DISTANCE ANALYSIS OF LARGE DATA SETS OF CATEGORICAL VARIABLES USING OBJECT WEIGHTS

Patrick J.F. Groenen
Jacques J.F. Commandeur
Jacqueline J. Meulman

Department of Data Theory
Leiden University

RR-96-01

DISTANCE ANALYSIS OF LARGE DATA SETS OF CATEGORICAL VARIABLES USING OBJECT WEIGHTS

Patrick J.F. Groenen*
Jacques J.F. Commandeur
Jacqueline J. Meulman

Department of Data Theory
Leiden University

This research was supported by The Netherlands Organization for Scientific Research (NWO) by grant nr. 575-67-053 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis'.

* Requests for reprints should be sent to Patrick J.F. Groenen, Department of Data Theory, Faculty of Social and Behavioral Sciences, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

Abstract

Categorical variables are often analyzed by multiple correspondence analysis (or homogeneity analysis), which places great emphasis on graphical representation. A drawback of this method is that regularly only a minor aspect of the data is displayed due to outliers, or, if a dominant first dimension exists, the horse-shoe effect occurs. Here, we elaborate on a competing approach to multiple correspondence analysis based on distance approximation. This method emphasizes the distance between objects; they are graphically displayed as points, where objects close together are considered more similar than objects farther apart. A limiting factor of this method is that the number of objects cannot be very large (say, no more than 500). We show how the majorization algorithm for distance approximation can be extended using frequency counts as object weights such that much larger data sets can be analyzed without a significant amount of additional computational effort. A second advantage of the use of object weights is that resampling methods, such as the bootstrap, are easily implemented. We present two illustrative examples, and investigate the stability in one through the bootstrap .

Keywords: object weights, multiple correspondence analysis, homogeneity analysis, multidimensional scaling, categorical data, distance approach, bootstrap, majorization.

1. Introduction

In social science research, data are often categorical in nature. Typically, objects are evaluated on a number of characteristics, each characteristic being separable into mutually exclusive categories. (The neutral term object refers to the entities on which the characteristics, or variables, are measured, such as respondents to a questionnaire, students on an achievement test, countries in a comparative economics study, stimuli in a physiological experiment, and so on.) For example, in an intelligence test the respondent either fails or a passes a test item, or in marketing research, the consumer belongs to one of several socio-economic classes. A popular graphical method for the analysis of categorical data is *multiple correspondence analysis* (MCA), also known as homogeneity analysis (see, e.g., Benzécri, 1973; Greenacre, 1984; Gifi, 1990; Nishisato, 1994). This method searches for groups of objects that are homogeneous in their response to the variables. MCA produces plots in which both objects and categories are represented as points in a low-dimensional space. Similar objects are represented close to each other in the object space, and objects that have different scores on the variables are represented farther apart. Also, categories being at close distance reveal that they have particular objects in common. Some drawbacks of MCA are that the technique only reconstructs a small part of the data (Greenacre, 1988), is sensitive to outliers, and may yield solutions that display objects and categories in two dimensions in a horse-shoe shaped form, which is also known as the Guttman effect. In such a solution, the second dimension is a quadratic function of the first dimension, and does not contribute substantially to the representation of objects and categories.

A different approach was pursued by Meulman (1986, 1992) who emphasized distance approximation between the objects. Unlike multiple correspondence analysis, the distance approach to multivariate analysis seeks to estimate distances between objects directly. Objects are assigned to points in the object space (by *object scores*) such that the Euclidean distance between object points approximates the difference in the variable scores of the objects as well as possible. This approach is modeled by a loss function closely related to the STRESS loss function used in multidimensional scaling (Kruskal, 1964a, 1964b).

It often happens in categorical data sets with a large number of objects that certain objects have the same combination of category scores, the same *profile*. Without loss of generality, such objects can be represented by their profile together with the frequency with which they occur in the data set. This yields a smaller data set than one where each object appears separately. Here, we exploit this reduction in size of the data set by incorporating *profile weights* in the loss function; these weights are equal to the (relative) frequency of occurrence of a profile. For example, if three objects score exactly the same on all variables, then only one profile is included in the analysis, together with an profile weight of 3. Of course, in general we may define an object weight that does not necessarily have to be a frequency count, but may be any positive value that indicates the importance of the object. In the present context, however, we will restrict ourselves, without loss of generality, to weights that represent profile counts. Until now the distance approach to MCA could handle up to about 500 objects; the use of profile weights allows for an almost limitless number of objects, as long as the maximum number of different profiles does not exceed about 500.

The minimization of the loss function for distance based MCA has to be done with an iterative method. Because the number of operations in the algorithm is of the order n^2 per iteration (with n the number of profiles) the reduction in problem size by using weights is extremely worthwhile. However, with profile weights updating the object scores requires the inversion of a matrix of order $n \times n$, which is a computational effort of the order n^3. We present an efficient solution for this problem. We also illustrate through examples that the distance approach does not suffer from the drawbacks of MCA mentioned above. Finally, using profile weights, we indicate how a stability study can be carried out with the bootstrap method (Efron, 1979).

2. The loss function for distance-based multiple correspondence analysis

In the distance approach the dissimilarity of objects i and j ($i, j = 1, ..., n$) is modeled by their Euclidean distance. If object i is at large distance from object j, then object i is different from

object j; if they are close together, the objects are quite similar. The overall dissimilarity between objects will be an average over the contributions of each variable, where variable k ($k = 1, ..., K$) contributes to dissimilarity δ_{ijk} between the object pair i and j. If objects i and j are in the same category for variable k the contribution to the dissimilarity is zero; if they are in different categories, the contribution is non-zero. We return to the exact definition of δ_{ijk} below. In our loss function for *distance-based multiple correspondence analysis* (DB-MCA) the dissimilarities are approximated by the Euclidean distance $d_{ij}(\mathbf{X})$ between points representing objects i and j, and the loss function is written as

$$\sigma(\mathbf{X}) = \sum_k \sum_{i<j} (\delta_{ijk} - d_{ij}(\mathbf{X}))^2, \tag{1}$$

which has to be minimized over the $n \times p$ matrix of object scores $\mathbf{X}$ (with dimensionality p), where $d_{ij}(\mathbf{X})$ is the Euclidean distance $(\sum_{s=1}^{p} (x_{is} - x_{js})^2)^{1/2}$ between the scores for objects i and j, and δ_{ijk} is a dissimilarity measure between objects i and j for variable k.

