j \in O_{u_j}, d_{Gij} = d_{Gi}^-\}$, es decir, el conjunto de clusters diferentes a u_i en la misma distancia mínima d_{Gi}^- desde o_i. Sobre la base de L_i, para $i = 1, \ldots, N$, podemos construir la matriz $N \times K$, F, cuyos elementos están dados por $\varphi_{il} = d_{Gi}^-$, si $l \in L_i$ y $d_{Gi}^- = \min\limits_{o_j \in O_{u_i}} \{d_{Gj}^- \mid l \in L_j\}$, y $\varphi_{il} = 0$, en otro caso.

Considerando cada estación o_i y su cluster correspondiente, O_{u_i}, podemos definir el conjunto $J_i = \{o_j \in O_{u_i} \mid \varphi_{jl} \neq 0, l = 1, \ldots, K\}$ de todas las estaciones en el mismo cluster O_{u_i}, que podrían ser movidas a cualquier otro cluster. A partir de J_i, para $i = 1, \ldots, N$, podemos definir el conjunto

asociado de restricción de amplitud, Φ_i, como

$$\Phi_i = \left\{ o_j \in J_i \mid \sum_{\substack{o_h, o_y \in O_{u_i} \\ h,y \neq j}} d_{Ghy} = \min_{o_w \in J_i} \sum_{\substack{o_h, o_y \in O_{u_i} \\ h,y \neq w}} d_{Ghy} \right\},$$

que consiste de las estaciones en el cluster O_{u_i} que pueden ser transferidas a otro cluster, sin dividir el cluster original, y finalmente, los elementos de la matriz de contigüidad $C_{\mathcal{P}}$, $N \times K$, se pueden definir como

$$c_{ik} = \begin{cases} k, & \text{si } \varphi_{ik} \neq 0,\ o_i \in \Phi_i \text{ y } |O_{u_i}| > 3 \\ 0, & \text{otro caso.} \end{cases}$$

Por lo tanto, esta matriz de contigüidad $C_{\mathcal{P}}$ tiene tantos vectores filas cero como estaciones que no puedan ser movidas de su partición actual $\mathcal{P}(O)$, los cuales debería ser removidos del modelo. Entonces, definiendo el vector columna $\beta = (b_1, ..., b_N)'$, con $b_i = i$, si la i-ésima fila de $C_{\mathcal{P}}$ no es el vector de ceros, y cero en otro caso, $i = 1, ..., N$, podemos considerar la matriz extendida $N \times (K+1)$, $\widetilde{C}_{\mathcal{P}} = [\beta | C_{\mathcal{P}}]$, resultante de la unión del vector columna β y la matriz $C_{\mathcal{P}}$. Entonces $r_{\mathcal{P}} = rango(diag(\beta))$, es decir, el rango de la matriz $K \times (K)$ dada, expresando el vector β como una matriz diagonal es el número de filas no cero en $C_{\mathcal{P}}$. Denotando por v_i, el i-ésimo vector unitario de longitud N, es decir, $v_{im} = 1$ si $m = i$, y cero en otro caso, $m = 1, ..., N$, y definiendo la matriz $R_{\mathcal{P}}$ cuyos filas está dadas por los $R_{\mathcal{P}}$ vectores unitarios $\{v_i | b_i \neq 0\}$, se puede mostrar que la vecindad de la matriz de contiguidad para la partición actual $\mathcal{P}(O)$ está dada por

$$\vartheta(C_{\mathcal{P}}) = R_{\mathcal{P}} \widetilde{C}_{\mathcal{P}}. \tag{4.13}$$

De esta forma, en cualquier vector fila de la matriz $r_{\mathcal{P}} \times (K+1)$, $\vartheta(C_{\mathcal{P}})$, el primer elemento indica la estación que puede ser reasignada o movida, y cualquier otro elemento distinto de cero representa un cluster candidato al que esta estación puede ser movida.

El algoritmo inicia con la partición inicial de bloques $\mathcal{P}(O)$ en la matriz de disimilaridades Δ, y los valores relacionados de Z, asociados a una partición inicial $\mathcal{P}^{(0)}(O)$ de las N estaciones en K clases. Para preservar las restricciones espaciales, esta clasificación inicial es obtenida usando un procedimiento de clasificación $k-means$ de las estaciones en el espacio geográfico G, bajo la restricción computacional de que $n_k > 2$, $k = 1, ..., K$, que se

mantiene en todo el algoritmo para evitar la presencia de columnas cero en la matriz Z y la presencia de bloques diagonales de varianza cero. Entonces, se inicializan el factor de enfriamiento γ que controla la taza de decremento de la temperatura, y la longitud de truncamiento de la cadena de Markov en cada nivel de temperatura LC, la cual se incrementa por un número fijo IC cada m iteraciones. La temperatura final del sistema $\mathcal{T}_f$ se elige muy cercana a cero, garantizando que el algoritmo termina por lo menos en un óptimo local, y la temperatura inicial $\mathcal{T}_0$ se calcula siguiendo un procedimiento de muestreo aleatorio adoptada en implementaciones previas, por ejemplo en Murillo, Vera y Heiser (2005), o en Vera, Heiser y Murillo (2007). Se fijan un valor de la probabilidad χ, y un máximo Ma_{max} de asignaciones aleatorias para promediar los posibles incrementos Ma de la solución que empeora la logverosimilitud condicional, para obtener una temperatura inicial tal que en las primeras iteraciones son aceptadas el $100\chi\,\%$ de las peores soluciones. En la s-ésima iteración, el algoritmo SA puede describirse esquematicamente como sigue:

1. A partir de una partición $\mathcal{P}^{(s)}(O)$, se calculan las probabilidades a posteriori $\hat{\pi}_{ij,kl}^{(s)}$, el resto de parámetros, λ, Σ y X (usando SMACOF) son estimados dada $\widehat{Z}^{(s)}$, y entonces se evalua la función de log-verosimilitud $log\ L$.

2. Se estiman la matriz de contigüidad $C_{\mathcal{P}^{(s)}}$ y la matriz de vecindad $\vartheta(C_{\mathcal{P}^{(s)}})$. Entonces, se selecciona aleatoriamente una partición de prueba $\mathcal{P}^{(s+1)}(O)$ usando el siguiente procedimiento: primero, se selecciona al azar una fila v de $\vartheta(C_{\mathcal{P}^{(s)}})$. Segundo, la estación o_{v_l} se selecciona para ser movida al cluster indicado por el valor v_j distinto de cero seleccionado aleatoriamente, $j = 2, ..., (K+1)$, si la condición no degenerada se sostiene. De otra forma, se selecciona una nueva fila de $\vartheta(C_{\mathcal{P}^{(s)}})$, y el proceso se repite hasta que se encuentra una nueva partición de prueba $\mathcal{P}^{(s+1)}(O)$.

3. Los nuevos parámetros son estimados de $\mathcal{P}^{(s+1)}(O)$, y se evalua el incremento en la log-verosimilitud, tal que si $\Delta logL > 0$, la partición es aceptada. Inversamente, la nueva partición es aceptada con una probabilidad $\exp(\triangle logL/\mathcal{T}^{(s)})$, usando la regla de aceptación de Metrópolis.

El proceso anterior se repite en un ciclo interno de longitud creciente LC, alcanzando una partición definitiva $\mathcal{P}^{(s+1)}(O)$. Entonces, la temperatura

$\mathcal{T} = \gamma \times \mathcal{T}$ se disminuye, y el proceso continua hasta que se alcanza un criterio de convergencia. Como en otras implementaciones de SA descritas por los autores, se probaron dos criterios de finalización. Uno, evaluando cuando se alcanza el número máximo de iteraciones dado por $It_{\text{máx}} = log((\mathcal{T}_f/\mathcal{T}_0) - \gamma)$, y el otro, cuando una solución permanece inalterada durante un número de iteraciones previamente establecidas, $R_{\text{máx}}$, asegurando que el parámetro de la temperatura final $\mathcal{T}_f$, está cercano a cero.

4.2.2. Una estrategia de selección del modelo

Dos decisiones principales se deben adoptar en la formulación del modelo. Una se refiere al número de clases latentes, siendo este en general un problema inherente al modelado de mezclas finitas. La otra es determinar el número apropiado de dimensiones en la representación MDS de los centros de los clusters. Aunque se prefiere normalmente una representación planar, el presente algoritmo nos permite determinar el número de dimensiones más apropiado.

El número de clases latentes se determina del número de componentes de la mezcla, siendo esto una tarea importante que todavía no se ha solucionado totalmente (Yang y Yang 2007). Es bien conocido, que el resultado clásico que proporciona una distribución asintótica chi cuadrada no es válido en este contexto. Una alternativa es usar un enfoque bootstrap como en Vera, Macías y Heiser (2007). Sin embargo, para evitar el incremento en el tiempo CPU asociado con la naturaleza misma del procedimiento bootstrap, empleamos el criterio de informacion Bayesiano BIC (Schwarz 1978), usando el ajuste en el tamaño de la muestra sugerido por Rissanen (1978) para mejorar su funcionamiento en este contexto (Yang y Yang 2007). De esta forma, el criterio de información propuesto, denotado por BIC*, está dado por,

$$BIC^* = -2logL_c(X, \Sigma, \lambda|\Delta, Z) + dlog((p+2)/24), \tag{4.14}$$

donde $p = N(N-1)/2$, y $d = (3K(K+1)/2) - 1$ denota el número de parámetros desconocidos en el modelo, si las condiciones geométricas no son consideradas. El criterio de información BIC* es usado de nuevo para probar la dimensionalidad de la solución MDS, en la cual $d = K(K+1+M) - (M(M+1)/2) - 1$ describe el número de parámetros desconocidos en el modelo cuando se estima la configuración MDS. Al igual que la prueba de razón de verosimilitud, las condiciones de regularidad no se cumplen para la

justificación teórica del BIC, según lo derivado originalmente. Sin embargo, varios estudios recientes han revelado la considerable ayuda de su uso en este contexto, y Rissanen (1986, 1989) incluso derivó la formulación del BIC en el contexto de selección del modelo, desde una perspectiva distinta basada en teoría de codificación de información (ver McLachlan y Peel 2001, Secc. 6.9.3, para más detalles).

Por lo tanto, basado en el criterio BIC* y bajo restricciones espaciales, fue aplicada la siguiente estrategia para seleccionar el modelo apropiado de clases latentes. Primero, se determina el número de clases latentes como el valor de K relacionado con el valor más bajo del estadístico BIC*, cuando no son tomadas en cuenta restricciones geométricas. Después de fijar el número de cluster, se imponen las restricciones geométricas y la dimensionalidad de la representación se determina de nuevo mediante el valor más bajo del criterio BIC* cuando los valores de M aumentan, condicionado al número de componentes de la mezcla fijado previamente. Como en Dasgupta y Raftery (1998), se propone una regla para determinar un valor significativamente mas bajo del BIC*; diferencias que excedan $0.005 \cdot \text{mín}(BIC^*)$ entre dos valores del BIC* son consideradas como una fuerte evidencia de que el modelo correspondiente al valor más bajo del BIC* es el mejor modelo ajustado.

4.3. Aplicaciones

Para ilustrar el algoritmo propuesto, se analizaron tres conjuntos de datos. Los primeros dos correponden al variograma muestral, bajo la suposición de que los componentes de la mezcla siguen una distribución normal, situando a las estaciones en un mallado o red regular y en un dominio distribuido irregularmente, respectivamente. El tercer ejemplo corresponde al análisis bajo la distribución lognormal de las distancias acuáticas analizadas por Løland y Høst (2003).

El algoritmo propuesto fue implementado en Fortran, trabajando en una PC Pentium(R) IV 3.00 GHz con 1 Gb de RAM bajo ambiente Microsoft Windows XP. Como en otras implementaciones previas de SA reportadas por los autores, se eligió el mejor óptimo local en 20 réplicas independientes como la mejor solución, usando para la fase de la temperatura inicial los valores de los parámetros: χ=0.95, $Ma = 50K$ y $Ma_{max} = 500$, y los valores de γ=0.95, $T_f = 10^{-7}$, $R_{max} = 10$, $LC = 5N$, $IC = 50N$ y $m = 10$ para los demás parámetros. Para el procedimiento SMACOF en la etapa de la estimación

de la configuración, se eligió la solución clásica como la configuración inicial, utilizandose como criterios de convergencia un maximo de 200 iteraciones así como una diferencia menor a 10^{-7} en valores consecutivos del STRESS.

El primer conjunto de datos analizados corresponde a la precipitación promedio mensual (en ug S/m^3) de dioxido de sulfuro(SO2) en al año 2002, medidas en 59 estaciones de Europa central. Los datos y la localización geográfica de las estaciones fueron proporcionados por *Cooperative programme for monitoring and evaluation of the long-range transmissions of air pollutants in Europe*(*EMEP*), en la dirección electrónica http://www.emep.int/. Aunque el periodo de tiempo considerado fue elegido específicamente debido a la buena disponibilidad de observaciones, todavía se encontraron datos faltantes para algunas estaciones, y fue diseñado un mallado de 5×6 nodos sobre la superficie geográfica analizada, como se muestra en la figura 4.1. La precipitación promedio registrada por estaciones que no presentaron valores faltantes en cada celda fue tomada para representar la precipitación mensual de dióxido de sulfuro en el área geográfica abarcada por la celda. Por lo tanto, obtenemos una matriz centrada $\mathcal{Z}$ de tamaño 30×12, tomando el centro geográfico de las celdas como las localizaciones geográficas, de lo cual la matriz raíz-variograma Δ, de tamaño 30×30 es dada como una medida de la dispersión espacial.

Suponiendo una distribución normal de los componentes de la mezcla (Davis y Borgman 1982), los datos son primero analizados sin tomar en cuenta las restricciones geométricas para $K = 1, ..., 7$ clases latentes, lo cual produce un valor mínimo del estadístico BIC* para $K = 3$ clusters, como se muestra en la tabla 4.1. Considerando las restricciones geométricas en dos dimensiones para una mezcla de 6 componentes, la tabla 4.2 muestra los valores estimados de los parámetros, donde los valores $\widehat{\mu}_{kl} = d_{kl}, k < l$ indican la distancia geométrica entre los centros de los clusters dados por el variograma.

En el panel izquierdo de la figura 4.2 se muestra la estructura de particion sobre el mallado muestral, donde el signo $\times$ representa el centro geográfico de los clusters, y en el panel derecho se presenta la configuración MDS de los centros de los cluster en dos dimensiones. Se puede ver que existen diferentes relaciones entre las distancias geográficas y geométricas de los centros de los clusters. El primer cluster corresponde aproximadamente al área desde Irlanda al Reino Unido, el segundo al noroeste del área analizada y el tercero a las estaciones de monitoreo en el suroeste.

Analizamos un segundo conjunto de datos agrupados naturalmente, tam-

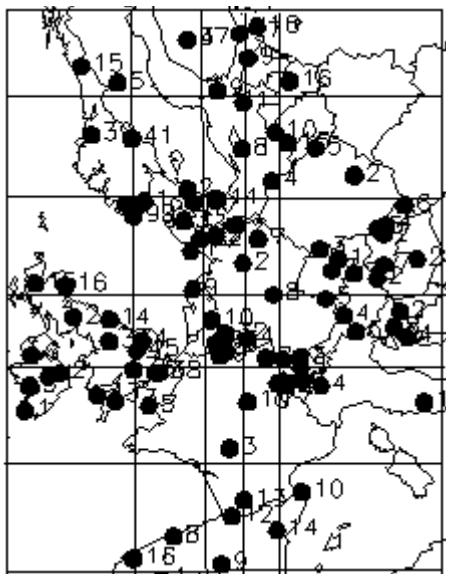

Figura 4.1: Mallado muestral para los datos de dióxido de sulfuro.

Tabla 4.1: Resultados del criterio BIC* para probar el número de clases para los datos de dióxido de sulfuro.

Análisis sin considerar restricciones geométricas			
No. de clases (K)	gl	log-verosimilitud	BIC*
1	2	-385.98	777.76
2	8	-334.09	691.40
3	17	-302.27	653.87
4	29	-295.58	675.31
5	44	-301.07	729.82
6	62	-264.94	709.79
7	83	-263.69	768.24

Tabla 4.2: Valores estimados de los parámetros en dos dimensiones para los datos de dióxido de sulfuro, donde μ_{kk} representa $\overline{\delta}_{kk}$.

Parámetros estimados			
Δ_{kl}	$\widehat{\lambda}_{kl}$	$\widehat{\mu}_{kl}$	$\widehat{\sigma}^2_{kl}$
Δ_{11}	0.023	0.516	0.038
Δ_{12}	0.126	0.443	0.045
Δ_{13}	0.161	0.884	0.324
Δ_{22}	0.126	0.386	0.032
Δ_{23}	0.354	0.853	0.312
Δ_{33}	0.210	1.186	0.473

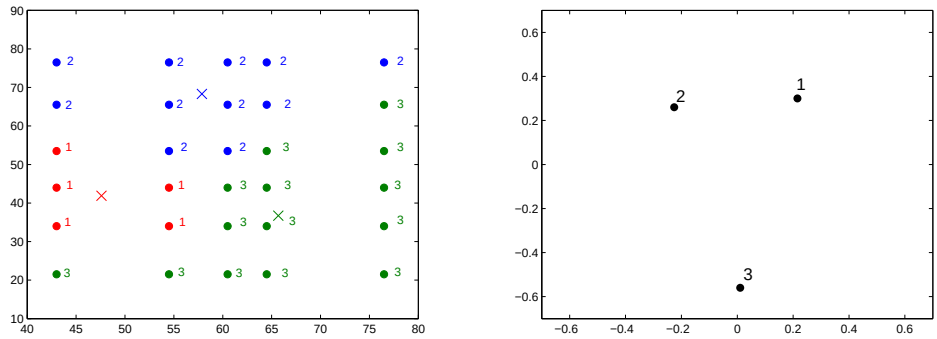

Figura 4.2: Estructura de partición para tres clusters con restricciones espaciales mostradas sobre el plano geográfico original (panel izquierdo), y la representación MDS de las clases latentes en dos dimensiones (panel derecho) para los datos de dióxido de sulfuro. El signo $\times$ en el panel izquierdo indica la localización geográfica de los centros de los clusters.

Tabla 4.3: Resultados del criterio BIC* para probar el número de clases para los datos de zinc.

Análisis sin condiciones geomátricas			
No. de Clases (K)	gl	log L	BIC*
1	2	-875.44	1757.53
2	8	-561.94	1150.49
3	17	-225.74	508.03
4	29	-27.89	152.24
5	44	45.62	55.11
6	62	92.62	20.99
7	83	91.03	94.02
8	107	141.78	72.34

bién de la base de datos de EMEP, basado en el variograma y en una mezcla de componentes normales. La matriz muestral $\mathcal{Z}$ de tamaño 37×12, corresponde a la precipitación mensual promedio (en ug/l) de zinc, medida por 37 estaciones de monitoreo EMEP distribuidas en Europa en el año 2004. La tabla 4.3 muestra los resultados de la bondad de ajuste de acuerdo al criterio de información BIC* cuando se aplica el modelo sin restricciones geométricas para $K = 1, \ldots, 8$ clases latentes, de lo cual se puede inferir que una mezcla de 21 componentes ($K = 6$) puede ser adecuada para el variograma.

Considerando las restricciones geométricas para el modelo de seis clases latentes, la tabla 4.4 muestra los valores BIC* para probar una adecuada dimensionalidad para la representación. De esta forma, un plano parece adecuado para explicar el modelo de seis clases latentes, de lo cual se obtienen los valores estimados de los parámetros en dos dimensiones, dados en la tabla 4.5. La inclusión de restricciones espaciales en el modelo de clases latentes incrementa la dificultad de obtener una clasificación óptima globalmente, con respecto a la qué se podría obtener sin restricciones. Así, y como en el ejemplo previo, algunas de las estimaciones $\hat{\mu}$ para los bloques diagonales presentan un valor más alto que las estimaciones de los bloques no diagonales, como es esperado.

La figura 4.3 muestra los clusters sobre el plano geográfico original (panel izquierdo) y la configuración MDS de las clases latentes en dos dimensiones para los centros de los clusters asociados (panel derecho). Existen diferencias evidentes entre las estaciones de monitoreo localizadas en el norte, centro y sur de Europa. En la parte superior del mapa podemos ver los clusters dos y

Tabla 4.4: Resultados del criterio de información para el modelo de $K = 6$ clases cuando los centros de los clusters son restringidos a ser escalados en dos y en tres dimensiones para los datos de zinc.

Análisis de 5-clases con restricciones sobre las medias de las clases				
No. de Clases (K)	dimensión	gl	log L	BIC*
6	2	50	84.68	-3.05
6	3	53	79.72	16.85

Tabla 4.5: Valores estimados de los parametros en dos dimensiones para los datos de zinc, donde μ_{kk} representa $\bar{\delta}_{kk}$.

Parámetros estimados			
Δ_{kl}	$\hat{\lambda}_{kl}$	$\hat{\mu}_{kl}$	$\hat{\sigma}^2_{kl}$
Δ_{11}	0.023	0.044	0.0001
Δ_{12}	0.036	0.492	0.4072
Δ_{13}	0.036	0.065	0.0008
Δ_{14}	0.081	0.059	0.0038
Δ_{15}	0.081	0.092	0.0069
Δ_{16}	0.045	0.995	2.1385
Δ_{22}	0.009	0.850	0.5317
Δ_{23}	0.024	0.434	0.3947
Δ_{24}	0.054	0.451	0.3993
Δ_{25}	0.054	0.521	0.3625
Δ_{26}	0.030	1.331	2.0991
Δ_{33}	0.009	0.094	0.0010
Δ_{34}	0.054	0.071	0.0039
Δ_{35}	0.054	0.101	0.0057
Δ_{36}	0.030	1.051	2.1071
Δ_{44}	0.054	0.112	0.0058
Δ_{45}	0.122	0.146	0.0077
Δ_{46}	0.068	0.986	2.1002
Δ_{55}	0.054	0.156	0.0080
Δ_{56}	0.068	1.060	2.0963
Δ_{66}	0.015	1.869	2.8978

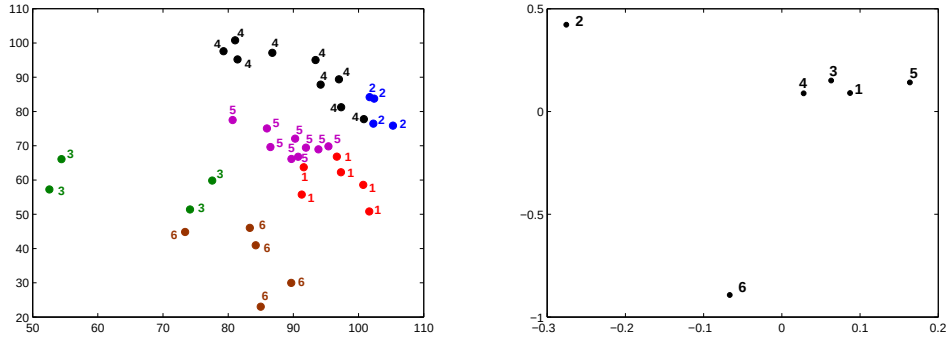

Figura 4.3: Estructura de partición asociada a $K = 6$ mostrada sobre el plano geográfico original (panel izquierdo) y la representación MDS de clusters latentes en dos dimensiones (panel derecho) para los datos de zinc.

cuatro, que están compuestos por estaciones localizadas en Noruega, Suecia, Finlandia, Estonia, Latvia y Polonia. En el centro del mapa se observan los clusters uno y cinco formado por estaciones localizadas en Alemania, Holanda y Dinamarca, seguido del cluster tres cuyas estaciones pertenecen a Islandia y al norte de Inglaterra. En la parte inferior se localiza el cluster seis, que está compuesto por estaciones que pertenecen a Francia, España, Portugal y al sur del Reino Unido.

Siguiendo la estructura de partición y la representación de los centros de los cluster descritas anteriormente, podemos aplicar la metodología de interpolación de Sampson y Guttorp (1992) o cualquier otro procedimiento para la estimación no paramétrica de la estructura de covarianza. El procedimiento propuesto asegura una estimación estacionaria e isotrópica de la estructura de covarianza.

Finalmente, usando la distribución lognormal para los componentes de la mezcla, analizamos las distancias acuáticas dadas por Løland y Høst (2003). La matriz de datos $\mathcal{Z}$ está constituida de mediciones del arenque noruego spring-spawning recogidas del sistema Vestfjord durante diciembre de 1996. Usando un mallado triangular de 123 nodos de los fiordos en el espacio geográfico, se calculó una matriz de distancias acuáticas mediante la *búsqueda gráfica* de la trayectoria mas corta de Dijkstra (ver Løland y Høst, 2003, para más detalles). Bajo restricciones geográficas, un modelo de mezclas de

Tabla 4.6: Resultados del criterio BIC* para probar el número de clases para los datos del sistema Vestfjord imponiendo restricciones geográficas.

Análisis sin restricciones geométricas			
No. de Clases (K)	gl	log L	BIC*
1	2	-5729.88	11471.25
2	8	-5464.17	10974.30
3	17	-5205.12	10507.91
4	29	-5144.10	10454.81
5	44	-5085.76	10424.31
6	62	-5071.61	10499.43
7	83	-5053.04	10582.94

15 componentes lognormales ($K = 5$) parece ser apropiado para los datos de distancias acuáticas como se muestra en la tabla 4.6. Sin embargo como se puede apreciar, existe una diferencia menor entre los valores del BIC* para $K = 4$ y $K = 5$ (mas pequeño que $5^{-3} \cdot \text{BIC}^*_{K=5}$).

La dimensionalidad de la representación MDS es probada considerando $\mu_{kl} = log\ d_{kl}$, para el modelo de cinco clases latentes. Como se muestra en la tabla 4.7, el valor más bajo del estadístico BIC* se obtiene en tres dimensiones. Sin embargo, desde un punto de vista práctico no existen diferencias significativas (más grande que 52) entre el valor más bajo del BIC* y los otros valores. De esta forma, un modelo de cinco clases latentes representado en dos dimensiones parece ser la mejor opción para las distancias acuáticas. Debido a que existen solamente diferencias menores entre las distancias acuáticas exactas y las aproximadas dadas por cualquier configuración MDS, en general se esperaría que las restricciones geográficas tuvieran solamente una influencia leve en la configuración MDS de los clusters. Quizas la excepción puede ser apreciada en clusters muy unidos como los clusters uno y cinco, los cuales presentan un valor pequeño del parámetro estimado $\exp(\hat{\mu}_{15})$ en la tabla 4.8. La consideración de restricciones geométricas determina la forma de las fronteras de estos clusters.

