Statistical Science
1998, Vol. 13, No. 4, 307–336

The Gifi System of Descriptive Multivariate Analysis

George Michailidis and Jan de Leeuw

Abstract. The Gifi system of analyzing categorical data through nonlinear varieties of classical multivariate analysis techniques is reviewed. The system is characterized by the optimal scaling of categorical variables which is implemented through alternating least squares algorithms. The main technique of homogeneity analysis is presented, along with its extensions and generalizations leading to nonmetric principal components analysis and canonical correlation analysis. Several examples are used to illustrate the methods. A brief account of stability issues and areas of applications of the techniques is also given.

Key words and phrases: Optimal scaling, alternating least squares, multivariate techniques, loss functions, stability.

1. A GEOMETRIC INTRODUCTION TO HOMOGENEITY ANALYSIS

Homogeneity analysis, also known as multiple correspondence analysis, can be introduced in many different ways, which is probably the reason why it was reinvented many times over the years (for a more detailed account on the history and the variations of the technique see Section 2). In this paper we motivate homogeneity analysis in graphical language, since complicated multivariate data can be made more accessible by displaying their main regularities in pictures (e.g., scatterplots).

Consider the following fairly typical situation that arises in practice in various fields in the physical, social and life sciences. Data on J categorical variables have been collected for N objects or individuals, where variable $j \in \mathbf{J} = \{1, 2, \ldots, J\}$ can take ℓ_j possible values (categories). The use of categorical variables is not particularly restrictive, since in every data application a continuous numerical variable can be thought of as a categorical variable with a very large number of categories. Given such a data matrix, one can represent all the available information by a bipartite graph, where the first set of N vertices corresponds to the objects and the second set of $\sum_{j\in\mathbf{J}} \ell_j$ vertices to the categories of the J variables. Each object is connected to the categories of the variables it belongs to; thus, the set of $N \sum_{j\in\mathbf{J}} \ell_j$ edges provides information about which categories an object belongs to, or alternatively which objects belong to a specific category. Thus, the N vertices corresponding to the objects all have degree J, while the $\sum_{j\in\mathbf{J}} \ell_j$ vertices corresponding to the categories have varying degrees, equal to the number of objects in the categories. We can then draw this graph and attempt to find interesting and useful patterns in the data. In Figure 1 the bipartite graph of a toy example corresponding to a 4×3 contingency table with seven objects is given. However, except for very small data sets (both in terms of objects and variables) such a representation is not very helpful. A better approach would be to try to find a low-dimensional space in which objects and categories are positioned in such a way that as much information as possible is retained from the original data. Hence, the goal becomes to construct a low-dimensional joint map of objects and categories in Euclidean space ($\mathbb{R}^p$). The choice of low dimensionality is because the map can be plotted, and the choice of Euclidean space stems from its nice properties (projections, triangle inequality) and our familiarity with Euclidean geometry. The problem of drawing graphs that are easy to understand and present has attracted a lot of attention in the computer science literature [25]. There are different approaches to

George Michailidis is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109 (e-mail: gmichail@umich.edu). Jan de Leeuw is Professor and Chair, Department of Statistics, Interdivisional Program in Statistics, University of California, Los Angeles, California 90095 (e-mail: deleeuw@stat.ucla.edu).

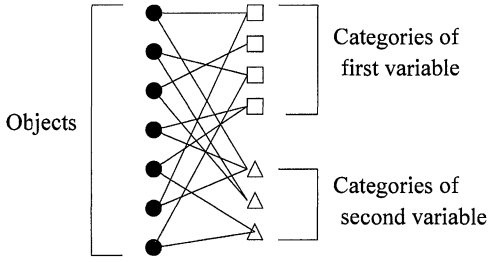

FIG. 1. *The bipartite graph of a toy example.*

drawing such maps and different ways of finding them; a particular set of criteria defines the former [23, 24] and the specific algorithm employed determines the latter.