For categorical data, δ_{ijk} is defined as follows. Let $\mathbf{G}_k$ be the indicator matrix of order $n \times L_k$ with L_k the number of categories of variable k. Matrix $\mathbf{G}_k$ consists of L_k columns of binary zero-one variables with element $g_{il} = 1$ if object i belongs to category l ($l = 1, ..., L_k$) and $g_{il} = 0$ otherwise. This implies that each row of $\mathbf{G}_k$ contains a single 1. The diagonal matrix $\mathbf{M}_k = \mathbf{G}_k'\mathbf{G}_k$ contains the marginal frequencies m_{kl} of variable k on the diagonal. Now, we define δ_{ijk} as $n^{1/2} d_{ij}(\mathbf{G}_k \mathbf{M}_k^{-1/2})$, that is, $n^{1/2}$ times the Euclidean distance between objects i and j of the coordinate matrix $\mathbf{G}_k \mathbf{M}_k^{-1/2}$ in L_k dimensions. Thus, every column of $\mathbf{G}_k$ is divided by the square root of the frequency of the corresponding category, so that the matrix $\mathbf{G}_k \mathbf{M}_k^{-1/2}$ is orthonormal, i.e., $\mathbf{M}_k^{-1/2} \mathbf{G}_k' \mathbf{G}_k \mathbf{M}_k^{-1/2} = \mathbf{I}$. If two objects fall in the same category, $d_{ij}(\mathbf{G}_k \mathbf{M}_k^{-1/2}) = 0$; if they are in different categories, then $d_{ij}(\mathbf{G}_k \mathbf{M}_k^{-1/2}) = (m_{kr}^{-1} + m_{ks}^{-1})^{1/2}$, where r and s are two different categories. To correct δ_{ijk} for the marginal frequencies, we use $\mathbf{G}_k \mathbf{M}_k^{-1/2}$ instead of $\mathbf{G}_k$. Then, δ_{ijk} is a so-called χ^2-distance (see, e.g., Meulman, 1992). Let C_{kr} be the set of objects belonging to category r of variable k. Then, we have that

$$\delta_{ijk} = \begin{cases} 0 & \text{if } i, j \in C_{kr}, \\ n^{1/2}(m_{kr}^{-1} + m_{ks}^{-1})^{1/2} & \text{if } i \in C_{kr},\ j \in C_{ks}, \text{ and } s \neq r. \end{cases} \tag{2}$$

Extending $\sigma(\mathbf{X})$ to include profile weights $w_i > 0$ (if an profile weight is zero, the profile can be removed entirely from the analysis), (1) becomes

$$\begin{aligned}\sigma(\mathbf{X}) &= \sum_k \sum_{i<j} w_i w_j (\delta_{ijk} - d_{ij}(\mathbf{X}))^2 \\ &= \sum_k \sum_{i<j} w_i w_j \delta_{ijk}^2 + K \sum_{i<j} w_i w_j d_{ij}^2(\mathbf{X}) - 2 \sum_{i<j} w_i w_j d_{ij}(\mathbf{X}) \sum_k \delta_{ijk} \\ &= \eta_\delta^2 + K\eta^2(\mathbf{X}) - 2\rho(\mathbf{X}). \qquad (3)\end{aligned}$$

Clearly, if all weights are unity then (3) reduces to (1). For unit weights it can be shown that $\eta_\delta^2 = \sum_{i<j} \delta_{ijk}^2 = n^2(L_k - 1)$ for each variable k. For $w_i \neq 1$ we let $\mathbf{M}_k = \mathbf{G}_k'\text{Diag}(\mathbf{w})\mathbf{G}_k$, with Diag($\mathbf{w}$) a diagonal matrix containing the elements of vector $\mathbf{w} = \{w_i\}$ on the main diagonal. Then, using

$$\delta_{ijk} = \begin{cases} 0 & \text{if } i, j \in C_{kr}, \\ (\sum_i w_i)^{1/2}(m_{kr}^{-1} + m_{ks}^{-1})^{1/2} & \text{if } i \in C_{kr}, j \in C_{ks}, \text{ and } s \neq r, \end{cases} \qquad (4)$$

yields $\eta_\delta^2 = (\sum_i w_i)^2(\sum_k L_k - K)$. It can be shown that dividing $\sigma(\mathbf{X})$ by η_δ^2 at the point of convergence gives $0 \leq \sigma(\mathbf{X})/\eta_\delta^2 \leq 1$ (see, for example, Commandeur, 1992). This normalization is convenient, because it results in a value of the loss function which is independent of the number of objects, and of the value of the profile weights. Moreover, the minimum of $\sigma(\mathbf{X})$ is the same as the minimum of $\sigma(\mathbf{X})/\eta_\delta^2$, because η_δ^2 is a constant. In the sequel we will report $\sigma(\mathbf{X})/\eta_\delta^2$ as the loss.

In matrix form, $\sigma(\mathbf{X})$ can be written as

$$\sigma(\mathbf{X}) = \eta_\delta^2 + K \text{ tr } \mathbf{X}'\mathbf{V}\mathbf{X} - 2\sum_k \text{tr } \mathbf{X}'\mathbf{B}(\mathbf{X};\mathbf{\Delta}_k)\mathbf{X} \qquad (5)$$

where the matrix $\mathbf{B}(\mathbf{X};\mathbf{\Delta}_k)$ has off-diagonal elements $b_{ijk} = -w_i w_j \delta_{ijk}/d_{ij}(\mathbf{X})$ if $d_{ij}(\mathbf{X}) \neq 0$, $b_{ijk} = 0$ if $d_{ij}(\mathbf{X}) = 0$, and diagonal elements $b_{iik} = -\sum_{j \neq i} b_{ijk}$. The matrix $\mathbf{V}$ has off-diagonal elements $v_{ij} = -w_i w_j$, and diagonal elements $v_{ii} = -\sum_{j \neq i} v_{ij}$. Since distances do not change under translation, without loss of generality we impose the restriction that $\mathbf{X}$ has a weighted column mean of zero, or $\mathbf{w}'\mathbf{X} = \mathbf{0}$.

3. Updating the object scores

No solution exists that minimizes $\sigma(\mathbf{X})$ analytically (Kruskal, 1964a). Therefore, we have to revert to an iterative method to minimize $\sigma(\mathbf{X})$. Here, we incorporate the theory of majorization for multidimensional scaling (De Leeuw & Heiser, 1980; De Leeuw, 1988). According to majorization theory, $\sigma(\mathbf{X})$ is replaced in each iteration by a more simple function, the *majorizing* function $\mu(\mathbf{X};\mathbf{Y})$, where $\mathbf{Y}$ is the previous configuration in the series. The majorizing function has to satisfy two conditions. First, $\sigma(\mathbf{X}) \leq \mu(\mathbf{X};\mathbf{Y})$, that is, the majorizing function must always be greater than (or at most equal to) $\sigma(\mathbf{X})$; second, $\sigma(\mathbf{Y}) = \mu(\mathbf{Y};\mathbf{Y})$, that is, at $\mathbf{Y}$ the majorizing function must touch $\sigma(\mathbf{X})$. Let $\mathbf{X}^+$ be a configuration that minimizes $\mu(\mathbf{X};\mathbf{Y})$ over $\mathbf{X}$ for fixed $\mathbf{Y}$, so that $\mu(\mathbf{X}^+;\mathbf{Y}) \leq \mu(\mathbf{X};\mathbf{Y})$. Using the latter inequality and the two conditions above, we obtain the chain $\sigma(\mathbf{X}^+) \leq \mu(\mathbf{X}^+;\mathbf{Y}) \leq \mu(\mathbf{Y};\mathbf{Y}) = \sigma(\mathbf{Y})$, which proves that the majorizing algorithm never increases $\sigma(\mathbf{X})$. For more details on majorization in special contexts, we refer to Heiser (1995) and Groenen and Heiser (in press). For $\sigma(\mathbf{X})$, the majorizing function is written as

$$\mu(\mathbf{X};\mathbf{Y}) = \eta_\delta^2 + K \text{ tr } \mathbf{X}'\mathbf{V}\mathbf{X} - 2\sum_k \text{tr } \mathbf{X}'\mathbf{B}(\mathbf{Y};\mathbf{\Delta}_k)\mathbf{Y}. \tag{6}$$