Los cinco convex hulls en el panel izquierdo de la figura 4.4 representan las áreas comprendidas por el modelo de clases latentes estimado, representando con el símbolo $\times$ el centro geográfico de los clusters. El panel derecho muestra la representación procrustes de la configuración Euclídea MDS de los centros de los clusters en el modelo de cinco clases latentes y la configuración de los centros geográficos. Como se mencionó antes, diferencias menores entre las

Tabla 4.7: Valores del criterio de información BIC* para el modelo de cinco clases cuando los centros de los clusters son restringidos a ser escalados en dos, tres y cuatro dimensiones, para los datos del sistema Vestfjord .

Resultados para el modelo restringido de cinco clases latentes				
No. de Clases (K)	dimensión	gl	log L	BIC*
5	2	36	-5110.1126	10427.05
5	3	38	-5090.8793	10400.08
5	4	39	-5090.8732	10405.81

distancias acuáticas exactas y aproximadas significa que, aunque pocas, son evidentes diferencias significantes entre los valores estimados $\exp(\hat{\mu}_{kl})$ y las distancias entre los centros geográficos de los clusters. Está situación también se puede inferir de la suma de cuadrados de errores de 0.1131 relacionada a la representación procrustes, cuando son comparadas las dos configuraciones usando el procedimiento de procrustes en MATLAB.

Además de la configuración MDS de los centros de los clusters, la clasificación dada de los nodos del mallado es óptima tanto en el espacio original de las distancias acuáticas como en el espacio MDS dado. Así, una clasificación óptima basada en distancias acuáticas es también obtenida para las estaciones, donde toda las estaciones pertenecientes al mismo mismo convex hull, y las obtenidas del procedimiento de cluster-MDS, están relacionadas. Por lo tanto, un valor estimado de la variable $\mathcal{Z}$ podría ser asignado a los centros de los clusters mediante interpolación, o por cualquier otro procedimiento de estimación apropiado basado en información relacionada a las estaciones agrupadas, de lo cual se puede estimar el variograma.

4.4. Conclusiones

Escalamiento Multidimensional proporciona una metodología no estacionaria para la estimación de la estructura de covarianza espacial en el marco del análisis de procesos ambientales espacio-temporales. Se puede obtener una configuración Euclídea no únicamente a partir del variograma basado en la muestra de observaciones repetidas de una red de monitoreo (Sampson y Guttorp, 1992), sino siendo calculadas directamente de una matriz dada de distancias aproximadas (Løland y Høst, 2003).

Sin embargo, para un número grande de localizaciones de la muestra o

Tabla 4.8: Valores estimados de los parámetros para los datos del sistema Vestfjord para $K = 5$ en dos dimensiones ($\hat{\mu}_{kk}$ es representado por $\bar{\delta}_{kk}$).

Estimación de parámetros			
Δ_{kl}	$\hat{\lambda}_{kl}$	$\exp(\hat{\mu}_{kl})$	$\hat{\sigma}^2_{kl}$
Δ_{11}	0.079	0.237	0.503
Δ_{12}	0.084	1.347	0.049
Δ_{13}	0.065	0.910	0.164
Δ_{14}	0.098	0.517	0.163
Δ_{15}	0.163	0.448	0.121
Δ_{22}	0.020	0.313	0.427
Δ_{23}	0.033	1.745	0.056
Δ_{24}	0.050	0.886	0.191
Δ_{25}	0.084	1.627	0.029
Δ_{33}	0.012	0.435	0.427
Δ_{34}	0.039	0.948	0.178
Δ_{35}	0.065	1.281	0.066
Δ_{44}	0.028	0.245	0.383
Δ_{45}	0.098	0.916	0.045
Δ_{55}	0.079	0.197	0.402

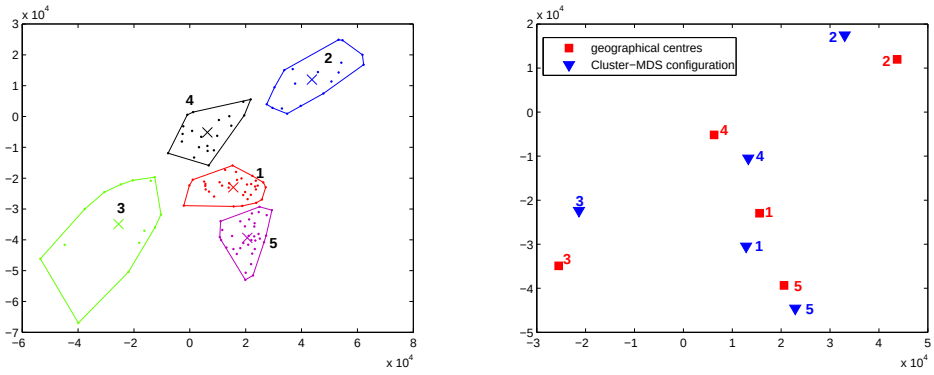

Figura 4.4: El panel izquierdo representa la estructura de partición para los cinco clusters con restricciones mostradas sobre el plano geográfico original para los datos del sistema Vestfjord, donde el signo $\times$ indica los centros geográficos. El panel derecho muestra la representación Procrustes en dos dimensiones de la configuración MDS de los centros de los clusters y la de los centros geográficos.

cuando es evidente una estructura natural de agrupamiento en las localizaciones de la muestra (Kovitz y Christakos, 2004), o aun en una estrecha red regular, también debe ser considerada una estructura espacial de clusters. Desde un punto de vista exploratorio, Vera, Macías y Angulo (2008) propusieron resolver el problema usando un modelo de cluster-MDS de mínimos cuadrados con restricciones geográficas, con un procedimiento puramente descriptivo para determinar la idoneidad de la suposición espacial y del número apropiado de clusters.

En este trabajo proponemos un modelo de cluster-MDS de clases latentes que, bajo la suposición de que los componentes de la mezcla siguen una distribución normal o lognormal, nos permite particionar la muestra de estaciones en clases y simultaneamente representar los centros de los clusters en un espacio de baja dimensionalidad, mientras que las estaciones y clusters mantienen sus relaciones espaciales. Este modelo extiende el modelo de escalamiento multidimensional de clases latentes de Vera, Macías y Angulo (2008) para datos continuos de disimilaridades unimodales a dos vías, incluyendo restricciones geográficas espaciales y considerando mezclas de distribuciones normales y lognormales.

Basado en el estadístico BIC, se propuso una estrategia para la selección del modelo para determinar objetivamente el número de clases latentes así como la dimensionalidad del modelo. A diferencia de la estrategia bootstrap propuesta en Vera, Macías y Heiser (2007), el uso del criterio de información BIC* ajustando el tamaño de la muestra (Rissanen, 1978) bajo restricciones espaciales hace la elección del número de componentes de la mezcla mucho más eficiente en términos del tiempo CPU. Además, el procedimiento de estimación condicional de máxima verosimilitud propuesto y la obtención del BIC* desde la perspectiva de la teoría de codificación de información, nos permite considerar ésto como una estrategia conveniente para la selección del modelo.

A diferencia del algoritmo EM, el uso de Annealing Simulado como procedimiento de estimación de parámetros evita la necesidad de seleccionar una solución inicial adecuada. Mas aún, en todas las pruebas ejecutadas y descritas en el presente trabajo, la solución obtenida produjo un alto índice de atracción, es decir, el porcentaje de veces que se encuentra el valor óptimo más pequeño durante las veinte réplicas, el cual es un indicador de la eficiencia del algoritmo en términos de la calidad de la solución, como se definió en Murillo, Vera y Heiser (2005) y en Vera, Heiser y Murillo (2007). De esta forma, el problema de mínimos locales intrínseco al algoritmo EM es menos

severo cuando se usa Annealing Simulado

El procedimiento propuesto ha sido analizado mediante tres situaciones diferentes. Las primeras dos están basadas en el procedimiento de estimación de Sampson y Guttorp (1992), aplicado a una red regular y a un dominio agrupado de manera natural. Mediante la configuración reducida de cluster-MDS el problema primero se transforma a uno en el cual la estructura de covarianza es estacionaria e isotrópica, a partir de lo cual se puede realizar la estimación no paramétrica de la estructura de covarianza. El tercer ejemplo ilustra la situación en la cual un dominio espacial métrico pero no euclídeo debe ser convertido a uno euclídeo mediante MDS antes que el variograma muestral sea estimado. Los datos usados están basados en las distancias acuáticas entre los nodos de una red triangular analizada por Løland y Høst (2003), pero nuestro procedimiento podría ser aplicado directamente a las distancias acuáticas entre estaciones. La red muestral geográfica es agrupada mediante el procedimiento propuesto de cluster -MDS en $K = 5$ clases, cuyas distancias representan las distancias acuáticas aproximadas entre los cinco clusters correspondientes en el mallado geográfico. Así, los valores estimados de $\mathcal{Z}$ sobre la configuración MDS de los centros de los clusters puede ser empleada para calcular el variograma de la muestra reducida K.

Se proponen dos distribuciones principales para los componentes de la mezcla, de acuerdo con Ramsay (1982). La distribución normal parece ser un candidato adecuado cuando las disimilaridades provienen del variograma muestral (Davis y Borgman, 1982). Sin embargo, la distribución lognormal parece ser más apropiada para datos no negativos con varianza proporcional a la media, como se espera de las disimilaridades medidas directamente. En cualquier caso, el procedimiento de estimación condicional basado en Annealing Simulado puede ser extendido a cualquier distribución de mezclas y una estrategia de selección del modelo para determinar la distribución de los componentes mas adecuada constituye un problema interesante para futuros estudios. La desventaja del procedimiento propuesto es que demanda un alto costo computacional, inherente a cualquier procedimiento de optimización de SA. Además, un enfoque exploratorio podría ser más adecuado en algunas situaciones prácticas donde no son apropiadas las suposiciones de(log)normalidad y /o independencia .

Capítulo 5

Un modelo dual de clases latentes Unfolding para preferencias bimodales a dos vías

5.1. Introducción

En el contexto de ciencias del comportamiento, la técnica de unfolding fue desarrollada por Coombs (1964) para el análisis de datos de preferencias. Este enfoque supone que existe una relación de proximidad entre los elementos de dos diferentes conjuntos, tal que cada elemento v_i, $i = 1, \ldots, R$ en el primer conjunto V, llamado individuos, proporciona un grado de preferencia s_{ij} sobre cada uno de los elementos o_j, $j = 1, \ldots, N$ en el segundo conjunto O llamado objetos. El modelo de distancia para datos de preferencia proporciona una representación de los individuos y objetos en el mismo espacio Euclídeo de dimensión M, asumiendo que la distancia del punto que representa un individuo v_i (llamado el punto ideal) al punto que representa un objeto o_j, está inversamente relacionado al correspondiente valor de la preferencia s_{ij}. Así, un valor grande de s_{ij} indica que una fuerte preferencia será asociada con una distancia pequeña entre los puntos que representan al individuo v_i y al objeto o_j, e inversamente una débil preferencia será relacionada a una distancia grande.

Desde un punto de vista computacional, unfolding puede considerarse un caso especial de Escalamiento Multidimensional (MDS) donde las proximidades entre individuos o entre objetos son datos faltantes (Heiser, 1981). La

formulación general del modelo de unfolding incluye la estimación de una transformación monótona de los datos de preferencias comparada con las distancias en la representación, la cual en el modelo métrico es generalmente restringida a ser lineal para tratar con datos de intervalo o escala de razón. Esta transformación varía para cada individuo en el modelo condicional por fila, o es la misma para todos los individuos en el modelo incondicional. La degeneración es quizás el problema más grande en unfolding, y especialmente para la situación no métrica. Varios procedimientos analíticos has sido propuestos para evitar el problema (ver Busing, Groenen y Heiser, 2005; Borg y Groenen, 2005; o Van Deun, Groenen, Heiser, Busing y Delbeke, 2005, para más detalles). El objetivo en este trabajo es obtener un ajuste del modelo métrico de unfolding para datos de grado de preferencias en el nivel de intervalo. Debido a que únicamente se permite una constante aditiva como en De Soete y Heiser (1993), no existe problema de degeneración en la presente situación (ver Busing et al., 2005).

Denotando por $\boldsymbol{S} = (s_{ij})$ a la $R \times N$ matriz de preferencias, y considendo la matriz $\boldsymbol{A}$ $(R \times M)$ y la matriz $\boldsymbol{B}$ $(N \times M)$, cuyos vectores fila $\boldsymbol{a}_i$, $i = 1 \ldots, R$, y $\boldsymbol{b}_j$, $j = 1, \ldots, N$, son las coordenadas de los R individuos y de los N objectos respectivamente en dimensión M. Entonces, el modelo métrico incondicional de unfolding encuentra $\boldsymbol{A}$ y $\boldsymbol{B}$, tal que para cada individuo v_i y objecto o_j, la proximidad $(\alpha - s_{ij})$ es tan cercana como sea posible a la distancia d_{ij}, donde α denota una constante aditiva para los datos de escala intervalo, asumiendo que la pendiente de la transformación lineal se incluye en la escala de la configuración, y donde d_{ij} es la distancia Euclídea entre los vectores $\boldsymbol{a}_i$ y $\boldsymbol{b}_j$, definida por

$$d_{ij} = d(\boldsymbol{a}_i, \boldsymbol{b}_j) = [(\boldsymbol{a}_i - \boldsymbol{b}_j)^{\mathsf{T}}(\boldsymbol{a}_i - \boldsymbol{b}_j)]^{1/2}.$$

Para mejorar la interpretación o cuando se debe analizar una gran cantidad de datos, el agrupamiento puede ser un procedimiento recomendable. Los datos son categorizados en un pequeño número de grupos de elementos similares tal que cada etiqueta del grupo resume la información requerida sobre el grupo. Conjuntamente con MDS, se han desarrollado procedimientos de cluster-MDS para datos de disimilaridades unimodales a dos vías en el marco clásico (Bock 1986), en un contexto de mínimos cuadrados (Heiser, 1993; Heiser y Groenen, 1997; Vera, Macías y Angulo, 2008) y en un marco de máxima verosimilitud para un modelo de clases latentes para datos continuos (Vera, Macías y Heiser, 2007). También se han desarrollado modelos

de clases latentes en el conjunto de individuos para datos bimodales, en el marco de MDS para disimilaridades a tres vías (Winsberg y De Soete, 1993) y en unfolding para datos de grado de preferencia de estímulos a dos vías (De Soete y Heiser, 1993).

Además de considerar grupos de individuos con patrones de preferencia similares, en este trabajo proponemos una manera adicional para categorizar la información dada en un conjunto de datos bimodal a dos vías, considerando grupos homogéneos de objetos tal que cada grupo contiene objetos percibidos con cualidades similares. En este contexto, se pueden emplear varios métodos de agrupamiento a dos modos (ver Van Mechelen, Bock y De Boeck, 2004 para una revisión exhaustiva de la literatura al respecto) y procedimientos de optimización (ver Van Rosmalen, Groenen, Trejos y Castillo, 2005 para una revisión práctica de comparaciones adicionales) para particionar una matriz de preferencias en su espacio original. Se podrían usar también procedimientos de dos pasos, es decir, reduciendo primero la dimensionalidad mediante unfolding y entonces agrupando las coordenadas, o realizando primero un agrupamiento a dos modos y entonces realizar unfolding a la matriz de datos agrupados. Sin embargo, debe observarse que el espacio reducido de unfolding es óptimo para los puntos representantes de los individuos y objetos, pero no para los clusters sobrepuestos, mientras que la estructura de cluster es óptima en el espacio original no reducido (ver Heiser y Groenen, 1997).

De esta forma, se propone un modelo dual de clases latentes para una matriz de datos de grado de preferencias, teniendo como objetivo particionar los individuos en $T(T << R)$ y los objetos en $C(C << N)$ clases, y simultaneamente representar los $(T + C)$ centros de los clusters en un espacio de dimensión baja, mientras que los individuos y objetos mantienen sus relaciones de preferencias. Asumiendo una distribución normal para los grados de preferencia (De Soete y Heiser, 1993), los parámetros son estimados mediante un procedimiento de estimación condicional de máxima verosimilitud, basado en un algoritmo de annealing simulado. Dada una clasificación de prueba de los individuos y objetos, obtenemos una partición en bloques de la matriz rectangular de preferencias, de la cual se pueden estimar los parámetros del modelo de mezclas, así como la configuración de los centros usando SMACOF (de Leeuw y Heiser, 1980), siendo entonces evaluada la log-verosimilitud y el proceso se repite hasta que se obtiene la convergencia del algoritmo. Se propone un procedimiento de selección del modelo basado en el estadístico BIC para determinar la mejor combinación de clases latentes para los individuos y objetos, así como para determinar la dimensión de la representación de los

centros de los clusters.

5.2. Un modelo dual de clases latentes unfolding

Consideremos una partición $\mathcal{P}(V)$ del espacio de individuos en T clases latentes V_t con r_t elementos, donde $r_1 + \cdots + r_T = R$, y una partición $\mathcal{P}(O)$ del espacio de objetos en C clases latentes O_c con n_c elementos, donde $n_1 + \cdots + n_C = N$. Se asumirá que un individuo u objeto pertenece a uno y solamente uno de los subconjuntos de su partición correspondiente, y que no conocemos de antemano a cual clase latente pertenece un individuo o un objeto particular. Los elementos en $\boldsymbol{S}$ son considerados para ser organizados permutando filas y columnas de acuerdo con la secuencia en el conjunto de índice de las clases latentes de individuos y objetos. Así, en términos de la matriz de preferencias $\boldsymbol{S}$, la situación se puede describir suponiendo una partición en forma de bloques $\mathcal{P}(\boldsymbol{S})$ de la matriz $\boldsymbol{S}$ en TC clases latentes $S_{tc} = (s_{ij})$ de $r_t n_c$ elementos, con $v_i \in V_t$ y $o_j \in O_c$, tal que $\sum_t \sum_c r_t n_c = RN$. Entonces, cada preferencia s_{ij} pertenece a uno y solamente uno de los TC subconjuntos S_{tc}, mientras que se preserva la condición de la partición en forma de bloques, pero no se conoce de antemano a cual bloque latente pertenece una preferencia particular.

En la formulación del modelo, no son los individuos ni los objetos sino los dos conjuntos de clusters los que son representados mediante los vectores filas $\boldsymbol{a}_t$ de la matriz $\boldsymbol{A}_T$ $(T \times M)$, para el conjunto de individuos V_t, $t = 1, \ldots, T$, y por los vectores fila $\boldsymbol{b}_c$ de la matriz $\boldsymbol{B}_C$ $(C \times M)$, para el conjunto de objetos O_c, $c = 1, \ldots, C$, es un espacio de dimensión M. Siguiendo la notación general para modelos de clases latentes, denotamos por λ_{tc} la probabilidad incondicional de que un valor de la preferencia s_{ij} pertenezca a una clase latente S_{tc}, mientras que se preserva la condición de la forma de bloques de $\mathcal{P}(\boldsymbol{S})$, es decir, mientras que $v_i \in V_t$ y $o_j \in O_c$, donde $0 \leq \lambda_{tc} \leq 1$, y

$$\sum_{t=1}^{T} \sum_{c=1}^{C} \lambda_{tc} = 1. \tag{5.1}$$

La hipótesis de distribución normal para los componentes de la mezcla es congruente con la formulación de unfolding de mínimos cuadrados, así como con las formulaciones previas del modelo de unfolding de clases latentes para

valores de preferencias (De Soete y Heiser, 1993). Entonces asumimos que las s_{ij} que pertenecen a una clase latente S_{tc} son observaciones de variables aleatorias independientes distribuidas normalmente de parámetros media y varianza, μ_{tc} y σ^2_{tc} respectivamente, es decir,

$$s_{ij} \sim \mathcal{N}(\mu_{tc}, \sigma^2_{tc}), \text{ para } s_{ij} \in S_{tc}, \tag{5.2}$$

donde las medias de los bloques μ_{tc} están geométricamente relacionadas al correspondiente par de centros de clusters de individuos y objectos por $\mu_{tc} = \alpha_t - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ en el modelo de unfolding condicional, mientras que en el modelo incondicional se considera la situación particular de $\alpha_t = \alpha$, $t = 1, \ldots, T$.

Aun cuando una varianza constante σ^2 es congruente con muchas de las formulaciones del modelo unfolding de mínimos cuadrados, la consideración de una varianza dependiente de la clase latente puede contribuir a obtener un modelo mucho mas parsimonioso. La reducción en el número de parámetros cuando los individuos y objetos son categorizados significa que, la consideración de la varianza dependiente de la clase latente no incrementa significativamente el número de parámetros a ser estimados, mientras que esto puede llevar a seleccionar un modelo con pocos componentes de la mezcla. Si las μ_{tc} no son geométricamente relacionadas a los centros de los clusters, el modelo propuesto tiene $3TC$ parámetros a estimar. Entonces, tomando en cuenta la condición dada por (5.1), los grados de libertad del modelo son $3TC - 1$. Cuando se estima la configuración de unfolding, y considerando la invarianza rotacional y traslacional de la solución de unfolding, los grados de libertad del modelo son $T(1 + 2C) + (T + C)M - (M(M + 1)/2) - 1$ para el modelo condicional y $2(TC) + (T + C)M - M(M + 1)/2$ si se considera un escalar α, lo cual nos permite establecer una cota superior para la dimensionalidad del modelo, de tal forma que los grados de libertad del modelo completo sean menores que los del modelo sin restricciones geométricas.

5.2.1. Procedimiento de estimación condicional de máxima verosimilitud

La formulación del problema de clases latentes desde la perspectiva de la matriz de valores de preferencia $\boldsymbol{S}$, además del hecho de que no se conoce de antemano a cual bloque latente pertenece una preferencia, nos conduce a definir la f.d.p de s_{ij} como una mezcla de densidades normales univariadas de la forma,

$$g(s_{ij} \mid \boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = \sum_{t=1}^{T} \sum_{c=1}^{C} \lambda_{tc} f_{tc}(s_{ij} \mid \boldsymbol{a}_t, \boldsymbol{b}_c, \alpha_t, \sigma_{tc}^2), \tag{5.3}$$

donde $\boldsymbol{\Sigma} = (\sigma_{tc}^2)$ denota la matriz $T \times C$ de varianzas *dentro* del bloque, $\boldsymbol{\alpha}$ es el vector $(\alpha_1, \ldots, \alpha_T)'$ en la situación condicional o un escalar en el modelo incondicional, y $\boldsymbol{\Lambda} = (\lambda_{tc})$ es la matriz $T \times C$ de probabilidades incondicionales bajo la restricción de la forma en bloques de $\mathcal{P}(\boldsymbol{S})$, y donde $f_{tc}(s_{ij})$ es la función de densidad de probabilidades normal de $s_{ij} \in S_{tc}$, dada por

$$f_{tc}(s_{ij} \mid \boldsymbol{a}_t, \boldsymbol{b}_c, \alpha_t, \sigma_{tc}^2) = \frac{1}{\sigma_{tc}(2\pi)^{1/2}} \exp\left[-\frac{(s_{ij} - \mu_{tc})^2}{2\sigma_{tc}^2}\right]. \tag{5.4}$$

La función de log-verosimilitud asociada del modelo de mezclas puede escribirse como

$$\log L(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda} \mid \boldsymbol{S}) = \sum_{i=1}^{R} \sum_{j=1}^{N} \log(g(s_{ij} | \boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})), \tag{5.5}$$

la cual, sin considerar las restricciones geométricas, presenta un valor máximo en los estimadores de los parámetros dados por

$$\widehat{\lambda}_{tc} = \frac{\sum_{i=1}^{R} \sum_{j=1}^{N} \pi_{ij,tc}}{RN}, \tag{5.6}$$

$$\widehat{\mu}_{tc} = \frac{\sum_{i=1}^{R} \sum_{j=1}^{N} \pi_{ij,tc} s_{ij}}{\sum_{i=1}^{R} \sum_{j=1}^{N} \pi_{ij,tc}}, \tag{5.7}$$

$$\widehat{\sigma}_{tc}^2 = \frac{\sum_{i=1}^{R} \sum_{j=1}^{N} \pi_{ij,tc} (s_{ij} - \widehat{\mu}_{tc})^2}{\sum_{i=1}^{R} \sum_{j=1}^{N} \pi_{ij,tc}}, \tag{5.8}$$

donde

$$\pi_{ij,tc} = \frac{\lambda_{tc} f_{tc}(s_{ij})}{\sum_{t=1}^{T}\sum_{c=1}^{C} \lambda_{tc} f_{tc}(s_{ij})}. \tag{5.9}$$

Bajo la restricción de la partición en forma de bloques en $\boldsymbol{S}$, los coeficientes $\pi_{ij,tc}$ representan la probabilidad a posteriori de que un valor de preferencia observado s_{ij} pertenezca a la clase latente S_{tc}, es decir, que s_{ij} proviene de una f.d.p normal $f_{tc}(s_{ij})$. Como es usual en el contexto de modelos de clases latentes, los valores de los estimadores de los parámetros dados por (5.6), (5.7) y (5.8) dependen de los valores estimados $\pi_{ij,tc}$, pero para obtener estos, son necesarios los valores de los parámetros. El algoritmo EM (Dempster, Laird y Rubin, 1977), conjuntamente con el teorema de Bayes, puede proporcionar una solución al problema, si se imponen condiciones adicionales en la estimación de parámetros, para asegurar que se preserve la partición $\mathcal{P}(\boldsymbol{S})$ en forma de bloques. Debido a que una partición $\mathcal{P}(V)$ en el espacio de individuos y una partición $\mathcal{P}(O)$ en el espacio de objetos conduce a una partición en forma de bloques $\mathcal{P}(\boldsymbol{S})$ en las preferencias, el inverso, en general no es verdad, lo cual desde un punto de vista computacional incrementa la dificultad en la estimación de parámetros. Alternativamente, se puede emplear un método de optimización de Monte Carlo conjuntamente con un procedimiento de estimación condicional de máxima verosimilitud para tratar con el problema de estimación bajo la condición de una partición en forma de bloques.

