RR 94-03

CLUSTER DIFFERENCES SCALING WITH A WITHIN-CLUSTERS LOSS COMPONENT AND A FUZZY SUCCESSIVE APPROXIMATION STRATEGY TO AVOID LOCAL MINIMA

Willem J. Heiser

Patrick J.F. Groenen

Department of Data Theory

University of Leiden

August 1994

The authors are indebted to Robert Tijssen for making available to co-citation data, and to Jacqueline Meulman for her useful and stimulating comments during the completion of this manuscript, which is an extended version of the paper presented at the Annual Meeting of the Psychometric Society at Berkeley, June 1993.

Abstract

Cluster differences scaling is a method for finding a low-dimensional spatial representation of K cluster points, to model a given square table of dissimilarities among n stimuli or objects. The least squares loss function of cluster differences scaling, originally defined only on the residuals of pairs of objects that are allocated to different clusters, is extended with a loss component for pairs that are allocated to the same cluster. It is shown that this extension makes the method equivalent to multidimensional scaling with cluster constraints on the coordinates. A breakdown of the sum of squared dissimilarities into contributions from several sources of variation is described, including the appropriate degrees of freedom for each source. After developing a convergent algorithm for fitting the cluster differences model, it is argued that the individual objects and the cluster locations can be jointly displayed in a configuration obtained as a by-product of the optimization. Finally, the paper introduces a fuzzy version of the loss function, which can be used in a successive approximation strategy for avoiding local minima. A simulation study demonstrates that this strategy significantly outperforms two other well-known initialization strategies, and that it has a success rate of 92 out of 100 in attaining the global minimum.

Keywords: multidimensional scaling, iterative majorization, *K*-means clustering, fuzzy clustering, local minima, constrained optimization, analysis of dispersion, co-citation analysis.

Cluster Differences Scaling with a Within-clusters Loss Component and a Fuzzy Successive Approximation Strategy to Avoid Local Minima

Introduction

Cluster analysis and multidimensional scaling (MDS) are often used in conjunction, since it is generally assumed that their combined use might lead to a better understanding of the data than what can be learned from each method separately (Kruskal, 1977). One aspect that MDS has to offer, which is lacking in most clustering methods, is an embedding in a low-dimensional spatial representation. Some clustering methods also provide a spatial representation of clusters. In *K*-means clustering, for example, groups of objects are represented as *cluster centers* in the space of the observations. But these centers will in general span a $(K - 1)$-dimensional subspace, which is low-dimensional only if either the original number of dimensions, m, or the chosen number of clusters, K, is small. When K and m are larger than two or three, or when the data have the form of some measure of (dis)similarity between pairs of objects, the dimensionality of the problem is prohibitive; the usual practice then is to reduce the number of dimensions by MDS, and to superimpose the results of a cluster analysis on the MDS representation.

Although this practice has much to recommend, it should be noted that two forms of optimality are confounded: the reduced space is optimal for the embedded points, but not for the superimposed clusters, while the cluster structure is optimal in the original, non-reduced space only, where a spatial representation, if any, misses its data analytic appeal. Therefore, it might be attractive to integrate the two objectives by introducing the assumption that not the original points, but the cluster centers are located in a low-dimensional metric space, whatever the value of K.

The metric space assumption forms the key concept of a group of methods that models the dispersion between clusters directly in reduced dimensionality, one of which is the *cluster differences scaling* (CDS) method proposed by Heiser (1993). The basic elements of CDS are as follows. Let Δ

be an $n \times n$ symmetric matrix with non-negative elements δ_{ij} (where $i = 1, \ldots, n$ and $j = 1, \ldots, n$), which denote the *dissimilarities* to be analyzed. The row and column entries of Δ may refer to stimuli, persons, test items, or some other labelled units of analysis, which are called *objects* here. Any symmetric measure of similarity among the objects is also a potential source of data, provided that the direction of the numerical scale is reversed in advance by a suitable transformation. Each object will be allocated to one and only one of $K \leq n$ classes or *clusters*, where K is a parameter with the status of an *option* in the analysis. It is convenient to have two ways to express the allocation of an object to a cluster: either an *indicator matrix* may be used, defined as the order $n \times K$ matrix $\mathbf{E} = \{e_{ik}\}$, where e_{ik} is a binary variable which is equal to one if object i is an element of cluster k, and equal to zero otherwise, or a collection of *index sets*, which are defined as $J_k = \{ i \mid e_{ik} = 1 \text{ for } k = 1, \ldots, K\}$. Thus the assumption that the classes form a partition can be expressed either as diagonality of the matrix $\mathbf{N} = \mathbf{E}'\mathbf{E}$, or as the equation $J_k \cap J_l = \varnothing$ for all k,l. The diagonal elements of $\mathbf{N}$ are n_k, the number of objects in cluster k.

Spatial modeling in cluster differences scaling consists of mapping the set of clusters onto $\mathbb{R}^p$, so that each cluster k is associated with a point $\mathbf{x}_k$, with coordinates $\{x_{ka}\}$, where $a = 1, \ldots, p$. The cluster coordinates are collected in the $K \times p$ matrix $\mathbf{X}$, called the *cluster configuration*, and the *Euclidean distance* between cluster points $\mathbf{x}_k$ and $\mathbf{x}_l$ (rows k and l of $\mathbf{X}$) is defined as

$$d_{kl}(\mathbf{X}) = \|\mathbf{x}_k - \mathbf{x}_l\|,$$

where the notation $\| \cdot \|$ denotes the Euclidean norm. It is useful to be able to assign a *weight* w_{ij} to each pair of objects; this double-indexed variable may be *binary* – for instance, in case of a (0,1)-coding scheme for missing data –, or *continuous* – for instance, in case of a differential weighting scheme based upon the inverse of the estimated standard error in replicated measurements of δ_{ij}. It is assumed that $w_{ii} = 0$ for all i, i.e. the self-dissimilarities are not modelled.

When the allocation leads to object $i \in J_k$ and object $j \in J_l$, their dissimilarity δ_{ij} will be represented in the model as the Euclidean distance $d_{kl}(\mathbf{X})$, which is constant for all other pairs of objects in which the first is chosen from cluster k and the second from cluster l. Thus cluster differences scaling

inherits an *equidistance property* from the hierarchical clustering model, in which all objects within one subclass have the same distance towards all objects in any other non-overlapping subclass. It is useful to consider the consequences of this property for the expected patterns in the dissimilarities. If the data matrix and the corresponding $n \times n$ distance matrix are partitioned into blocks (k,l) by permuting the row and column indices according to the sequence in the index sets $J_1, \ldots, J_k, \ldots, J_K$, then the dissimilarities are supposed to vary randomly within a block, while the correponding distance is constant within the same block. Between blocks, differences in distance will reflect the tendency of the corresponding dissimilarities to vary systematically.

Several least squares loss functions can be built up from the natural block-wise components

$$\sigma_{kl}^2(\mathbf{E},\mathbf{X}) = \sum_{i \in J_k} \sum_{j \in J_l} w_{ij}(\delta_{ij} - d_{kl}(\mathbf{X}))^2, \tag{1}$$

which is the loss due to the fact that the $n_k \times n_l$ measurements $\{\delta_{ij} \mid i \in J_k, j \in J_l\}$ deviate from the model distance between cluster k and cluster l. Indicator matrix $\mathbf{E}$ is included in the notation $\sigma_{kl}^2(\mathbf{E},\mathbf{X})$ to keep the dependence of the loss function on the allocation of objects to clusters (in the summation) explicit. Generally, cluster differences scaling aims at finding a partitioning of the n objects into K classes and at a simultaneous embedding of the classes into a p-dimensional Euclidean space. Different varieties of the method are obtained by different choices in building up the overall loss function. Even if we restrict attention to varieties based on the simple sum (called *Total Stress*)

$$\sigma^2(\mathbf{E},\mathbf{X}) = \sum_{k \le l} \sigma_{kl}^2(\mathbf{E},\mathbf{X}), \tag{2}$$

there is a major consideration of including or excluding a *Within-clusters loss component*, i.e., to extend the summation over all pairs including $k = l$, as in (2), or not. Each Within-clusters component necessarily has $d_{kk}(\mathbf{X}) = 0$, so that (1) becomes $\sigma_{kk}^2(\mathbf{E}) = \sum_{i \in J_k} \sum_{j \in J_k} w_{ij}\delta_{ij}^2$, which is a function of $\mathbf{E}$ only. It was argued in Heiser (1993) that inclusion of Within-clusters loss would favor the selection of hyperspherically shaped clusters, so that long, straggling sets of points would tend to be broken up into several different groups. Particularly in situations where dissimilarity is defined on several strongly correlated variables, or where nonlinearities exist, sphericity is a well-known

disadvantage of sum of squares partitioning methods (e.g., Gordon, 1981). However, as noted in Groenen (1993) and in Heiser and Groenen (1993), there are also advantages in including the Within component, and these will be described in this paper.