Because $\mu(\mathbf{X};\mathbf{Y})$ is a quadratic function in $\mathbf{X}$, it can be minimized in one step by computing the so-called Guttman transform, i.e.,

$$\mathbf{X}^+ = K^{-1}\mathbf{V}^-\sum_k \mathbf{B}(\mathbf{Y};\mathbf{\Delta}_k)\mathbf{Y}, \tag{7}$$

where $\mathbf{V}^-$ is a generalized inverse of $\mathbf{V}$. One such inverse is the Moore-Penrose inverse which is defined as $\mathbf{V}^+ = (\mathbf{V} + \mathbf{1}\mathbf{1}')^{-1} - n^{-2}\mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ denotes an n vector with ones. However, the computation of the inverse of $\mathbf{V} + \mathbf{1}\mathbf{1}'$ is of the order n^3 operations. When n is large (say, larger than 200) the computation of this inverse becomes prohibitive. The main purpose of the present section is to show how the computation of $\mathbf{V}^+$ can be avoided, and replaced by a much more efficient computation that is linear in n.

An important observation is that in the present application matrix $\mathbf{V}$ has the simple structure

$$\mathbf{V} = \mathbf{D} - \mathbf{w}\mathbf{w}',$$

where $\mathbf{D}$ is a diagonal matrix containing $(\mathbf{1}'\mathbf{w})\mathbf{w}$ on the diagonal. Thus, if the weights sum to one (that is, if $\mathbf{1}'\mathbf{w} = 1$), then a 3×3 example of matrix $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} w_1 - w_1^2 & -w_1 w_2 & -w_1 w_3 \\ -w_1 w_2 & w_2 - w_2^2 & -w_2 w_3 \\ -w_1 w_3 & -w_2 w_3 & w_3 - w_3^2 \end{bmatrix}.$$

The simple structure of $\mathbf{V}$ allows us to write the quadratic part of (6) as

$$\text{tr } \mathbf{X}'\mathbf{V}\mathbf{X} = \text{tr } \mathbf{X}'(\mathbf{D} - \mathbf{w}\mathbf{w}')\mathbf{X} = \text{tr } \mathbf{X}'\mathbf{D}\mathbf{X} - \text{tr } \mathbf{X}'\mathbf{w}\mathbf{w}'\mathbf{X}. \tag{8}$$

The additional requirement that $\mathbf{X}$ has a weighted column mean of zero (i.e., that $\mathbf{w}'\mathbf{X} = \mathbf{0}$, see Section 2), implies that tr $\mathbf{X}'\mathbf{w}\mathbf{w}'\mathbf{X} = 0$. Inserting this result into (6) gives

$$\mu(\mathbf{X};\mathbf{Y}) = \eta_\delta^2 + K \text{ tr } \mathbf{X}'\mathbf{D}\mathbf{X} - 2\textstyle\sum_k \text{tr } \mathbf{X}'\mathbf{B}(\mathbf{Y};\Delta_k)\mathbf{Y}. \tag{9}$$

Setting the gradient of (9) with respect to $\mathbf{X}$ equal to $\mathbf{0}$, yields the update

$$\mathbf{X}^+ = K^{-1}\mathbf{D}^{-1}\textstyle\sum_k \mathbf{B}(\mathbf{Y};\Delta_k)\mathbf{Y}, \tag{10}$$

which amounts to first computing $\mathbf{Z} = K^{-1}\sum_k \mathbf{B}(\mathbf{Y};\Delta_k)\mathbf{Y}$, and then $\mathbf{D}^{-1}\mathbf{Z}$ (or, equivalently, $z_{ij}/(w_i\sum_i w_i)$). Finally, letting $\mathbf{P} = (\mathbf{I} - (\mathbf{1}'\mathbf{w})^{-1}\mathbf{1}\mathbf{w}')$, we have to compute $\mathbf{P}\mathbf{X}^+$ to guarantee that the update has a weighted column mean of zero. In this way, update (10) replaces the order n^3 inversion of $\mathbf{V} + \mathbf{1}\mathbf{1}'$ needed in (7) by the order n inversion of the diagonal matrix $\mathbf{D}$, which is indeed a major improvement in efficiency.

4. Decomposition of the loss

The total loss in $\sigma(\mathbf{X})$ is a rather crude measure which only indicates the overall goodness-of-fit. Here, we discuss two decompositions of the total loss. First, the loss can be decomposed into *loss of homogeneity* σ^H, measuring the sum of squared differences between

the δ_{ijk}'s and $\bar{\delta}_{ij} = K^{-1}\sum_k \delta_{ijk}$ (which is the average dissimilarity), and *proper loss* $\sigma^P(\mathbf{X})$, measuring the approximation of $\bar{\delta}_{ij}$ by the distances in the configuration $d_{ij}(\mathbf{X})$. That is,

$$\begin{aligned}\sigma(\mathbf{X}) &= \sum_k \sum_{i<j} w_i w_j (\delta_{ijk} - d_{ij}(\mathbf{X}))^2 \\ &= \sum_k \sum_{i<j} w_i w_j (\delta_{ijk} - \bar{\delta}_{ij})^2 + K \sum_{i<j} w_i w_j (\bar{\delta}_{ij} - d_{ij}(\mathbf{X}))^2 \\ &= \sigma^H + \sigma^P(\mathbf{X}). \end{aligned} \tag{11}$$

Loss of homogeneity indicates to what extent the variables are different from each other. If, for example, all variables are equal, then $\sigma^H = 0$. The proper loss shows the (in)ability of distances in the configuration to approximate $\bar{\delta}_{ij}$.

With the second partitioning of the loss in $\sigma(\mathbf{X})$, the relative contribution of each separate variable and object to the total loss can be identified. This decomposition is obtained by splitting the loss per object i for variable k, i.e.,

$$\sigma_{ik}(\mathbf{X}) = 1/2 \sum_{j \neq i} w_i w_j (\delta_{ijk} - d_{ij}(\mathbf{X}))^2. \tag{12}$$

Summing $\sigma_{ik}(\mathbf{X})$ over k gives the loss of object i, $\sigma_{i+}(\mathbf{X}) = \sum_k \sigma_{ik}(\mathbf{X})$, and summing $\sigma_{ik}(\mathbf{X})$ over i gives the loss for variable k, $\sigma_{+k}(\mathbf{X}) = \sum_i \sigma_{ik}(\mathbf{X})$. Note that $\sigma_{++}(\mathbf{X}) = \sum_i \sum_k \sigma_{ik}(\mathbf{X})$ is equal to the total loss $\sigma(\mathbf{X})$. A complete overview of this decomposition is given in Table 1. Finally, the loss in (12) can be further decomposed into loss of homogeneity σ^H_{ik}, and proper loss, σ^P_i, that is,

$$\begin{aligned}\sigma^H_{ik}(\mathbf{X}) &= 1/2 \sum_{j \neq i} w_i w_j (\delta_{ijk} - \bar{\delta}_{ij})^2 \\ \sigma^P_i(\mathbf{X}) &= 1/2 \sum_{j \neq i} w_i w_j (\bar{\delta}_{ij} - d_{ij}(\mathbf{X}))^2 \\ \sigma_{ik}(\mathbf{X}) &= \sigma^H_{ik}(\mathbf{X}) + \sigma^P_i(\mathbf{X}). \end{aligned} \tag{13}$$

The term $\sigma^P_i(\mathbf{X})$ does not contain a subscript k and hence is constant for each variable k.