De esta forma, definimos la matriz $\boldsymbol{Z} = (z_{ij,tc})$ $(RN \times TC)$ de variables indicadoras de clases latentes por

$$z_{ij,tc} = \begin{cases} 1, & \text{si } s_{ij} \in S_{tc},\ i = 1,\ldots,R,\ j = 1,\ldots,N,\ t = 1,\ldots,T, \overline{=}1,\ldots,C \\ 0, & \text{otro caso,} \end{cases}$$

donde

$$\sum_{t=1}^{T}\sum_{c=1}^{C} z_{ij,tc} = 1, \quad \text{y} \quad \sum_{i=1}^{R}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C} z_{ij,tc} = RN.$$

Dada una clasificación de individuos y objetos, la partición en forma de bloque $\mathcal{P}(\boldsymbol{S})$ es conocida y los vectores fila de $\widehat{\boldsymbol{Z}}$, denotados por $\hat{\boldsymbol{z}}_{ij} = (\hat{z}_{ij,11},\ldots,\hat{z}_{ij,TC})^{\mathsf{T}}$, tienen todos los elementos igual a cero excepto para $\hat{z}_{ij,tc} =$

1, si $s_{ij} \in S_{tc}$, sin asumir alguna distribución de probabilidad para $\boldsymbol{z}_{ij}$. De esta forma, los valores de la probabilidad condicional a posteriori están dados por $\hat{\pi}_{ij,tc} = \hat{z}_{ij,tc}$, de lo cual, cuando son sustituidos en (5.6), obtenemos el estimador condicional de λ_{tc} dado por

$$\widehat{\lambda}_{tc} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{N}\widehat{z}_{ij,tc}}{RN}. \tag{5.10}$$

Entonces, si los valores $\boldsymbol{Z}$ son conocidos, la f.d.p condicional de s_{ij}, dada $\hat{\boldsymbol{z}}_{ij}$, se puede expresar como

$$g(s_{ij} \mid \hat{\boldsymbol{z}}_{ij}, \boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_t \prod_c f_{tc}(s_{ij} \mid \boldsymbol{a}_t, \boldsymbol{b}_c, \alpha_t, \sigma_{tc}^2)^{\hat{z}_{ij,tc}}, \tag{5.11}$$

y la función de log-verosimilitud condicional sigue la expresión,

$$\mathcal{Q}(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \mid \boldsymbol{S}, \widehat{\boldsymbol{Z}}) = \sum_{i=1}^{R}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{z}_{ij,tc}\log f_{tc}(s_{ij} \mid \boldsymbol{a}_t, \boldsymbol{b}_c, \alpha_t, \sigma_{tc}^2). \tag{5.12}$$

Imponiendo las restricciones geométricas $\mu_{tc} = \alpha_t - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$, las coordenadas de los centros de los clusters asociadas $\boldsymbol{A}_T$, $\boldsymbol{B}_C$, y el vector $\boldsymbol{\alpha}$ (el cual es un escalar α en el modelo incondicional), pueden ser estimados maximizando (5.12), o equivalentemente minimizando

$$q(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}) = \sum_{i=1}^{R}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{z}_{ij,tc}(s_{ij} - \mu_{tc})^2. \tag{5.13}$$

Sin embargo, $q(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha})$ puede ser descompuesto ortogonalmente en un componente *dentro* de clases y un componente *entre* clases

$$q(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}) = \sum_{i=1}^{R}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{z}_{ij,tc}(s_{ij} - \overline{s}_{tc})^2 + \sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\overline{s}_{tc} - \mu_{tc})^2, \tag{5.14}$$

donde

$$\overline{s}_{tc} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{N}\hat{z}_{ij,tc}s_{ij}}{\sum_{i=1}^{R}\sum_{j=1}^{N}\hat{z}_{ij,tc}}, \quad \text{y} \quad \gamma_{tc} = \sum_{i=1}^{R}\sum_{j=1}^{N}\hat{z}_{ij,tc},$$

de lo cual únicamente el último término en (5.14) debería ser minimizado para estimar $\boldsymbol{A}_T$, $\boldsymbol{B}_C$ y $\boldsymbol{\alpha}$. El último término de (5.14) es denotado por

$$\phi(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}) = \sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\overline{s}_{tc} - \mu_{tc})^2. \tag{5.15}$$

Definiendo las disparidades $\hat{d}_{tc}$ como

$$\hat{d}_{tc} = \begin{cases} \alpha_t - \overline{s}_{tc}, & \text{(modelo condicional)} \\ \alpha - \overline{s}_{tc}, & \text{(modelo incondicional),} \end{cases} \tag{5.16}$$

la función de mínimos cuadrados ponderada $\phi(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha})$, se puede escribir como

$$\phi(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}) = \sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\hat{d}_{tc} - d(\boldsymbol{a}_t, \boldsymbol{b}_c))^2, \tag{5.17}$$

la cual puede ser minimizada usando cualquier algoritmo de unfolding. En este trabajo usamos la adaptación de SMACOF a unfolding (Heiser 1981, 1987), tomando en cuenta las modificaciones propuestas por Heiser (1991) y De Soete y Heiser (1993) para evitar el problema de valores negativos en $\hat{d}_{tc}$, estimando apropiadamente los valores de $\widehat{\boldsymbol{A}}_T$, $\widehat{\boldsymbol{B}}_C$, y donde el nuevo valor de $\hat{\boldsymbol{\alpha}}$ en el modelo condicional de unfolding está dado por

$$\hat{\alpha}_t = \frac{\sum_{c=1}^{C}\gamma_{tc}(\overline{s}_{tc} + d(\hat{\boldsymbol{a}}_t, \hat{\boldsymbol{b}}_c))}{\sum_{c=1}^{C}\gamma_{tc}}, \tag{5.18}$$

mientras que para el modelo incondicional adopta la expresión,

$$\hat{\alpha} = \frac{\sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\bar{s}_{tc} + d(\hat{\boldsymbol{a}}_t, \hat{\boldsymbol{b}}_c))}{RN}. \tag{5.19}$$

Así, los valores estimados de $\widehat{\boldsymbol{A}}_T$, $\widehat{\boldsymbol{B}}_C$ y $\hat{\boldsymbol{\alpha}}$ son obtenidos al final del procedimiento modificado de SMACOF, condicionados a los valores dados de $\widehat{\boldsymbol{Z}}$. Finalmente, cuando los valores $\hat{\pi}_{ij,tc}$, y las restricciones geométricas $\hat{\mu}_{tc} = \hat{\alpha}_t - d(\hat{\boldsymbol{a}}_t, \hat{\boldsymbol{b}}_c)$ (o $\hat{\mu}_{tc} = \hat{\alpha} - d(\hat{\boldsymbol{a}}_t, \hat{\boldsymbol{b}}_c)$ en la situación incondicional) son sustituidas en (5.8), los valores $\widehat{\boldsymbol{\Sigma}}$ adoptan la expresión

$$\widehat{\sigma}^2_{tc} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{N}\hat{z}_{ij,tc}(s_{ij} - \hat{\mu}_{tc})^2}{\sum_{i=1}^{R}\sum_{j=1}^{N}\hat{z}_{ij,tc}}. \tag{5.20}$$

5.3. Algoritmo de Annealing para propósitos de estimación

Puesto que en la práctica, $\boldsymbol{Z}$ son variables indicadoras no observadas, el procedimiento de estimación condicional se convierte en parte de un algoritmo iterativo en una estructura de annealing simulado. De esta forma, para una partición de prueba en forma de bloques $\mathcal{P}(\boldsymbol{S})$, derivada de las particiones $\mathcal{P}(V)$ y $\mathcal{P}(O)$ elegidas aleatoriamente, los valores de la matriz $\boldsymbol{Z}$ están dados y se estiman los otros parámetros, y el procedimiento se repite hasta que se alcanza un criterio de convergencia.

Annealing simulado (SA) es un procedimiento de optimización Monte Carlo intoducido por Kirkpatrick, Gelatt y Vecchi (1983), y por Černy (1985). El término proviene de una analogía con el proceso físico de calentamiento y luego de un lento enfriamiento de una sustancia para obtener una estructura cristalina libre de impurezas. Un sólido es calentado a una temperatura inicial alta $\mathcal{T}_0$, fijando la energía del sistema a un estado inicial $\mathcal{E}_0$. Entonces, el sistema se enfría lentamente hasta alcanzar la mejor configuración cristalina posible, es decir, hasta que el sistema alcanza su estado de energía mínima $\mathcal{E}_f$. El esquema de enfriamiento es de gran importancia debido a que si el sistema es enfriado demasiado rápido, podrían aparecer impurezas en

el enrejado cristalino. En optimización, el proceso es simulado mediante la generación de una secuencia de cadenas de Markov en temperaturas decrecientes, proporcionando recursos para escapar de óptimos locales permitiendo movimientos de *subir colinas* en la busqueda del óptimo global, aceptando puntos que pueden tener más energía que los previos mediante la regla de aceptación de Metropolis (Metropolis et al, 1953). Durante el proceso, en la temperatura $\mathcal{T}$ del sistema, la energía del sistema $\mathcal{E}$ es avaluada en un punto elegido aleatoriamente. El nuevo punto es aceptado si la energía ha disminuido con respecto al punto previamente seleccionado. Por el contrario, si la energía se incrementa por $\triangle\mathcal{E}$, el nuevo punto puede ser aceptado, con una probabilidad de $\exp(-\triangle\mathcal{E}/\mathcal{T})$, la cual es llamada la regla de aceptación de Metropolis. El proceso se repite hasta que el sistema alcanza un equilibrio, de tal forma que al final, valores pequeños de $\mathcal{T}$ aseguran que únicamente serán aceptadas buenas soluciones.

Annealing simulado ha demostrado su utilidad en el tratamiento de la optimización directa de la función de pérdida en escalamiento unidimensional métrico exploratorio (Murillo, Vera y Heiser, 2005) así como en escalamiento multidimensional (Vera, Heiser y Murillo, 2007). En el marco de estimación condicional de máxima verosimilitud, SA ha sido usado recientemente, por ejemplo, en el problema general de estimación de parámetros en una distribución lognormal de tres parámetros (Vera y Díaz-García, 2008) y en escalamiento multidimensional para datos a dos vías un modo (Vera, Macías y Heiser, 2007).

El procedimiento de optimización parte de una partición inicial en forma de bloques $\mathcal{P}^{(0)}(\boldsymbol{S})$ derivada de las particiones iniciales elegidas aleatoriamente $\mathcal{P}^{(0)}(V)$ y $\mathcal{P}^{(0)}(O)$. Los valores de los parámetros son estimados, la logverosimilitud inicial es evaluada (energía inicial), y el factor de enfriamiento γ es inicializado, junto con la longitud de la cadena de Markov LC y el factor IC, el cual incrementa la longitud de la cadena de Markov cada m iteraciones. Para un valor de probabilidad dado χ, la temperatura inicial $\mathcal{T}_0$ es estimada en un procedimiento de muestreo aleatorio como se describió en implementaciones previas de SA (ver Murillo et al., 2005 para más detalles), promediando un número de M_a posibles incrementos de las soluciones que empeoran la logverosimilitud. Así, se obtiene un valor de la temperatura inicial tal que en las primeras iteraciones son aceptadas el $100\chi\,\%$ de las peores soluciones. La temperatura final $\mathcal{T}_f$ se elige muy cercana a cero para asegurar que, eventualmente, el algoritmo se detiene en por lo menos un mínimo local. El esquema de enfriamiento constituye un ciclo iterativo principal de

$It_{max} = \log(\mathcal{T}_f/\mathcal{T}_0 - \eta)$ iteraciones, en el cual la temperatura actual $\mathcal{T}$ decrece gradualmente. En cada iteración principal, se selecciona una nueva partición óptima en forma de bloques para S, maximizando la logverosimilitud, en un ciclo iterativo secundario de longitud de incremento LC. Así, en la p-ésima iteración secundaria, el algoritmo SA se puede describir como sigue:

1. Dada una partición en forma de bloques $\mathcal{P}^{(p)}(\boldsymbol{S})$, los parámetros son estimados condicionalmente como $\widehat{\boldsymbol{Z}}^{(p)}$ y se calcula la log-verosimilitud del modelo de mezclas $\log L^{(p)}$.

2. Basados en $\mathcal{P}^{(p)}(\boldsymbol{S})$, se obtiene una nueva partición aleatoriamente mediante un procedimiento de dos pasos. Primero, el conjunto de individuos o el conjunto de objetos es elegido aleatoriamente. Segundo, un elemento elegido aleatoriamente del conjunto previamente seleccionado es movido a un nuevo cluster elegido aleatoriamente, produciendo una partición de prueba $\mathcal{P}^{(p+1)}(\boldsymbol{S})$, bajo la condición de que la cardinalidad de cualquier bloque S_{tc} es mayor o igual a dos, para evitar soluciones degeneradas. Entonces, se calcula la nueva matriz indicadora $\widehat{\boldsymbol{Z}}^{(p+1)}$.

3. De $\widehat{\boldsymbol{Z}}^{(p+1)}$, se estiman las probabilidades a posteriori $\hat{\pi}_{ij,tc}^{(p+1)}$, y los coeficientes de la mezcla $\hat{\lambda}_{tc}$, los parámetros con restricciones geométricas $\boldsymbol{A}_T^{(p+1)}$, $\boldsymbol{B}_C^{(p+1)}$ y $\hat{\boldsymbol{\alpha}}^{(p+1)}$ (usando la versión modificada de SMACOF) y los parámetros de la dispersión $\boldsymbol{\Sigma}^{(p+1)}$ son calculados condicionalmente. La log-verosimilitud del modelo de mezclas, $\log L^{(p+1)}$, dada por (5.5) es entonces evaluada y se calcula el incremento en la log-verosimilitud $\triangle \log L = \log L^{(p+1)} - \log L^{(p)}$.

4. Usando la regla de aceptación de Metropolis, si la log-verosimilitud se incrementa, la partición de prueba $\mathcal{P}^{(p+1)}(\boldsymbol{S})$ es seleccionada; de lo contrario, se selecciona la partición de prueba con una probabilidad $\exp(\triangle \log L/\mathcal{T})$.

El proceso se repite en un ciclo interno de longitud LC, obteniendo la nueva partición definitiva $\mathcal{P}(\boldsymbol{S})$. Entonces la temperatura se disminuye a $\mathcal{T} = \gamma\mathcal{T}$, y el proceso continua hasta que se alcanza un criterio de convergencia, es decir, se cumple el número máximo de iteraciones, o el valor de la log-verosimilitud es repetido un número de iteraciones principales R_{max} previamente fijado, conjuntamente con un valor pequeño de la temperatura.

5.3.1. Consideraciones prácticas

Aunque el procedimiento de optimización de annealing simulado no depende de la solución inicial, varias ejecuciones del algoritmo y un adecuado esquema de enfriamiento debería ser empleado para mantener un buen balance entre la calidad del óptimo encontrado y el costo computacional. De esta forma, para un valor dado de T y C, la partición inicial $\mathcal{P}(\boldsymbol{S})$ está dada mediante la asignación aleatoria de individuos y objetos en T y C clases respectivamente, bajo la condición de que $r_t n_c \geq 2$, para evitar el problema de bloques con varianza cero, y para prevenir la presencia de una columna cero en la matriz $\boldsymbol{Z}$. La restricción computacional anterior introduce la condición de que si $T \geq R/2$, entonces $C < N/3$, e inversamente, si $C \geq N/2$, entonces $T < R/3$, lo cual restringe las posibles combinaciones de clases latentes para el propósito de selección del modelo.

En el siguiente paso, para estimar la configuración de los centros de los clusters, la matriz $(T + C) \times M$, $\boldsymbol{X} = [\boldsymbol{A}_T^{'} \boldsymbol{B}_C^{'}]^{'}$, que une las coordenadas de los centros de los clusters de los individuos y objetos, es considerada en SMACOF. Entonces, interpretando el procedimiento de unfolding de mínimos cuadrados como un problema de escalamiento multidimensional, la matriz de preferencias $\boldsymbol{S}$ comprende los valores de similaridades *entre-conjuntos*, mientras que los valores de las proximidades *dentro-conjuntos* son valores faltantes que deben estimarse para la solución inicial. El procedimiento de imputación empleado para los valores faltantes es descrito en la sección 2.4 de Heiser (1981), eligiendo un valor inicial de $\boldsymbol{\alpha}$ para asegurar que todos los valores $\alpha_t - \overline{s}_{tc}$ son positivos. Entonces, la solución clásica de MDS es usada como la configuración inicial en SMACOF.

Desde un punto de vista teórico, SA es un procedimiento de optimización global. Desde un punto de vista práctico, como cualquier procedimiento de optimización Monte Carlo, éste se basa en prueba y error necesitando generalmente un tiempo considerable de procesamiento, debido a su naturaleza aleatoria. Por lo tanto, aun cuando el método de clases latentes unfolding es en principio adecuado para conjuntos de datos grandes, el heurístico SA es especialmente adecuado para conjuntos de datos de tamaño pequeño y mediano debido a su costo CPU inherente. Por otro lado, aun cuando el problema de óptimos locales persiste en el algoritmo de unfolding, especialmente en la situación unidimensional, para la estimación heurística global el problema de mínimos locales es menos severo cuando se utiliza SA.

5.4. Selección del modelo

Además del problema de estimación de parámetros, uno de los principales objetivos de los modelos de clases latentes es determinar el número de componentes de la mezcla. En el modelo de clases latentes propuesto, se debe adoptar una decisión independiente en relación al número de clusters para los conjuntos de individuos y de objetos. Entonces, en lugar de determinar el número final TC de componentes en la mezcla, deben ser identificados los factores que conducen a este valor final, es decir, el número de clases latentes T en los individuos y el número de clases de clases latentes C en los objetos. El procedimiento de estimación Monte Carlo propuesto en este trabajo proporciona una solución al problema, debido a que el número de componentes de la mezcla en los datos de preferencia es determinado condicionalmente por la clasificación dada en ambos conjuntos originales.

El algoritmo propuesto basado en SA ofrece una estimación del modelo de mezclas condicionada a la restricción en forma de bloques en la matriz de preferencias $\boldsymbol{S}$, a diferencia de los estimadores incondicionados que pueden obtenerse con la forma general del algoritmo GEM. Así, cualquier modelo de clases latentes correspondiente a una factorización en T y C clases de un modelo de mezclas con TC componentes fijos, tiene el mismo número de parámetros (si las condiciones geométricas no son tomadas en cuenta) y puede ser un candidato para describir el modelo con mejor ajuste. Así, por ejemplo, para una mezcla de $TC = 12$ componentes, cualquier modelo de clases latentes correspondiente a los (T, C) pares de $(2, 6)$, $(6, 2)$, $(3, 4)$, o $(4, 3)$, pueden ser un candidato. Por lo tanto, es investigada la mejor factorización que conduce al número de componentes de la mezcla, en lugar del número directo de componentes de la mezcla.

El modelo también provee la posibilidad de determinar la dimensionalidad de la representación unfolding usando el criterio de información. Sin embargo, es conocido el hecho de que las condiciones de regularidad no se sostienen para la prueba de razón de verosimilitud cuando se comparan mezclas con diferente número de distribuciones componentes. El enfoque bootstrap (Hope, 1968) es un procedimiento alternativo extensamente usado. Aunque este método ha sido empleado en el contexto de modelos de clases latentes para MDS y unfolding (ver De Soete y Heiser, 1993 para más detalles), el tiempo CPU se incrementa considerablemente.

Para propósitos de selección del modelo, proponemos la utilización del criterio de información Bayesiana (BIC) (Schwarz, 1978). Su aplicación en

este contexto está apoyado en varios estudios recientes, siendo derivado por Rissanen (1986, 1989) en un contexto de selección de modelos, desde una perspectiva diferente basada en la teoría de codificación de información (ver Sección 6.9.3 de McLachlan y Peel, 2001 para más detalles). Para mejorar el criterio de información BIC en este contexto, incluimos el ajuste del tamaño de muestra sugerido por Rissanen (1978), donde el número de datos RN es ajustado por $(RN + 2)/24$ (Yang y Yang, 2007). Bajo este ajuste, el criterio BIC adopta la siguiente expresión,

$$\text{BIC}^* = -2 \log L + l \log h,$$

donde $h = (RN+2)/24$, y $l = 3TC-1$ para el modelo incondicional. Cuando se imponen restricciones geométricas, el número de parámetros desconocidos en el modelo condicional está dado por $l = T(1+2C)+(T+C)M-(M(M+1)/2)-1$, mientras que considerando $\alpha_t = \alpha$, $t = 1\ldots,T$, el número de parámetros está dado por $l = 2(TC) + (T + C)M - M(M + 1)/2$.

Cuando no se imponen restricciones geométricas, el número adecuado de clases latentes en el espacio de individuos y de objetos se indica como el correspondiente al valor más pequeño del estadístico BIC*. Después de determinar el número de clases latentes para individuos y para objetos, se puede emplear el criterio BIC* para establecer la dimensión de la representación unfolding. Entonces, bajo los valores previamente seleccionados de T y C, e imponiendo condiciones geométricas, la dimensión M correspondiente al valor más bajo del BIC* se selecciona como la mejor representación del modelo.

5.5. Aplicaciones ilustrativas

Para ilustrar el algoritmo propuesto, analizamos dos conjuntos de datos. El modelo fue aplicado primero a un conjunto de datos de preferencias agrupados artificialmente, después analizamos un conjunto de datos empíricos. Además, los resultados obtenidos para este conjunto de datos reales de preferencias son comparados con los obtenidos por un procedimiento de dos pasos, donde primero se derivan los clusters y luego la representación unfolding de las clases.

Uno de los aspectos más importantes de cualquier algoritmo de Annealing Simulado se refiere a la eficiencia de su implementación computacional, para minimizar el tiempo CPU. Como en aplicaciones previas de SA realizadas por

el autor, el procedimiento propuesto fue implementado en Fortran, trabajando en un ordenador Pentium IV 3.00 GHz con 2 Gb de RAM bajo Microsoft Windows XP. Para incrementar la eficiencia del algoritmo propuesto, el mejor óptimo local en 20 réplicas independientes se elegió como la mejor solución, con el índice de atracción siendo definido como el porcentaje de veces que se obtiene el mejor óptimo local. En todos los conjuntos de datos examinados, los valores adecuados de los parámetros fueron χ=0.95, $Ma = 50(T + C)$ para la fase de la temperatura inicial, con valores de γ =0.95, $T_f = 10^{-7}$, $R_{\text{máx}} = 10$, $LC = 2(TR+CN)$, $IC = (TR+CN)$ y $m = 20$, para los parámetros restantes. Para el procedimiento SMACOF en la etapa de estimación de la configuración, empleamos criterios de convergencia de un máximo de 300 iteraciones con una diferencia en valores subsiguientes del STRESS menor que 10^{-7}.

Se generó una matriz rectangular de preferencias artificiales, después de localizar en un plano 20 individuos agrupados en 5 clusters, y 12 objetos clasificados en 3 clases. Entonces, fue derivada una matriz de preferencias 5×3 de las distancias Euclídeas entre las coordenadas de los centros de los clusters localizados, usando la condición geométrica $\mu_{tc} = \alpha - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$. La varianza entre-clusters σ^2_{tc} fue calculada de tal forma que aproximadamente el 25 % del total de la varianza de los datos generados fue error de varianza (De Soete y Heiser, 1993). De cada componente de la mezcla, fueron generados 49 valores, correspondientes a un bloque 7×7 S_{tc}, de esta forma generamos una matriz de datos correspondiente a 35 individuos y 21 objetos.

Para determinar la combinación apropiada de clase latentes, probamos el modelo correspondiente a todas las combinaciones admisibles de pares de valores T y C, considerando por lo menos dos clusters en las filas y en las columnas de S para evitar soluciones triviales. El valor más bajo del estadístico BIC* (4059.6) se encontró para los valores de $T = 5$ y $C = 3$, como se esperaba. La tabla 5.1 muestra los resultados correspondientes a todas las combinaciones de valores para T y C hasta 6, siendo limitados por razones de espacio. Como puede ser apreciado, diferentes valores del BIC* y de la logverosimilitud están presentes en la comparación de combinaciones de clases latentes correspondientes al mismo número de componentes de la mezcla.

Siguiendo el procedimiento de selección del modelo, la restricción geométrica $\mu_{tc} = \alpha - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ fue considerada en el modelo de $T = 5$, $C = 3$ clases latentes. Como se muestra en la tabla 5.2, se seleccionan dos dimensiones para representar los centros de los clusters, lo cual corresponde al valor

Tabla 5.1: Resultados del criterio BIC* para probar el número de clases latentes para el conjunto de datos artificiales sin considerar restricciones geométricas.

Resultados del análisis sin restricciones geométricas

(T,C)	TC	gl	$\log L$	BIC*
2,2	4	11	-2340.73	4719.1
2,3	6	17	-2245.65	4549.5
3,2	6	17	-2236.65	4531.5
2,4	8	23	-2245.50	4569.8
4,2	8	23	-2146.86	4372.5
3,3	9	26	-2092.81	4274.7
2,5	10	29	-2245.50	4590.3
5,2	10	29	-2079.81	4258.9
2,6	12	35	-2245.51	4610.9
3,4	12	35	-2090.13	4300.1
4,3	12	35	-2004.82	4129.5
6,2	12	35	-2079.55	4279.0
3,5	15	44	-2088.63	4328.0
5,3	15	44	-1954.45	4059.6
4,4	16	47	-2004.23	4169.4
3,6	18	53	-2088.20	4357.9
6,3	18	53	-1953.33	4088.2
4,5	20	59	-2003.82	4209.7
5,4	20	59	-1953.16	4108.4
4,6	24	71	-2003.48	4250.1
6,4	24	71	-1951.07	4145.3
5,5	25	74	-1952.19	4157.8
5,6	30	89	-1951.57	4207.9
6,5	30	89	-1947.56	4199.9
6,6	36	107	-1945.36	4257.1

Tabla 5.2: Resultados del criterio de información para el modelo $T = 5$ y $C = 3$, cuando los centros de los clusters están restringidos a ser escalados incondicionalmente en una, dos y en tres dimensiones, para el conjunto de datos artificiales.

Resultados para el modelo incondicional de clases latentes

No. de Clases (T,C)	No. de dimensiones	gl Modelo	log-Likelihood	BIC*
5 , 3	1	37	-2122.51	4371.73
5 , 3	2	43	-1954.59	4056.43
5 , 3	3	48	-1954.45	4073.28

más pequeño del estadístico BIC*. La tabla 5.3 muestra los valores estimados de los parámetros para el modelo incondicional resultante; se puede ver que la estructura artificial de 7 elementos por cluster es recuperada, y que obtenemos valores iguales de la proporción de mezcla (0.0666), lo cual está de acuerdo con el modelo. Distancias más pequeñas entre los centros de los clusters se encuentran para valores grandes de μ_{tc}, como se muestra en la figura 5.1, donde existe una proximidad evidente entre los datos verdaderos y los recuperados. Un bajo valor de la varianza σ_{tc} indica un alto grado de precisión entre los centros de los clusters de preferencias $\alpha - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ y las preferencias entre los individuos y objetos en las respectivas clases V_t y O_c.