The next section will show how the weighted loss component in (1) can be decomposed into various parts that do not depend on **X** and one part that does depend on **X**. Also, orthogonality of least squares estimates with their residuals allows a breakdown of the sum of squared dissimilarities into contributions from several sources of variation, leading to an Analysis of Dispersion analogous to the familiar Analysis of Variance. These decompositions are illustrated with an example from co-citation analysis. Then a convergent iterative algorithm for fitting the cluster differences model is sketched, with special emphasis on the phase that handles the allocation of objects to clusters, which resembles a rudimentary K-means algorithm. Consideration of the dissimilarity variation within clusters is next shown to make cluster differences scaling technically equivalent to a constrained multidimensional scaling task that forces the n points to occupy at most K locations. An attractive joint display of the individual objects and the cluster locations is possible, in which the latter are centers of gravity of the former, by plotting the implicit object configuration, which is a simple matrix function of the cluster configuration with a specific local optimality property. The joint display is illustrated with a well-known example from the literature, the Iris data. Before closing with the Discussion section, we develop a fuzzy version of cluster differences scaling, primarily to be used in a successive approximation strategy for avoiding local minima. A small but effective Monte Carlo study shows very clearly that the proposed strategy outperforms two sensible competing strategies, and markedly alleviates the local minimum problem.

Loss Decomposition and Breakdown of Sum of Squares

For a good insight in the CDS method we must study the orthogonal decomposition of the sum of squared residuals in each block. As a consequence of the least squares optimality of the (weighted) mean, each block-wise loss component (1) can be additively decomposed into two parts:

$$\sigma^2_{kl}(\mathbf{E},\mathbf{X}) = \sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kl})^2 + \tilde{w}_{kl}(\tilde{\delta}_{kl} - d_{kl}(\mathbf{X}))^2. \quad (3)$$

Here $\tilde{w}_{kl}$ is the sum of all weights in block (k,l), and $\tilde{\delta}_{kl}$ is the weighted mean dissimilarity; explicitly, these quantities are defined as:

$$\tilde{w}_{kl} = \sum_{i\in J_k}\sum_{j\in J_l} w_{ij}, \quad (4)$$

$$\tilde{\delta}_{kl} = \sum_{i\in J_k}\sum_{j\in J_l} w_{ij}\delta_{ij} \;/\; \tilde{w}_{kl}. \quad (5)$$

The block-wise mean is called the *Sokal-Michener dissimilarity*, because it plays a central role in the group-average method of Sokal and Michener (1958). The first part in (3) is a subcomponent measuring the variability in the dissimilarities that is not accounted for, neither by the clustering, nor by the spatial model; because this subcomponent is a sum of squared deviations of the dissimilarities associated with cluster k and l around their mean, it will be called the *Among-clusters Error Sum of Squares (SSQ)*. Introduction of the terminology "Among-clusters" is necessary to avoid possible confusion which could arise when the seemingly more natural "Between-clusters" would be used, because the quantity dealt with is actually a "Within"-type SSQ, not a "Between"-type. The second part of (3) measures *Lack of fit* of the spatial model, i.e., the weighted deviation of the distance in the cluster configuration from the Sokal-Michener dissimilarity. The residuals making up the Among-clusters Error SSQ are always zero in case $K = n$, and if $K < n$ they are zero only if all dissimilarities are equal within blocks.

The inclusion of a Within-clusters loss component yields two additional subcomponents. For the kth diagonal block, insertion of $d_{kk}(\mathbf{X}) = 0$ into (3) leads to

$$\sigma^2_{kk}(\mathbf{E}) = \sum_{i\in J_k}\sum_{j\in J_k} w_{ij}(\delta_{ij} - \tilde{\delta}_{kk})^2 + \tilde{w}_{kk}\tilde{\delta}^2_{kk}, \quad (6)$$

it being understood that the summation is over $i < j$. The first part of (6) is the residual variability of the dissimilarities in block k around their mean, and will be called the *Within-clusters Error Sum of Squares*. The second part reflects the relative tightness of cluster k, to be called *Lack of homogeneity*, because the Sokal-Michener dissimilarity of a homogeneous cluster will tend to be small, while it

becomes larger with increasing heterogeneity. Note that neither of these subcomponents depend on **X**; it is only the *k*th column of **E** that is directly involved in (6).

Summarizing the four components that can be distinguished in the Total Stress loss function (2), in which (3) and (6) are pooled across blocks and pairs of blocks, we obtain:

$$\sigma^2(\mathbf{E},\mathbf{X}) = \sum_{k\le l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - d_{kl}(\mathbf{X}))^2 \qquad \underline{\text{Total Stress}}$$

$$= \sum_{k<l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kl})^2 \qquad \text{Among-clusters Error SSQ}$$

$$+ \sum_{k}\sum_{i\in J_k}\sum_{j\in J_k} w_{ij}(\delta_{ij} - \tilde{\delta}_{kk})^2 \qquad \text{Within-clusters Error SSQ}$$

$$+ \sum_{k<l} \tilde{w}_{kl}(\tilde{\delta}_{kl} - d_{kl}(\mathbf{X}))^2 \qquad \text{Lack of spatial fit}$$

$$+ \sum_{k} \tilde{w}_{kk}\tilde{\delta}^2_{kk} \qquad \text{Lack of homogeneity}$$

This decomposition allows us to evaluate two aspects of a CDS solution against an internal standard: the adequacy of the spatial assumption given the chosen number of clusters by comparing the Lack of spatial fit to the Among-clusters Error SSQ, and the adequacy of the number of clusters by comparing the Among-clusters Error SSQ to the Within-clusters Error SSQ. The Among-clusters component should be approximately $(K - 1)$ times larger than the Within-clusters component, since it involves that many more (or less, if $K < 3$) blocks of dissimilarities (a more precise statement is possible by dividing the SSQ's by the appropriate degrees of freedom, see below). To evaluate the Lack of spatial fit and the Lack of homogeneity of the clusters, it is better to use a breakdown of the total (weighted) sum of squared dissimilarities, to which we turn next.

Two independent least squares estimates are involved in the CDS problem: given the allocation of the objects into K clusters, the Sokal-Michener dissimilarities are least squares estimates of the expected dissimilarity among the clusters without assuming a spatial model, and the distances among cluster points in turn form an optimal spatial representation of the Sokal-Michener dissimilarities. A basic property of least squares estimates is that they are orthogonal to the residuals. With respect to the expected dissimilarity among the clusters, orthogonality implies that for any block (k,l), including

$k = l$, the weighted cross product of $\tilde{\delta}_{kl}$ and $(\delta_{ij} - \tilde{\delta}_{kl})$ vanishes, so that the equation $\delta_{ij} = (\delta_{ij} - \tilde{\delta}_{kl}) + \tilde{\delta}_{kl}$, when squared, multiplied by w_{ij}, and summed, yields

$$\sum_{k\leq l} \sum_{i\in J_k} \sum_{j\in J_l} w_{ij}\delta_{ij}^2 = \sum_{k\leq l} \sum_{i\in J_k} \sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kl})^2 + \sum_{k\leq l} \tilde{w}_{kl}\tilde{\delta}_{kl}^2 . \qquad (7)$$

Here the Total SSQ of the dissimilarities is broken down into an Error SSQ and a Between SSQ. The Between SSQ measures the total dispersion between the clusters, and can be further broken down by using orthogonality a second time. The usual necessary condition (e.g., Guttman, 1968) for a minimum of least squares multidimensional scaling implies that the Lack of spatial fit will be minimal for any $\tilde{\mathbf{X}}$ satisfying $d_{kl}(\tilde{\mathbf{X}}) \neq 0$ and

$$\sum_{k<l} \tilde{w}_{kl}(\tilde{\delta}_{kl} - d_{kl}(\tilde{\mathbf{X}}))d_{kl}(\tilde{\mathbf{X}}) = 0,$$

and therefore, taking weighted sums of squares of both sides of the equation $\tilde{\delta}_{kl} = (\tilde{\delta}_{kl} - d_{kl}(\tilde{\mathbf{X}})) + d_{kl}(\tilde{\mathbf{X}})$, we must have

$$\sum_{k<l} \tilde{w}_{kl}\tilde{\delta}_{kl}^2 = \sum_{k<l} \tilde{w}_{kl}(\tilde{\delta}_{kl} - d_{kl}(\tilde{\mathbf{X}}))^2 + \sum_{k<l} \tilde{w}_{kl}\tilde{d}_{kl}^2(\tilde{\mathbf{X}}) \qquad (8)$$

at the optimum. Inserting (8) in (7), simplifying the summation notation in the Total SSQ, separating the summation over $k \leq l$ in the Error SSQ and Between SSQ again into $k < l$ and $k = l$, and rearranging terms, we obtain an *Analysis of Dispersion* (in analogy to Analysis of Variance) collected in Table 1. Included in this Table are the *degrees of freedom* (*df*) for each sum of squares, so that a

--- Insert Table 1 about here ---

Mean Square (MS) can be calculated as MS = SSQ / *df*. The Between component includes K degrees of freedom for Lack-of-homogeneity, originating from the Within-clusters loss component, and an *Among-clusters Dispersion-Accounted-For (D.A.F.)*, corresponding to the second term of (8). The degrees of freedom of the Among-clusters D.A.F. are equal to the number of free parameters in $\tilde{\mathbf{X}}$, which is $K \times p$ minus the number of identification constraints to fix the invariances of the distance function: p constraints to identify the center of $\tilde{\mathbf{X}}$, and $p(p - 1)/2$ constraints to identify its orientation. Analogous to the common way to present Variance-Accounted-For, the Dispersion-Accounted-For

could be expressed as a percentage of the Total SSQ of the dissimilarities.