INSERT TABLE 1 ABOUT HERE.

5. Example

To illustrate our method we analyzed data concerning the occupational status of fathers and sons in 3497 British families (Table 2). These data have been analyzed by many methods, e.g., Glass (1954), Goodman (1965, 1969), Haberman (1974), Bishop, Fienberg, and Holland (1975), and Gifi (1990), among others. An analysis of $n = 3497$ individuals is hardly feasible in the distance-based multivariate analysis, since this would involve the calculation of $n(n - 1)/2 = 6112756$ distances between all object pairs i, j at every iteration. Even on fast modern computers this would be a demanding task. However, there are only 47 different combinations of the categories of the two variables in this data set, and thus, 47 profiles. By attaching the frequencies of co-occurence to profiles (see Table 3), the full information is retained. Thus, these frequencies serve as profile weights for the profiles in the DB-MCA. Since there are 47 different profiles, we only need to consider the 1081 distances between all pairs of profiles, which is clearly an enormous reduction.

INSERT TABLE 2 ABOUT HERE.

INSERT TABLE 3 ABOUT HERE.

In Gifi (1990) these data were analyzed by correspondence analysis of the 7×7 contingency table, which gives the same results as MCA of the profile-frequency matrix, up to a normalization of the category points (the centroids of the object scores belonging to a category). The upper panel in Figure 1 shows the first two dimensions of the MCA solution, and the lower panel shows the solution obtained by DB-MCA. The category points are depicted as open circles and the objects scores as dots. In Gifi (1990) only the category points are displayed, and in the absence of object points, the solution is interpreted as a horse-shoe. The first dimension orders occupational status class from high (left) to low (right). The second dimension is regarded as a quadratic function of the first dimension, so that on the second dimension, the

extreme occupational classes ('prof', unsk', and 'semi') have high values, whereas the middle occupational classes have low values. However, we see in Figure 1 that the object points do not show the horse-shoe shape, which leads us to conclude that the data are not essentially one-dimensional. To be more precise, the profile representing professional fathers and professional sons is very different from the other profiles, and is positioned in the upper left corner. This profile attracts the subset of profiles with either fathers with a professional son or sons with a professional father. We conclude that the analysis of the profile frequency matrix shows that the result is not a horse-shoe representation, but a solution with a strong outlier in the profile that represents families with professional fathers and sons. The particular shape of the joint configuration of profile and category points would seem to deserve additional study, which is, however, beyond the scope of the present paper.

INSERT FIGURE 1 ABOUT HERE.

INSERT FIGURE 2 ABOUT HERE.

The distance-based MCA solution (the lower panel in Figure 1) distributes the objects and the category centroids much more evenly over the two-dimensional space than does the original MCA solution (upper panel). In Figure 2 we find that the weights are also spread more evenly in the DB-MCA solution (lower panel) than in the MCA solution (upper panel). The most frequent profiles are located near the center and the less frequent profiles more to the outside. The categories 'skill' for fathers and for sons, by far the largest categories, are located in the center of the solution. In contrast to the MCA solution, the DB-MCA solution does not display `outlier' profiles.

Apart from the scatter of the object points, how does the DB-MCA solution describe the data? Consider the distances between the category centroids in the lower panel of Figure 1. The category professional father (f-prof) is closest to professional sons (s-prof), indicating that the combination professional father–professional son has the highest frequency over all

combinations including a professional father. If the son of a professional father is not a professional himself, then it is most likely that he has a higher supervisory job (s-hsup), since it is located second closest to f-prof. In this way, for every father-category a rank-order of the son-categories can be made, predicting the occurence by distances between category points. The points for the occupations of sons closely located to skilled father are 's-skil', 's-hsup', and 's-lsup' indicating that for skilled fathers, sons are most likely in any of these three occupational categories. To compare the reconstructed rank-order of the category points in Figure 1 with the observed rank-order in the data, we use the Spearman rank-order correlation. For professional fathers the correlation is .714 in the DB-MCA solution, and .607 in the MCA analysis solution. These rank-order correlations were computed for all the father-categories and all the son-categories, giving an average rank-order correlation of .492 for the MCA solution and .592 for the DB-MCA solution. This indicates that the DB-MCA solution describes the data better. Due to the better distribution of the category points, the distances between the categories reveal more information in the DB-MCA solution than in the MCA solution. The latter solution merely shows that professional fathers have professional sons, but does not show easily in what other jobs their sons are employed.

A second example concerns all marriages in the Netherlands from 1850 to 1993 (with the exception of the period 1911–1936) reported by Van Poppel, Post, and Groenen (1995). Apart from the year in which the marriage took place, the age of the spouses was noted. From these data we can see how the preference for spouses in terms of age changes over time. In the recorded period, 7,149,650 couples were married. In our analysis, each marriage takes the place of a single object. The total number of distances between the pairs of marriages is 51×10^{12}, which is evidently too many to be considered. However, our data were recoded into five age groups for each of the spouses and six periods, yielding a total of $5 \times 5 \times 6 = 150$ different profiles. Using these profiles with the frequency as profile weight, only 11175 distances have to be considered. The profiles were analyzed by DB-MCA, and the configuration of category centroids is shown in Figure 3.

INSERT FIGURE 3 ABOUT HERE.

In Figure 3 several relations in the data can be found by inspecting the distances between the category points. For younger age groups (younger than 35), the male and female category points are located relatively near each other, indicating that marrying within the same age group ("homogeneous" marriages) happens more often than marrying among different age groups ("heterogeneous" marriages). However, for the older age groups, the male and female categories are located at some distance. This signifies that for younger age groups, homogeneous marriages are more likely, and that the older age groups tend to marry in different age groups. The somewhat older age groups tend to marry within younger age groups, such as males of 45+ marrying females of 35–44 and females of 45+ marrying males of 35-44.