La segunda aplicación corresponde al análisis de la evaluación sobre una escala likert de 7 puntos, variando desde 1 - en completo desacuerdo a 7 - completamente de acuerdo, de 22 declaraciones acerca de internet por 193 respondientes en la Universidad Erasmus de Rotterdam, después de eliminar los valores faltantes. Los datos fueron recopilados alrededor del 2002 antes de que el internet de banda ancha tuviera una amplia cobertura en Holanda. La matriz $\boldsymbol{S}$ de los datos originales analizados fue recopilada de *http://people.few.eur.nl/groenen*. Esta matriz primero fue centrada y entonces analizada en el contexto de agrupamiento a dos modos por Van Rosmalen, Groenen, Trejos y Castillo (2005), para comparar varios procedimientos. El propósito principal de analizar el conjunto de datos empíricos es ilustrar el funcionamiento del modelo propuesto, teniendo como objetivo particionar la matriz de preferencias a dos vías dos modos y simultáneamente representar los centros de los clusters en un espacio de dimensión baja mediante unfolding. Para demostrar que el espacio reducido de unfolding es óptimo para

Tabla 5.3: Valores estimados de los parámetros en dos dimensiones para el modelo de $T = 5$, $C = 3$ clases latentes con restricciones geométricas incondicionales de $\mu_{tc} = \alpha - d(a_t, b_c)$ para el conjunto de datos artificiales.

Estimación de parámetros			
S_{tc}	λ_{tc}	μ_{tc}	σ^2_{tc}
S_{11}	0.0666	27.03	0.87
S_{12}	0.0666	26.84	0.56
S_{13}	0.0666	5.09	0.63
S_{21}	0.0666	22.47	1.75
S_{22}	0.0666	7.50	1.36
S_{23}	0.0666	22.97	0.67
S_{31}	0.0666	7.57	0.98
S_{32}	0.0666	22.50	1.63
S_{33}	0.0666	6.71	0.62
S_{41}	0.0666	5.55	1.82
S_{42}	0.0666	5.65	1.44
S_{43}	0.0666	26.36	2.87
S_{51}	0.0666	23.10	1.67
S_{52}	0.0666	24.78	1.35
S_{53}	0.0666	23.50	1.26

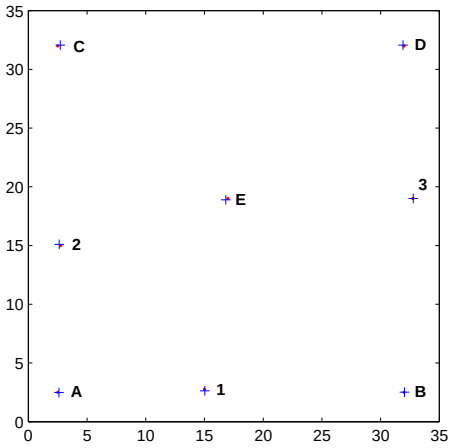

Figura 5.1: Representación procrustes de la solución verdadera ($\cdot$) y la recuperada ($+$) para el conjunto de datos artificiales.

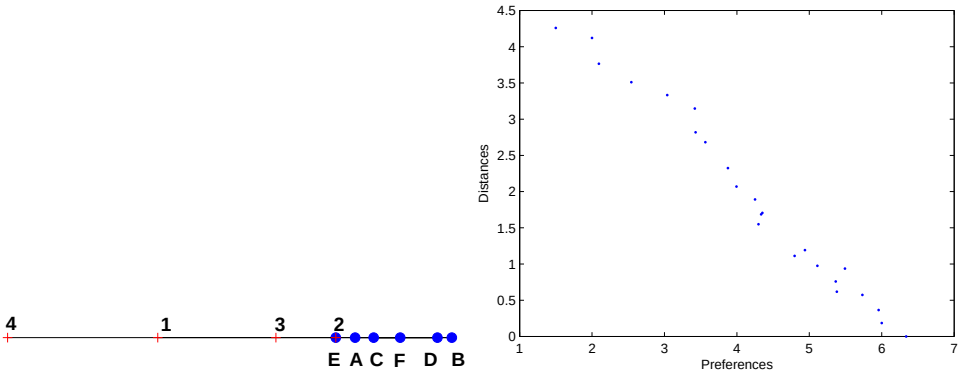

Figura 5.2: Representación óptima de los centros de los clusters en una dimensión para los datos de internet (panel izquierdo) y su diagrama de Shepard (panel derecho). Los respondientes son representados por letras mientras que las declaraciones son representadas por números.

los puntos representantes de los clusters, pero no para los clusters sobrepuestos, mientras que la estructura de clusters es óptima en el espacio original no reducido, el conjunto de datos es también analizado usando un procedimiento que primero realiza agrupamiento a dos modos y entonces obtiene la representación unfolding de los centros de los clusters.

Primero, fue probado el número de clases latentes para los grupos de respondientes T y los grupos de declaraciones de internet C. El valor más bajo del estadístico BIC* (15798.5) se encontró asociado con un valor de la logverosimilitud de -7715.51, para una mezcla de 24 componentes correspondiente a $T = 6$ y $C = 4$ clases latentes. Para el modelo de clases latentes obtenido, se encontró un valor más pequeño del estadístico BIC* (15440.8) con el modelo incondicional de una dimensión comparado con el modelo de dos dimensiones (15663.7) y con el de tres dimensiones (15966.8), y también comparado con el mejor resultado para el modelo condicional, encontrado en tres dimensiones (15719.2). Por lo tanto, el modelo incondicional de clases latentes en una dimensión fue ajustado a los datos, obteniendose los valores de los parámetros que se muestran en la tabla 5.4, donde el valor estimado de $\hat{\alpha}$ =6.1528 es muy cercano a μ_{52}, lo cual hace que la distancia d_{52} sea muy cercana a cero, como es evidente en el panel izquierdo de la figura 5.2.

La tabla 5.5 muestra la conformación de las declaraciones en la estructura de partición O. El cluster O_1 está compuesto de perfiles de declaraciones

Tabla 5.4: Valores estimados de los parámetros en una dimensión para el modelo de $T = 6$, $C = 4$ clases latentes con restricciones geométricas incondicionales de $\mu_{tc} = \alpha - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ para el conjunto de datos de internet.

Estimación de parámetros			
S_{tc}	λ_{tc}	μ_{tc}	σ^2_{tc}
S_{11}	0.0918	4.2614	2.2343
S_{12}	0.0122	5.9675	0.0010
S_{13}	0.0244	5.3938	1.2326
S_{14}	0.0061	2.8213	2.5456
S_{21}	0.1130	3.3346	2.8497
S_{22}	0.0150	5.0408	2.5338
S_{23}	0.0301	4.4671	2.1777
S_{24}	0.0075	1.8945	0.9056
S_{31}	0.1236	4.0832	2.9316
S_{32}	0.0164	5.7894	0.8405
S_{33}	0.0329	5.2157	1.5553
S_{34}	0.0082	2.6432	1.8582
S_{41}	0.1271	3.4717	2.5026
S_{42}	0.0169	5.1779	1.5754
S_{43}	0.0339	4.6042	1.9556
S_{44}	0.0084	2.0316	0.0010
S_{51}	0.1519	4.4473	2.7364
S_{52}	0.0202	6.1535	0.8386
S_{53}	0.0405	5.5798	0.9750
S_{54}	0.0101	3.0073	2.6916
S_{61}	0.0741	3.8289	2.6249
S_{62}	0.0098	5.5351	1.3072
S_{63}	0.0197	4.9614	1.9135
S_{64}	0.0049	2.3888	1.6009

de internet correspondientes a consideraciones relacionadas con la fiabilidad [*Pagar usando Internet es seguro*, *Internet no es fiable*, *El envío de datos personales usando el Internet es inseguro*, *El contenido de los sitios web debería ser regulado*], calidad-costo [*Internet es lento*, *Los precios de las suscripciones del Internet son altos*, *Los costos por navegar son altos*, *Los costos de internet vía telefono son altos*], frecuencia de uso [*A menudo hablo con los amigos sobre el Internet*, *Me gusta estar informado de nuevas cosas importantes*, *Visito regularmente sitios web recomendados por otros*, *Sé mucho sobre el Internet*] y facilidad [*Internet es rápido*, *Internet es de uso amigable*, *Internet es adictivo*]. El cluster O_2 se refiere al entusiasmo [*Internet es el medio de comunicación a futuro*, *Me gusta navegar*]. El cluster O_3 se refiere el riesgo de su facilidad de uso [*Navegar por internet es fácil*, *Internet ofrece muchas posibilidades de abuso*, *Internet ofrece oportunidades ilimitadas*, *Internet es fácil de usar*], y el último cluster está relacionado a la naturaleza online de internet [*Siempre intento nuevas cosas en el internet primero*]. Por lo tanto, la solución de unfolding obtenida sugiere el siguiente orden de declaraciones preferidas: los individuos entrevistados están relacionados al entusiasmo, riesgo de uso, consideraciones prácticas y la naturaleza online, como se puede apreciar en el panel izquierdo de la figura 5.2. En este caso, el modelo de unfolding se reduce a un modelo aditivo de efectos principales, lo cual se muestra en la figura 5.2 por el hecho de que prácticamente todos los clusters de declaraciones están a la izquierda de todos los clusters de respondientes.

Las diferencias entre los clusters de respondientes podrían ser explicadas más fácilmente si estuviera disponible información auxiliar. En general, debido a la naturaleza métrica del modelo unfolding, tales diferencias demuestran el grado con el cual los clusters de respondientes perciben a los clusters de declaraciones. Así, el orden de preferencia puede sugerir un respondiente más conservador para los clusters B y D (quizas de los padres de hijos jovenes), y un perfil de usuario más experimentado para los clusters E, A y C, seguido por el cluster F.

Para ilustrar la utilidad del procedimiento simultáneo de cluster-unfolding propuesto, los datos de internet fueron analizados con un procedimiento de dos pasos que primero determina el mejor modelo de clases latentes a dos-modos y entonces representa los centros de los clusters obtenidos mediante unfolding, considerando los valores de γ_{tc} como las ponderaciones entre conjuntos. Así, primero se considera la clasificación resultante a dos-modos correspondiente al modelo de $T = 6$, $C = 4$ clases latentes asociado al mejor valor del BIC* 15798.5 encontrado en el procedimiento de selección del mo-

Tabla 5.5: Clasificación de declaraciones óptima para el modelo de unfolding incondicional de $C = 4$ clases latentes en una dimensión para los datos de internet.

Grupo de declaraciones	O_c
Pagar usando Internet es seguro	1
Internet no es fiable	1
Internet es lento	1
Internet es de uso amigable	1
Internet es adictivo	1
Internet es rápido	1
Envío de datos personales usando Internet es inseguro	1
Los precios de las suscripciones del Internet son altos	1
Los costos de navegar son altos	1
Los costos de internet vía telefono son altos	1
El contenido de los sitios web debería ser regulado	1
A menudo hablo con los amigos sobre el Internet	1
Me gusta estar informado de nuevas cosas importantes	1
Regularmente Visito sitios web recomendados por otros	1
Sé mucho sobre el Internet	1
Internet es el medio de comunicación a futuro	2
Me gusta navegar	2
Navegar por internet es fácil	3
Internet ofrece muchas posibilidades de abuso	3
Internet ofrece oportunidades ilimitadas	3
Internet es fácil de usar	3
Siempre intento nuevas cosas en el internet primero	4

delo. Entonces, los centros de los clusters resultantes son representados con el modelo incondicional de unfolding en una y en dos dimensiones, obteniéndose los valores del STRESS normalizado de 0.1191, y 0.002, respectivamente.

En términos del agrupamiento a dos-modos, cuando el procedimiento de dos pasos se compara con la clasificación obtenida con el algoritmo de cluster-unfolding combinado en una dimensión, se encontró la misma clasificación para el conjunto de declaraciones pero una diferente para el conjunto de individuos. En términos de la calidad de la solución, el procedimiento combinado obtuvo un valor pequeño del STRESS de 0.0040 en el modelo de una dimensión y de 0.0005 en el modelo de dos dimensiones (este último está relacionado a una clasificación diferente), así como el valor pequeño del BIC* de 15440.8 en una dimensión, y de 15663.7 para el modelo de dos dimensiones. Aunque el valor más pequeño global del BIC* se encontró únicamente en una de veinte réplicas independientes del procedimiento heurístico propuesto para el modelo unidimensional, el valor grande del BIC* (16678.93) que se encontró en el proceso de réplicas en una dimensión corresponde al valor del STRESS para el modelo incondicional de 0.0049, de nuevo un valor del STRESS más pequeño que el encontrado con un procedimiento de dos pasos en una dimensión. Con el procedimiento de dos pasos y considerando una dimensión se encontró una representación ligeramente distinta de unfolding con respecto a la obtenida con el procedimiento heurístico propuesto de SA, con la solución en dos dimensiones siendo más comparable con la que obtuvimos en una dimensión mediante el procedimiento combinado, como puede verse en la figura 5.3. Por lo tanto, en un procedimiento de dos pasos parecen ser necesarias dos dimensiones para representar adecuadamente los centros de los clusters.

Es bien conocido en unfolding que los cambios a través de la escala de disimilaridades pueden influenciar de manera significativa la solución obtenida e incluso a la dimensionalidad estimada de la configuración de puntos (ver por ejemplo Heiser, 1991). Así, en este contexto se analiza la matriz de preferencias originales únicamente para propósitos de ilustración, aunque el unfolding condicional puede ser influenciado por un efecto del estilo de la respuesta (ver Van Rosmalen et al., 2005, para una revisión de procedimientos de optimización para agrupamiento a dos-modos para la matriz de datos doblemente centrada).

Cuando el procedimiento propuesto es aplicado al conjunto de datos originales, se encuentra un valor mínimo global del BIC* correspondiente al modelo de $T = 6$ y $C = 4$ clases, como se puede ver en la figura 5.4. La

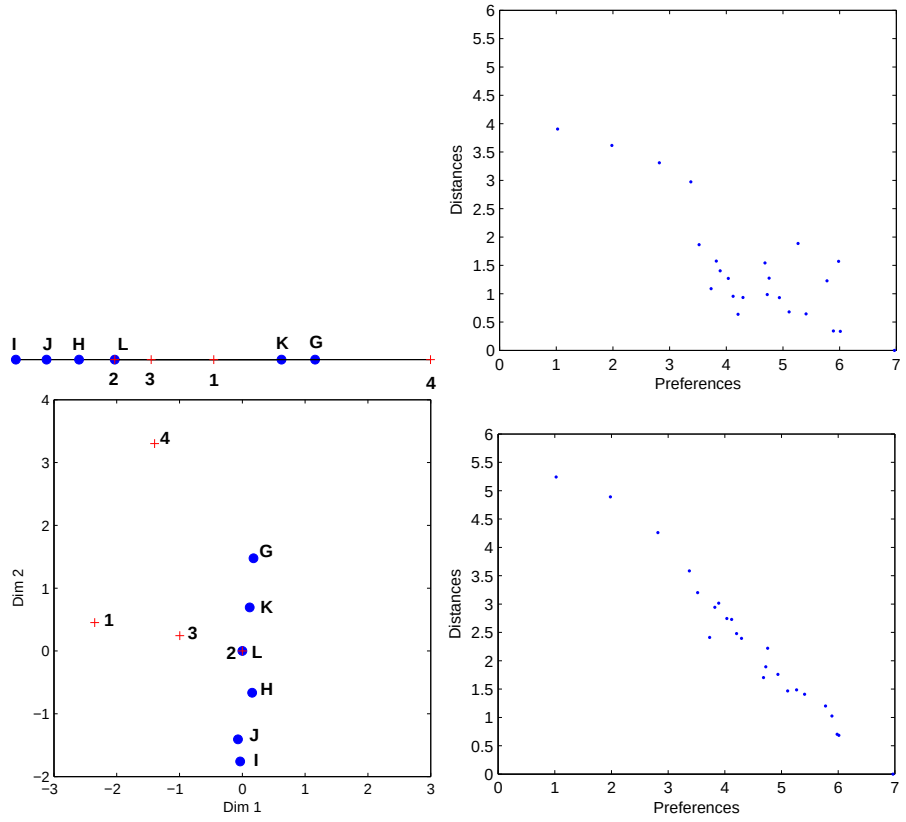

Figura 5.3: Representación de los centro de los clusters para el procedimiento de dos pasos en una y en dos dimensiones para los datos de internet (panel izquierdo) y su diagrama de Shepard (panel derecho). Los respondientes están representados por letras mientras que las declaraciones son representadas por números.

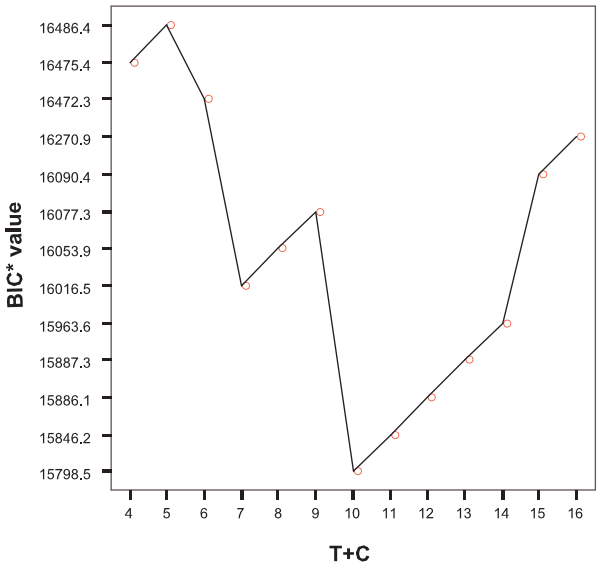

Figura 5.4: Valores ajustados de BIC* para $T+C = 4, \ldots, 16$ para el conjunto de datos de Internet, cuando no se imponen restricciones geométricas en el modelo de clases latentes.

desventaja es que éste demanda un alto costo computacional, inherente a cualquier procedimiento de optimización SA, como se aprecia en la figura 5.5, donde se presentan los tiempos CPU promedio para los datos simulados y para los datos de internet para todo T y C tal que $T + C = m$, para $m = 4, \ldots, 16$, (ver Van Rosmalen et al., 2005), lo cuál sugiere que el procedimiento propuesto puede ser recomendable solamente hasta conjuntos de datos de tamaño mediano.

5.6. Conclusiones y extensiones

En este trabajo proponemos un modelo dual de clases latentes unfolding para datos de preferencias de dos-modos a dos vías. El modelo propuesto puede ser visto como un procedimiento que nos permite categorizar un conjunto de individuos y de objetos mientras que simultáneamente proporciona una representación unfolding de los centros de las categorías, en base a una matriz de preferencias. Las categorías obtenidas y la configuración unfolding

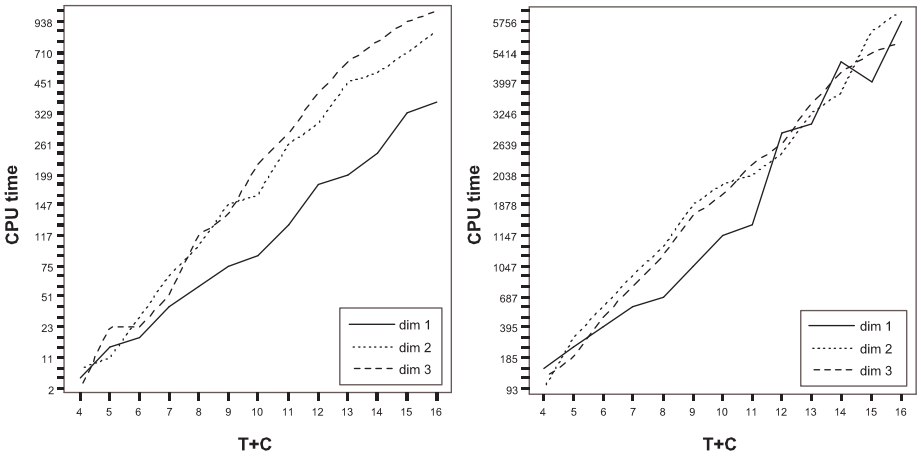

Figura 5.5: Tiempo CPU promedio (en segundos) para $T + C = 4, \ldots, 16$ para los datos Simulados (panel izquierdo) y para los datos de Internet (panel derecho), cuando los datos son representados en una, dos y tres dimensiones.

son estimadas de tal forma que son simultáneamente óptimas en un marco de máxima verosimilitud, permitiendonos tomar una decisión estadística sobre los parámetros en el modelo. En términos de los valores del STRESS, los resultados experimentales demuestran claramente la superioridad del método combinado, comparado con un procedimiento de dos pasos que primero determina la clasificación a dos-modos y entonces representa los centros de los clusters mediante unfolding en un espacio de baja dimensión.

En el análisis de los datos de internet con nuestro modelo, notamos que la solución unidimensional óptima que se encontró para estas similaridades, tiene la propiedad de que $a_t \geq b_c$ para todo t y c. Entonces la distancia unfolding toma la forma específica $d_{tc} = \mid a_t - b_c \mid = a_t - b_c$, y las similaridades promedio reconstruidas bajo esta condición se convierten en

$$\bar{s}_{tc} = \alpha - a_t + b_c.$$

Esta ecuación demuestra que el procedimiento de unfolding propuesto con sus pruebas del modelo nos permite identificar un modelo lineal con estrucutura aditiva simple sin términos de interacción como un caso especial. Esta es la estructura de unfolding no degenerada más simple posible, la cuál describe los efectos principales de la respuesta de los respondientes y los efectos princi-

pales de la atracción de las declaraciones. La conclusión substancial debe ser que en este conjunto de datos no hay evidencia para respuestas diferenciadas de los grupos de respondientes a las declaraciones. Aunque el resultado es posiblemente decepcionante desde un punto de vista de la aplicación, es un resultado satisfactorio desde un punto de vista estadístico, debido a que la aplicación combinada de cluster y unfolding nos permite resumir $193 \times 22 = 4246$ puntos de referencia mediante un modelo con $2 \times (6 \times 4) + (6 + 4) - 1 = 57$ parámetros.

El problema se aborda desde la perspectiva de un modelo de clases latentes, no en los espacios originales de los individuos y de los objetos, sino en la matriz de preferencias bajo una partición en forma de bloques. Por lo tanto, este enfoque introduce una complicación adicional en los procedimientos generales de estimación usados en modelos de clases latentes, debido a que la estimación directa en el modelo de mezclas finitas derivado puede no estar asociada con cualquier partición en los espacios originales. Las mezclas finitas de TC componentes estarán asociadas con todas las factorizaciones en T y C clases latentes en los espacios originales; un método de estimación condicional basado en un procedimiento de optimización Monte Carlo, puede proporcionar una solución fácil y natural al problema, a expensas de invertir mayor tiempo CPU.

Una de las hipótesis del modelo es la consideración de la varianza sin restricciones en los componentes de la mezcla. Aunque esta consideración es congruente con el principio de parsimonia debido a que el empleo de varianzas distintas tiende a reducir el número estimado de clases latentes, la suposición de una varianza constante en todos los componentes de la mezcla pueden ser un procedimiento recomendable, especialmente si se emplean valores grandes de T y C. Además, debido a que SA maneja tanto la restricción de la forma en bloque como el problema de máximos locales inherente al algoritmo EM (a expensas de un incremento en el tiempo CPU), están implicados dos procedimientos de optimización en este algoritmo y el problema del óptimo local persiste, especialmente para conjuntos de datos grandes. Así, el algoritmo de SA propuesto se puede considerar adecuado para conjuntos de datos de tamaño moderado, aunque se pueden investigar otros procedimientos de optimización para reducir el tiempo CPU pero de tal manera que se preserve la calidad de la solución.

Para el tema de la selección del modelo, el estadístico BIC* nos proporciona una aproximación al problema, tanto desde la perspectiva de estimación de la densidad verdadera cuando se comparan modelos igualmente

parametrizados, y cuando el número de clases latentes se establece mediante la determinación del número de componentes en el modelo de mezcla. Sin embargo, también se pueden investigar otros criterios, especialmente aquellos relacionados al principio de descripción de longitud mínima de Rissannen. Desafortunadamente, el problema de determinar el número de componentes en modelos de mezcla no se ha resuelto totalmente (McLachlan y Peel, 2001, Sección 6; Yang y Yang, 2007).

La ocurrencia de degeneraciones es un problema importante en unfolding cuando se permiten transformaciones de los datos que incluye por lo menos un intercepto y una pendiente, tales como las transformaciones de intervalo y ordinales. Aun cuando el procedimiento de SMACOF modificado produce buenos resultados en todos los datos examinados, un procedimiento PREFSCAL (Busing et al. 2005) u otro procedimiento (como el propuesto por Van Deun, Marchal, Heiser, Engelen y Van Mechelen, 2007) podría ser empleado cuando se imponen las restricciones geométricas, y en particular para extender el método para transformaciones ordinales. Nuestro modelo fue desarrollado para el modelo de unfolding simple, pero puede ser desarrollado también para el modelo vectorial, y también en MDS a dos vías, unimodal; esta última posibilidad es el tema de un próximo trabajo del autor.

Capítulo 6

Cluster Differences Unfolding para datos de preferencias bimodales a dos vías

6.1. Introducción

Unfolding es una importante técnica desarrollada por Coombs (1964) originalmente para el análisis de datos de elección de preferencias. La representación conjunta de los individuos $v_i \in V$, $i = 1, \ldots, R$ y de los objetos $o_j \in O$, $j = 1, \ldots, N$ mediante $(R + N)$ puntos en un espacio euclídeo de dimensión M, se determina de las similaridades s_{ij}, $i = 1, \ldots, R$, $j = 1 \ldots, N$, de una matriz de preferencias $\boldsymbol{S}$, que representa una relación de proximidad entre conjuntos.

Se encuentra una matriz $\boldsymbol{A}$ $(R \times M)$ y una matriz $\boldsymbol{B}$ $(N \times M)$, cuyos vectores filas $\boldsymbol{a}_i$, $i = 1 \ldots, R$, y $\boldsymbol{b}_j$, $j = 1, \ldots, N$, representan las coordenadas de los R individuos y de los N objetos respectivamente en dimensión M, de tal manera que las distancias entre las filas de $\boldsymbol{A}$ y $\boldsymbol{B}$ están inversamente relacionadas a las similaridades correspondientes entre los V y O elementos. En este trabajo tratamos con la situación métrica en la cual las preferencias son relacionadas a las distancias mediante una transformación lineal en la cual únicamente se permite una constante aditiva. Esta transformación varía para cada individuo en el modelo condicional por fila, o es la misma para todos los individuos en la situación incodicional. Por lo tanto, el problema de degeneración, que es quizás el problema más grande en unfolding (ver Busing,

Groenen y Heiser, 2005; Borg y Groenen, 2005; o Van Deun, Groenen, Heiser, Busing y Delbeke, 2005, donde se proponen varios procedimientos analíticos para evitar el problema), no existe en la situación actual (ver Vera, Macías y Heiser, 2008).