When it is reasonable to assume that the dissimilarities are normally distributed, the obtained Mean Squares could be evaluated against a non-central F-distribution with the appropriate degrees of freedom, to establish the significance of the model. It should be noted, however, that such comparisons are strictly valid only for fixed **E**, i.e., for a fixed grouping of the objects (as in Gower, 1989). Nevertheless, when we have been optimizing over **E**, the obtained Mean Squares in cluster differences scaling can be useful as estimated upper bounds for these ratio's.

As an illustration of the model and the associated Analysis of Dispersion, a table of co-citations among journals in the earth sciences was analyzed with the algorithm to be discussed below (Note). The co-citation data consist of symmetrized frequencies c_{ij} with which journal i is cited by journal j, or vice versa. Frequency of co-citation is usually supposed to be inversely related to inter-discipline distance (Tijssen, 1992). By extension of the classic gravity model to situations of non-physical interaction (Tobler, 1976), it may be assumed that $c_{ij} = m_i m_j / d_{ij}^2(\mathbf{X})$, so that not only inverse inter-discipline distance, but also the size of the disciplines contributes to mutual citations, in terms of masses $m_i = c_{i+}$ and $m_j = c_{+j}$. Accordingly, the co-citation frequencies were transformed to dissimilarities by $\delta_{ij} = (m_i m_j / c_{ij})^{1/2}$, thus correcting in each pair the raw co-citations for the total number of citations of both journals. If some $c_{ij} = 0$, then δ_{ij} is not defined, and $w_{ij} = 0$ (otherwise, $w_{ij} = 1$); the diagonal entries c_{ii} (number of self-citations) were excluded from the analysis.

For the CDS analysis with $K = 8$ clusters and $p = 2$ dimensions, the configuration of clusters is given in Figure 1, together with the allocation of journals to clusters. Some clusters have a clear

--- Insert Figure 1 about here ---

interpretation, like cluster 7, which contains journals involved in oceanography, and cluster 8, which consists of journals in palaeontology. Cluster 3 also contains journals from oceanography, marine biology, as well as some multi-disciplinary journals. Clusters 1, 2, and 4 consist of journals from geology, geosciences and geography, which may be considered to be the "harder" geology sciences. The interpretation of other clusters is sometimes difficult, like in case of cluster 5, which seems to

contain journals that cite in an interdisciplinary fashion. One advantage of CDS is that from the positioning of the clusters in the fitted space something can be said about the relationships between the clusters, while in conventional non-hierarchical cluster analysis such information is frequently not available. From Figure 1 it appears, for example, that along the first (horizontal) dimension, the neighbours of cluster 4 at the left, apart from the minor cluster 6, are clusters 1 and 2, confirming the above remarks on the basis of cluster memberships alone. Furthermore, Figure 1 also shows the clusters involved in oceanography and palaeontology (clusters 3, 7 and 8) to be located together at the lower-right side of the plot, while cluster 5 (the interdisciplinary journals) is in the middle.

The various Stress components and the Analysis of Dispersion for the co-citation example are

--- Insert Table 2 about here ---

given in Table 2. Clearly, the high Among-clusters Dispersion-Accounted-For (75.6 %) and the low Lack of spatial fit for the cluster configuration (2.8 %) are signs that this analysis is satisfactory. When we compare the SSQ due to Lack of spatial fit (raw Stress) with the Among-clusters Error SSQ, we see that the Error is about 6.5 times larger than the Stress, indicating that the spatial assumption is adequate. The two Error components (in terms of their Mean Squares: 0.210 and 0.045) are in the same order of magnitude, but not the same, indicating that we might have chosen the number of clusters K too low. The figures in the MS column show that the total Between component (57.695) is big enough compared to the Error (0.190) to feel confident that the variation of the Sokal-Michener dissimilarities is systematic. However, the Mean Squares for Lack of homogeneity (8.995) and Lack of spatial fit (4.805) are perhaps a bit too high, and might also profit from a higher K.

Algorithm

A convergent algorithm for minimizing the cluster differences scaling loss function will now be described. It alternates between an incomplete K-means clustering phase, called *allocation phase*, and an unconstrained multidimensional scaling phase. The incomplete K-means clustering procedure involves n allocations, while the size of the multidimensional scaling problem is of order K. The allocation phase is treated first.

For any given, fixed configuration $\underline{\mathbf{X}}$, the cluster differences loss function (2) can be rewritten, by switching to indicator matrix notation and rearranging some terms, as

$$\begin{aligned}\sigma^2(\mathbf{E},\underline{\mathbf{X}}) &= \sum_i \sum_k e_{ik} \left[\sum_{j\neq i} w_{ij} \sum_l e_{jl}(\delta_{ij} - d_{kl}(\underline{\mathbf{X}}))^2\right] \\ &= \sum_i \sum_k e_{ik}\gamma_{ik}^2(\mathbf{E}_{(i)},\underline{\mathbf{X}}). \end{aligned} \tag{9}$$

Here the loss function has been extended with terms involving $k > l$ for the Among-clusters loss, and $i > j$ for the Within-clusters loss. This extended loss is always exactly twice the original one, due to the symmetry of w_{ij}, δ_{ij}, and $d_{kl}(\underline{\mathbf{X}})$, and its use has the advantage of simplifying the presentation of the allocation phase. It is important to note that the quantity $\gamma_{ik}^2(\mathbf{E}_{(i)},\underline{\mathbf{X}})$, implicitly defined in (9), does not involve the ith row of the indicator matrix, which is expressed by writing it as a function of $\mathbf{E}_{(i)}$, i.e., the indicator matrix with the ith row replaced by a row of zeros. The summation over j has been explicitly written as $j \neq i$, even though we already tacitly have $w_{ii} = 0$, to underline the assumption that the self-dissimilarities are not modelled, which is also the reason why the ith row of $\mathbf{E}$ is skipped in the summation over j, and $\mathbf{E}_{(i)}$ suffices. Since we can write

$$\gamma_{ik}^2(\mathbf{E}_{(i)},\underline{\mathbf{X}}) = \sum_{j\neq i} w_{ij} \sum_{l\neq k} e_{jl}(\delta_{ij} - d_{kl}(\underline{\mathbf{X}}))^2 + \sum_{j\in J_k} w_{ij}\delta_{ij}^2 , \tag{10}$$

the only difference between the present version and a treatment that would exclude the Within-clusters loss component is, that on top of the comparison of rows of dissimilarities with rows of the corresponding model distances in the first term of (10), the present version also includes the second term $\sum_{j\in J_k} w_{ij}\delta_{ij}^2$, the sum of squared dissimilarities of object i towards the objects in cluster k. All other things being equal, the inclusion of this component, as expected, implies that object i will be

joined with close objects, even when it more naturally would belong to a non-compact cluster, and that the cluster sizes n_k will tend to become equal (because a sum is involved, not an average).

The first term of (10) is a weighted squared Euclidean distance, to be read as follows: when considering object i in regard to cluster k, calculate a (weighted) sum over all *other* objects in the *other* clusters, each term of which is the deviation of δ_{ij} from the model distance between the cluster to which j is presently allocated (cluster l) and the one we are contemplating for i (cluster k). Clearly, we would like to (re)allocate object i to the cluster k that would give it the best fitting distance(s). In the present version of CDS, $\gamma^2_{ik}(\cdot)$ combines the aspect of best Among-clusters fit with the objective to obtain maximal Within-clusters compactness.

Returning to (9), we may observe that this double-indexed inner product criterion is equivalent to the *minimum sum of squares* criterion of K-means clustering, frequently called – not quite correctly, since it does not involve division by the cluster sizes (Späth, 1985) – the *minimum variance* criterion. However, the equivalence is formal only, because usual K-means algorithms work with a squared distance (or sum of squares) between data points and cluster centers, while here we work with a squared distance between rows of Δ and rows of $D(\underline{\mathbf{X}})$. In convergent algorithms for K-means clustering (Gordon and Henderson, 1977; Hartigan and Wong, 1979; Selim and Ismail, 1984), two substeps are distinguished: the allocation or reallocation of objects among clusters, using a specific allocation or exchange rule, and the updating of cluster centers. In the present case, optimal updating of cluster centers (conditional upon the current allocation) implies recalculating the Sokal-Michener dissimilarities. Note that this substep – averaging rows of Δ within the current clusters – is only an intermediary one here, because we have the additional requirement of a Euclidean embedding of these averages (which are the Sokal-Michener dissimilarities) into p dimensions, to be executed in the multidimensional scaling phase. Therefore, the allocation phase cycles over objects with the cluster points and their distances fixed, adjusts $\mathbf{E}$ whenever a reallocation decreases (9), and finishes with the calculation of new $\tilde{w}_{kl}$ and $\tilde{\delta}_{kl}$.