The first dimension separates the age of the spouses (with young on the left, and old on the right). The second dimension shows the difference in marriage patterns between the periods. Considering triples of category points, we find that the period 1961–1980 can be characterized by marriages between young spouses (males and females younger than 25). Large distances between triples of categories indicate that the profile does not occur very often. However, considering completely heterogeneous marriages, we see that the situation where males older than 45 marry females younger than 25 happens particularly in the period 1881-1910, albeit with a low frequency. Females older than 45 tend to marry males of 30 to 34, especially in the period 1850-1880. Since the periods 1945-1960 and 1981-1993 are located close to the centroid of the male and female categories, and because the distance between the homogeneous couples is small, we conclude that these periods can be characterized by homogeneous marriages. By considering other triples of distances, many more relations in the data could be described.

6. Stability study

To study the stability of the DB-MCA solution, we applied the bootstrap (Efron, 1979). In an extensive Monte Carlo study, Markus (1994) concluded that the bootstrap is valid when applied to MCA. Without making parametric assumptions about distributions, the bootstrap assesses stability by repeatedly sampling, say N times, with replacement from the original data,

analyzing these N samples, and comparing the solutions. For an introduction to the bootstrap, we refer to Efron and Tibshirani (1993). Our method is defined on profiles, so the bootstrap concentrates on the profile points. For each profile, the bootstrap generates N points. This cloud is an empirical description of the distribution of the profile. If the position of a profile is stable, the cloud of points will show little dispersion. The bootstrap is particularly easy to implement in our approach, since it is only the profile weights (frequencies) that differ between the bootstrap samples.

We applied the bootstrap to the occupational mobility data; for each one of 100 bootstrap samples, 3497 profiles were drawn randomly with replacement. Then, using the profile frequencies in the bootstrap sample as profile weights, a DB-MCA solution was computed that yields the profile points for the bootstrap sample, and this was repeated a 100 times.

The configuration obtained by DB-MCA is unique up to a translation, rotation, and reflection, since distances are invariant under these transformations. Therefore, if we compare multiple configurations, as we do in the bootstrap study, this indeterminacy has to be removed. The scale of the configuration is not very important since it merely reflects the goodness-of-fit (De Leeuw and Heiser, 1980). Moreover, when comparing two configurations, it is the ratio of distances that determines the congruence, not the actual scale. Therefore, it is deemed appropriate to apply a weighted Procrustes analysis (Everitt & Gower, 1981) that optimally translates, rotates, and dilates a bootstrap configuration to the target configuration obtained for the original data. For every profile, the bootstrap produces 100 points; instead of displaying all 100 bootstrap points for every profile, we computed an ellipse that shows the position of the cloud, and contains 95% of the bootstrap sample points (Meulman & Heiser, 1983).

Figure 4 shows the 95% regions for every profile based on the 100 bootstrap samples from the occupational mobility data. All object scores of the original sample fall well within the 95% regions, showing that the DB-MCA solution is very stable. The ellipses of profiles with a large frequency are very small. For example, for the profile "55" (skilled father – skilled son, with a frequency of 714), the ellipse is so small that it cannot be distinguished from the point representing profile "55" in the original analysis. However, profile "21" (executive father – professional son, with a frequency of 16) has a very large ellipse. The bootstrap points for this

profile are located between point "21" and point "11", filling only the lower half of the ellipse. For profile "21", the method for generating the ellipses largely overestimates the instability of the profile.

INSERT FIGURE 4 ABOUT HERE.

For every bootstrap sample we computed the weighted centroid for father and son categories. This resulted in the 95% regions of the category centroids in Figure 5. Just as for the profile scores, we find that all category centroids of the original analysis are located within the 95% ellipses generated by the bootstrap. Moreover, there is no overlap between any of the category centroids.

INSERT FIGURE 5 ABOUT HERE.

7. Discussion

We have proposed a new method for the analysis of categorical variables based on distances between profiles. This method can be viewed as a special case of multidimensional scaling. By using profile weights, DB-MCA is extended to deal efficiently with large data sets, if the number of profiles is much smaller than the number of objects. In the examples shown it was demonstrated that DB-MCA solutions compare favorably to the popular method of MCA. In particular, we found that more aspects of the data are retrieved in the representation, and that distance-based MCA prevents horse-shoe solutions. In addition, a bootstrap study showed that the DB-MCA results were stable .

The DB-MCA solution can also be computed by standard multidimensional programs that allow for dissimilarity weights, such as KYST (Kruskal, Young, & Seery, 1977) and PROXSCAL (Heiser, 1988; Commandeur & Heiser, 1993). The input dissimilarity weights w_{ij} should be equal to $w_i w_j$ if $i \neq j$, and 0 if $i = j$, and the dissimilarities defined as in Section 2.

However, because our computational approach is tailor-made for DB-MCA it can be expected that these programs are (much) less efficient. Upon request, a FORTRAN program can be obtained from the authors in which the algorithm for distance-based MCA has been implemented.

In contrast to classical MCA, DB-MCA can be subject to local minima, just as any multidimensional scaling method that minimizes STRESS (Groenen & Heiser, in press). Further study would be needed to obtain an indication how severe the local minimum problem really is. The use of object weights in a more general context is currently studied by the present authors as well.

References

Benzécri, J.P. et al. (1973). *L'Analyse des données*. Paris: Dunod.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Commandeur, J.J.F. (1992). *Missing data in the distance approach to Principal Component Analysis* (Research Report No. RR-92-07). Leiden: Department of Data Theory.

Commandeur, J.J.F. & Heiser, W.J. (1993). Mathematical derivations in the proximity scaling PROXSCAL of symmetric data matrices. (Research Report No. RR-93-03). Leiden: Department of Data Theory.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, **5**, 163-180.

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.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 172–184.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Everitt, B.S. & Gower, J.C. (1981). Plotting the optimum positions of an array of cortical electrical phosphenes. In: V.D. Barnett (Ed.), *Interpreting Multivariate Data* (pp. 279-287). Chichester: Wiley.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Glass, D.V. (1954). *Social mobility in Britain*. Glencoe: Free Press.

Goodman, L.A. (1965). On the statistical analysis of social mobility tables. *American Journal of Sociology*, **70**, 564-585.

Goodman, L.A. (1969). On the measurement of social mobility: an index of status persistence. *American Sociological Review*, **34**, 832-850.

Greenacre, M.J. (1984). Theory and applications of correspondence analysis. New York: Academic Press.

Greenacre, M.J. (1988). Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika*, **75**, 457-467.

Groenen, P.J.F. & Heiser, W.J. (in press). The tunneling method for global optimization in multidimensional scaling. Accepted for publication in: *Psychometrika*.

Haberman, S.J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Heiser, W.J. (1988). PROXSCAL, multidimensional scaling of proximities. In: A. Di Ciaccio and G. Bove (Eds.), *International meeting on the analysis of multiway data matrices, Software guide* (pp. 77-81). Rome: C.N.R.

Heiser, W.J. (1995). Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In: W.J. Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis*. (pp. 51-89). Oxford: Oxford University Press.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-28.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, **29**, 115-129.

Kruskal, J.B., Young, F.W. & Seery, J.B. (1977). *How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding*. Murray Hill, NJ: AT&T Bell Laboratories.

Markus, M. Th. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis*. Leiden: DSWO Press, Leiden University.

Meulman, J.J. (1986). *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press, Leiden University.

Meulman, J.J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika*, **57**, 539-565.