El modelo métrico de unfolding busca $\boldsymbol{A}$ y $\boldsymbol{B}$, tal que para cada individuo v_i y objeto o_j, la similaridad s_{ij} está inversamente relacionada a la distancia d_{ij}, que generalmente denota la distancia Euclídea entre los vectores $\boldsymbol{a}_i$ y $\boldsymbol{b}_j$, definidos por

$$d_{ij} = d(\boldsymbol{a}_i, \boldsymbol{b}_j) = [(\boldsymbol{a}_i - \boldsymbol{b}_j)'(\boldsymbol{a}_i - \boldsymbol{b}_j)]^{1/2}.$$

Para datos de escala intervalo, el problema general de unfolding métrico se puede formular en un esquema de mínimos cuadrados como la minimización de una forma incompleta del STRESS (Heiser, 1981). Entonces asumiendo que las similaridades entre individuos y entre objetos son datos faltantes, la función de pérdida está dada por,

$$\sigma^2(\boldsymbol{\alpha}, \boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{R} \sum_{j=1}^{N} (\alpha_i - s_{ij} - d_{ij})^2,$$

donde $\boldsymbol{\alpha}$ es el vector $(\alpha_1, \ldots, \alpha_R)'$ en la situación condicional o un escalar α en el modelo incondicional denotando la constante aditiva para los datos de escala intervalo, asumiendo que la pendiente de la transformación lineal está incluida en la escala de la configuración. Así, SMACOF (de Leeuw y Heiser, 1980), o cualquier otro procedimiento general se pueden utilizar para resolver el problema de estimación.

Se ha propuesto un número considerable de métodos de clasificación y espaciales de manera conjunta para representar la información de preferencias similares resumida por medio de grupos, lo cual reduce significativamente el número de parámetro a estimar en el modelo. Para datos a dos vías, se han propuesto varios de estos procedimientos combinados en un marco determinístico, para estimar grupos de individuos en el contexto de MDS (Heiser, 1993; Heiser y Groenen, 1997; Kiers, Vicari y Vichi, 2005; Vera, Macías y Angulo, 2008a), escalamiento óptimo (Van Buuren y Heiser, 1989) y análisis de componentes principales (De Soete y Carroll, 1994; Vichi y Kiers, 2001). Desde la perspectiva de un modelo de clases latentes, se han propuesto varios procedimientos para datos de preferencias para estimar vectores (DeSarbo, Howard y Jedidi, 1990; DeSoete y Winsberg, 1993; Chintaguna, 1994; DeSarbo, Ramaswamy y Chatterjee, 1995) o puntos ideales (DeSarbo, Jedidi, Cool

y Schendel, 1991; De Soete y Heiser, 1993; Böckenholt y Böckenholt, 1991) de clusters individuales (ver DeSarbo, Manrai y Manrai, 1994 o Wedel y DeSarbo, 1996, para una revisión exhaustiva de tales procedimientos).

En el contexto de modelos de clases latentes para datos a dos vías unimodales Vera, Macías y Heiser, (2007) y Vera, Macías y Angulo (2008a) han propuesto un modelo de Cluster-MDS para disimilaridades particionando los objetos en T clusters , $T << N$, mientras que los centros de los clusters son representados por MDS. Así, la naturaleza probabilística de estos enfoques permite la utilización de un criterio estadístico de selección para determinar el número de clusters y la dimensionalidad de la representación. Para datos a dos vías, bimodales, Vera, Macías y Heiser, (2008) propusieron un procedimiento de cluster-unfolding basado en Annealing Simulado para datos de preferencias de escala intervalo, que simultáneamente particiona los individuos en $T << R$ clusters y los objetos en $C << N$ clusters, mientras que ambos conjuntos de centros de clusters son representados mediante unfolding. Se propone un criterio de selección basado en el estadístico BIC* (Yang y Yang, 2007) para la elección de la mejor combinación del número de cluster en el conjunto de individuos y en el conjunto de objetos, así como para determinar la dimensionalidad de la representación. También, se demuestra la superioridad de la metodología combinada de clustering a dos modos y espacial con respecto a un procedimiento de dos pasos que primero reduce los individuos y objetos mediante un agrupamiento a dos modos y entonces representa los centros de los clusters mediante unfolding.

Aunque el método de Vera, Macías y Heiser (2008) funciona bien, en muchas situaciones prácticas las hipótesis de independencia y normalidad de los datos de preferencias pueden resultar restrictivas, y entonces debería ser más recomendable un enfoque determinístico . El procedimiento de estimación propuesto basado en Annealing simulado (SA) también hace menos severo el problema de óptimos locales. La desventaja es que demanda un alto costo computacional, que hace del procedimiento SA únicamente recomendable para conjuntos de datos pequeños o medianos. Primero, deben ser probados TC modelos con $T = 1, \ldots, R$ y $C = 1, \ldots, N$, para determinar el número apropiado de clusters. Entonces, es necesaria otra ejecución imponiendo restricciones geométricas para la estimación final de los parámetros. Por lo tanto, desde un punto de vista exploratorio, debería ser recomendable el procedimiento más eficiente en términos de tiempo CPU, especialmente para conjuntos de datos grandes, debido a que el tiempo CPU crece en el modelo de clases latentes cuando R, N, T y C aumentan (ver Vera, Macías y Heiser,

2008 para más detalles).

Para una matriz de datos de preferencias, en este trabajo se propone un modelo de mínimos cuadrados cuyo objetivo es particionar los individuos y los objetos en $T(T << R)$ y $C(C << N)$ clases, respectivamente, mientras que simultáneamente se representan los $T+C$ centros de los clusters en un espacio de baja dimensión tal que los individuos y objetos mantienen sus relaciones de preferencias. Los parámetros son estimados por medio de un procedimiento condicional alternante de mínimos cuadrados basado en una extensión del procedimiento de distancia mínima (**M**inimal **D**istance) para datos a dos modos propuesto en Heiser y Groenen (1997). Debido a que el procedimiento de distancia mínima es muy dependiente de la solución inicial, se propuso también un algoritmo de Annealing simulado modificado en el contexto de mínimos cuadrados para tratar con el problema de óptimos locales. De esta forma, dada una clasificación de prueba de los individuos y de los objetos, obtenemos una partición de la matriz de preferencias rectangular en bloques, a partir de la cual se estima la configuración de los centros usando SMACOF (de Leeuw y Heiser, 1980). La función de pérdida es entonces evaluada y el proceso se repite hasta que se obtiene la convergencia del algoritmo. Se emplea un procedimiento de selección del modelo basado en la adaptación del criterio de información de Sugar y Gareth (2003) para determinar la mejor combinación del número de clusters para los individuos y objetos. Además, la dimensión de la representación de los centros de los clusters es elegida usando el criterio BIC (Lee, 2001).

6.2. Modelo

Denotamos por $\boldsymbol{E}_T = (e_{it})$, $i = 1, \ldots, R$, $t = 1, \ldots, T$, la matriz indicadores $R \times T$ definida mediante una partición $\mathcal{P}(V)$ del espacio de individuos en T clases V_t con r_t elementos, $t = 1, \ldots, T$, tal que $r_1 + \cdots + r_T = R$, donde $e_{it} = 1$ si $v_i \in V_t$ y cero en cualquier otro caso. De la misma forma, denotamos por $\boldsymbol{E}_C = (e_{jc})$, $j = 1, \ldots, N$, $c = 1, \ldots, C$, la matriz indicadora $N \times C$ definida por una partición $\mathcal{P}(O)$ del espacio de objetos en C clases separadas O_c con n_c elementos, $c = 1, \ldots, C$, tal que $n_1 + \cdots + n_C = N$, donde $e_{jc} = 1$ si $o_j \in O_c$ y cero en cualquier otro caso. Así, $V_t \cap V_{t'} = \varnothing$ para todo t, t' y $O_c \cap O_{c'} = \varnothing$ para todo c, c', y los elementos en $\boldsymbol{S}$ son ordenados permutando las filas y las columnas de acuerdo con la secuencia en los conjuntos de índices de las clases de individuos y de objetos; entonces, la

partición dual provoca una partición en forma de bloques $\mathcal{P}(\boldsymbol{S})$ de la matriz de preferencias $\boldsymbol{S}$ en TC clases separadas $\boldsymbol{S}_{tc}$.

El objetivo en el modelo propuesto es representar los centros de los clusters para el conjunto de individuos V_t, $t = 1, \ldots, T$, mediante los vectores fila $\boldsymbol{a}_t$ de la matrix $\boldsymbol{A}_T$ $(T \times M)$, y los centros de los clusters para el conjunto de objetos O_c, $c = 1, \ldots, C$ por los vectores fila $\boldsymbol{b}_c$ de la matrix $\boldsymbol{B}_C$ $(C \times M)$, en un espacio de dimension M. Denotando por $d_{tc} = d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ la distancia Euclídea entre los vectores representantes de los centros de los clusters, y considerando el vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)'$ en la situación condicional, o un escalar α en el modelo incondicional, las disparidades $\hat{s}_{tc}$ en términos de una medida de similaridad se pueden escribir como

$$\hat{s}_{tc} = \begin{cases} \alpha_t - d_{tc}, & \text{(modelo condicional)} \\ \alpha - d_{tc}, & \text{(modelo incondicional),} \end{cases} \tag{6.1}$$

y el STRESS puede ser formulado como,

$$\sigma^2(\boldsymbol{E}_T, \boldsymbol{E}_C, \boldsymbol{\alpha}, \boldsymbol{A}_T, \boldsymbol{B}_C) = \sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}(s_{ij} - \hat{s}_{tc})^2. \tag{6.2}$$

Considerando la descomposición ortogonal de la suma de cuadrados residuales en cada bloque, el STRESS se puede descomponer aditivamente en dos partes, una dependiente únicamente de la clasificación y la otra dependiente de la clasificación y de las coordenadas de los centros de los clusters $\boldsymbol{A}_T$, y $\boldsymbol{B}_C$ de la siguiente forma:

$$\sigma^2(\boldsymbol{E}_T, \boldsymbol{E}_C, \boldsymbol{\alpha}, \boldsymbol{A}_T, \boldsymbol{B}_C) = \sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}(s_{ij}-\overline{s}_{tc})^2+\sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\overline{s}_{tc}-\hat{s}_{tc})^2 \tag{6.3}$$

donde

$$\overline{s}_{tc} = \frac{\displaystyle\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}s_{ij}}{\displaystyle\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}}, \quad \text{y} \quad \gamma_{tc} = \sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}. \tag{6.4}$$

Se debe notar que el STRESS es simultáneamente minimizado con respecto a la partición óptima en $\boldsymbol{S}$ y a la configuración de los centros de los clusters $\boldsymbol{A}_T$, y $\boldsymbol{B}_C$. Sin embargo, para un valor dado de $\boldsymbol{E}_T$, y $\boldsymbol{E}_C$, se puede encontrar una partición en forma de bloque $\mathcal{P}(\boldsymbol{S})$ y entonces los centros de los clusters pueden ser estimados usando SMACOF. Por lo tanto, se puede emplear un procedimiento de estimación condicional alternante para el procedimiento de estimación global.

6.3. Algoritmo condicional alternante de clustering a dos modos y unfolding

El procedimiento de estimación global es un algoritmo de optimización alternante en el cual en una iteración principal, se obtiene primero una clasificación $\boldsymbol{E}_T$, mientras que $\boldsymbol{E}_C$ y el resto de los parámetros permanecen fijos, entonces, en un procedimiento iterativo secundario los parámetros del modelo son estimados condicionalmente a las clasificaciones actuales de individuos y objetos. En la siguiente iteración principal, se obtiene $\boldsymbol{E}_C$ mientras que $\boldsymbol{E}_T$ y el resto de parámetros permanecen fijos, después los parámetros del modelo son también estimados condicionalmente a la clasificación actual de los individuos y objetos en un procedimiento iterativo secundario. El procedimiento continua hasta que se minimiza el STRESS total (6.2). Por lo tanto, el algoritmo asegura que no únicamente al final, sino cada vez a lo largo del procedimiento de optimización, la relación en el espacio de individuos y objetos entre los elementos y sus correspondientes clases latentes, se preserva.

Para describir el procedimiento de estimación condicional, primero suponemos que la clasificación de objetos (con respecto a los individuos) es fijada en una iteración principal. Entonces, se emplea un procedimiento iterativo secundario para minimizar el STRESS global. Éste está compuesto de dos pasos; primero, la clasificación óptima de los individuos (con respecto a los objetos), se obtiene en una fase de asignación, y entonces los parámetros del modelo son estimados condicionalmente usando SMACOF en una fase de unfolding.

6.3.1. Fase de asignación

En la p-ésima iteración, para un valor dado de $\boldsymbol{\alpha}^{(p-1)}$, $\boldsymbol{A}_T^{(p-1)}$, $\boldsymbol{B}_C^{(p-1)}$ y $\boldsymbol{E}_C^{(p-1)}$, denotamos por $\varsigma_{it}^2(\boldsymbol{S}, \boldsymbol{L}_T^{(p-1)}) = \|\boldsymbol{s}_i - \boldsymbol{l}_t^{(p-1)}\|^2$ la distancia Euclídea entre la i-ésima fila de $\boldsymbol{S}$ y la t-ésima fila de la matriz matrix $\boldsymbol{L}_T^{(p-1)}$ $(T \times N)$ de elementos $l_{tj}^{(p-1)} = \sum_c e_{jc}^{(p-1)} \hat{s}_{tc}^{(p-1)}$. Entonces, denotando por $\kappa^2(\boldsymbol{E}_T) = \sigma^2(\boldsymbol{E}_T, \boldsymbol{E}_C^{(p-1)}, \boldsymbol{\alpha}^{(p-1)}, \boldsymbol{A}_T^{(p-1)}, \boldsymbol{B}_C^{(p-1)})$, el STRESS total se puede minimizar condicionalmente en términos de $\boldsymbol{E}_T^{(p)}$ minimizando,

$$\kappa^2(\boldsymbol{E}_T) = \sum_{i=1}^{R} \sum_{t=1}^{T} e_{it} \varsigma_{it}^2(\boldsymbol{S}, \boldsymbol{L}_T^{(p-1)}). \tag{6.5}$$

Este resultado fue demostrado por Heiser y Groenen (1997) para datos de disimilaridades a dos vías unimodales. En este contexto,

$$\sigma^2(\boldsymbol{E}_T, \boldsymbol{E}_C^{(p-1)}, \boldsymbol{\alpha}^{(p-1)}, \boldsymbol{A}_T^{(p-1)}, \boldsymbol{B}_C^{(p-1)}) = \sum_{i=1}^{R} \sum_{t=1}^{T} e_{it} \sum_{j=1}^{N} \sum_{c=1}^{C} e_{jc}^{(p-1)} (s_{ij} - \hat{s}_{tc}^{(p-1)})^2,$$

y

$$\sum_{c=1}^{C} e_{jc}^{(p-1)} s_{ij}^2 = s_{ij}^2 \quad \text{y} \quad \sum_{c=1}^{C} e_{jc}^{(p-1)} (\hat{s}_{tc}^{(p-1)})^2 = \left(\sum_{c=1}^{C} e_{jc}^{(p-1)} \hat{s}_{tc}^{(p-1)} \right)^2.$$

Así, tomando en cuenta los valores de $\boldsymbol{E}_C^{(p-1)}$, se obtiene el siguiente resultado,

$$\sum_{j=1}^{N} \sum_{c=1}^{C} e_{jc}^{(p-1)} (s_{ij} - \hat{s}_{tc}^{(p-1)})^2 = \sum_{j=1}^{N} \left(s_{ij} - \sum_{c=1}^{C} e_{jc}^{(p-1)} \hat{s}_{tc}^{(p-1)} \right)^2 = \varsigma_{it}^2(\boldsymbol{S}, \boldsymbol{L}_T^{(p-1)}).$$

Por lo tanto, minimizar (6.2) en términos de $\boldsymbol{E}_T$ es equivalente al criterio de clasificación k-means para obtener el valor de $\boldsymbol{E}_T^{(p)}$ que minimiza (6.5),

$$\sum_{i=1}^{R} \sum_{t=1}^{T} e_{it}^{(p)} \varsigma_{it}^2(\boldsymbol{S}, \boldsymbol{L}_T^{(p-1)}) = \kappa^2(\boldsymbol{E}_T^{(p)}).$$

Este resultado también se puede enunciar en términos de una partición de objetos. Así, para un valor dado de $\boldsymbol{\alpha}^{(p-1)}$, $\boldsymbol{A}_T^{(p-1)}$, $\boldsymbol{B}_C^{(p-1)}$ y de

$\boldsymbol{E}_T^{(p-1)}$, denotamos por $\varsigma_{jc}^2(\boldsymbol{S}, \boldsymbol{L}_C^{(p-1)}) = \|\boldsymbol{s}_j - \boldsymbol{l}_c^{(p-1)}\|^2$ la distancia Euclídea entre la j-ésima columna de $\boldsymbol{S}$ y la c-ésima fila de la matriz $\boldsymbol{L}_C^{(p-1)}$ $(C \times R)$ de elementos $l_{ci}^{(p-1)} = \sum_t e_{it}^{(p-1)} \hat{s}_{tc}^{(p-1)}$. Entonces, denotando por $\kappa^2(\boldsymbol{E}_C) = \sigma^2(\boldsymbol{E}_C, \boldsymbol{E}_T^{(p-1)}, \boldsymbol{\alpha}^{(p-1)}, \boldsymbol{A}_T^{(p-1)}, \boldsymbol{B}_C^{(p-1)})$ el STRESS total puede ser minimizado condicionalmente en términos de $\boldsymbol{E}_C^{(p)}$ minimizando,

$$\kappa^2(\boldsymbol{E}_C) = \sum_{j=1}^{N} \sum_{c=1}^{C} e_{jc} \varsigma_{jc}^2(\boldsymbol{S}, \boldsymbol{L}_C^{(p-1)}).$$

De esta forma, dada una clasificación en el espacio de objetos (con respecto al espacio de individuos), primero se encuentra una clasificación de prueba de los individuos (con respecto a los objetos) minimizando el STRESS total. Para este propósito, se emplea una extensión del método de *distancia mínima* de Heiser y Groenen (1997). De $\boldsymbol{E}_c^{(p-1)}$ y $\hat{s}_{tc}^{(p-1)}$, un valor de $\boldsymbol{E}_t^{(p)}$ se puede encontrar minimizando $\kappa^2(\boldsymbol{E}_T)$ como sigue:

1. Para el i-ésimo individuo, se elige la fila $\boldsymbol{e}_i$ de $\boldsymbol{E}_T$, mientras que se fijan el resto de las filas. Entonces, son probados todos los posibles reagrupamientos del individuo v_i en las T clases, siendo asignado al cluster V_t de tal forma que se minimice

$$\min_{e_i} \sum_{t=1}^{T} e_{it} \|s_i - l_t^{(p-1)}\|^2.$$

2. El proceso se repite hasta que todos los individuos son reagrupados y entonces se obtiene $\boldsymbol{E}_T^{(p)}$, mientras que la clasificación de objetos, $\boldsymbol{\alpha}^{(p-1)}$, $\boldsymbol{A}_T^{(p-1)}$ y $\boldsymbol{B}_C^{(p-1)}$ no se actualizan

3. De $\boldsymbol{E}_C^{(p-1)}$ y $\boldsymbol{E}_T^{(p)}$, los valores $\overline{s}_{tc}$ y γ_{tc} , $t = 1, \ldots, T$, $c = 1, \ldots, C$, se actualizan usando (6.4).

Si los objetos (en lugar de los individuos) son elegidos para reagruparse en la fase de asignación, el algoritmo descrito antes puede ser adaptado usando los valores de $\boldsymbol{\alpha}^{(p-1)}$, $\boldsymbol{A}_T^{(p-1)}$, $\boldsymbol{B}_C^{(p-1)}$ y de $\boldsymbol{E}_T^{(p-1)}$, y buscando $\boldsymbol{E}_C^{(p)}$ que minimice,

$$\min_{e_j} \sum_{c=1}^{C} e_{jc} \|s_j - l_c^{(p-1)}\|^2.$$

La fase de asignación consiste de un ciclo de longitud RT para la clasificación de individuos o de longitud NC para la clasificación de objetos. Para cualquier clasificación de prueba en el espacio de individuos y de objetos, el resto de los parámetros en el modelo son estimados imponiendo la restricciones geométricas y usando SMACOF en la fase de unfolding.

6.3.2. Fase de unfolding

Dada cualesquiera clasificaciones $\boldsymbol{E}_T$ y $\boldsymbol{E}_C$ en el espacio de individuos y en el espacio de objetos, se calculan los valores correspondientes $\overline{s}_{tc}$ y γ_{tc}, $t = 1, \ldots, T$, $c = 1, \ldots, C$, y los parámetros $\boldsymbol{\alpha}^{(p)}$, $\boldsymbol{A}_T^{(p)}$, $\boldsymbol{B}_C^{(p)}$ pueden ser actualizados minimizando el segundo término en (6.3), el cual puede ser escrito en términos de las medidas de disimilaridad como

$$\phi(\boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}) = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_{tc} (\hat{d}_{tc} - d_{tc})^2, \tag{6.6}$$

donde las disparidades $\hat{d}_{tc}$ son definidas como

$$\hat{d}_{tc} = \begin{cases} \alpha_t - \overline{s}_{tc}, & \text{(modelo condicional)} \\ \alpha - \overline{s}_{tc}, & \text{(modelo incondicional).} \end{cases} \tag{6.7}$$

Entonces, (6.6) se puede minimizar usando cualquier algoritmo de unfolding ponderado, y en particular mediante la adaptación de SMACOF a unfolding (Heiser 1981, 1987), tomando en cuenta las modificaciones propuestas por Heiser (1991) y De Soete y Heiser (1993) para evitar el problema de que cualquier valor $\hat{d}_{tc}$ puede ser negativo. De esta manera se estiman los valores apropiados de $\boldsymbol{A}_T^{(p)}$, $\boldsymbol{B}_C^{(p)}$, y donde el nuevo valor de $\boldsymbol{\alpha}^{(p)}$ en el modelo de unfolding condicional está dado por

$$\alpha_t^{(p)} = \frac{\sum_{c=1}^{C} \gamma_{tc} (\overline{s}_{tc} + d(\boldsymbol{a}_t^{(p)}, \boldsymbol{b}_c^{(p)}))}{\sum_{c=1}^{C} \gamma_{tc}}, \tag{6.8}$$

mientras que para el modelo incondicional se adopta la expresión,

$$\alpha^{(p)} = \frac{\sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}(\overline{s}_{tc} + d(\boldsymbol{a}_t^{(p)}, \boldsymbol{b}_c^{(p)}))}{RN}. \tag{6.9}$$

El ciclo iterativo principal alterna entre la clasificación de individuos y de objetos. En cada iteración principal, se aplica un procedimiento de asignación de dos fases y un procedimiento de estimación condicional unfolding, y el proceso principal continua hasta que el STRESS total es minimizado. Luego se utilizan dos criterios de optimización para la convergencia del procedimiento de estimación global. El primero es de naturaleza computacional, evaluando el proceso hasta que se alcanza un número máximo de iteraciones previamente establecido. El otro criterio es el de la convergencia, cuando una solución permanece inalterada durante un número $It_{\text{máx}}$ de iteraciones, establecido previamente por el investigador.

Desde un punto de vista práctico, se tomaron en cuenta algunas restricciones. Para valores dados de T y C, la partición inicial $\mathcal{P}(\boldsymbol{S})$ se obtiene asignando aleatoriamente individuos y objetos en T y C clases respectivamente, bajo la condición de que $r_t n_c \geq 2$, para evitar el problema de obtener un bloque con varianza cero, y la presencia de cualquier $S_{tc} = \emptyset$. La restricción computacional anterior introduce la condición de que si $T \geq R/2$, entonces $C < N/3$, e inversamente, si $C \geq N/2$, entonces $T < R/3$, lo cual restringe las posibles combinaciones de clases latentes para el propósito de selección del modelo.

6.4. Criterio de selección del modelo

Para el propósito de estimación de parámetros, los valores de T y C se suponen conocidos. Aunque en algunas situaciones experimentales el número de clusters se fija previamente por el investigador, uno de los principales objetivos del procedimiento de cluster-unfolding es determinar el número de clusters así como la dimensión para la representación de los centros de los clusters. Para este objetivo se proponen dos procedimientos independientes.

6.4.1. Selección del número de clusters. Una extensión del método *jump*

Para cada par de valores de T y C, una partición de la matriz $\boldsymbol{S}$ en TC clases es asociada a la clasificación de clustering a dos-modos. Tal partición se encuentra a partir de una clasificación independiente dada de los individuos, $\boldsymbol{E}_T$ y de los objetos $\boldsymbol{E}_C$, obtenidas sin imponer restricciones geométricas en la formulación del STRESS, mediante la minimización aplicando el procedimiento de distancia mínima de la ecuación

$$\sigma^2(\boldsymbol{E}_T, \boldsymbol{E}_C) = \sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}(s_{ij} - \overline{s}_{tc})^2, \tag{6.10}$$

donde $\hat{s}_{tc} = \overline{s}_{tc}$, $t = 1, \ldots, T$, $c = 1, \ldots, C$, se fija en (6.2).

En la literatura se han propuesto varios criterios para determinar el número de clusters (ver por ejemplo Milligan y Cooper, 1985 y Sugar y James, 2003 para una descripción adicional). Algunos de los criterios más eficientes se basan en el análisis del modelo de dispersión como en Heiser y Groenen (1997), siendo el índice de Calinski y Harabasz (1974), denotado por CH, uno de los más eficientes, dado por

$$CH = \frac{\sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_{tc}\overline{s}_{tc}^2/TC}{\sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}(s_{ij} - \overline{s}_{tc})^2/(RN - TC)}. \tag{6.11}$$

Sin embargo, en el contexto de variables aleatorias, Sugar y James (2003), demostraron que el índice CH puede subestimar el número verdadero de clusters cuando un gran número de dimensiones está asociado con la variable. Así, propusieron un criterio no paramétrico basado en el concepto de *distorsión*, una cantidad que mide la distancia promedio, por dimensión, entre cada observación y su centro de cluster más cercano. Formalmente, sea X una variable aleatoria p-dimensional con una mezcla de distribuciones de G componentes, cada uno con covarianza Γ. Sea $f_1, \ldots, f_K$ un conjunto candidato de centros de cluster, y sea f_X el más cercano a X. Entonces, la distorsión mínima obtenible asociada con el ajuste de los κ centros a los datos es el promedio de distancia Mahalanobis , por dimensión, entre X y f_X, dada por

$$d_\kappa = \frac{1}{p} \min_{f_1,\ldots,f_K} E[(X - f_X)^\mathsf{T} \Gamma^{-1} (X - f_X)]. \tag{6.12}$$

Cuando Γ es la matriz identidad, esta cantidad es simplemente el cuadrado medio del error, y en práctica, d_κ se estima mediante la distorsión mínima obtenida aplicando el algoritmo de clustering k-means a los datos observados. Para elegir el número de clusters, Sugar y James (2003) propusieron graficar la distorsión transformada d_κ^{-q} versus κ, para todos los valores candidatos de κ, donde $q > 0$ es un valor adecuado tal que la gráfica podría exhibir un salto abrupto en el número verdadero de clusters. Para seleccionar el valor de q, se pueden examinar todos los valores posibles de manera decreciente desde $p/2$, hasta que el salto máximo ocurre (ver Sugar y James, 2003, Sección 5).