Let X be the $N \times p$ matrix containing the coordinates of the object vertices in $\mathbb{R}^p$, and let Y_j, $j \in \mathbf{J}$, be the $\ell_j \times p$ matrix containing the coordinates of the ℓ_j category vertices of variable j. We call X the object scores matrix and Y_j's the category quantifications matrices. If we assign random values to X and the Y_j's and plot the vertices and the corresponding edges, we will typically get a picture similar to the one shown in Figure 2 for the mammals' dentition data set that is analyzed in Section 1.1 (the data are given in Appendix A). It can be seen that very little is gained by this two-dimensional representation. The picture has too much "ink" and no interesting patterns emerge. A more informative picture would emerge if the edges were short, or in other words if objects were close to the categories they fall in, and categories were close to the objects belonging in them [24]. Hence, our goal becomes that of making a *graph plot* that minimizes the total *squared* length of the edges. This criterion is chosen because it leads to an eigenvalue problem, and thus is nicely related to many classical multivariate analytic techniques.

[In this paper we employ the following notational conventions. Uppercase letters are used for matrices (e.g., A), and lowercase letters for vectors (e.g., a). The (s,t)th element of a matrix is denoted by $A(s,t)$, the sth row by $A(s,\cdot)$ and the tth column by $A(\cdot,t)$. Analogously, the sth element of a vector is denoted by $a(s)$. Finally, u_p denotes a p-dimensional column vector composed of only ones, and I_p denotes the identity matrix of order p.]

The data are coded by using *indicator* matrices G_j, with entries $G_j(i,t) = 1$ if object i belongs to category t, and $G_j(i,t) = 0$ if it belongs to some other category, $i = 1,\ldots,N$, $t = 1,\ldots,\ell_j$. The matrix $G = (G_1,\ldots,G_J)$ is simply the *adjacency* matrix of the bipartite graph. The average squared edge length (over all variables) is then given by

$$\begin{aligned}\sigma(X;Y_1,\ldots,Y_J) &= J^{-1}\sum_{j=1}^{J}\mathrm{SSQ}(X-G_jY_j)\\ &= J^{-1}\sum_{j=1}^{J}\mathrm{tr}(X-G_jY_j)'(X-G_jY_j),\end{aligned} \tag{1.1}$$

where $\mathrm{SSQ}(H)$ denotes the sum of squares of the elements of the matrix H. We want to minimize (1.1) simultaneously over X and the Y_j's. The loss function (1.1) is at the heart of the Gifi system [33], and the entire system is mainly about different versions of the above minimization problem. By imposing various *restrictions* on the category quantifications Y_j, and in some cases on the coding of the data, different types of analysis can be derived.

In order to avoid the trivial solution corresponding to $X = 0$, and $Y_j = 0$ for every $j \in \mathbf{J}$, we require in addition

$$X'X = NI_p, \tag{1.2}$$

$$u_N'X = 0. \tag{1.3}$$

The second normalization restriction basically requires the graph plot to be centered around the origin. The first restriction standardizes the squared length of the object scores (to be equal to N), and in two or higher dimensions also requires the columns of X to be orthogonal. Although this is computationally convenient, in many respects it is not completely satisfactory, a fact already noted by Guttman [50]. It is worth noting that the criterion defined by (1.1) expresses how well the categorical variables can be reduced to one score vector (in case $p = 1$) by appropriate rescaling, thus tying it to ideas from principal components analysis.

Let us examine the solution to our minimization problem (1.1) subject to the normalization constraints (1.2) and (1.3). An *alternating least squares* (ALS) algorithm is employed that, in addition, reveals some of the properties of the solution discussed below (a different presentation of the solution to the problem is discussed in Section 2.2). In the first step, (1.1) is minimized with respect to Y_j for fixed X. Essentially, we are fitting the multivariate linear model $X = G_jY_j + \text{error}$, for each $j \in \mathbf{J}$ to the data and the solution is given by

$$\hat{Y}_j = D_j^{-1}G_j'X, \quad j \in \mathbf{J}, \tag{1.4}$$

where $D_j = G_j'G_j$ is the $\ell_j \times \ell_j$ diagonal matrix containing on its diagonal the relative frequencies of the categories of variable j. In the second step of

the algorithm, (1.1) is minimized with respect to X for fixed Y_j's. The optimal $\hat{X}$ is given by

$$\hat{X} = J^{-1} \sum_{j=1}^{J} G_j Y_j. \tag{1.5}$$

In the third step of the algorithm the object scores X are column centered by setting $W = \hat{X} - u_N(u_N' \hat{X}/N)$, and then orthonormalized by the modified Gram–Schmidt procedure [39] $X = \sqrt{N}\,\mathrm{GRAM}(W)$, so that both normalization constraints (1.2) and (1.3) are satisfied. The ALS algorithm cycles through these three steps until it converges (this algorithm is identical to Bauer–Rutishauser simultaneous iteration, the natural generalization of the power method for computing some of the dominant eigenvalues along with their corresponding eigenvectors).