Meulman, J.J. & Heiser, W.J. (1983). *The display of bootstrap solutions in multidimensional scaling*. Unpublished manuscript.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Van Poppel, F., Post, W. & Groenen, P.J.F. (1995). *Changing age preferences of spouses: The Netherlands 1811-1993.* Paper presented at the Twentieth Social Science History Association Meeting, Chicago.

Table 1. A decomposition of loss $\sigma(\mathbf{X})$ over objects and variables. Loss for object i due to variable k is denoted by σ_{ik}. Summation of loss over objects for variable k is denoted by σ_{+k}; summation of loss over variables for object i is denoted by σ_{i+}. σ_{++} denotes the total loss over objects and variables.

object	Loss per variable						object
1	σ_{11}	σ_{12}	...	σ_{1k}	...	σ_{1K}	σ_{1+}
2	σ_{21}	σ_{22}	...	σ_{2k}	...	σ_{2K}	σ_{2+}
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
i	σ_{i1}	σ_{i2}	...	σ_{ik}	...	σ_{iK}	σ_{i+}
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
n	σ_{n1}	σ_{n2}	...	σ_{nk}	...	σ_{nK}	σ_{n+}
	σ_{+1}	σ_{+2}	...	σ_{+k}	...	σ_{+K}	σ_{++}

Table 2. Occupational mobility of fathers and sons. (Source: Gifi, 1990.)

		Occupation son							
Occupation father		prof	exec	hsup	lsup	skil	semi	unsk	Total
1. prof:	professional and high administrative	50	19	26	8	18	6	2	129
2. exec:	managerial and executive	16	40	34	18	31	8	3	150
3. hsup:	higher supervisory	12	35	65	66	123	23	21	345
4. lsup:	lower supervisory	11	20	58	110	223	64	32	518
5. skil:	skilled manual and routine non-manual	14	36	114	185	714	258	189	1510
6. semi:	semi-skilled manual	0	6	19	40	179	143	71	458
7. unsk:	unskilled manual	0	3	14	32	141	91	106	387
	Total	103	159	330	459	1429	593	424	3497

Table 3. The data of Table 2 expressed as profiles and their weights.

	occupation				occupation		
profile	father	son	weight	profile	father	son	weight
1	1	1	50	25	4	4	110
2	1	2	190	26	4	5	223
3	1	3	26	27	4	6	64
4	1	4	8	28	4	7	32
5	1	5	18	29	5	1	14
6	1	6	6	30	5	2	36
7	1	7	2	31	5	3	114
8	2	1	16	32	5	4	185
9	2	2	40	33	5	5	714
10	2	3	34	34	5	6	258
11	2	4	18	35	5	7	189
12	2	5	31	36	6	2	6
13	2	6	8	37	6	3	19
14	2	7	3	38	6	4	40
15	3	1	12	39	6	5	179
16	3	2	35	40	6	6	143
17	3	3	65	41	6	7	71
18	3	4	66	42	7	2	3
19	3	5	123	43	7	3	14
20	3	6	23	44	7	4	32
21	3	7	21	45	7	5	141
22	4	1	11	46	7	6	91
23	4	2	20	47	7	7	106
24	4	3	58				

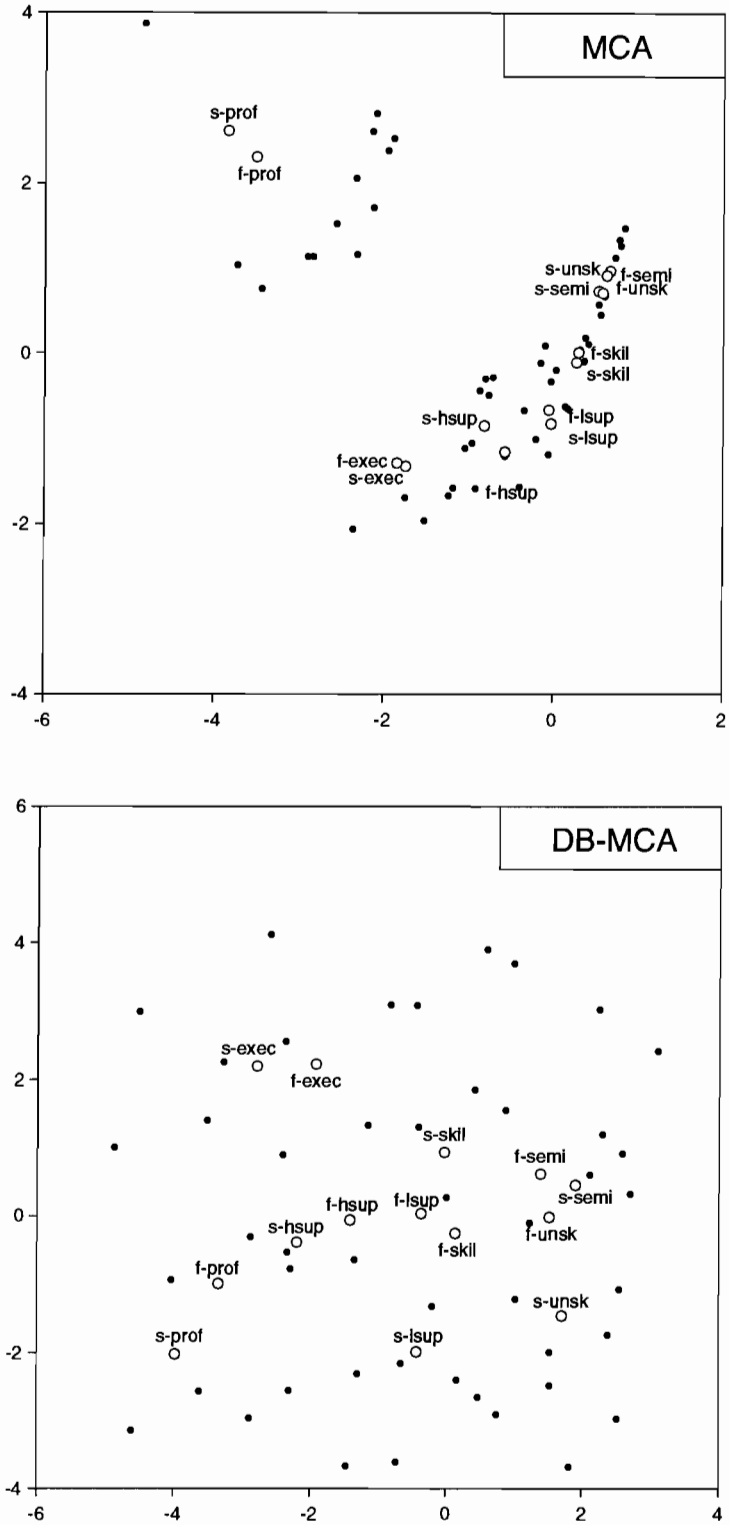

Figure 1. Two solutions in the analysis of the occupational mobility data. Multiple correspondence analysis of the profile-frequency matrix (upper panel) and distance-based MCA (lower panel). The category points are denoted by circles, profile points are displayed as dots.