Para la implementación del criterio de distorsión, es necesario conocer los valores de X. Así, para los datos de preferencias, el criterio *del salto* de Sugar y James (2003) no puede ser utilizado directamente, entonces proponemos una extensión del criterio de distorsión de tal forma que nos permita aplicar el criterio del salto en este contexto. Para este objetivo, la matriz de datos de preferencias $\boldsymbol{S}$ es representada primero mediante unfolding en un espacio de dimensión completa M^*, y la configuración asociada $\boldsymbol{X} = [\boldsymbol{A}^\mathsf{T} | \boldsymbol{B}^\mathsf{T}]^\mathsf{T}$ se emplea para el criterio de distorsión. Denotando por $M^*_{(0)}$ el número de valores propios positivos asociados a la solución clásica, cuando el problema de unfolding se considera un caso especial de MDS donde las proximidades *dentro* del mismo conjunto (de individuos y objetos) son valores faltantes (Heiser, 1981). Entonces, desde un punto de vista práctico, la dimensión M^* se puede determinar por la dimensión menor desde $1 \leq M^* \leq M^*_{(0)}$, asociada al valor más grande del STRESS por debajo de un valor $MINSTRESS$ previamente establecido por el investigador. De esta forma, el procedimiento SMACOF también se puede usar para obtener X en M^* dimensiones.

La aplicación del método del salto en este contexto puede describirse como sigue:

1. De $\boldsymbol{S}$, se determina el número inicial de dimensiones $M^*_{(0)}$ por el número de valores propios positivos asociados en el contexto de MDS clásico, cuando el problema de unfolding es considerado un caso especial de MDS. Entonces la configuración auxiliar X se obtiene usando SMACOF, examinando todas las configuraciones desde 1 a $M^*_{(0)}$ dimensiones, y eligiendo la configuración X en M^* dimensiones asociada al valor mas grande del STRESS por debajo de $MINSTRESS = 10^{-6}$.

2. Para cada κ, desde $\kappa = 4$ a $\kappa = H$ (considerando por lo menos dos clusters en las filas y en las columnas de S para evitar soluciones triviales), donde H es cualquier valor entero establecido por el investigador, el modelo es ejecutado sin imponer restricciones geométricas para todas las combinaciones posibles del número de clusters de individuos y de objetos tal que $T + C = \kappa$. Entonces, la combinación de los valores de T y C que minimicen (6.10) se selecciona como la asociada al valor κ.

3. Para cada valor κ, el índice de distorsión d_κ se calcula usando (6.12), donde $\Gamma = I$, y $f_1, \ldots, f_T, f_{T+1}, \ldots, f_{T+C}$ son los $T + C = \kappa$ centros de cluster asociados a la correspondiente partición en X dada por,

$$f_i = \frac{1}{M^*} \sum_{m=1}^{M^*} x_{im}, \quad i = 1, \ldots, \kappa.$$

   Este procedimiento se repite para todos los valores κ.

4. La transformación de potencia $q > 0$ para el índice de distorsión es elegida de tal manera que ocurrirá un salto máximo (un valor típico es $q = M^*/2$, o algún valor menor, como propusieron Sugar y James, 2003), y los valores de los saltos $J_\kappa = d_\kappa^{-q} - d_{\kappa-1}^{-q}$ se calculan, definiendo $d_\kappa^{-q} = 0$, para $\kappa < 4$, para evitar menos de cuatro clusters en el modelo.

5. El número de clusters en el conjunto de individuos y en el de objetos son estimados de tal forma que estén asociados a $\kappa^* = \arg\max_\kappa J_\kappa$, es decir, el valor de κ asociado al salto mas grande.

6.4.2. Selección de la dimensionalidad de la representación

El modelo también proporciona la posibilidad de determinar la dimensionalidad de la representación unfolding usando un criterio de información BIC en una aproximación de clases latentes al modelo de cluster-unfolding. Según lo precisado por Lee (2001), el problema de determinar la dimensionalidad apropiada de una representación unfolding debe ser tratado por el principio de parsimonia, es decir, se debe elegir el modelo que mejor se adapte a los datos pero que además exhiba la complejidad mínima. De esta forma, el estadístico ajustado BIC* (Yang y Yang, 2007) puede adoptarse para la selección

del modelo. Sin embargo, para su aplicación es necesario proporcionar una formulación probabilística de los datos fijos mostrados por diferentes configuraciones espaciales. Para este fin, el modelo de clases latentes unfolding de Vera, Macías y Heiser (2008) nos proporciona el marco apropiado para la formulación del BIC*.

En la formulación del modelo, se asume que cada preferencia pertenece a uno y solamente a uno de los TC subconjuntos S_{tc}, mientras que se preserva la condición de la partición en forma de bloques, pero no es conocido de antemano a cual bloque latente pertenece una preferencia particular. Denotando por λ_{tc} la probabilidad incondicional de que un valor de la preferencia s_{ij} pertenezca a un bloque latente S_{tc}, mientras $v_i \in V_t$ y $o_j \in O_c$, donde $0 \leq \lambda_{tc} \leq 1$, y

$$\sum_{t=1}^{T}\sum_{c=1}^{C} \lambda_{tc} = 1. \tag{6.13}$$

Entonces asumimos que las s_{ij} que pertenecen a una clase latente S_{tc} son observaciones de variables aleatorias independientes y normalmente distribuidas de parámetros media y varianza μ_{tc} y σ_{tc}^2 respectivamente, es decir,

$$s_{ij} \sim \mathcal{N}(\mu_{tc}, \sigma_{tc}^2), \text{ para } s_{ij} \in S_{tc}, \tag{6.14}$$

donde las medias de los bloques μ_{tc} están geometricamente relacionadas a los correspondientes pares de centros de clusters de individuos y objetos mediante $\mu_{tc} = \alpha_t - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$ en el modelo condicional de unfolding , mientras que en el modelo incondicional se considera la situación particular $\alpha_t = \alpha, t = 1, \ldots, T$. La hipótesis de la distribución normal para los componentes de la mezcla es congruente con la formulación de mínimos cuadrados de unfolding. Cuando la configuración unfolding es estimada, y considerando la invarianza rotacional y translacional de la solución unfolding, los grados de libertad del modelo son $T(1+2C)+(T+C)M-(M(M+1)/2)-1$ para el modelo condicional y $2(TC)+(T+C)M-M(M+1)/2$ si un escalar α es considerado, lo cual nos permite establecer una cota superior para la dimensionalidad del modelo, tal que los grados de libertad del modelo completo sean menores que los del modelo sin restricciones.

La f.d.p. de s_{ij} como una mezcla de densidades normales univariadas de la forma,

$$g(s_{ij} \mid \boldsymbol{A}_T, \boldsymbol{B}_C, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = \sum_{t=1}^{T}\sum_{c=1}^{C} \lambda_{tc} f_{tc}(s_{ij} \mid \boldsymbol{a}_t, \boldsymbol{b}_c, \alpha_t, \sigma_{tc}^2), \tag{6.15}$$

donde $\boldsymbol{\Sigma} = (\sigma_{tc}^2)$ denota la matrix $T \times C$ de varianzas dentro de bloques, $\boldsymbol{\alpha}$ es el vector $(\alpha_1, \ldots, \alpha_T)'$ en la situación condicional o un escalar en el modelo incondicional, y $\boldsymbol{\Lambda} = (\lambda_{tc})$ es la matriz $T \times C$ de probabilidades incondicionales bajo la restricción en forma de bloques de $\mathcal{P}(\boldsymbol{S})$, y donde $f_{tc}(s_{ij})$ es la f.d.p normal de $s_{ij} \in S_{tc}$.

Entonces, el criterio del BIC ajustado puede escribirse como

$$BIC^* = \sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}\left(\log(\widehat{\sigma}_{tc}^2) + \frac{(s_{ij} - \widehat{s}_{tc})^2}{\widehat{\sigma}_{tc}^2}\right) + \eta \log h \tag{6.16}$$

donde

$$\widehat{\sigma}_{tc}^2 = \frac{\displaystyle\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}(s_{ij} - \widehat{s}_{tc})^2}{\displaystyle\sum_{i=1}^{R}\sum_{j=1}^{N} e_{it}e_{jc}}, \tag{6.17}$$

y donde $h = (RN+2)/24$, y η es el número de parámetros desconocidos dados por $\eta = 1 + (TC) + (T + C)M - M(M + 1)/2$ en el modelo incondicional, mientras que considerando $\alpha_t = \alpha$, $t = 1 \ldots, T$, este valor está dado por $\eta = T + (TC) + (T + C)M - M(M + 1)/2$.

6.5. Aplicaciones ilustrativas

Para ilustrar el desempeño del algoritmo propuesto primero se consideraron dos conjuntos de datos previamente analizados por Vera, Macías y Heiser (2008), y los resultados dados son comparados con los obtenidos mediante el modelo de clases latentes. El procedimiento extendido de **jump** se emplea para determinar el número de clusters de individuos y de objetos, y los resultados obtenidos se comparan con los alcanzados por el criterio de Calinski

y Harabasz (1974) para probar el funcionamiento del criterio propuesto de selección del modelo .

El procedimiento propuesto fue implementado en Fortran, trabajando en un ordenador Pentium IV 3.00 GHz con 1 Gb de RAM bajo Microsoft Windows XP. Debido a que es conocido que el método de distancia mínima depende de la solución inicial (ver Heiser y Groenen, 1997), se eligió el mejor óptimo local obtenido en 3000 replicas independientes como la mejor solución. Para el procedimiento de estimación global y para el algoritmo SMACOF en el paso de la estimación de la configuración, empleamos un criterio de convergencia de un maximo de 300 iteraciones con una diferencia en los valores subsecuentes de STRESS menor a 10^{-7}.

Primero, fue analizada la matriz rectangular de preferencias artificiales de Vera, Macías y Heiser (2008). Los datos fueron generados después de localizar en un plano 20 individuos agrupados en 5 clusters, y 12 objetos agrupados en 3 clusters. Entonces, se derivó una matriz de preferencias 5×3 de las distancias Euclídeas entre las coordenadas de los centros de los clusters localizados, usando la condición geométrica $\hat{s}_{tc} = \alpha - d(\boldsymbol{a}_t, \boldsymbol{b}_c)$, donde el valor de α fue escogido arbitrariamente mayor que la distancia máxima. La varianza entre-clusters σ_{tc}^2 se calculó de tal forma que aproximadamente el 25 % de la varianza total de los datos generados fue error de varianza. De cada componente de la mezcla en el modelo de clases latentes, se generaron 49 valores, correspondientes a una matriz en bloque 7×7, S_{tc}, generando así una matriz de datos correspondiente a 35 individuos y 21 objetos.

Sin imponer restricciones geométricas, primero se determinó el número de clusters de individuos y objetos analizando el conjunto de datos artificiales para toda combinación de los valores de T y C, tal que $T + C = \kappa$, para $\kappa = 4, \ldots, 12$. Para cada valor de κ, se seleccionó el par de valores T and C correspondientes al STRESS más bajo, para lo cual se calcularon el criterio CH (6.11), y el método *jump* usando la configuración auxiliar X en $M^* = 20$ dimensiones y el valor de $q = 3$ para la transformación de la distorsión. El valor más grande del *jump* se obtuvo para $T + C = 8$, correspondiente a la combinación del número de clusters de $T = 5$ y $C = 3$, la misma solución se obtuvo con el método CH. La tabla 6.1 muestra los resultados correspondientes a todas las combinaciones de valores de T y C hasta 6, siendo limitados por razones de espacio. En la figura 6.1 se representan la curva de la distorsión transformada (multiplicada por $6 \cdot 10^6$ para propósitos comparativos) y el índice CH, contra los valores de κ.

Para $T = 5$ y $C = 3$ clusters, fue aplicado el modelo considerando la

Tabla 6.1: Resultados obtenidos por los criterios *jump* y CH para determinar el número de clusters de individuos y de objetos para el conjunto de datos artificiales. Para cada κ, $4 \leq \kappa \leq 12$, se presenta el par (T, C) asociado al valor más bajo del STRESS, y para el criterio jump se emplearon los valores de $q = 3$ y $M^* = 20$.

Resultados del Análisis sin restricciones geométricas					
κ	(T, C)	d_κ	d_κ^{-3}	JUMP	CH
4	2,2	14.39	0.00033	0.00021	1296.7783
5	3,2	12.18	0.00055	0.00021	1351.7747
6	4,2	9.52	0.00115	0.00060	1494.3847
7	4,3	7.65	0.00223	0.00107	1946.2259
8	5,3	6.04	0.00453	0.00230	13862.4130
9	5,4	5.91	0.00484	0.00030	10555.3104
10	6,4	5.68	0.00545	0.00061	9002.5017
11	6,5	5.54	0.00588	0.00042	7311.4605
12	6,6	5.48	0.00607	0.00019	6146.6227

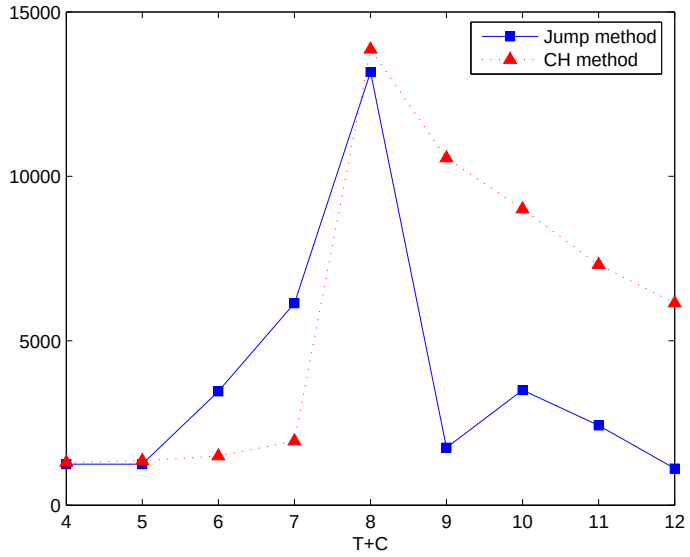

Figura 6.1: Gráfico del *jump* ($\times 6 \cdot 10^6$) usando $M^* = 20$ y $q = 3$ (cuadrados), y gráfico de CH (triángulos) para el conjunto de datos artificiales.

Tabla 6.2: Resultados del criterio BIC^* para el modelo de $T = 5$ y $C = 3$ clases, cuando los centros de los cluster se restringen a ser escalados incondicionalmente en una, dos y en tres dimensiones, para el conjunto de datos artificiales.

Resultados para el modelo incondicional de cluster-unfolding				
No. de Clases (T,C)	No. de dimensioness	No. parámetros	Stress Total	BIC*
5 , 3	1	23	13436.15	4047.94
5 , 3	2	29	957.31	2302.65
5 , 3	3	34	956.03	2318.76

Tabla 6.3: Preferencias promedio estimadas por cluster $\hat{s}_{st}$ y tamaño del cluster para el conjunto de datos artificiales.

	O_1	O_2	O_3	tamaño
V_1	27.04	26.85	5.09	7
V_2	22.47	7.50	22.97	7
V_3	7.57	22.50	6.71	7
V_4	5.55	5.65	26.36	7
V_5	23.10	24.78	23.50	7
tamaño	7	7	7	

restricción geométrica $\hat{s}_{tc} = \alpha - d(a_t, b_c)$. Se seleccionaron dos dimensiones para la representación de los centros de los clusters, que corresponde al valor más pequeño del estadístico BIC (2302.65), como se puede apreciar en la tabla 6.2. En la tabla 6.3 se muestran los valores promedios resultantes de las preferencias, $\hat{s}_{tc}$, y el tamaño de los centros de los cluster de individuos y objetos. Las distancias más pequeñas entre los centros de los clusters se corresponden con grandes valores de $\hat{s}_{tc}$, encontrándose siete elementos por cluster, como se esperaba. La configuración verdadera es recuperada satisfactoriamente por el modelo, como puede ser apreciado en la representación Procrustes de la figura 6.2.

Para ilustrar el desempeño del modelo para conjuntos de datos reales, fue considerado el conjunto de datos de internet analizado previamente por Vera, Macías y Heiser (2008). El conjunto de datos de preferencias corresponden a la evaluación sobre una escala Likert de siete puntos, variando de 1 - discrepa totalmente a 7 - de acuerdo totalmente, de 22 declaracio-

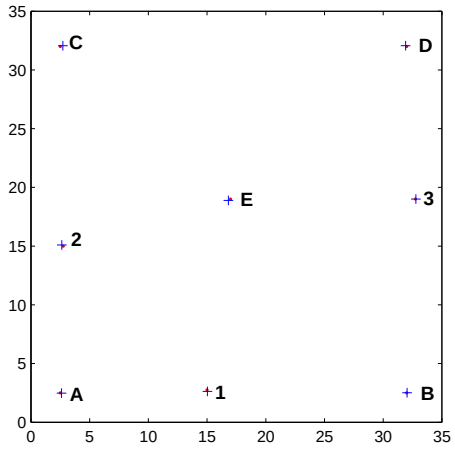

Figura 6.2: Representación Procrustes de la solución verdadera (puntos) y de la recuperada (cruces) para el conjunto de datos artificiales. Los centros de los clusters para individuos son representados por letras, mientras que los clusters para los objetos por números.

nes sobre internet obtenidas de 193 respondientes en la Universidad Erasmus de Rotterdam, después de eliminar los valores faltantes (ver Borg y Groenen, 2005). La matriz $\boldsymbol{S}$ de los datos originales de internet fue recopilada de *http://people.few.eur.nl/groenen*, la cual (previamente centrada) fue analizada por Van Rosmalen, Groenen, Trejos y Castillo (2005), en el contexto de clustering a dos-modos, para comparar varios procedimientos.

Siguiendo el procedimiento de selección del modelo, se emplearon los criterios *jump* y CH para probar el número de clusters para $4 \leq \kappa \leq 16$ (ver Van Rosmalen et al., 2005, o Vera, Macías y Heiser, 2008). Para el método *jump*, primero se analizó la matriz de preferencias completa 193×22 usando SMACOF y se consideró la configuración dada en dimensión $M^* = 49$ como la configuración auxiliar. Usando el valor de $q = 20$ para la transformación de la distorsión, el valor más grande del *jump* se obtuvo para $T + C = 13$ clases, correspondiente a la combinación del número de clusters $T = 7$ y $C = 6$, como se aprecia en la tabla 6.4. Para el criterio CH, se encontró la combinación $T = 2$ y $C = 2$ correspondiente a $\kappa = 4$, la cual comparada a la solución *jump* parece subestimar el número verdadero de clusters, como

Tabla 6.4: Resultados obtenidos por los criterios *jump* y CH para determinar el número de clusters de individuos y objetos para el conjunto de datos de Internet. Para cada κ, $4 \leq \kappa \leq 16$, se muestra el par (T, C) asociado al STRESS más pequeño, y para el criterio *jump* se usaron los valores de $q = 20$ y $M^* = 49$.

Resultados del Análisis sin restricciones geométricas					
κ	(T, C)	d_κ	d_κ^{-20}	JUMP	CH
4	2,2	1.930	1.94E-06	6.93E-07	8609.050
5	2,3	1.922	2.11E-06	1.68E-07	6100.249
6	3,3	1.872	3.58E-06	1.46E-06	4229.143
7	3,4	1.851	4.48E-06	9.06E-07	3285.104
8	4,4	1.807	7.25E-06	2.77E-06	2533.941
9	4,5	1.759	1.24E-05	5.17E-06	2092.115
10	5,5	1.735	1.63E-05	3.93E-06	1721.102
11	6,5	1.715	2.06E-05	4.27E-06	1465.931
12	6,6	1.703	2.37E-05	3.11E-06	1242.147
13	7,6	1.671	3.47E-05	1.09E-05	1084.677
14	7,7	1.663	3.82E-05	3.49E-06	945.855
15	7,8	1.655	4.20E-05	3.86E-06	844.335
16	8,8	1.651	4.41E-05	2.05E-06	749.237

se aprecia en la figura 6.3, lo cual coincide con lo demostrado por Sugar y James (2003).

Cuando se consideran las restricciones geométricas para $T = 7$ y $C = 6$, el valor más pequeño del BIC^* para el modelo incondicional y condicional, en una, dos y tres dimensiones, se encontró para el modelo incondicional en tres dimensiones (14659.41), como se aprecia en la tabla 6.5. A partir de este modelo se obtiene la partición de preferencias de los clusters mostrada en la tabla 6.6. Como se puede observar en la figura 6.4, se presenta una mayor interacción entre los centros de los clusters de individuos y objetos en tres dimensiones que la encontrada por Vera, Macías y Heiser (2008) en el contexto de clases latentes. Para la solución obtenida, se encontró un valor de 0.05328 para el índice de entremezclado de DeSarbo et al. (1997).

La composición de los $C = 6$ clusters de declaraciones se presenta en la tabla 6.7. El cluster O_1 está compuesto de las declaraciones de internet que se refieren a la frecuencia de su uso [*Internet es adictivo, Frecuentemente hablo con los amigos sobre Internet, Me gusta estar informado de cosas importantes nuevas, Visito regularmente sitios Web recomendados por otros,*

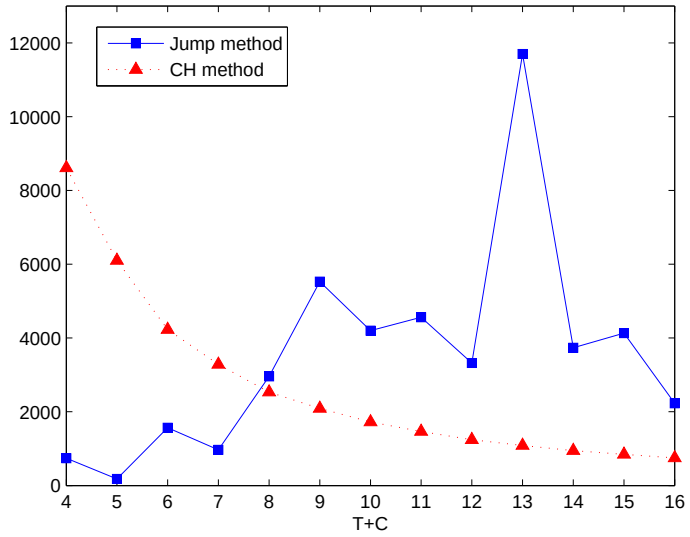

Figura 6.3: Gráfico de Jump ($\times 10^9$) usando $M^* = 49$ y $q = 20$ (cuadrados), y gráfico de CH (triángulos) para el conjunto de datos de internet.

Tabla 6.5: Resultados del criterio de información para el modelo de $T = 7$ y $C = 6$ clusters, cuando los centros de los clusters son escalados en una, dos y en tres dimensiones, para el conjunto de datos de internet.

Resultados para el modelo incondicional de cluster-unfolding				
No. de Clases (T,C)	No. de dimensiones	No. parámetros	Stress Total	BIC*
7 , 6	1	55	8382.18	15002.03
7 , 6	2	66	7671.71	14698.78
7 , 6	3	76	7505.73	14659.41
Resultados para el modelo condicional de cluster-unfolding				
No. de Clases (T,C)	No. de dimensiones	No. parámetros	Stress Total	BIC*
7 , 6	1	61	7869.23	14749.82
7 , 6	2	72	7589.80	14695.82
7 , 6	3	82	7450.50	14664.77

Tabla 6.6: Preferencias estimadas entre clusters $\hat{s}_{tc}$ y tamaño de los clusters, para el conjunto de datos de internet, cuando los centros de los clusters están restringidos a ser escalados incondicionalmente en tres dimensiones.

Clusters de individuos	Clusters de objetos						
	O_1	O_2	O_3	O_4	O_5	O_6	Tamaño
V_1	3.93	5.37	3.35	4.25	2.62	5.88	34
V_2	5.08	5.99	2.99	4.84	3.71	2.48	34
V_3	4.84	5.11	3.53	3.46	3.98	1.84	16
V_4	2.73	4.24	3.07	3.60	1.70	1.92	27
V_5	2.48	4.73	3.33	4.44	1.14	5.39	29
V_6	3.55	5.28	2.73	5.15	2.21	2.50	31
V_7	3.91	5.82	3.12	5.19	2.44	5.36	22
Tamaño	5	6	3	6	1	1	

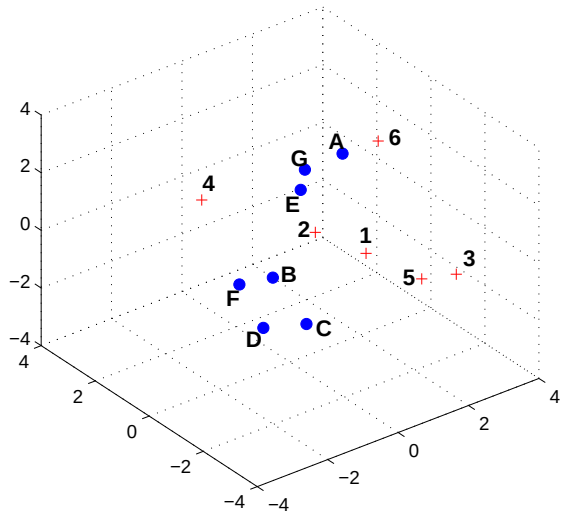

Figura 6.4: Representación óptima de los centros de los cluster en tres dimensiones para los datos de internet. Los respondientes están representados por letras mientras que las declaraciones por números.