It is usually regarded as not very harmful to let the (re)allocation rule involve all kinds of heuristic procedures for avoiding local minima, such as cluster lumping and cluster splitting in the well-known

ISODATA algorithm of Ball and Hall (1967), simultaneous exchange of two objects (Banfield and Bassill, 1977), or of several objects (Kernighan and Lin, 1970). But in contrast to the usual situation, criterion (9) involves $\mathbf{E}_{(i)}$ within the squared distance function, and this feature is a fundamental complicating factor for such heuristics. Fortunately, however, the quantity $\gamma_{ik}^2(\cdot)$ does not depend on the ith row of $\mathbf{E}$, so that a simple convergent algorithm is possible if we proceed row by row. Later, an approach of successive approximations will be described that turns out to have good chances to find the global minimum.

For any particular row i, minimizing (9) over $\{e_{ik} \mid k = 1, \ldots, K\} = \mathbf{e}_i$ with the allocations of the other objects fixed implies that the argument $\mathbf{E}_{(i)}$ of the weighted squared Euclidean distance can be dropped, and therefore the problem becomes one of finding

$$\min_{\mathbf{e}_i} \; \Sigma_k \, e_{ik} \gamma_{ik}^2(\underline{\mathbf{X}}) \tag{11}$$

over all feasible binary vectors $\mathbf{e}_i$. Since only the K rows of the $K \times K$ identity matrix are feasible for $\mathbf{e}_i$, clearly the minimal inner product is attained for row $\tilde{k}$, where $\tilde{k}$ indicates the index of the minimal element of $\{\gamma_{ik}^2(\underline{\mathbf{X}}) \mid k = 1, \ldots, K\}$. Thus, given the allocation of the other objects and the current cluster locations, the optimal allocation of object i is cluster $\tilde{k}$. If $\tilde{k}$ is equal to the current allocation of object i, we immediately move to the next object. Otherwise, object i is reallocated and the ith row of $\mathbf{E}$ is adjusted first. It is known that a reallocation scheme based on repeated use of (11), called the *minimal distance method*, may generate empty classes (Späth, 1985), but in all test runs of the present version of CDS so far, this problem did not appear. However, it turns out that cluster differences scaling without the Within-clusters loss component does encounter the empty classes problem, so that it needs extension with an *exchange method* of reallocation.

The multidimensional scaling phase regards only the lack-of-fit component of loss function (2), which is (*cf.* Table 1)

$$\sigma^2(\mathbf{X}) = \Sigma_{k<l} \, \tilde{w}_{kl} (\tilde{\delta}_{kl} - d_{kl}(\mathbf{X}))^2. \tag{12}$$

Any standard algorithm for finding stationary values of the weighted least squares function (12) could

be used to update the cluster configuration, conditional upon the current weights $\{\tilde{w}_{kl}\}$ and the current Sokal-Michener dissimilarities $\{\tilde{\delta}_{kl}\}$; the present implementation uses the SMACOF algorithm based on a strategy called *iterative majorization* (De Leeuw and Heiser, 1980). Some more comments on this approach will be made in the next section, which discusses a connection with restricted multidimensional scaling of the individual points.

Summarizing, the proposed CDS algorithm alternates between one phase in which the clusters are updated with a minimal distance method, and another phase in which cluster configuration is updated with an iterative majorization method, until successive values of the Total Stress reach a predetermined stop criterion. The whole sequence can be shown to be convergent, but because all operations are merely conditionally optimal, while the loss function is known to have multiple stationary points, only convergence to a local minimum is guaranteed. As remarked earlier, initialization is an important issue that will be addressed in a separate section after the next one.

Equivalence to Constrained MDS, and Use of the Implicit Object Configuration

An interesting property of CDS with inclusion of the Within-clusters component is its equivalence to an order-n multidimensional scaling task in which the $n \times p$ object configuration $\mathbf{Z}$ is constrained to be of the form $\mathbf{Z} = \mathbf{EX}$, with $\mathbf{E}$ and $\mathbf{X}$ defined as before. To show that this property holds, it is convenient to first re-express the loss function in yet another way, by switching from summation over $k \leq l$ with both i and j ranging over 1, ... , n to an equivalent summation over $i < j$ with both k and l ranging over 1, ... , K. Explicitly, we have

$$\begin{aligned}\sigma^2(\mathbf{E},\mathbf{X}) &= \sum_{k<l} \sum_i \sum_j e_{ik} e_{jl} w_{ij} (\delta_{ij} - d_{kl}(\mathbf{X}))^2 + \sum_k \sum_{i<j} e_{ik} e_{jk} w_{ij} \delta_{ij}^2 \\ &= \sum_{i<j} w_{ij} [\sum_k \sum_l e_{ik} e_{jl} (\delta_{ij} - d_{kl}(\mathbf{X}))^2] = \sum_{i<j} w_{ij} \sigma_{ij}^2(\mathbf{E},\mathbf{X}). \qquad (13)\end{aligned}$$

The correctness of this switch can be checked by more closely considering the summation

$$\sum_{k<l} \sum_i \sum_j e_{ik} e_{jl} = \sum_{k<l} \sum_{i<j} e_{ik} e_{jl} + \sum_{k<l} \sum_{i>j} e_{ik} e_{jl} + \sum_{k<l} \sum_{i=j} e_{ik} e_{jl}\,, \qquad (14)$$

remembering that the weights and the residuals are symmetric. The last term of (14) is zero, because the classes are assumed to be exclusive, and the middle term remains the same if we switch both unequal signs simultaneously. Thus $\sum_{k<l} \sum_i \sum_j e_{ik}e_{jl} = \sum_{i<j} \sum_{k\neq l}, e_{ik}e_{jl}$ and for the Within-clusters part $\sum_k \sum_{i<j} e_{ik}e_{jk} = \sum_{i<j} \sum_{k=l} e_{ik}e_{jl}$, which establishes that the same residuals are selected in (13).

Looking further into (13), we note that, for any pair of objects (i,j), the term $\sigma_{ij}^2(\mathbf{E},\mathbf{X})$ may be rewritten as

$$\begin{aligned}\sigma_{ij}^2(\mathbf{E},\mathbf{X}) &= \sum_k \sum_l e_{ik}e_{jl}\delta_{ij}^2 + \sum_k \sum_l e_{ik}e_{jl}d_{kl}^2(\mathbf{X}) - 2\delta_{ij} \sum_k \sum_l e_{ik}e_{jl}d_{kl}(\mathbf{X}) \\ &= \delta_{ij}^2 + d_{ij}^2(\mathbf{EX}) - 2\delta_{ij}d_{ij}(\mathbf{EX}) = (\delta_{ij} - d_{ij}(\mathbf{EX}))^2,\end{aligned}$$

where the following properties of $\mathbf{E}$ were used: $\sum_k \sum_l e_{ik}e_{jl} = 1$ for all i and j, which simplifies the first term of $\sigma_{ij}^2(\cdot)$, and, for any fixed particular pair of objects, the $K \times K$ matrix with elements $\{e_{ik}e_{jl}\}$ has a one at a single position, say position (s,t), and zeros elsewhere. Thus row i of the indicator matrix satisfies $\sum_k e_{ik}\mathbf{x}_k = \mathbf{x}_s$, while row j satisfies $\sum_l e_{il}\mathbf{x}_l = \mathbf{x}_t$, so that

$$\sum_k \sum_l e_{ik}e_{jl}d_{kl}^2(\mathbf{X}) = d_{st}^2(\mathbf{X}) = \|\mathbf{x}_s - \mathbf{x}_t\|^2 = \|\sum_k e_{ik}\mathbf{x}_k - \sum_l e_{jl}\mathbf{x}_l\|^2 = d_{ij}^2(\mathbf{EX}),$$

which accounts for the simplication of the second term of $\sigma_{ij}^2(\cdot)$, while the third term can be handled analogously. It follows that $\sigma^2(\mathbf{E},\mathbf{X}) = \sum_{i<j} w_{ij}\, \sigma_{ij}^2(\mathbf{E},\mathbf{X})$ is a weighted multidimensional scaling loss function with configuration of the form $\mathbf{Z} = \mathbf{EX}$, as asserted.