Equation (1.4) expresses the so-called *first centroid principle* [3] (a category quantification is in the centroid of the object scores that belong to it), while (1.5) shows that an object score is the average of the quantifications of the categories it belongs to. Hence, this solution accomplishes the goal of producing a graph plot with objects close to the categories they fall in and categories close to the objects belonging in them. It should also be noted that the use of indicator matrices makes the above ALS procedure equivalent to the method of reciprocal averaging (which can already be found in the works of Fisher [31] and Guttman [50]). This solution is known in the literature [33, 18, 22] as the *Homals solution* (homogeneity analysis by means of alternating least squares) and has been implemented in computer software in various platforms (program Homals in SPSS [90], and by Bond and Michailidis in Lisp-Stat [7]).

Once the ALS algorithm has converged, by using the fact that $\hat{Y}_j' D_j \hat{Y}_j = \hat{Y}_j' D_j (D_j^{-1} G_j' \hat{X}) = \hat{Y}_j' G_j' \hat{X}$, we can write the loss function as

$$\begin{aligned} & J^{-1} \sum_{j=1}^{J} \mathrm{tr}(\hat{X} - G_j \hat{Y}_j)'(\hat{X} - G_j \hat{Y}_j) \\ & \quad = J^{-1} \sum_{j=1}^{J} \mathrm{tr}(\hat{X}'\hat{X} + \hat{Y}_j' D_j \hat{Y}_j - 2\hat{Y}_j' D_j \hat{Y}_j) \\ & \quad = J^{-1} \sum_{j=1}^{J} \mathrm{tr}(N I_p - \hat{Y}_j' D_j \hat{Y}_j) \\ & \quad = Np - J^{-1} \sum_{j=1}^{J} \mathrm{tr}(\hat{Y}_j' D_j \hat{Y}_j). \end{aligned} \tag{1.6}$$

The sum of the diagonal elements of the matrices $\hat{Y}_j' D_j \hat{Y}_j$ is called the *fit* of the solution. Furthermore, the *discrimination measures* of variable j in dimension s are given by

$$\eta_{js}^2 \equiv \hat{Y}_j'(\cdot, s) D_j \hat{Y}_j(\cdot, s)/N, \qquad j \in \mathbf{J},\ s = 1, \ldots, p. \tag{1.7}$$

The discrimination measures give the average squared distance (weighted by the marginal frequencies) of the category quantifications to the origin of the p-dimensional space. Hence, variables that discriminate well between categories have their category points further apart, which in turn better separates the objects. It can be shown that (assuming there are no missing data) the discrimination measures are equal to the squared correlation between an optimally quantified variable $G_j \hat{Y}_j(\cdot, s)$ in dimension s, and the corresponding column of object scores $\hat{X}(\cdot, s)$ (see [33], Chapter 3). Hence, the loss function can also be expressed as

$$N\left(p - \frac{1}{J} \sum_{j=1}^{J} \sum_{s=1}^{p} \eta_{js}^2\right) = N\left(p - \sum_{s=1}^{p} \gamma_s\right), \tag{1.8}$$

where the *eigenvalues* $\gamma_s = J^{-1} \sum_{j=1}^{J} \eta_{js}^2$, $s = 1, \ldots, p$, correspond to the average of the discrimination measures and give a measure of the fit of the Homals solution in the sth dimension.

Next we summarize some basic properties of the Homals solution:

- Category quantifications and object scores are represented in a joint space (see Figure 4 for the mammals' dentition data set).
- A category point is the centroid of objects belonging to that category, a direct consequence of (1.4) (see Figure 1 for two variables from the mammals' dentition data set).
- Objects with the same response pattern (identical profiles) receive identical object scores [follows from (1.5)] (see Figure 7). In general, the distance between two object points is related to the "similarity" between their profiles.
- A variable discriminates better to the extent that its category points are further apart [follows from (1.7)].
- If a category applies uniquely to only a single object, then the object point and that category point will coincide.
- Objects with a "unique" profile will be located further away from the origin of the joint space, whereas objects with a profile similar to the "average" one will be located closer to the origin (direct consequence of the previous property).
- The category quantifications of each variable $j \in \mathbf{J}$ have a weighted sum over categories equal to zero. This follows from the employed normal-

ization of the object scores, since $u'_{l_j} D_j \hat{Y}_j = u'_{l_j} D_j D_j^{-1} G'_j \hat{X} = u'_{l_j} G'_j \hat{X} = u'_N \hat{X} = 0$.