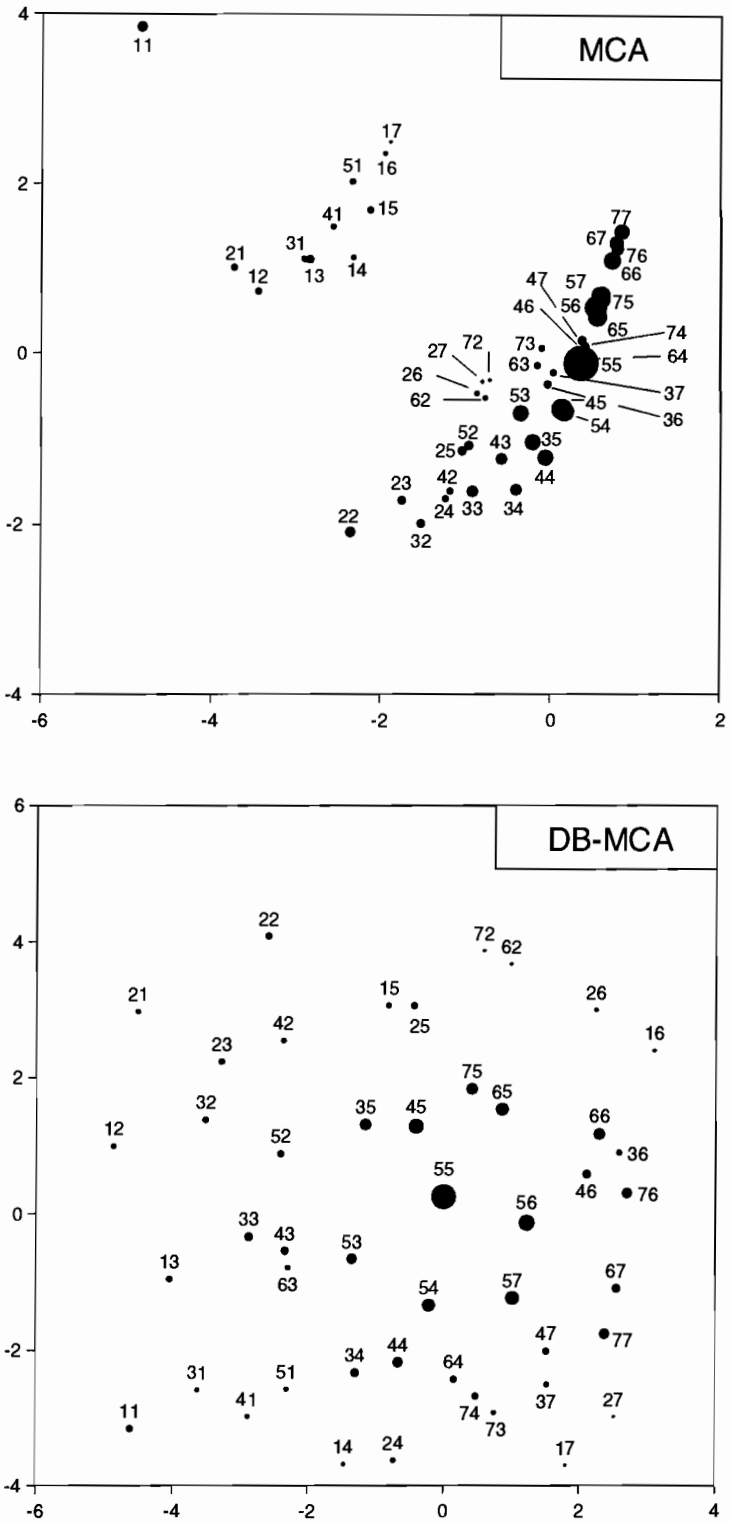

Figure 2. Plot of profile points with weights as masses and labeled by the father-son variables (for the labels, see Table 2) for multiple correspondence analysis (upper panel) and distance-based MCA (lower panel) .

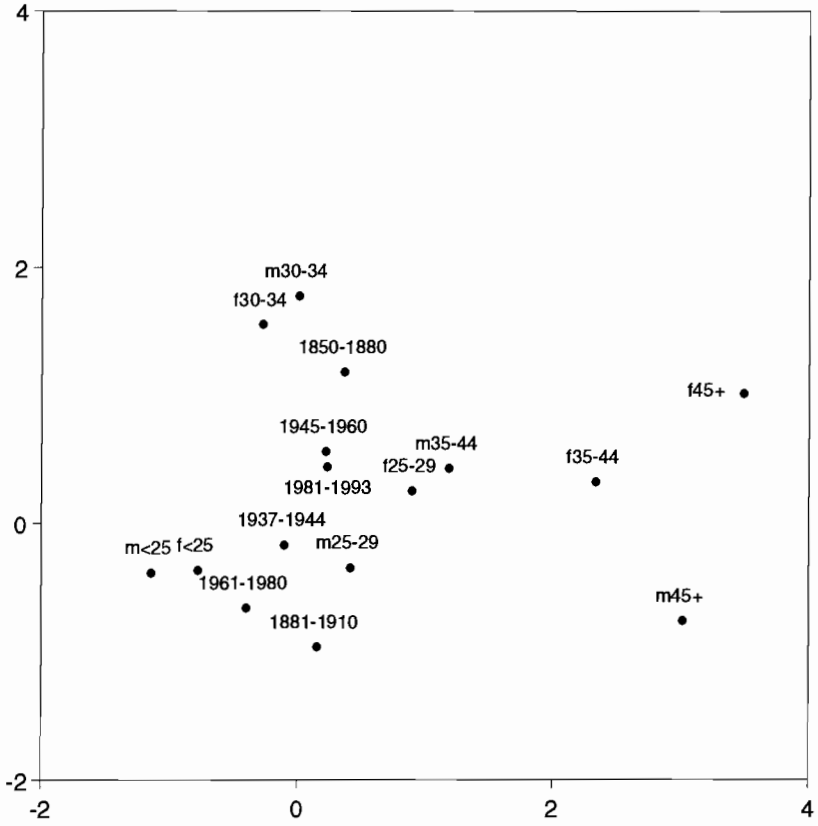

Figure 3. Distance-based MCA solution of spouse preference of all marriages in the Netherlands during the period 1850–1993.