The equivalence to a constrained multidimensional scaling task enables us to locate the individual objects in the cluster configuration in a natural way. To understand why, we need some further results from SMACOF theory. Suppose that we consider any individual object configuration $\mathbf{Z}$, not necessarily one satisfying the constraints $\mathbf{Z} = \mathbf{EX}$. Let the associated loss function be expressed as

$$\begin{aligned}\sigma^2(\mathbf{Z}) &= \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}(\mathbf{Z}))^2 \\ &= \sum_{i<j} w_{ij}\delta_{ij}^2 + \text{tr } \mathbf{Z}'\mathbf{V}\mathbf{Z} - 2 \text{ tr } \mathbf{Z}'\mathbf{B}(\mathbf{Z})\mathbf{Z}\,, \qquad (15)\end{aligned}$$

where the $n \times n$ matrix $\mathbf{V}$ has off-diagonal elements $-w_{ij}$ and diagonal elements $\sum_{j\neq i} w_{ij}$, and where

the off-diagonal values of the (order $n \times n$) matrix-valued function $B(\mathbf{Z})$ are $b_{ij}(\mathbf{Z}) = -w_{ij}\delta_{ij} / d_{ij}(\mathbf{Z})$ if $d_{ij}(\mathbf{Z}) > 0$ and zero otherwise, while the diagonal values of $B(\mathbf{Z})$ are such that its rows and columns sum to zero. In the SMACOF approach, the *Guttman transform* of $\mathbf{Z}$ is the (order $n \times p$) matrix-valued function $G(\mathbf{Z})$ defined as

$$G(\mathbf{Z}) = \mathbf{V}^{+}B(\mathbf{Z})\mathbf{Z}, \tag{16}$$

with $\mathbf{V}^{+}$ the Moore-Penrose inverse of $\mathbf{V}$. Given any fixed configuration $\underline{\mathbf{Z}}$, the Guttman transform $G(\underline{\mathbf{Z}})$ provides the locally best improvement of $\underline{\mathbf{Z}}$. At the point of convergence, the optimal, unconstrained configuration $\mathbf{Z}_{\circ}$ has the *fixed point* property $G(\mathbf{Z}_{\circ}) = \mathbf{Z}_{\circ}$, indicating that further improvements are not possible. In cluster differences scaling we may now define the *implicit object configuration* as $G(\mathbf{Z}_{*})$, where $\mathbf{Z}_{*}$ is the optimal *constrained* configuration $\mathbf{Z}_{*} = \mathbf{E}_{*}\mathbf{X}_{*}$, with $\mathbf{E}_{*}$ the optimal partitioning of the objects, and $\mathbf{X}_{*}$ the optimal cluster configuration.

The implicit object configuration has the following optimality property. Generally, the loss function $\sigma^2(\cdot)$ is approximated from above by a quadratic called the *majorizing function* (De Leeuw and Heiser, 1980). At the constrained local optimum $\mathbf{Z}_{*}$, the majorizing function has the form

$$\mu(\mathbf{Z} \mid \mathbf{Z}_{*}) = \sum_{i<j} w_{ij}\delta_{ij}^{2} - \text{tr } G(\mathbf{Z}_{*})'\mathbf{V}G(\mathbf{Z}_{*}) + \text{tr } (G(\mathbf{Z}_{*}) - \mathbf{Z})'\mathbf{V}(G(\mathbf{Z}_{*}) - \mathbf{Z}) , \tag{17}$$

which satifies $\mu(\mathbf{Z}_{*} \mid \mathbf{Z}_{*}) = \sigma^2(\mathbf{Z}_{*})$, as can be readily verified by comparing (17) to (15), using (16). The unconstrained minimum of $\mu(\cdot \mid \mathbf{Z}_{*})$ is obtained by annihilating the last term of (17); it follows that $G(\mathbf{Z}_{*})$ is the unconstrained object configuration that minimizes the best local approximation of $\sigma^2(\cdot)$ at the constrained optimum. The constrained minimum of $\mu(\cdot \mid \mathbf{Z}_{*})$ is obtained by projection of $G(\mathbf{Z}_{*})$, in the metric $\mathbf{V}$, onto the constrained region. Under the present type of constraints, where the constrained region is $\mathbf{Z} = \mathbf{EX}$, the result of the projection of $G(\mathbf{Z}_{*})$, denoted by $P(G(\mathbf{Z}_{*}))$, is $\mathbf{E}_{*}(\mathbf{E}_{*}'\mathbf{V}\mathbf{E}_{*})^{-1}\mathbf{E}_{*}'\mathbf{V}G(\mathbf{Z}_{*})$. Since $\mathbf{Z}_{*}$ is optimal, we must have $P(G(\mathbf{Z}_{*})) = \mathbf{Z}_{*}$, i.e., $\mathbf{Z}_{*}$ is a fixed point of the compound function $P(G(\cdot))$. Operation $P(\cdot)$ collapses the points into the cluster-wise (weighted) centers of gravity, and operation $G(\cdot)$ inversely relocates the points so that their increased scatter better accommodates the dissimilarities. Thus the implicit object configuration is our current best guess of the location of the individual objects, and represents them in such a way that the cluster

points are the centers of gravity of the corresponding individual points.

A mechanical interpretation of the operation that moves the clustered configuration into the implicit object configuration can be given (for the unweighted case, in which $\mathbf{V}^+$ amounts to division by n) by inserting the constraint structure of $\mathbf{Z}_*$ into (16), and simplifying; the ith row of the Guttman transform becomes

$$g_i(\mathbf{Z}_*) = ((n_k + \beta_i) / n) \{ \mathbf{x}_{k*} - \frac{1}{n_k + \beta_i} [n_k \mathbf{x}_{k*} + \sum_{l \neq k} \sum_{j \in J_l} (\delta_{ij} / d_{kl}(\mathbf{X}_*)) \mathbf{x}_{l*}] \},$$

where $\mathbf{x}_{k*}$ and $\mathbf{x}_{l*}$ are the kth and the lth row of $\mathbf{X}_*$, respectively, and the total correction factor β_i is equal to $\sum_{l \neq k} \sum_{j \in J_l} (\delta_{ij} / d_{kl}(\mathbf{X}_*))$. When the $n - n_k$ coefficients $\delta_{ij} / d_{kl}(\mathbf{X}_*)$ in the linear combination of the $\mathbf{x}_{l*}$ are on average equal to one, the scale factor outside the braces becomes unity. Interpreting the terms inside the braces, we see that object i will be moved from position $\mathbf{x}_{k*}$ into a direction that is the resultant of n_k unit forces exerted from cluster point k, and $n - n_k$ fit-dependent forces exerted from the $K - 1$ other cluster points. Each $\mathbf{x}_{l*}$ is associated with n_l forces. Object i is pulled or pushed away in proportion to the extent that $d_{kl}(\mathbf{X}_*)$, the distance between the cluster to which object i has been allocated and the lth cluster, overestimates or underestimates δ_{ij}, the dissimilarity between object i and an object j in cluster l. Thus the move from the cluster point $\mathbf{x}_{k*}$ to the implicit object point $g_i(\mathbf{Z}_*)$ amounts to a translation in the direction of a misfit correcting, convex combination of the cluster points, plus a rescaling if the average correction deviates from one.

The use of implicit object configuration is now illustrated with an analysis of the Anderson (1935) Iris data, a widely used test data set since Fisher (1936) used it to illustrate linear discriminant analysis. The Iris data describe the floral leaves of 150 irises, 50 from each of 3 varieties – *setosa (a)*, *versicolor (b)*, and *virginica (c)* – , in terms of 4 descriptive variables: petal width, petal length, sepal width, and sepal length. Cluster differences scaling was applied with dissimilarity between the objects defined as the Euclidean distance in the original 4-dimensional space. Since the variables are clearly correlated, dimensionality was chosen low ($p = 2$), but the number of clusters was chosen high ($K = 25$), to accomodate possible non-spherical shapes of the varieties. The analysis of dispersion for this example is given in Table 3. Almost all dispersion is preserved in the 25-cluster configuration,

--- Insert Table 3 and Figure 2 about here ---

which accounts for 99.6% of the sum of squared dissimilarities (in fact, of the *total variance*, as Euclidean distances are used here). Because the Lack of spatial fit (6.87) is a factor 4 smaller than the Among-clusters Error SSQ (28.02), and its Mean square (0.027) is very close to zero, it seems entirely justified to go from four to two dimensions. The mean squared Among-clusters Error (0.003) and Within-clusters Error (0.005) are very low and about equal, and the Lack of homogeneity (0.1% of the total dispersion) is so small that the number of clusters is certainly not too low.

Figure 2 displays the implicit object configuration for the Iris data, labelled by varieties, and connected with small solid lines to their centroids (the cluster points). At the right we see 9 clusters containing the 50 *setosa (a)* cases, with an average of 5.5 objects per cluster, and well separated from the rest. At the left there is a band of 7 clusters with an average of 6 objects per cluster, all of which consist entirely of *virginica (c)*, and in the middle a band of 7 clusters that consist purely of 44 *versicolor (b)* cases. In between these two bands, we see two mixed clusters, one with 7 *virginica (c)* and 3 *versicolor (b)*, and the other, at the lower part of the plot, containing three *b*'s and one *c*. The implicit object configuration enables us to assess the shape of the clusters (slightly elongated at the left upper part and the lower part of the plot), and shows that the individual irises within the mixed clusters are located at the side of their own variety. On the whole, the solution seems to agree with the conjecture that *Iris versicolor* is a *setosa-virginica* hybrid, and even with Fisher's assertion, cited in Bezdek (1981, p.106), that *Iris versicolor* is "twice as similar" to *Iris virginica* as it is to *Iris setosa*.