- The Homals solutions are *nested*. This means that if one requires a p_1-dimensional Homals solution and then a second ($p_2 > p_1$)-dimensional solution, then the first p_1 dimensions of the latter solution are identical to the p_1-dimensional solution.
- The solutions for subsequent dimensions are *ordered*. This means that the first dimension has the absolute maximum eigenvalue. The second dimension has the maximum eigenvalue subject to the constraint that $X(\cdot, 2)$ is uncorrelated to $X(\cdot, 1)$, and so forth.
- The object scores are uncorrelated in subsequent dimensions [follows from (1.3)]. However, the category quantifications need not necessarily be uncorrelated; in fact, their correlation patterns might be rather unpredictable.
- The solution is *invariant* under *rotations* of the object scores in p-dimensional space and of the category quantifications. To see this, suppose we select a different basis for the column space of the object scores X; that is, let $X^* = X \times R$, where R is a rotation matrix satisfying $R'R = RR' = I_p$. We then get from (1.4) that $Y_j^* = D_j^{-1} G'_j X^* = \hat{Y}_j R$.

1.1 An Illustration: Mammals' Dentition Example

In this section we discuss the results we obtained by applying the Homals algorithm to the mammals dentition data set (see Figures 2–9). The complete data together with the coding of the variables are given in Appendix A. Dental characteristics are used in the classification of mammals. One of the reasons is that teeth constitute one of the most common remains (together with bones of the jaw and the skull) of mammals, since they are highly resistant to chemical and physical weathering [27]. Because of the abundance of teeth in deposits of fossil mammals, dentition has been stressed in the interpretation of mammalian phylogeny and relationships. Moreover, dental characteristics provide information about the food habits of mammals and hence to some extent about their natural habitats. The main question is whether the technique managed to produce a fairly clean picture and uncover some interesting features of this data set.

A two-dimensional analysis gives an adequate fit, with eigenvalues 0.73 and 0.38 for the two dimen-

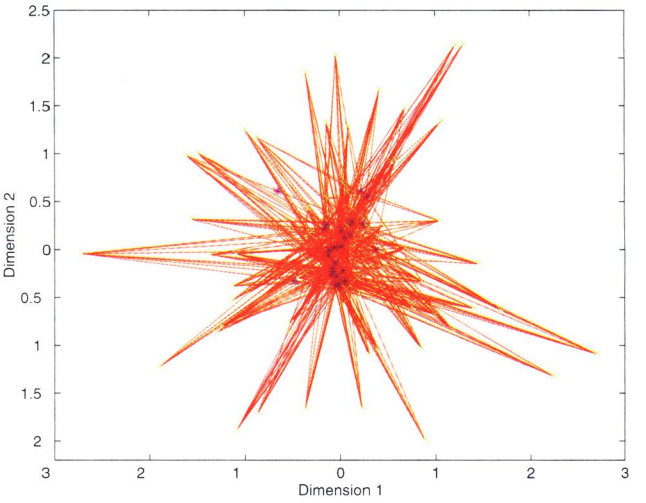

FIG. 2. *An arbitrary two-dimensional graph plot of the mammals' dentition data set* ($\times$ = *objects*, $*$ = *categories*).