Sé mucho sobre Internet]. El cluster O_2 agrupa declaraciones que reflejan un entusiasmo por internet pero también un riesgo debido su facilidad de uso [*Internet es el medio de comunicación del futuro, Me gusta navegar, Navegar en internet es fácil, Internet ofrece muchas posibilidades de abuso, Internet ofrece oportunidades ilimitadas, Internet es fácil de usar*]. El cluster O_3 esta compuesto por declaraciones relacionados a la confiabilidad [*Pagar mediante internet es seguro, Internet no es fiable, Internet es rápido*]. El cluster O_4 se refiere a declaraciones que reflejan pesimismo [*Internet es lento, Internet es de uso amigable, Enviar datos personales usando Internet es inseguro, los precios de suscripciones a Internet son altos, los costos por navegar son altos, los costos de Internet vía teléfono son altos*]. El cluster O_5 está relacionado a la naturaleza online [*siempre intento nuevas cosas en Internet primero*] y el cluster O_6 a la libertad de su uso [*El contenido de los sitios Web debe ser regulado*]. La solucion de unfolding podría sugerir que los respondientes en los clusters V_1, V_5 y V_7 se pronuncian principalmente por una regulacion de internet, ponderan la relacionan calidad-costo del servicio y aunque muestran interes tienen poca disponibilidad para explorar todo su contenido, lo que los perfila como gente adulta o padres de usuarios jovenes. Los respondientes en los clusters V_2 y V_3 parecen ser expertos usuarios de internet que advierten sin embargo de su riesgo debido a la facilidad de uso, aunque paradójicamente no consideran su regulacion, que los perfila como usuarios jovenes. Los respondientes en los clusters V_4 y V_6 manifiestan entusiasmo por internet y advierten de sus riegos pero tambien muestran muy poca predisposición para explorar todo su contenido, lo que indica un perfil de usuarios noveles.

Para ilustrar la utilidad del procedimiento simultáneo de cluster-unfolding propuesto, los datos de internet fueron también analizados con un procedimiento de dos pasos que primero determina el mejor modelo de clases a dos-modos y entonces representa los centros de los clusters obtenidos mediante unfolding, considerando los valores γ_{tc} como las ponderaciones entre-conjuntos. De esta forma, se considera primero la clasificación a dos-modos resultante correspondiente al modelo $T = 7$, $C = 6$ asociada con el valor más grande del jump. Entonces, los centros de los clusters resultantes fueron representados con el modelo incondicional de unfolding en tres dimensiones, obteniendo el valor del STRESS normalizado de 0.0021. En términos de agrupamiento a dos-modos, una clasificación diferente se encontró para el conjunto de individuos y para el conjunto de objetos, cuando el procedimiento de dos pasos se compara con la clasificación obtenida con el algoritmo combinado de cluster-unfolding en tres dimensiones. En términos de la cali-

Tabla 6.7: Clasificación óptima de las declaraciones para el modelo incondicional de unfolding métrico de mínimos cuadrados para $C = 6$ en tres dimensiones para los datos de internet.

Grupo de declaraciones	O_c
Internet is adictivo	1
Frecuentemente hablo con los amigos sobre Internet	1
Me gusta estar informado de cosas importantes nuevas	1
Visito regularmente sitios Web recomendados por otros	1
Sé mucho sobre Internet	1
Internet es el medio de comunicación del futuro	2
Me gusta navegar	2
Navegar en internet es fácil	2
Internet ofrece muchas posibilidades de abuso	2
Internet ofrece oportunidades ilimitadas	2
Internet es fácil de usar	2
Pagar mediante internet es seguro	3
Internet no es fiable	3
Internet es rápido	3
Internet es lento	4
Internet es de uso amigable	4
Enviar datos personales usando Internet es inseguro	4
Los precios de suscripciones a Internet son altos	4
Los costos por navegar son altos	4
los costos de Internet vía teléfono son altos	4
Siempre intento nuevas cosas en Internet primero	5
El contenido de los sitios Web debe ser regulado	6

dad de la solución, el procedimiento combinado obtuvo el valor más pequeño del STRESS de 0.00062 para el modelo en tres dimensiones.

6.6. El problema del mínimo local

El algoritmo de optimización basado en la distancia mínima es muy eficiente en términos del tiempo CPU comparado con los procedimientos usuales de optimización Monte Carlo. La desventaja es que la solución obtenida es muy dependiente de la solución inicial. Experimentalmente, este problema se puede apreciar, por ejemplo, en el conjunto de datos artificiales, para los cuales el valor más pequeño del STRESS se obtuvo 729 veces en 3000 réplicas, lo que indica un índice de atracción (es decir, el porcentaje de veces que se obtiene la mejor solución) de 24.3 %. Por lo tanto, se puede emplear un procedimiento de estimación condicional basado en un algoritmo de optimización de Monte Carlo, para el problema de estimación global.

6.6.1. Annealing simulado para propósitos de estimación

Annealing simulado (SA) es un método estocástico de optimización global introducido por Metropolis et al. (1953) para la minimización de una función en un conjunto finito muy grande, y después aplicado para optimización sobre un conjunto continuo (Duflo, 1996). Kirkpatrick, Gelatt y Vecchi (1983), e independientemente Černy (1985), demostraron su utilidad para encontrar óptimos globales en problemas de optimización combinatorios. Annealing está relacionado al principio de termodinámica de cristalización y es un proceso usado para revelar el estado de la baja temperatura de un material, en el cual el material primero se calienta a un estado de energía atómica alto, y subsecuentemente la temperatura se reduce lentamente hasta que el material se ha enfriado de tal forma que el cristal resultante es perfecto, es decir, existe un estado de energía mínima. El ritmo de enfriamiento es un aspecto fundamental del procedimiento; si el material se enfría muy rápido, existirán impurezas y no se obtendrá un óptimo.

En annealing simulado, el interés radica en encontrar el estado más probable k de una cadena de markov que modela el comportamiento del algoritmo. Si usamos $\mathcal{P}_k$ para denotar la probabilidad del estado k, la distribución del equilibrio $\mathcal{P}$ es reemplazada por una distribución fijando una probabilidad

dependiente sobre un parámetro positivo pequeño fundamental que tradicionalmente es llamado temperatura, $\mathcal{T}$, y la cadena es ejecutada con $\mathcal{T}$ decreciendo gradualmente a cero. Si $\mathcal{T}$ es grande, se aceptan casi todos los pasos propuestos, y la cadena muestrea ampliamente el espacio de estados. Como T declina gradualmente, se toman pocos pasos desfavorables, y la cadena finalmente se estabiliza en k o en un estado óptimo cercano.

Aarts y Korst (1989) probaron asintóticamente la existencia de una distribución estacionaria asociada con la cadena de Markov modelando una cadena de markov aperiódica, ergódica por medio de una trayectoria aleatoria, mostrando que el método de SA encuentra un óptimo global. Hàjek (1988) estableció las condiciones sobre $\mathcal{T}$ para la convergencia del algoritmo SA a un óptimo global en un espacios finitos. Winkler (1995) extendió la demostración de Hàjek y Andrieu y Doucet (1998) reportaron una demostración de la convergencia del algoritmo SA en el establecimiento de modelos de Markov ocultos.

Para su aplicación, Metropolis et al.(1953) introdujo un procedimiento iterativo conocido como la *regla de aceptación de Metropolis*, proponiendo una modificación del estado actual del sistema en los siguientes términos:

- Si la energía del sistema, E, decrece, acepta la modificación.
- Si la energía incrementa en $\triangle E$, la modificación puede ser aceptada con una probabilidad de $\exp(-\triangle E/\mathcal{T})$, donde $\mathcal{T}$ es la temperatura.

En práctica, la determinación del parámetro de control $\mathcal{T}$ ha sido estudiado por autores tales como Geman y Geman (1984), Mitra, Romeo y Sangiovanni-Vincentelli (1986), Van Laarhoven y Aarts (1987) y Aarts y Korst (1989), entre otros. Aunque SA obtiene el óptimo asintóticamente, en cualquier implementación, el algoritmo es un tipo de aproximación que garantiza que éste se detendrá en por lo menos un óptimo local, debido a los valores pequeños de $\mathcal{T}$ (Winkler, 1995). El procedimiento de busqueda en la vecindad es también de gran importancia para un adecuado esquema de enfriamiento para problemas de optimización continua como mostraron Murillo, Vera y Heiser (2005) o Vera y Díaz-García (2008).

El procedimiento de optimización SA propuesto parte de una partición inicial en forma de bloque $\mathcal{P}^{(0)}(\boldsymbol{S})$ derivada de unas particiones iniciales elegidas aleatoriamente $\mathcal{P}^{(0)}(V)$ y $\mathcal{P}^{(0)}(O)$. Entonces se estiman los valores de los parámetros, se evalua el STRESS inicial (energía inicial), y se inicializa

el factor de enfriamiento η, junto con la longitud de la cadena de Markov LC y el factor factor IC, el cual incrementa la longitud de la cadena cada m iteraciones. Para un valor de probabilidades dado χ, la temperatura inicial $\mathcal{T}_0$ se estima en un procedimiento de muestreo aleatorio como el descrito en las implementaciones previas de SA (ver Vera, Heiser y Murillo, 2007 para más detalles en el contexto de MDS), promediando un número de M_a posibles incrementos de las soluciones que empeoran el STRESS. De esta forma, se obtiene un valor de la temperatura inicial tal que en las primeras iteraciones se aceptan el $100\chi\,\%$ de las peores soluciones. La temperatura final $\mathcal{T}_f$ elegida es muy cercana a cero para asegurar que, eventualmente, el algoritmo se detiene en por lo menos un mínimo local. El esquema de enfriamiento constituye un ciclo iterativo principal de $It_{max} = \log(\mathcal{T}_f/\mathcal{T}_0 - \eta)$ iteraciones, en el cual la temperatura actual $\mathcal{T}$ decrece cada vez. En cada iteración principal, se selecciona una nueva partición óptima en forma de bloque para S, minimizando el STRESS, en un ciclo iterativo secundario de longitud de incremento LC. Así, en la p-ésima iteración secundaria, el algoritmo SA se puede describir como sigue:

1. Dada una partición en forma de bloques $\mathcal{P}^{(p)}(\boldsymbol{S})$, los parámetros son estimados condicionalmente y se calcula el STRESS asociado, $\widehat{\sigma^2}^{(p)}$ usando (6.2).

2. Basados en $\mathcal{P}^{(p)}(\boldsymbol{S})$, se obtiene aleatoriamente una nueva partición mediante un procedimiento de dos pasos. Primero, se elige aleatoriamente un conjunto de individuos o de objetos. Segundo, un elemento elegido aleatoriamente del conjunto previamente seleccionado se mueve a un nuevo cluster elegido aleatoriamente, produciendo una partición de prueba $\mathcal{P}^{(p+1)}(\boldsymbol{S})$, bajo las condiciones de que la cardinalidad de cualquier bloque S_{tc} es por lo menos de dos, para evitar soluciones degeneradas. Entonces, se calculan las nuevas matrices indicadoras $\widehat{\boldsymbol{E}}_T^{(p+1)}$ y $\widehat{\boldsymbol{E}}_C^{(p+1)}$.

3. De $\widehat{\boldsymbol{E}}_T^{(p+1)}$ y $\widehat{\boldsymbol{E}}_C^{(p+1)}$, los parámetros $\widehat{\boldsymbol{\alpha}}^{(p+1)}$, $\widehat{\boldsymbol{A}}_T^{(p+1)}$, y $\widehat{\boldsymbol{B}}_C^{(p+1)}$ (usando la versión modificada de SMACOF) son estimados condicionalmente. Se evalua el STRESS $\widehat{\sigma^2}^{(p+1)}$ dado por (6.2) y entonces se calcula el incremento en el STRESS total , $\triangle\sigma^2 = \widehat{\sigma^2}^{(p)} - \widehat{\sigma^2}^{(p+1)}$.

4. Usando la regla de aceptación de Metropolis, si el STRESS decrece se selecciona la partición de prueba $\mathcal{P}^{(p+1)}(\boldsymbol{S})$; de lo contrario, se selecciona esta partición de prueba con una probabilidad $\exp(\triangle\sigma^2/\mathcal{T})$.

El proceso anterior se repite en un ciclo interno de longitud LC, y se obtiene la nueva partición definitiva $\mathcal{P}(\boldsymbol{S})^{(p+1)}$. Entonces la temperatura es disminuida a $\mathcal{T} = \eta\mathcal{T}$, y el proceso general continua hasta que se alcanza un criterio de convergencia, es decir, se alcanza el número máximo de iteraciones, o el valor del STRESS se repite un número R_{max} de iteraciones principales previamente establecido, conjuntamente con un valor bajo de la temperatura.

6.6.2. Resultados experimentales para el desempeño de SA versus el procedimiento DM

Para ilustrar el desempeño del procedimiento SA comparado con el algoritmo DM para el problema del mínimo local, se han considerado los datos transformados del modelo de la escala BTL de mínimos cuadrados y que consisten en el grado de preferencia sobre 12 candidatos presidenciales observada por 21 grupos de individuos obtenidos de Wang, Schönemann, y Rusk (1975). La matrix S de los datos analizados está dada por la transformación $\log v_{ij}$, $i = 1, \ldots, 12$, $j = 1, \ldots, 21$, para relacionar los valores de escala BTL v_{ij} con las distancias unfolding (Luce, 1961; Krantz, 1967). Los datos v_{ij} se recopilaron de Borg y Groenen, 2005, Sección 16.4 (http://people.few.eur.nl/groenen/mmds). Para el algoritmo SA, se eligió el mejor óptimo local en 20 réplicas independientes como la mejor solución. Para el conjunto de datos de tamaño moderado probado, los valores adecuados de los parámetros son χ=0.70, $Ma = 50(T + C)$ para la fase de la temperatura inicial, con valores de γ =0.95, $T_f = 10^{-7}$, $R_{\text{máx}} = 10$, $LC = (T + C)$, $IC = 2(T + C)$ y $m = 20$, para los parámetros restantes .

Primero, tomando en cuenta las restricciones computacionales, se probó el número de clases para los grupos de entrevistados T y políticos C para $T, C = 2, \ldots, 6$. Ambos algoritmos, DM y SA, obtuvieron el valor menor del STRESS para $\kappa = 11$ clases correspondiente a la combinación $T = 5$ y $C = 6$ cuando se emplea el criterio *jump*, como se aprecia en la figura 6.5, usando los valores de $M^* = 13$ para la configuración auxiliar y $q = 3$ para la transformación de la distorsión. Para el algoritmo SA, el método CH de nuevo parece subestimar el número de clusters cuando se compara con el método

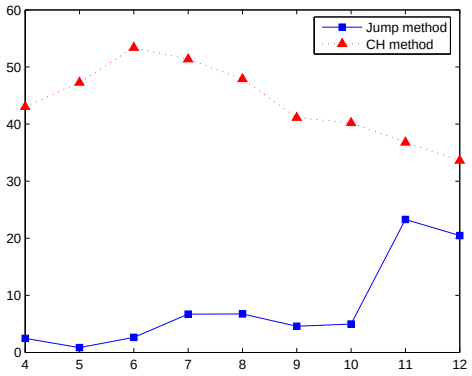

Figura 6.5: Gráfico *Jump* usando $M^* = 13$ y $q = 3$ (cuadrados), y gráfico CH (triángulos) para el conjunto de datos BTL.

jump, obteniendo el valor óptimo de $T = 2$ y $C = 4$ para la combinación de cluster, como tambien se puede ver en la tabla 6.8.

Para $T = 5$ y $C = 6$ clases, se aplicaron ambos algoritmos usando las restricciones geométricas, los resultados obtenidos se presentan en la tabla 6.9. El valor más bajo del estadístico BIC* (289.31) se encontró asociado con un valor del STRESS de 46.35, para el algoritmo SA con el modelo incondicional en dos dimensiones. En términos del tiempo CPU, el algoritmo SA es también más eficiente para este conjunto de datos de tamaño mediano. Sin embargo, para conjunto de datos de gran tamaño examinados, se debe emplear un esquema de enfriamiento diferente para el procedimiento heurístico SA, de tal manera que en esta situación el algoritmo DM parece ser mas eficiente.

La tabla 6.10 muestra los grupos de entrevistados y de políticos que conforman la estructura de partición en V y O clases, respectivamente. El cluster V_1 está compuesto de grupos de entrevistados correspondientes a: [blanco, demócrata débil, sur, educación alta], [blanco, demócrata débil, sur, educación baja], [blanco, independiente, sur, educación alta], [blanco, independiente, sur, educación baja], [blanco, independiente, no del sur, ed. alta.], [blanco, independiente, no del sur, ed. baja]. El cluster V_2 está compuesto de [negro, sur], [negro, no-sur], [blanco, demócrata fuerte, no del sur, ed. alta], [blanco, demócrata débil, no del sur, ed. alta]. El cluster V_3 está compuesto de [blanco,

Tabla 6.8: Resultados obtenidos para los criterios jump y CH para determinar el número de clusters de entrevistados y políticos para el conjunto de datos BTL. Para cada κ, $2 \leq \kappa \leq 12$, se presenta el par (T, C) asociado al STRESS más bajo, y para el criterio jump, se emplearon los valores de $q = 3$ y $M^* = 13$.

Resultados del análisis sin restricciones geométricas					
κ	(T, C)	d_κ	d_κ^{-3}	JUMP	CH
4	2,2	0.557	5.786	2.461	43.042
5	2,3	0.532	6.641	0.854	47.283
6	2,4	0.476	9.272	2.630	53.394
7	3,4	0.397	15.981	6.709	51.357
8	3,5	0.353	22.734	6.752	47.895
9	4,5	0.332	27.326	4.592	41.122
10	4,6	0.314	32.300	4.974	40.216
11	5,6	0.262	55.602	23.302	36.229
12	6,6	0.237	75.119	19.517	33.762

Tabla 6.9: Resultados obtenidos para el modelo de $T = 5$ y $C = 6$ clases, cuando los centros de los clusters son escalados en una, dos y en tres dimensiones, para el conjunto de datos BTL.

Resultados para el modelo incondicional de cluster-unfolding								
	Distancia Mínima				Annealing Simulado			
Dim	STRESS	BIC*	Tiempo	Atr(%)	STRESS	BIC*	Tiempo	Atr(%)
1	69.04	381.98	100	0.03	69.04	381.98	88	5
2	46.38	290.63	622	0.03	46.35	289.31	385	10
3	39.94	294.57	620	0.06	39.94	294.57	472	10
Resultados para el modelo condicional de cluster-unfolding								
	Distancia Mínima				Annealing Simulado			
Dim	STRESS	BIC*	Tiempo	Atr(%)	STRESS	BIC*	Tiempo	Atr(%)
1	69.93	366.04	96	0.76	69.93	366.04	75	10
2	43.59	329.34	771	0.03	43.59	329.34	653	5
3	38.64	313.63	674	0.03	38.34	313.32	522	5

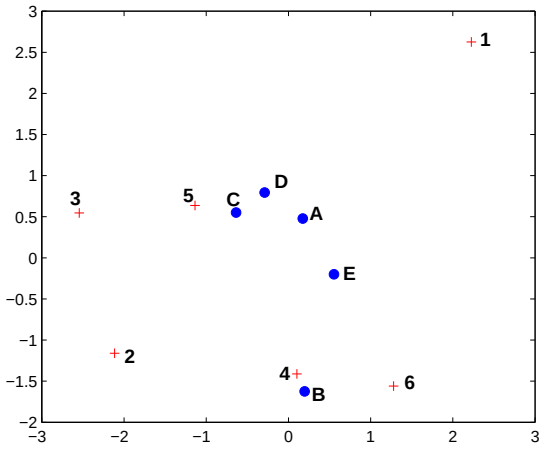

Figura 6.6: Representación óptima de los centros de los clusters en dos dimensiones para los datos de escala BTL. Los clusters de los grupos de entrevistados son representados por letras mientras que los clusters de políticos son representados por números.

republicano fuerte, sur, ed. baja], [blanco, republicano fuerte, no del sur, ed. alta], [blanco, republicano fuerte, no del sur, ed. baja], [blanco, republicano débil, no sur, ed. alta]. El cluster V_4 está compuesto de [blanco, republicano débil, sur, ed. alta], [blanco, republicano débil, sur, ed. baja], [blanco, republicano débil, no del sur, ed. baja], y el cluster V_5 está compuesto de [blanco, demócrata fuerte, sur, ed. alta], [blanco, demócrata fuerte, sur, ed. baja], [blanco, demócrata fuerte, no del sur, ed. baja], [blanco, demócrata débil, no del sur, ed. baja].

Sobre el espacio de los políticos, el O_1={Wallace, LeMay}, O_2={McCarthy, Rockefeller, Romney}, O_3={Reagan, Agnew}, O_4={Kennedy}, O_5={Nixon}, y O_6={Johnson, Muskie, Humphrey}. De esta forma, el grupo más liberal se siente muy identificado al grupo de Kennedy, Johnson, Muskie y Humphrey, mientras que el grupo más conservador se siente más cercano a Nixon, como se puede apreciar en la figura 6.6.

Para el algoritmo SA, los datos de escala BTL fueron también analizados con un procedimiento de dos pasos que primero obtiene el mejor modelo de cluster a dos-modos y entonces representa los centros de los clusters obte-

Tabla 6.10: Clasificación óptima en dos dimensiones para el modelo incondicional de unfolding de $T = 5$ y $C = 6$ clases para los datos de la escala BTL transformada.

	Grupo de entrevistados	V_t		Políticos	O_c
1	BS	2	1	Wal	1
2	BN	2	2	Hum	6
3	SDSH	5	3	Nix	5
4	SDSL	5	4	McC	2
5	WDSH	1	5	Rea	3
6	WDSL	1	6	Roc	2
7	SDNH	2	7	Joh	6
8	SDNL	5	8	Rom	2
9	WDNH	2	9	Ken	4
10	WDNL	5	10	Mus	6
11	ISH	1	11	Agn	3
12	ISL	1	12	LeM	1
13	INH	1			
14	INL	1			
15	SRSL	3			
16	SRNH	3			
17	SRNL	3			
18	WRSH	4			
19	WRSL	4			
20	WRNH	3			
21	WRNL	4			

nidos mediante unfolding. Usando el método jump, se selecciona el modelo $T = 5$ y $C = 6$ asociado al valor más grande del jump (23.302). Cuando los centros de los clusters son representados, se obtiene un valor del STRESS de 0.0086. En términos del agrupamiento a dos-modos, se encontraron diferentes clasificaciones para los conjuntos de grupos de entrevistados y de políticos, cuando el procedimiento de dos pasos se compara con la clasificación obtenida del algoritmo de cluster-unfolding combinado en dos dimensiones. En términos de la calidad de la solución, el procedimiento combinado obtuvo el valor del STRESS de 0.0029 en el modelo de dos-dimensiones. El procedimiento de cluster-unfolding SA simultáneo, parece ser también más eficiente que un procedimiento de dos pasos, como mostraron primero Vera, Macías y Heiser (2008).

6.7. Conclusiones y extensiones

En este trabajo se propone un modelo de mínimo cuadrados para una matriz de preferencias, cuyo objetivo es particionar los individuos en $T(T << R)$ y los objetos en $C(C << N)$ clases, mientras que simultáneamente son representados los the $T+C$ centros de los clusters en un espacio de baja dimensión. Aunque el enfoque de clases latentes anterior tiene un buen desempeño, la hipótesis de independencia y de normalidad sobre los datos de preferencias pueden resultar restrictivos en muchas situaciones prácticas, por tanto un enfoque determinístico debería ser más adecuado. Desde un punto de vista computacional, los anteriores métodos basados en una optimización Monte Carlo demandan un tiempo CPU alto, lo cual hace que el modelo sea recomendable únicamente para conjuntos de datos de tamaño mediano. Un procedimiento de estimación condicional alternante de mínimos cuadrados basada en una extensión para datos bimodales del procedimiento de distancia mínima propuesto en Heiser y Groenen (1997) se emplea para la estimación de los parámetros. El procedimiento garantiza que no únicamente al final sino cada vez a lo largo del proceso de estimación en el modelo, la partición en forma de bloques de la matriz de preferencias se puede asociar a una clasificación en los individuos y objetos. De esta forma, la utilización del método de distancia mínima permite al algoritmo ser eficiente también en tiempo CPU por lo que se puede emplear en conjuntos de datos grandes.

La determinación del número verdadero de grupos en un conjunto de datos es un problema fundamental sin resolver en análisis de cluster. En el

contexto de cluster-unfolding, se han sugerido varios enfoques a este problema. En un marco de mínimos cuadrados, el análisis de la dispersión del modelo ha sido utilizada para este fin, y en este contexto, el criterio de información de Sugar y Gareth (2003) se ha extendido para tratar con datos de preferencias bimodales. Usando la configuración unfolding de los conjuntos de datos originales en una dimensionalidad tal que el STRESS llegue a ser más pequeño que 10^{-6}, se emplea el *salto* maximo exhibido en la gráfica de la distorsión transformada para determinar el verdadero número de clusters. Como fue demostrado por Sugar y Gareth (2003), los resultados dados confirman que el criterio de Calinski y Harabasz (1974) parece subestimar el número de clusters cuando están implicadas un número grande de dimensiones en la configuración, como sucede usualmente cuando el modelo propuesto se aplica a conjuntos de datos medianos o grandes.

Para determinar la dimensionalidad de la representación unfolding, se emplea el criterio de información ajustado BIC en un enfoque de clases latentes al modelo de cluster-unfolding, siguiendo el principio de parsimonia como propuso Lee (2001). De esta forma, la formulación del criterio BIC* en un marco de modelos de clases latentes como propusieron Vera, Macías y Heiser (2008) se emplea para determinar la dimensión de la representación de los centros de los clusters, condicionada al número de clusters dado.

Para ilustrar el funcionamiento del modelo, se utilizaron los conjuntos de datos artificiales y de internet previamente analizados por Vera, Macías y Heiser (2008). Para los datos de internet, el método del *salto* proporciona un cluster más de respondientes y dos más de declaraciones que cuando se aplica el criterio BIC* para el modelo de clases latentes empleando varianzas desiguales. El algoritmo de optimización basado en la distancia mínima es muy eficiente en términos del tiempo CPU comparado con el procedimiento usual de optimización de Monte Carlo, pero la solución depende de la solución inicial. Este problema se aprecia también para el conjunto de datos de internet en el cual el valor del STRESS más bajo fue obtenido una sola vez en las 300 réplicas, obteniendose diferentes soluciones en todas las réplicas. Un procedimiento de estimación condicional de mínimos cuadrados basado en annealing simulado (SA) se propone para tratar con el problema de mínimos locales. El análisis del conjunto de datos BTL escalados muestra la eficiencia del procedimiento SA comparado con el algorimo DM, así como con un procedimiento de cluster y unfolding a dos pasos. Desafortunadamente, el procedimiento SA es recomendable únicamente para conjuntos de datos de tamaño pequeño y mediano porque el tiempo CPU crece cuando $R + N$

se incrementa. Por ejemplo, para los datos de internet se encontró un valor del STRESS de 7531.44 en 8715 segundos, mientras que con el procedimiento DM se se obtuvo un valor menor del STRESS de 7505.73 en únicamente 4515 segundos. Por lo tanto, un procedimiento condicional de estimación basado en un procedimiento de estimación Monte Carlo que también podría ser eficiente en términos del tiempo CPU podría ser recomendable para el problema global de estimación, un asunto que está siendo desarrollado actualmente por los autores.