Fuzzy Version of Cluster Differences Scaling and Its Use in Avoiding Local Minima by Successive Approximation

One of the advantages of the concept of an indicator matrix, not shared by a formulation using elements of index classes, is that a natural transition to *fuzzy clustering* (Zadeh, 1977) is possible. The rows of the indicator matrix **E** can be viewed as a set of n binary *indicator functions* $e_i(\cdot)$ defined

on the set of clusters $J = \{J_1, \ldots, J_k, \ldots, J_K\}$, with values $\mathbf{e}_i = \{e_{ik} = e_i(J_k)$ for $k = 1, \ldots, K\}$, each of which assigns to some object i a membership of one and only one class. In fuzzy clustering, the objective is to assign *grades of membership* with respect to *several* classes. This extension is expressed by regarding *membership functions*, in the notation here denoted as $\mathbf{f}_i = \{ f_{ik} = f_i(S_k)$ for $k = 1, \ldots, K\}$, collected in the $n \times K$ matrix $\mathbf{F}$, for which we require $0 \leq f_{ik} \leq 1$; we say that f_{ik} is the *(grade of) membership* of object i in *fuzzy subset (cluster)* S_k.

The fuzzy clusters $\{S_1, \ldots, S_k, \ldots, S_K\}$ form a *fuzzy K-partition* of the set of objects if and only if $f_i(S_1) + \ldots + f_i(S_k) + \ldots + f_i(S_K) = 1$ for all i. Note that fuzzy clusters are just a set of K mathematical objects; the only thing that makes them special is the rule by which individual objects are assigned to them, which has its range in the interval [0,1], and is normalized to have a sum of one. Nor does fuzziness force us to use a different *representation* of clusters; we can still represent fuzzy clusters as centers of gravity. However, a fuzzy K-partition can no longer be listed as a set of subsets, as can be done with an ordinary partition (called *hard partition* in this context); a fuzzy K-partition is equal to some matrix $\mathbf{F}$ satisfying the above conditions.

In the context of cluster differences scaling, the introduction of a fuzzy K-partition implies a different treatment of the residuals $\{\delta_{ij} - d_{kl}(\mathbf{X})\}$. While in ordinary cluster differences scaling with indicator matrix $\mathbf{E}$ we would take into consideration, for each pair of objects (i,j), the single residual selected by the non-zero element of the $K \times K$ matrix $\mathbf{e}_i\mathbf{e}_j' = \{e_{ik}e_{jl}\}$, when using the membership matrix $\mathbf{F}$ we consider *all* residuals, but weighted with the $K \times K$ matrix $\mathbf{f}_i\mathbf{f}_j' = \{f_{ik}f_{jl}\}$, which satifies $\sum_k \sum_l f_{ik}f_{jl} = \sum_k f_{ik} \sum_l f_{jl} = 1$. Thus the mass used in weighting the residuals is spread out over a whole matrix, rather than being concentrated in a single cell.

A well-known property of fuzzy clustering in a sums of squared residuals framework (Bezdek and Dunn, 1975), called *fuzzy c-means*, is the fact that just replacing $\mathbf{E}$ with $\mathbf{F}$ does not yield a fuzzy solution at all, but again a hard partition, despite the relaxation of the constraints (Bezdek, 1981, p. 70). The optimum is always found at one of the corners of the feasible region, a difficulty that is usually resolved by involution of the membership function with an exponent. Following this practice, the loss function for fuzzy cluster differences scaling becomes, for any $1 \leq q < \infty$ fixed in advance,

$$\sigma^2(\mathbf{F},\mathbf{X}) \;= \sum_{k\le l} \sum_i \sum_j f_{ik}^q f_{jl}^q w_{ij} (\delta_{ij} - d_{kl}(\mathbf{X}))^2 \;. \tag{18}$$

Equivalence of (18) with some form of constrained multidimensional scaling does not hold here, since the proof for the hard partition case rested critically upon the matrix $\mathbf{e}_i\mathbf{e}_j'$ having a single nonzero cell. An orthogonal decomposition of the residuals – as in (3) – does hold, however, when $\tilde{w}_{kl}$ and $\tilde{\delta}_{kl}$ defined in (4) and (5) are replaced by

$$\tilde{\tilde{w}}_{kl} = \sum_i \sum_j f_{ik}^q f_{jl}^q w_{ij}, \tag{19}$$

$$\tilde{\tilde{\delta}}_{kl} = \sum_i \sum_j f_{ik}^q f_{jl}^q w_{ij} \delta_{ij} \;/\; \tilde{\tilde{w}}_{kl}. \tag{20}$$

Consequently, the lack-of-fit function remains the same after insertion of these new quantities, and so the MDS phase for minimizing (18) is identical to the process used earlier. For a given spatial representation with fuzzy cluster points $\underline{\mathbf{X}}$, we may again proceed row after row in the allocation phase, since as before the quantity $\gamma_{ik}^2(\underline{\mathbf{X}})$, properly redefined for this case, does not depend on the ith row of $\mathbf{F}$. Analogous to (11), we have to solve

$$\min_{\mathbf{f}_i} \;\; \sum_k f_{ik}^q \gamma_{ik}^2(\underline{\mathbf{X}}) \tag{21}$$

over $0 \le f_{ik} \le 1$, with $\sum_k f_{ik} = 1$. The unique solution of subproblem (21), which involves a convex objective function, because it consists of a nonnegative combination of convex components, is obtained for (Bezdek, 1981, Theorem 11.1)

$$\tilde{\tilde{f}}_{ik} = 1 \Big/ \Big[\sum_l \Big(\frac{\gamma_{ik}(\underline{\mathbf{X}})}{\gamma_{il}(\underline{\mathbf{X}})} \Big)^{2/(q-1)} \Big], \tag{22}$$

showing that the scale of the residuals has no influence. In view of the behavior of ordinary fuzzy c-means, it is expected that for large q the fuzzy weighting will reach a maximum degree of fuzziness, with $f_{ik} = 1 / K$ for each pair of i and k. However, it should be noted that fuzzy cluster differences scaling poses the additional requirement that the quantities defined in (20) have to be embeddable in p-dimensional space. When all weights (19) and all fuzzy Sokal-Michener dissimilarities (20) become more alike, perfect embedding is only possible in high-dimensional space, in the limit only by

choosing $p = K - 1$: the cluster points have to be placed at the corners of a regular polyhedron with all sides equal. Total Stress will be maximal for this case. When q approaches 1, only the $\tilde{\tilde{f}}_{ik}$ corresponding to the smallest $\gamma_{ik}(\underline{\mathbf{X}})$ survives in (22), so that it becomes equivalent to the necessary condition for problem (11), allocation to one cluster with minimal distance (Dunn, 1974).

No theoretically sound basis for an optimal choice of the fuzziness exponent q has emerged to date, although some interesting heuristics have been proposed, e.g., minimizing the pooled Within-clusters standard deviation of the memberships, calculated across all $\tilde{\tilde{f}}_{ik}$ above $1 / K$ (Wedel, 1990). Moreover, it does not seem attractive to use definition (20) with membership values substantially deviating from zero and one, in view of the fact that a low-dimensional Euclidean embedding needs a large variance of the dissimilarities (Shepard, 1962), while (20) is variance reducing. But the approach has a major practical advantage: it seems to work well for the whole range of q, except for $q = 1.0$ (and a small neighborhood), where local minima abound. This property suggests a *successive approximation* procedure for hard cluster differences scaling, in which we exploit the fact that in the limit the two approaches become equivalent. Starting with a rather large q (e.g., $q = 3.0$, which makes (22) equal to one over a sum of ratios of distances), we compute a series of solutions for fuzzy cluster differences scaling with a gradually decreasing value of the exponent, each time using the optimal $\mathbf{F}_*$ and $\mathbf{X}_*$ of the previous step as the initial fuzzy partition and cluster configuration of the next step. When the exponent approaches 1.0, $\mathbf{F}_*$ will approach a hard partition $\mathbf{E}_*$, and the corresponding allocation will define our final solution.

A Monte Carlo experiment was performed to test the successive approximation strategy (called *Fuzzy steps*, for short) against two other strategies. The first competing strategy (called *MDS/K-means*) is to use an initial cluster configuration and cluster allocation obtained by ordinary K-means clustering of an individual object configuration from a multidimensional scaling analysis without cluster constraints. The second alternative strategy (called *Multistart-10*) is to start 10 times with a random cluster configuration and a random cluster allocation, and to keep the best result. These three strategies were compared in conditions where the number of objects was chosen as either 20 or 40, and the number of clusters as either 5 or 10. For each of the $2 \times 2 \times 3$ combinations in this design, 25

replications were created by repeatedly drawing random dissimilarities from a uniform distribution.