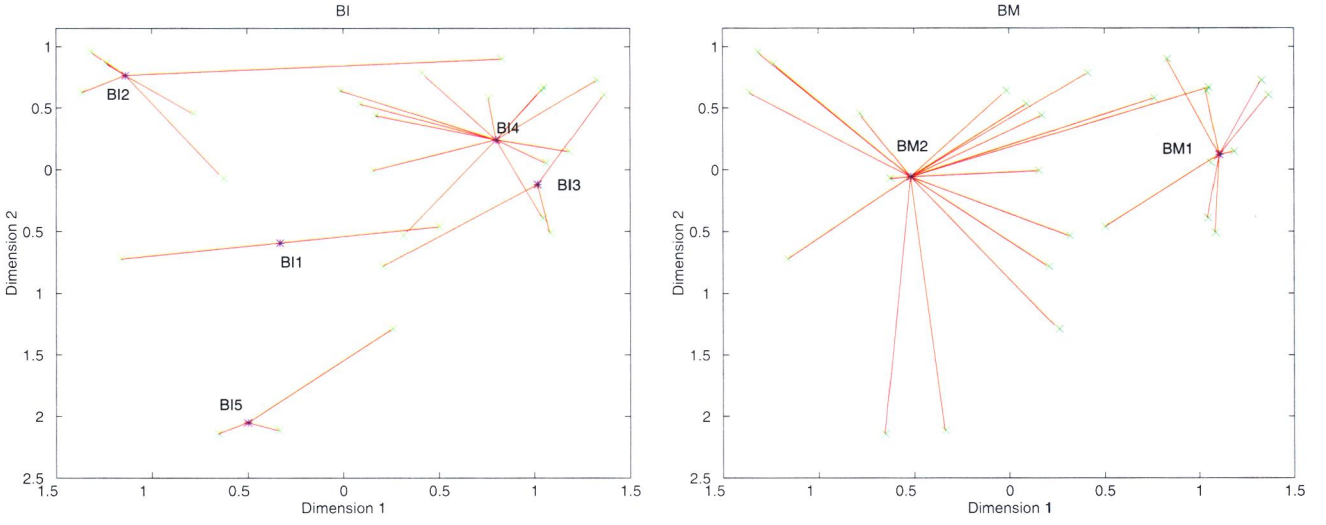

FIG. 3. *Star plots of variables bottom incisors (BI) and bottom molars (BM) from the mammals' dentition example.*

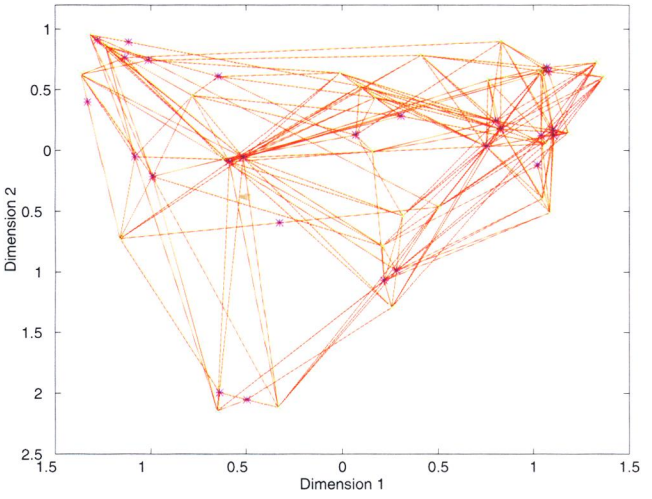

FIG. 4. *A two-dimensional graph plot of the mammals data set produced by Homals* ($\times$ = *objects*, $*$ = *categories*).

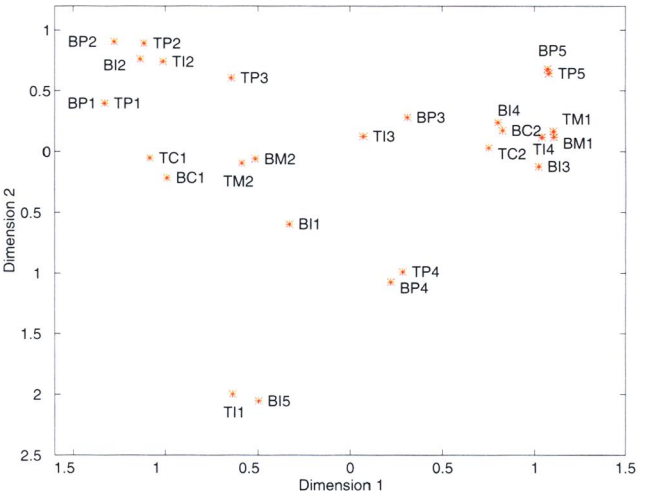

FIG. 5. *Category quantifications of the variables in the mammals' dentition example.*