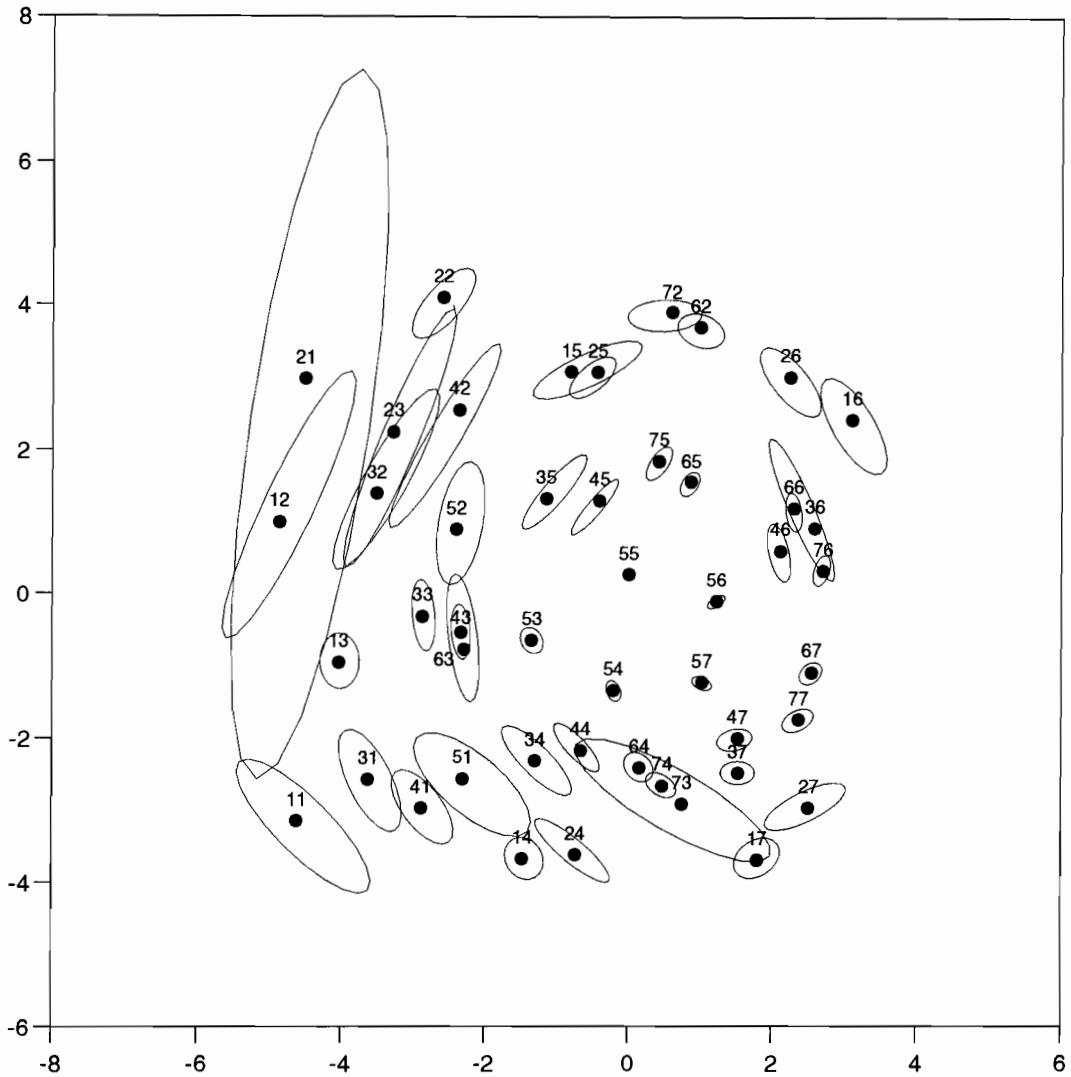

Figure 4. Bootstrap results of objects points of the distance-based MCA analysis on the occupational mobility data reported in Table 2. The ellipses are the 95% confidence regions of the bootstrap sample points.

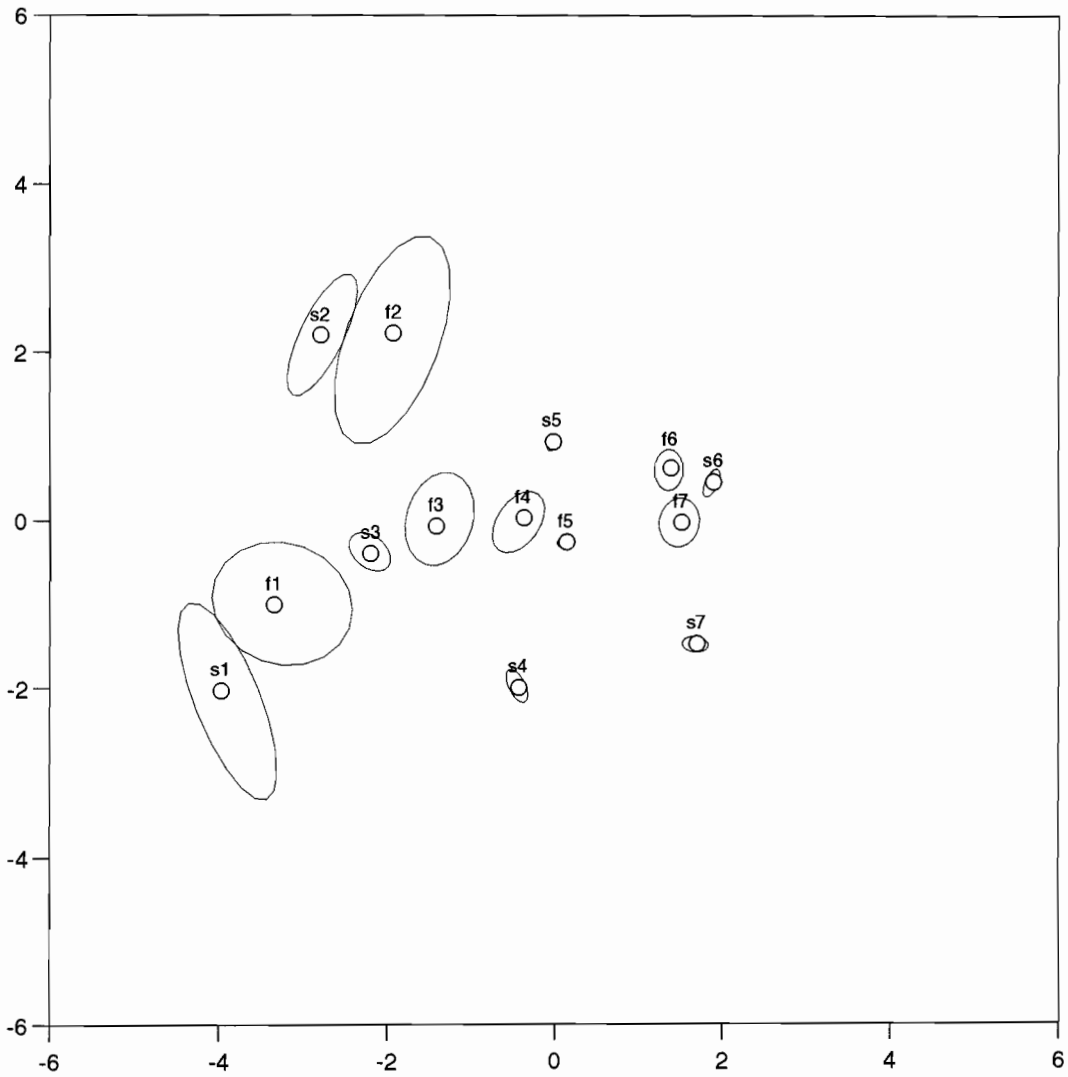

Figure 5. Bootstrap results of the category centroids of the distance-based MCA analysis on the occupational mobility data reported in Table 2. The ellipses are the 95% confidence regions of the bootstrap sample points.