Capítulo 7

General conclusions and possible extensions

The combined application of Cluster Analysis and Multidimensional Scaling is a very advisable methodology when one has large data sets and it is desirable to summarize their structure by means of a standardization of their behavior through homogenous groups, and by means of the representation of these groups formed in a space of low dimensionality.

Several methodologies have been considered to realize this. A common practice is to consider the application of both techniques separately, reducing first the number of dimensions by means of MDS and then, superposing on the MDS representation, obtain a grouping of the objects by some method of classification. However the objectives of both techniques are of different nature. The MDS solution is optimal in the sense of representing the objects in a reduced space, whereas CA constructs optimal groups on the original space from the dissimilarity data come. Therefore, the application of both techniques in this way is not very recommendable. A unique methodology that represents all the objects and their associations in a space of low dimension, using the original data, is needed.

In many situations, the application of the cluster-MDS models requires the consideration of certain restrictions in the phase of classification. In the context of spatiotemporal processes underlying environmental studies, the use of the cluster-MDS models in the estimation of the structure of nonstationary space covariance is very suitable. However, the space factor implies that the grouping of the objects must be realized using the geographical spatial location of the objects and their characteristics as fundamental criteria.

In this work, several models that apply jointly cluster analysis and multidimensional scaling were developed. The methodology can be used to analize a $(n \times n)$ matrix of two-way one-mode data, representing continuous dissimilarities between n objects, or a $(n \times p)$ matrix of two-way two-mode data that represent the degree of preference of n individual over p stimuli or objects. These models provide a classification of the objects (for two-way one-mode data) or of individuals and objects (two-way two-mode data) and simultaneously provide a representation of the centers of these classes in a space of low dimension. Thus, this implies a MDS and Unfolding solution, respectively, but on the basis of a number of points significantly smaller to the original ones. The estimation of the partition in the constructed models was obtained using least squares and maximum likelihood, whereas the MDS configuration was realized by means of SMACOF.

For two-way one-mode continuous rating dissimilarity data, a model of cluster-MDS was developed proposing a modification to the *Least squares metric cluster diffences scaling* procedure (Heiser and Groenen, 1997). This model can be used to estimate spatial dispersion in spatiotemporal processes. The model, called the CDC-R model, is a minimal distance method (equivalent to a k means algorithm) and is based on a new concept of geographical contiguity on the dissimilarities. The model simultaneously estimates a classification of the objects and the configuration of the cluster centers in a space of low dimension, such that the spatial relationships between the objects and the clusters are retained. Although generally, the inclusion of geographical restrictions on classification increases the error in terms of total STRESS function, the results obtained with the CDS-R model were very satisfactory. Furthermore the problem of empty classes did not appear, at least in the analyzed data sets.

The CDS model was formulated from a probabilistic approach. The search for the optimal object classification is raised as a block partition problem in the dissimilarity matrix, assuming that dissimilarities in a block are independent and normally distributed, and establishing a univariate normal mixture as the distribution for a previous unclassified dissimilarity. Then the optimal classification, the associated dispersion of each cluster and the cluster centers configuration are estimate by means of maximum likelihood applied an algorithm based on simulated annealing and for the representation of clusters using SMACOF. This was done in an iterative process. This methodology was called the LACSSCAL model. For the latent classes MDS model, spatial restrictions were considered by introducing the contiguity concept analogous

to the model CDS-R. By considering both normal and lognormal mixture distributions, the new model enables us to partition the sample objects into classes and simultaneously represent the cluster centers in a low dimensional space, while the objects and clusters maintain their spatial relationship.

An important problem with the mixture models is that the likelihood equation usually has multiple local maxima. In addition, partitioning the proximity matrix into blocks suffers from an implicit indetermination problem. In this sense, a partition achieved by means of the EM algorithm does not guarantee a direct relation with the classification in the objects space. Because the final solution of the EM algorithm depends strongly on the initial solution, we are forced to consider an alternative estimation strategy. Application of simulated annealing in this context is a very recommendable strategy, because it constructs a partition of the dissimilarity matrix into blocks that is directly related to the partition in the objects space, which diminishing the problem of local optimal.

A fundamental aspect of the cluster-MDS/Unfolding models is the selection of the suitable model, that is, the appropriate number of clusters and the suitable dimensionality for its representation. It is important to note that the finding the number of clusters is a problem that has not been solved satisfactorily until now. In the least squares models the procedure for the model selection is completely descriptive, whereas in the probabilistic models more objective strategy for the suitable selection based on the modified BIC statistic or by means of the bootstrap procedure is proposed. In addition, in probabilistic models, one of the established hypotheses is the nonrestrictive condition of the variances in the mixture components. This situation contributes to a more objective explanation of the cluster structure in the data, and is in accordance with the parsimony principle, because using unequal variances for each block tends to reduce the estimated number of latent classes.

The principal disadvantage of the proposed latent class MDS/Unfolding procedure based on the SA algorithm for the estimation of the parameters, is the high CPU time compared with other optimization algorithms. This limits its application to data sets of small or medium size. Nevertheless, in some practical situations where the independence and (log) normality assumptions of the proximities are not appropriate, an exploratory approach using for example the minimal distance algorithm that demands a time considerably smaller in the estimation of the parameters, could be an suitable alternative, that can be applied to large size data sets. The problem of the

high computer cost of the SA algorithm can be solved partially using reduction formulas that simplify the calculations of the estimators in each iteration of the algorithm. This possibility is being explored by the author at the moment. The Simulated Annealing based conditional estimation procedure can be extended to include any other mixture distributions. Another interesting possibility for future studies is to develop a new strategy of model selection to determine the most adequate distribution of the mixture components. The developed methodologies can also be extended to three-way two-mode data sets, to simultaneously determine groups of objects and/or of similar subjects, while the cluster centers are represented in a low dimensional space. This topic is also currently investigated by the author.

The problem of the degeneration appears frequently in unfolding models when transformations of the data are allowed that include at least an intercept and a slope, that is, ordinal and interval transformations. Even though the modified SMACOF procedure produces good results in all the data examples, other algorithms, for example, PREFSCAL (Busing et al., 2005) could be used when geometric constraints are imposed, and to extend the method for ordinal transformations. The two models for preference data were developed for the simple unfolding model, but it may be developed for the vector model as well.

The new methodologies developed in this thesis for the cluster MDS/unfolding models provided encouraging results in all the data sets analyzed. However, from the discussion it is clear that still much work can be done for example, considering new strategies for the estimation of the parameters and for model selection, as well as extending the proposed strategies to more general models.

Bibliografía

Aarts, E., Korst, J. 1989. Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. Wiley, Chichester [UK]; New York.

Aitkin, M., & Anderson, D., & Hinde, J. 1981. Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society Series A* , 144, 419-461.

Akaike, H. 1977. On entropy maximization. In P.R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27-41). Amsterdam: North-Holland.

Andrieu, C. and Doucet, A. 1998. *Simulated annealing for bayesian estimation of hidden Markov models.* Tr 317, Department of Engineering, University of Cambridge.

Andrews, R.L. & Manrai, A.K. 1999. MDS maps for product attributes and market response: An application to scanner panel data. *Marketing Science, 18*, 584-604.

Angulo JM, Bueso MC, Alonso FJ. 2000. A study on sampling design for optimal prediction of space-time stochastic processes. *Stochastic Environmental Research and Risk Assessment* **14**; 412-427.

Angulo JM, Ruiz-Medina MD, Alonso FJ, Bueso MC. 2005. Generalized approaches to spatial sampling design. *Environmetrics* **16**; 523-534.

Arbia G, Lafratta G. 1997. Evaluating and updating the sample design in repeated environmental surveys: monitoring air quality in Padua. *Journal of Agricultural, Biological, and Environmental Statistics* **2**: 251-466.

Arbia G, Lafratta G. 2002. Anisotropic spatial sampling designs for urban pollution. *Applied Statistics* **51**: 223-234.

Ball, G.H. and Hall, D.J. 1967. A clustering technique for summarizing multivariate data. *Behavioral Science*, *12*, 153-5

Banfield, J.D. and Raftery, A.E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803-21

Beale, E.M.L. 1969. Euclídean cluster analysis. *Bulletin of the International Statistical Institute*, **43 (2)**, 92-4.

Bock HH. 1986. Multidimensional scaling in the framework of cluster analysis. In *Studien zur Klassifikation: [Classification and its Environmet]*, Hermes HJ, Optiz O, Degens PO (eds.); INDEKS-Verlag: Frankfurt; **17**: 247-258.

Bock HH. 1987. On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In *Multivariate Statistical Modeling and Data Analysis*, Bozdogan H, Gupta AK (eds.); Reidel: New York; 17-34.

Bock HH. 1997. Simultaneous visualization and clustering methods as an alternative to Kohonen maps. In *Learning, Networks and Statistics*, Riccia GD, Lenz HJ, Kruse R (eds.); Springer Wien New York.

Borg I, Groenen PJF. 2005. Modern Multidimensional Scaling. Theory and Applications. *Springer Series in Statistics*; Springer: second edition.

Bockenholt, U., & Bockenholt, I. 1990. Modeling individual differences in unfolding preference data: A restricted latent class approach. *Applied Psychological Measurement, 14*. 257-269.

Bockenholt, U., & Bockenholt, I. 1991. Constrained latent class analysis: Simultaneous classification and scaling of discrete choice data. *Psychometrika, 56*, 699-716.

Brusco, M.J. 2001. A simulated annealing heuristic for unidimensional and multidimensional (city-block) scaling of symmetric proximity matrices. *Journal of Classification, 18*, 3-33.

Busing, F.M.T.A., Groenen, P.J.F., Heiser, W., 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika* 70(1), 71-98.

Calinski, R.B., Harabasz, J. 1974. A denrite method for cluster analysis. *Communications in Statistics* 3, 1-27.

Carroll JD, Chang JJ. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, **35**; 283-320.

Chintagunta, P.K. 1994. Heterogeneous logit model implications for brand positioning. *Journal of Marketing Research, 31*, 304-311.

Černy, V. 1985. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications, 45*, 41-51.

Coombs, C.H., 1964. A theory of data. New York. Wiley.

Cox TF, Cox MAA. 2001. Multidimensional Scaling. *Monographs on Statistics and Applied Probability 88*; Chapman Hall/CRC: second edition.

Cressie N, Huang H. 1999. Classes of nonseparable spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**: 1330-1340.

Dasgupta, A., Raftery, A., 1998. Detecting features in spatial point processes with cluster via model-based clustering. *Journal of the American Statistical Association* 93:294-302

Davis BM, Borgman LE. 1982. A note on the asymptotic distribution of the sample variogram. *Mathematical Geology* 14(2):189-193

Day, N.E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463-474.

De Leeuw J. 1977. Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133-145). Amsterdam, The Netherlands; North-Holland.

De Leeuw J. 1988. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, **5**; 163-180.

De Leeuw J, Heiser WJ. 1977. Convergence of correction-matrix algorithms for multidimensional scaling. In J.C. Lingoes, E.E Roskam, I. Borg (Eds.),*Geometric representations of relational data* (pp. 735-752). Ann Arbor, MI: Mathesis Press.

De Leeuw J, Heiser WJ. 1980. Multidimensional scaling with restrictions on the configuration. In *Multivariate analysis, Vol. V*, P.R. Krishnaiah (Ed.); Amsterdam: North-Holland; 501-522.

Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

DeSarbo, W., & Howard, D.J., & Jedidi, K. 1991. MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika, 56*, 121-136.

DeSarbo, W. S., Jedidi, K.J., Cool, K., Schendel, O. 1991. Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters 2*, 129-146.

DeSarbo, W., & Ajay, K.M., & Lalita, A.M. 1994. Latent class multidimensional scaling: a review of recent developments in the marketing and psychometric literature. In Richard P.B (Ed.) *Advanced Methods of Marketing Research* (pp. 190-222). Blackwell Publishers: Cambridge, MA.

DeSarbo, W. S., Ramaswamy, V., Chatterjee, R. 1995. Analyzing constant-sum multiple criterion data: A segment-level approach. *Journal of Marketing Research* 32, 222-232.

DeSarbo, W.S., Young, M.R., Rangaswamy, A. 1997. A parametric Multidimensional Unfolding Procedure for Incomplete Nonmetric Preference/Choice Set Data in Marketing Research (Tech. Rep.). The Pennsylvania State University.

De Soete, G., Hubert, L.J., and Arabie, P. 1988. The comparative performance of simulated annealing on two problems of combinatorial data

analysis. In Diday, E., editor, *Data Analysis and Informatics (Vol. 5)*, pp. 489-496. Amsterdam: North Holland.

De Soete, G. 1990. A latent class approach to modeling pairwise preferential choice data. In M. Schader & W. Gaul (Eds.), *Knowledge, data and computer-assisted decisions* (pp. 103-113). Berling: Springer-Verlag.

De Soete, G. 1992. Using latent class analysis in categorization research. In I. Van Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 309-330). London: Academic Press.

De Soete, G., & DeSarbo, W.S. 1991. A latent class probit model for analyzing pick any/N data. *Journal of Classification, 8*, 45-63.

De Soete, G., & Heiser, W.J. 1993. A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika, 58*, 545-565.

De Soete, G., & Winsberg, S. 1993a. A Thurstonian pairwise choice model with univariate and multivariate spline transformations. *Psychometrika, 58*, 233-256.

De Soete, G., & Winsberg, S. 1993b. A latent class vector model for analyzing preference ratings. *Journal of Classification, 10*, 195-218.

De Soete, G., Carroll, J.D., 1994. K-means clustering in a low dimensional Euclídean space, In E. Diday et al. (Eds)., New Approaches in Classification and Data Analyisis. Heidelberg. Springer Verlag, 212-219.

Diday, E. & Govaert, G. 1977. Classification automatique avec distances adaptatives. *RAIRO Informatique / Computer Sciences, 11*, 329-349.

Dhrymes PJ. 1978. Mathematics for Econometrics. New York: Springer-Verlag.

Everitt BS. 1981. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, **16**, 171-180.

Everitt BS, Landau S, Leese M. 2001. *Cluster Analysis.* Arnold: fourth edition.

Ferligoj A, Batagelj V. 1982. Some types of clustering with relational constraints. *Psychometrika* **47**; 541-552.

Formann, A. K. 1989. Constrained latent class models: Some further applications. *British Journal of Mathematical and Statistical Psychology, 42*, 37-54.

Friedman, H.P and Rubin, J. 1967, On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159-1178

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

Gordon, A.D. 1999. Classification. Second Edition. *Monographs on statistics and applied probability; 82.* Chapman and Hall/CRC.

Graef J, Spence I. 1979. Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, **86**; 60-66.

Hàjek, B., 1988. Cooling schedules for optimal annealing. *Mathematics of Operations Research* 13, 311-329.

Hansen, P., Jaumard, B. and Sanlaville, E. 1994. Partitioning problems in cluster analysis: A review of mathematical programming approaches, in *New Approaches in Classification and data Analysis* (E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and P. Burtschy, eds), pp. 228-240. Springer-Verlag, Berlin.

Hansen, P., Jaumard, B. 1997. Cluster analysis and mathematical programming. *Mathematical Programming*, **79**, 191-215.

Hartigan, J.A. and Wong, M.A. 1979, Algorithm AS 136. A k-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

Hasselblad, V. 1966. Estimation of parameters for a mixture of normal distributions, *Technometrics*, **8**, 431-444.

Hasselblad, V. 1969. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64, 1459-1471.

Heiser, W.J., 1981. Unfolding analysis of proximity data. Unpublished doctoral dissertation. University of Leiden. The Netherlands.

Heiser, W.J., 1987. The unfolding technique. In P. Legendre & L. Legendre (Eds.) Developments in numerical ecology. 189-221. Berlin: Springer-Verlag.

Heiser, W.J., 1991. A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika*, 56, 7-27.

Heiser WJ. 1993. Clustering in low-dimensional space. In *Information and Classification: Concepts, Methods and Applications.*, Lausen B, Klar R, Opitz O (eds.); Springer Verlag: Heidelberg; 162-173.

Heiser WJ, Groenen PJF. 1997. Cluster differences scaling with a within-clusters loss component and a fuzzy succesive approximation strategy to avoid local minima. *Psychometrika* **62(1)**; 63-83.

Hernández-Avilia, A. 1979. Problems in cluster analysis. Unpublished D.Phil. thesis. University of Oxford

Hope, A.C. 1968. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B, 30*, 582-598.

Ingrassia, S 1991. Mixture decomposition via the simulated annealing algorithm. *Applied Stochastic Models and Data Analysis, 7*, 317-325.

Ingrassia, S 1992. A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing, 2*, 203-211.

Ismail, M.A. and Kamel, M.S. 1989. Multidimensional data clustering utilizing hybrid search strategies. *Pattern Recognition*, **22**, 75-89.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data. An Introduction to Cluster Analysis.* Wiley-Interscience, New York.

Kiers, H.A.L., Vicari, D., Vichi, M. 2005. Simultaneous classification and multidimensional scaling with external information. *Psychometrika 70*, 433-460.

Kirkpatrick, S., Gelatt, D., & Vecchi, M.P. 1983. Optimization by simulated annealing. *Science, 220*, 671-680.

Klein, R.W., & Dubes, R.C. 1989. Experiments in projection and clustering by Simulated Annealin. *Pattern Recognition, 22*, 231-220.

Kovitz JL, Christakos G. 2004. Spatial statistics of clustered data. *Stochastic Environmental Research and Risk Assessment* **18**; 147-166.

Krantz, D.H., 1967. Rational distance functions for multidimensional scaling. *Journal of Mathematical Psychology 4*, 226-245

Kruskal JB. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**;1-27.

Kruskal JB. 1964b. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**; 28-42.

Kruskal JB. 1977. The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 17-44). New York: Academic Press.

Lee, D.M. 2001. Determining the dimensionality of Multidimensional Scaling representations for cognitive modeling. *Journal of Mathematical Psychology 45*, 149-166.

Levi, S. 1983. A cross-cultural analysis of the structure and levels of attitudes towards acts of political protest. *Social Indicators Research, 12*, 281-309.

Lland A, Hst G. 2003. Spatial covariance modelling in a complex coastal domain by multidimensional scaling. *Environmetrics* **10**; 307-321.

Luce, R.D., 1959. Individual choice behavior. New York. Wiley.

Luce, R.D., 1961. A choice theory analysis of similarity judgments. *Psychometrika 26*, 151-163.

MacQueen, J. 1967, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. LeCam and J. Neymen, eds), Vol 1, pp.281-297. University of California Press, Berkeley.

Mardia. KV, Kent JJ, Bibby JM. 1980. *Multivariate Analysis.* Academic Press.

Marriott, F.H.C. 1971. Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501-514.

McLachlan, G. J. 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixtures. *Applied Statistics*, **36**, 318-324.

McLachlan, G. J. & Basford, K.E. 1988. *Mixture Models.* New York: Marcel Dekker.

McLachlan, G. J. & Krishnan, T. 1997. The EM Algorithm and Extensions. *Wiley series in probability and statistics.* John Wiley & Sons, Inc. New York.

McLachlan, G. J. & Peel, D. 2001. Finite Mixture Models. *Wiley series in probability and statistics.* John Wiley & Sons, Inc. New York.

Maravalle M, Simeone B, Naldini R. 1997. Clustering on trees. *Computational Statistics and Data Analysis* **24**; 217-234.

Metropolis, N.A., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

Milligan, G.W., Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159-179.

Mitra, D., Romeo, F., Sangiovanni-Vincentelli, A., 1986. Convergence and finite-time behaviour of simulated annealing. *Advances in Applied Probability* 18, 747-771.

Murillo A, Vera JF, Heiser W. 2005. A permutation-translation simulated annealing algorithm for L_1 and L_2 unidimensional scaling. *Journal of Classification* **22**; 119-138.

Murtagh FD. 1995. Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters* **16**; 399-408.

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.

Pearson, K. 1894. Contribution to the mathematical theory of evolution. *Philosophical Transactions* A, **185**, 71-110.

Ramsay, J.O. 1973. The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values. *Psychometrika, 38*, 513-532.

Ramsay, J.O. 1977. Maximum likelihood estimation in multidiemsnional scaling. *Psychometrika, 42*, 241-266.

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

Ramsay, J.O. 1997. *MULTISCALE manual (Extended version)*. Montreal: McGill University.

Rissanen, J., 1978. Modeling by shortest data description. Automatica, 14:465-471.

Rissanen, J., 1986. Stochastic complexity. Annals of Statistics, 14, 1080-1100.

Rissanen, J., 1989. Stochastic complexity. Singapore: World Scientific Publihing.

Rubin, J., 1967. Optimal classification into groups: An approach for solving the taxonomy problem. *Journal of Theoretical Biology*, **15**,103-144.

Ruiz-Medina MD, Alonso FJ, Angulo JM, Bueso MC. 2003. Functional stochastic modeling and prediction of spatio-temporal processes. *Journal of Geophysical Research - Atmospheres* **108**, No. D24, 9003, doi:10.1029/2003JD003416, 2003, Special Section Application of Recent Advances in Space- Time Statistics to Atmospheric Data (SPCTME1).

Sampson PD, Guttorp P. 1992. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistial Association* **87**; 108-119.

Schoenberg I.J. 1935. Remarks to maurice fréchet's article "sur la definition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert". *Ann. Math*, **36**; 724-732.

Schwarz, G. 1978. Estimation the dimensions of a model. *Annals of Statistics, 6*, 461-464.

Scott, A.J. and Symons, M.J. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387-398.

Seidel, W., & Mosler, K., & Alker, M. 2000. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics, 52*, 481-487.

Selim,S.Z., & Asultan, K. 1991. A Simulated Annealing algorithm for the clustering problem. *Pattern Recognition, 24*, 1003-1008.

Sokal RR, Michener CD. 1958. A statiscial method for evaluating systematic relationships. *The University of Kansas Science Bulletin* **38**; 1409-1438.

Späth H. 1985. Cluster Dissection and Analysis. Chichester: Ellis Horwood.

Suckling PW, Hay JE. 1978. On the use of synoptic weather map typing to define solar radiation regimes. *Monthly Weather Review* **106**; 1521-1531.

Sugar, C.A., James, G.M. 2003. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Asssociation*. 98, 750-762.

Sun, L.X., & Xie, Y.L., & Song, X.H., & Wang, J.H., & Yu, R.Q. 1994. Cluster analysis by simulated annealing. *Computers and Chemistry, 18*, 103-108.

Takane Y, Young FW, de Leeuw J. 1977. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika* **42**; 7-67.

Takane Y, Carroll JD. 1981. Nonmetric metric maximum likelihood multidimensional scaling from directional rankings of similarities. *Psychometrika*,**46**; 389-405.

Thode, H.C, Finch, S.J. and Mendel, N.R. 1989. Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics*, **44**,1195-1201.

Tibshirani, R., Walther, G. and Hastie, T. 2001, Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B.* **63**, 411-423.

Van Buuren, S., Heiser, W.J. 1989. Clustering objects into groups under optimal scaling of variables. *Psychometrika 54*, 699-706.

Van Deun, K., Groenen, P.J.F., Heiser, W.J., Busing, F.M.T.A., Delbeke, L. 2005. Interpreting degenerate solutions in unfolding by use of the vector model and the compensatory distance model. *Psychometrika*, 70(1), 45-69

Van Deun, K., Marchal, K., Heiser, W.J., Engelen, K., Van Mechelen, I. 2007. Joint mapping of genes and conditions via multidimensional unfolding analysis, BMC *Bioinformatics*, 8:181

Van Laarhoven, P.J., Aarts, E.H.L., 1987. Simulated Annealing: Theory and Applications. D. Reidel Publishing Company, Dordrecht. The Netherlands.

Van Mechelen, I., Bock, H-H., De Boeck, P., 2004. Two-mode clustring methods: a strutured overview. *Statistical Methods in Medical Research*, 13, 363-394.

Van Rosmalen, J., Groenen, P.J.F., Trejos, J., Castillo, W., 2005. Global Optimization Strategies for Two-Mode clustering. *Econometric Institute Report.* EI-2005-33.

Vera JF, Heiser W, Murillo A. 2007. Global optimization in any Minkowski metric: A permutation-translation simulated annealing algorithm for Multidimensional Scaling. *Journal of Classification* 24:277-301.

Vera JF,& Macías R., Heiser, WJ, 2007. A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika.* In process.

Vera J.F., & Macías R., & Angulo J.M. 2008. Nonstationary Spatial Covariance Structure Estimation in Oversampled Domains by Cluster Differences Scaling with Spatial Constraints. *Journal of Stochastic Environmental Research and Risk Assessment* (SERRA) 22:95-106 .

Vera, J.F., Díaz-García, J.A., 2008. A global simulated annealing heuristic for the three-parameter lognormal maximum likelihood estimation. *Computational Statistics and Data Analysis.* doi:10.1016/j.csda.2008.04.033.

Vera, J.F., Macías, R., Heiser, W.J., 2008. A Dual Latent Class Unfolding Model for Two-Way Two-Mode Preference Rating Data. *Computational Statistics and Data Analysis.* doi:10.1016/j.csda.2008.07.019.

Vera, J.F., Macías, R., Angulo, J.M., 2008a. A latent class MDS model with spatial constraints for non-stationary spatial covariance estimation. *Stochastic Environmental Research and Risk Assessment.* In press.

Vichi, M., Kiers, H.A.L., 2001. Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis* 37, 49-64.

Wang, M.M, Schönemann, P.H., Rusk, J.B., 1975. A conjugate gradient algorithm for the multidimensional analysis of preference data. *Multivariate Behavioral Research.* 10, 45-79.

Ward JH. 1963. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, **58**; 236-244.

Wedel, M., & DeSarbo, W. 1996. An exponential-family multidimensional scaling mixture methodology. *Journal of Business and Economical Statistic, 14*, 447-459.

Winkler, G. 1995. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer-Verlag, New York.

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

Wolfe, J.H. 1970. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329-350.

Wolfe, J.H. 1971. A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. *Naval Personnel and Training Research Laboratory Technical Bulletin*, STB 72-2, San Diego, CA.

Yang, C-C., Yang, C-C., 2007. Separating latent vlasses by information criteria. *Journal of Classification*. 24:183-203

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