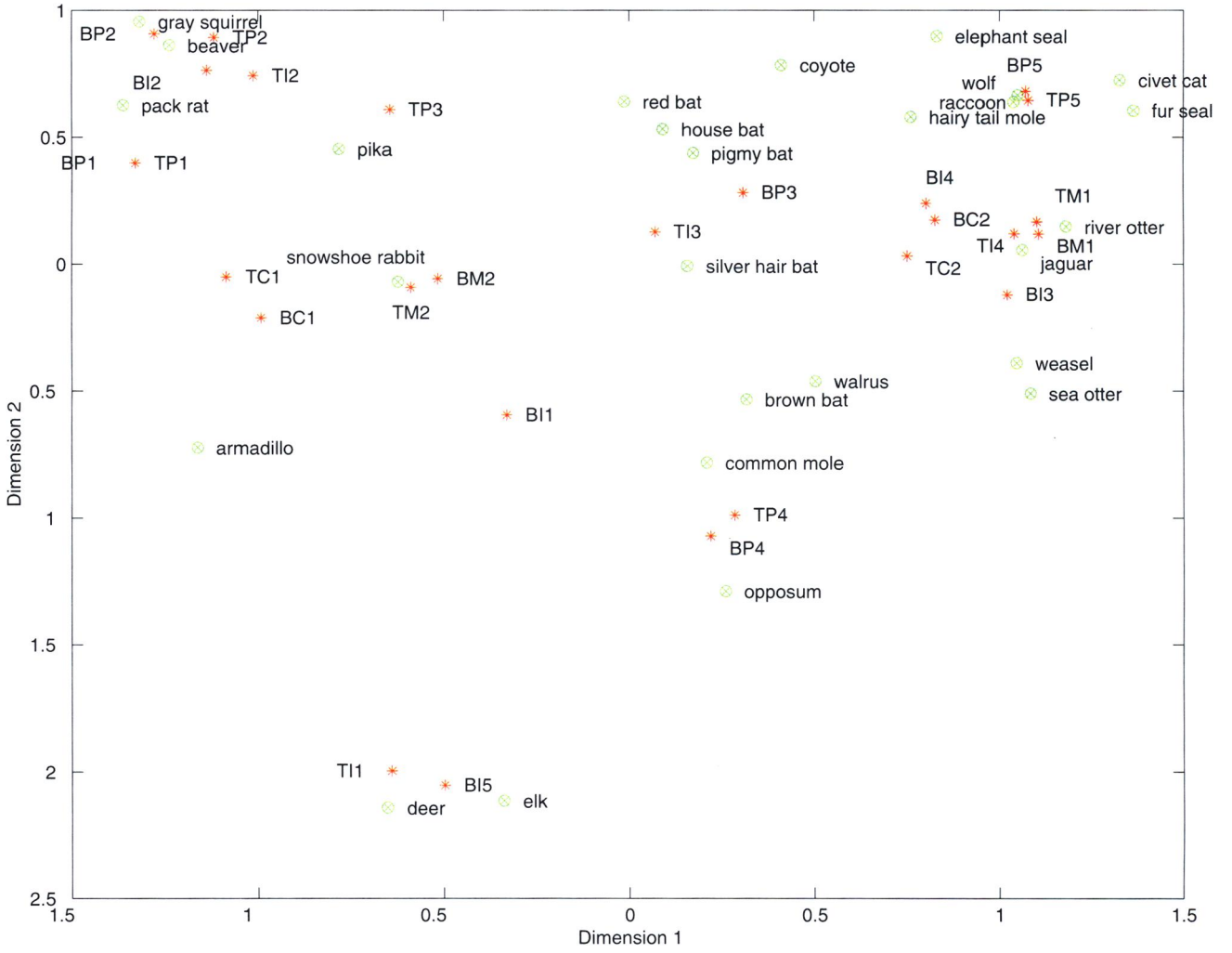

FIG. 6. *Category quantifications (red) and object scores (green) (omitting mammals which have dentition identical to the ones shown on the graph).*

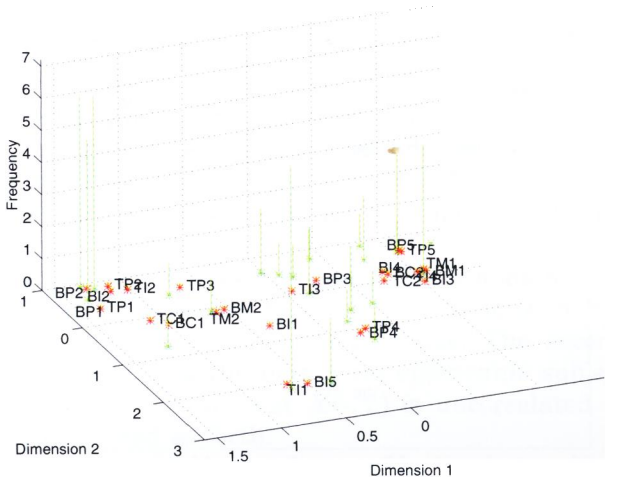

FIG. 7. *Category quantifications (red) and object scores (green) (height of the object scores shows how many mammals share the particular teeth profile).*