--- Insert Table 4 about here ---

All analyses were performed in two dimensions; q was varied from 3.0 to 1.0, in steps of 1, 1/2, 1/4, 1/8, 1/16, 1/32, and twice 1/64. Table 4 gives the *number of successes* (number of times a strategy yielded the best Stress value for the subsequent CDS), the *mean raw Stress* across all replications, and, in brackets, the *distance of the mean to the minimal Stress* obtained in each condition. Strategy *Fuzzy steps* yielded the lowest raw Stress value in 92% of the cases, *MDS/K-means* in 5 % of the cases, and *Multistart-10* in 3% of the cases. Table 4 also shows that in all 50 cases with 10 clusters, *Fuzzy steps* performed perfectly, while *MDS/K-means* and *Multistart-10* never attained the lowest Stress. It outperformed the other two strategies in the $K = 5$ case too, not only in terms of successes, but also in terms of the distance toward the lowest Stress when it failed (distances of 0.03 and 0.10, for Stress values of 17.80 and 91.73, respectively). Strategies *MDS/K-means* and *Multistart-10* did not seem to differ much from each other.

We may conclude that the successive approximation strategy using complete fuzzy clustering steps with a geometrically decreasing series of q values is far superior over the alternative initialization strategies. Although it is reassuring to know that the local minimium problem seems to be under control, it should be noted that *Fuzzy steps* involves extensive computations, especially for large n and large K. However, the extra effort will almost always be worthwhile.

Discussion

Three major structural representations for clusters are commonly used in cluster analysis: (1) the *partition*, associated with K cluster points in $K - 1$ dimensions; (2) the *hierarchical tree*, in which there are $n - 1$ cluster points forming the internal nodes of an additive tree; (3) the *covering*, i.e., a set of K overlapping clusters associated with some specific graph-theoretical structure (*cf.* Gordon, 1981). Except when K is small, they all tend to be high-dimensional.

The contribution of this paper is in the further development of another type of representation: the

spatial embedding of clusters, which can be low-dimensional even if K is large. Heiser (1993) coined the name cluster differences scaling for methods that fit such spatial models. When a least squares loss function is chosen and the residuals within clusters are included, cluster differences scaling is equivalent to multidimensional scaling with cluster constraints on the configuration, as suggested in Groenen (1993). The possibility of constraining multidimensional scaling as $\mathbf{Z} = \mathbf{EX}$, with $\mathbf{E}$ an unknown indicator matrix, was indicated in De Leeuw and Heiser (1980). It was more thoroughly studied in Bock (1987), who gave an alternating least squares algorithm for fitting the associated scalar products (rather than the dissimilarities themselves), and introduced a penalty term of the form $\text{tr}\ \mathbf{Z}'[\mathbf{I} - \mathbf{E}(\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}']\mathbf{Z}$ to control the extent to which $\mathbf{Z}$ satisfies the constraints. Ruspini (1970) pioneered multidimensional scaling with fuzzy clustering constraints on the configuration; he required $\mathbf{Z} = \alpha\mathbf{F}$ with α some scalar and $\mathbf{F}$ an unknown grade-of-membership matrix, and used a least squares loss function defined on the squared Euclidean distances (also see Bezdek, 1981, section S9).

It was shown in this paper that the least squares approach allows a decomposition of the sum of squared dissimilarities into contributions from several sources of variation, collected in an Analysis of Dispersion table that is the multidimensional analogue of an Analysis of Variance table for a one-way classification. The Between component is more refined than the usual one in an analysis of variance. This difference is due to the fact that the Sokal-Michener dissimilarities are embedded in lower dimensionality; the refinement is comparable to the breakdown of the Residual SSQ into a Pure Error and Lack of fit, when the effects of the one-way classification are restricted to be located on a line or on some polynomial (Draper & Smith, 1966). The split of the Total SSQ into a Between SSQ and an Error SSQ seems similar to the one discussed in Gower (1989), who used the terms "Between Groups" and "Within Groups", respectively. However, the two decompositions cannot be the same, as Gower appears to be able to further decompose his Within-Groups component into dimension-wise contributions. It is not clear how these were obtained and with how many degrees of freedom they are associated.

Cluster differences scaling is a *parsimonious* method, because the number of parameters fitted (the degrees of freedom for the Among-clusters D.A.F.) depends on K, the number of classes, not on n,

the number of objects, and it reduces the dimensionality of the class representation to $p \leq K - 1$, so that the number of classes may be relatively large. Both the journal co-citation example and the Iris example showed excellent D.A.F.'s in two dimensions with K as large as 8 and 25.

Due to its optimality properties, the implicit object configuration proposed in this paper seems to be an attractive alternative to an earlier suggestion (Heiser, 1993) for the *a posteriori* display of the original objects, which was to add supplementary points to the fitted space on the basis of object-to-cluster dissimilarities defined as $\tilde{\delta}_{ik} = \sum_{j \in J_k} w_{ij}\delta_{ij} \; / \sum_{j \in J_k} w_{ij}$, analogous to (5). For it turns out that aggregation at this level often reduces the variability of the $\{\tilde{\delta}_{ik}\}$ considerably, an effect of regression to the mean. Reduced variability of the object-to-cluster dissimilarities implies that the individual objects are to be located at about equal distance from the cluster points, a requirement that is often hard to realize in low dimensionality. By contrast, the implicit object configuration has a clear interpretation in low-dimensional space, and needs no further optimization.

Viewed from the point of view of a clustering technique, the present method can be characterized as K-means with restrictions, because it involves means that are restricted to be Euclidean distances. But note that in cluster differences scaling, we must distinguish between the existence of K clusters – represented as K cluster points – and $K(K - 1)/2$ means, which are not the cluster centers, but average dissimilarities. The complications caused by the double occurrence in the loss function of the quantities $\{e_{ik}\}$, and by their binary character, seem to have been circumvented satisfactorily by starting the process in a well-chosen point, using a regular succession of fuzzy approximations. In the fuzzy version, we have used n membership functions defined on the set of clusters, while it is more common to define K functions on the set of objects (Zadeh, 1977). However, the requirement that the memberships sum to one is more naturally defined in terms of a normalized function, similar to a density function.

The presence of weights in the loss function provides a flexibility that has not been systematically explored yet. In combining the cluster-wise loss components (1), variations of (2) easily come to mind, e.g. one that divides by the product of the number of objects in clusters k and l, but these should be studied more closely, as they pose new theoretical and computational challenges.

REFERENCES

Anderson, E. (1935). The Irises of the Gaspe peninsula. *Bulletin of the American Iris Society, 59*, 2-5.

Ball, G.H. & Hall, D.J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science, 12*, 153-155.

Banfield, C.F. & Bassill, L.C. (1977). Algorithm AS113. A transfer algorithm for non-hierarchical classification. *Applied Statistics, 26*, 206-210.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum.

Bezdek, J.C. & Dunn, J.C. (1975). Optimal fuzzy partitions: a heuristic for estimating the parameters in a mixture of normal distributions. *IEEE Transactions on Computers, C-24*, 835-838.

Bock, H.-H. (1987). On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In H. Bozdogan & A.K. Gupta (Eds.), *Multivariate Statistical Modeling and Data Analysis*, pp. 17-34. New York: Reidel.

De Leeuw, J. & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol. V*, pp. 501-522. Amsterdam: North-Holland.

Draper, N.R. & Smith, H. (1966). *Applied Regression Analysis*. New York: Wiley.

Dunn, J.C. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *Journal of Cybernetics, 3*, 32-57.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Gordon, A.D. (1981). *Classification: Methods for the Exploratory Analysis of Multivariate Data*. London: Chapman and Hall.

Gordon, A.D. & Henderson, J.T. (1977). An algorithm for Euclidean sum of squares classification. *Biometrics, 33*, 355-362.

Gower, J.C. (1989). Generalised canonical analysis. In R. Coppi & S. Bolasco (Eds.), *Multiway*

Data Analysis, pp. 221-232. Amsterdam: North-Holland.

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

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika, 33*, 469-506.

Hartigan, J.A. & Wong, M.A. (1979). Algorithm AS 136: A *K*-means clustering algorithm. *Applied Statistics, 28*, 100-108.

Heiser, W.J. (1993). Clustering in low-dimensional space. In O. Opitz, B. Lausen & R. Klar (Eds.), *Information and Classification: Concepts, Methods and Applications*, pp. 162-173. Heidelberg: Springer Verlag.

Heiser, W.J. & Groenen, P.J.F. (1993, June). *Stress decomposition and use of fuzzy memberships in cluster differences scaling*. Paper presented at the annual meeting of the Psychometric Society, Berkeley, California.

Kernighan, B.W. & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal, 49*, 291-307.

Kruskal, J.B. (1977). The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and Clustering*, pp. 17-44. New York: Academic Press.

Ruspini, E. (1970). Numerical methods for fuzy clustering. *Information Science, 2*, 319-350.

Selim, S.Z. & Ismail, M.A. (1984). *K*-Means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*, 81-87.

Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function I & II. *Psychometrika, 27*, 125-140 & 219-246.

Sokal, R.R., & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409-1438.

Späth, H. (1985). Cluster Dissection and Analysis. Chichester: Ellis Horwood.

Tijssen, R. (1992). Cartography of Science: Scientometric Mapping with Multidimensional scaling methods. Doctoral Dissertation. Leiden: DSWO Press.

Tobler, W. (1976). Spatial interaction patterns. *Journal of Environmental systems, 6*, 271-301.

Wedel, M. (1990). *Clusterwise Regression and Market Segmentation: Developments and Applications*. Unpublished doctoral dissertation, University of Wageningen.

Zadeh, L.A. (1977). Fuzzy sets and their application to pattern classification and clustering analysis. In J. Van Ryzin (Ed.), *Classification and Clustering*, pp. 251-299. New York: Academic Press.

NOTE

The co-citation data were kindly made available by dr. Robert Tijssen of the Centre for Science and Technology Studies (CWTS), Leiden, The Netherlands.

TABLE 1

Analysis of Dispersion for the Cluster Differences Model

Source	*SSQ*	*df*
Between	$\sum_{k\leq l} \tilde{w}_{kl}\tilde{\delta}^2_{kl}$	$1/2\ K(K+1)$
Lack of homogeneity	$\sum_{k} \tilde{w}_{kk}\tilde{\delta}^2_{kk}$	K
Lack of spatial fit	$\sum_{k<l} \tilde{w}_{kl}(\tilde{\delta}_{kl} - d_{kl}(\tilde{\mathbf{X}}))^2$	$(K-1)(K/2-p) + p(p-1)/2$
Among-clusters D.A.F.	$\sum_{k<l} \tilde{w}_{kl}\tilde{d}^2_{kl}(\tilde{\mathbf{X}})$	$Kp - p(p+1)/2$
Error SSQ	$\sum_{k\leq l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kl})^2$	$[n(n-1) - K(K+1)]/2$
Among-clusters Error	$\sum_{k<l}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kl})^2$	$\sum_{k<l}(n_k n_l - 1)$
Within-clusters Error	$\sum_{k}\sum_{i\in J_k}\sum_{j\in J_l} w_{ij}(\delta_{ij} - \tilde{\delta}_{kk})^2$	$\sum_{k}(n_k(n_k-1)/2 - 1)$
Total	$\sum_{i<j} w_{ij}\delta^2_{ij}$	$n(n-1)/2$

TABLE 2

Analysis of Dispersion for the Journal Co-citation Example

Source	*SSQ*	*%*	*df*	*MS*
Between	2077.01	81.3	36	57.695
Lack of homogeneity	71.96	2.8	8	8.995
Lack of spatial fit	72.07	2.8	15	4.805
Among-clusters D.A.F.	1932.97	75.6	13	148.690
Error SSQ	478.99	18.7	2520	0.190
Among-clusters Error	464.88	18.2	2210	0.210
Within-clusters Error	14.11	0.5	310	0.045
Total	2556.00	100.0	2556	

TABLE 3

Analysis of Dispersion for the CDS Analysis of the Iris Data

Source	*SSQ*	*%*	*df*	*MS*
Between	11145.21	99.7	325	34.293
Lack of homogeneity	10.32	0.1	25	0.413
Lack of spatial fit	6.87	0.1	253	0.027
Among-clusters D.A.F.	11128.02	99.6	47	236.766
Error SSQ	29.79	0.3	10850	0.003
Among-clusters Error	28.02	0.3	10466	0.003
Within-clusters Error	1.77	0.0	384	0.005
Total	11175.00	100.0	11175	

TABLE 4

Results of Monte Carlo Experiment on Avoidance of Local Minima

		number of successes			*mean Stress (dist. to min.)*		
clusters	*objects*	Fuzzy steps	MDS/ *K*-means	Multi-start-10	Fuzzy steps	MDS/ *K*-means	Multi-start-10
$K = 5$	$n = 20$	21	2	2	17.80 (0.03)	18.99 (1.22)	18.73 (0.96)
	$n = 40$	21	3	1	91.73 (0.10)	94.67 (3.07)	94.95 (3.33)
$K = 10$	$n = 20$	25	0	0	13.46 (0.00)	15.06 (1.59)	14.37 (0.91)
	$n = 40$	25	0	0	74.49 (0.00)	78.15 (3.66)	78.18 (3.69)
Total		92	5	3			

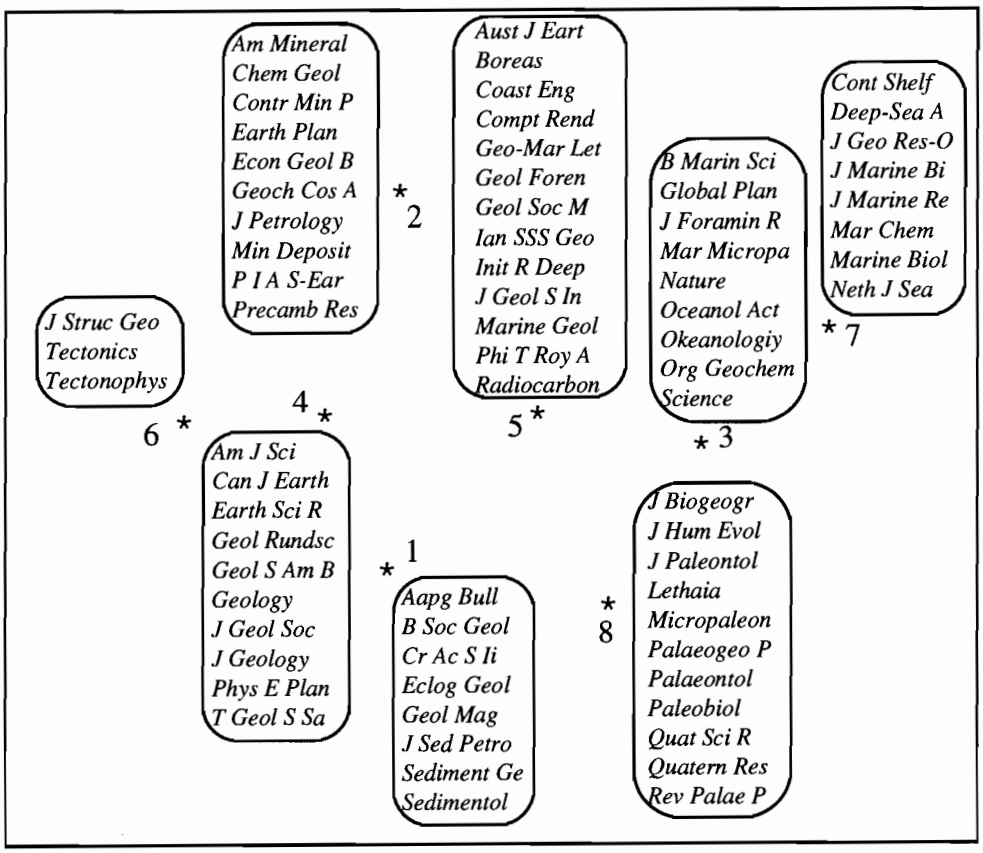

Figure 1. Cluster configuration for the journal co-citation data

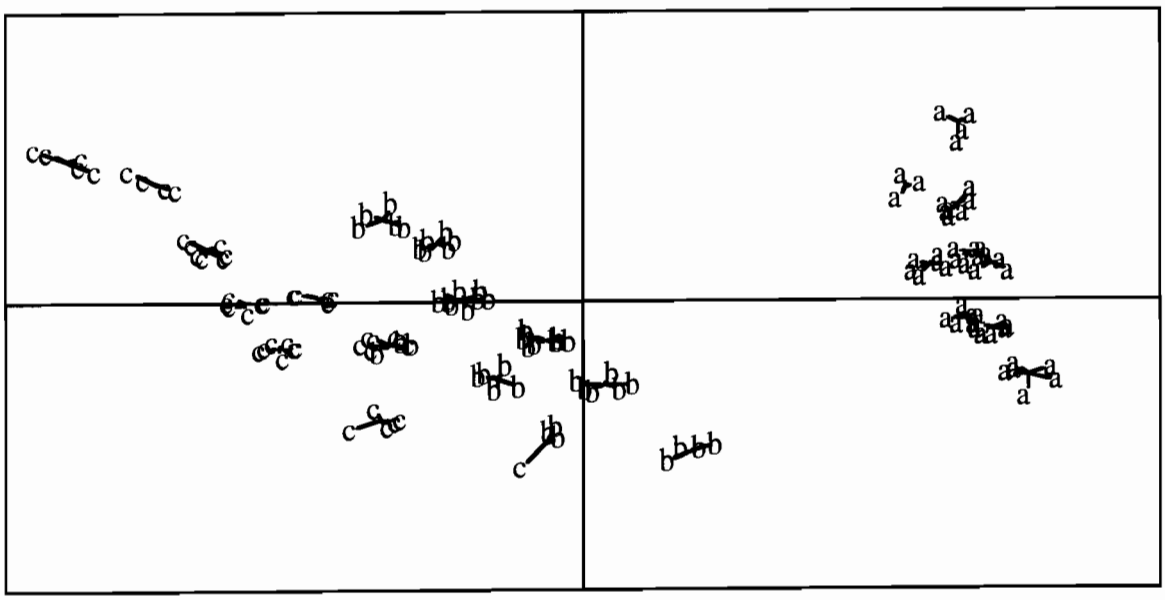

Figure 2. Implicit object configuration for the Iris data, labeled by varieties (a: *Iris setosa*; b: *Iris versocolor*; c: *Iris virginica*) and connected to their cluster centroids