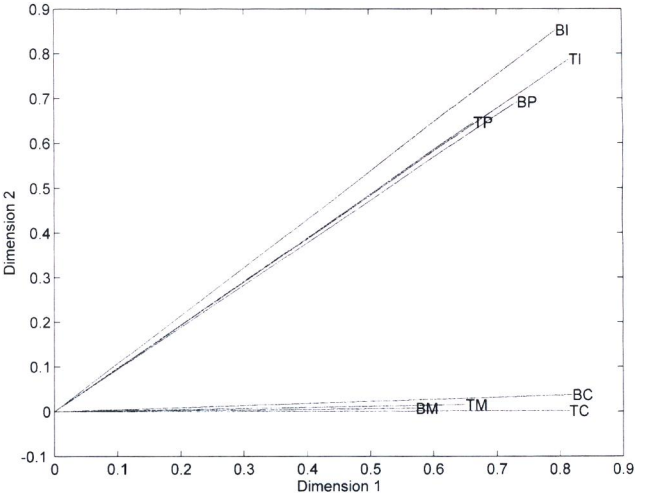

FIG. 8. *Discrimination measures of the variables in the mammals example.*

sions, respectively. The graph plot of this solution is given in Figure 4. Comparing it with Figure 2, we immediately notice that the objects and the categories have been arranged in such a way so that the amount of ink on the graph is minimized. Moreover, several patterns have emerged. To study those patterns more closely we turn our attention to the arrangement of only the category vertices on the map (see Figure 5). It can be seen that they form four groups. In the upper right corner we can find categories BM1, TM1, TC2, BC2, TI4, BI3, BI4, TP5 and BP5. Thus, objects located in this area of the map are associated with these categories, that is, the presence of canines and molars and a large number of incisors and premolars. Such a dental pattern characterizes the carnivores (meat eaters) that use their incisors for nipping, the canines to grab and hold the prey and the molars and premolars for grinding the food. In the upper left corner we find categories BP1, BP2, TP1, TP2, TP3, BI2 and TI2, while in the center the categories TC1, BC1, TM2, BM2, BI1, TI3, BP3, BP4 and TP4. However, the latter group can be split further into subgroups. For example, we see that categories TP4 and BP4 are close together, thus suggesting that objects with three top premolars, usually have three bottom premolars as well; similarly, for the subgroup TC1, BC1, TM2 and BM2, we have that animals with no top and bottom canines have three or more top and bottom molars. Finally, in the lower and slightly to the left area of the map we find a group of objects mainly characterized by the categories TI1 and BI5. At this point, it would be interesting to include in the picture the objects themselves (see Figure 6) along with their respective frequencies (see Figure 7). We see that the objects are located in the periphery of the map, which is a consequence of the first centroid principle. Moreover, the majority of the objects form four separate clusters located in the upper right, upper left, lower left and middle parts of the picture. For example, in the lower left area we find the following animals: elk, deer, moose, reindeer, antelope, bison, mountain goat, musk-ox and mountain sheep—that belong to the ruminants, which are particularly characterized by the absence of top incisors and the presence of a large number of bottom incisors (see in Figure 6 the proximity of these mammals to TI1 and BI5). Similarly, in the upper left area we find, among others, various types of squirrels (ground, gray, fox) and rats (kangaroo, pack, field mouse, black etc.), all belonging to the rodents (or rodentia group). The lack of canines and molars indicates that the dentition of these animals is adapted to a broadly herbivorous diet, but the varying number of incisors and premolars distinguishes to a large degree the rodents from the ruminants. In the upper right of the graph we find various groups of the carnivores, such as jaguar, cougar, lynx (all belonging to the felidae—i.e., cat—family), but also various types of seals (fur, sea lion, grey, elephant, etc.). Finally, in the middle upper part of the picture we find a group of various types of bats (all insectivorous mammals). It is worth noting that the Homals solution isolates the armadillo in the left part of the graph, which is a consequence of the fact that it belongs to the edentates (mammals that basically lack teeth). Moreover, it positions the opossum somewhere between the carnivores and the ru-

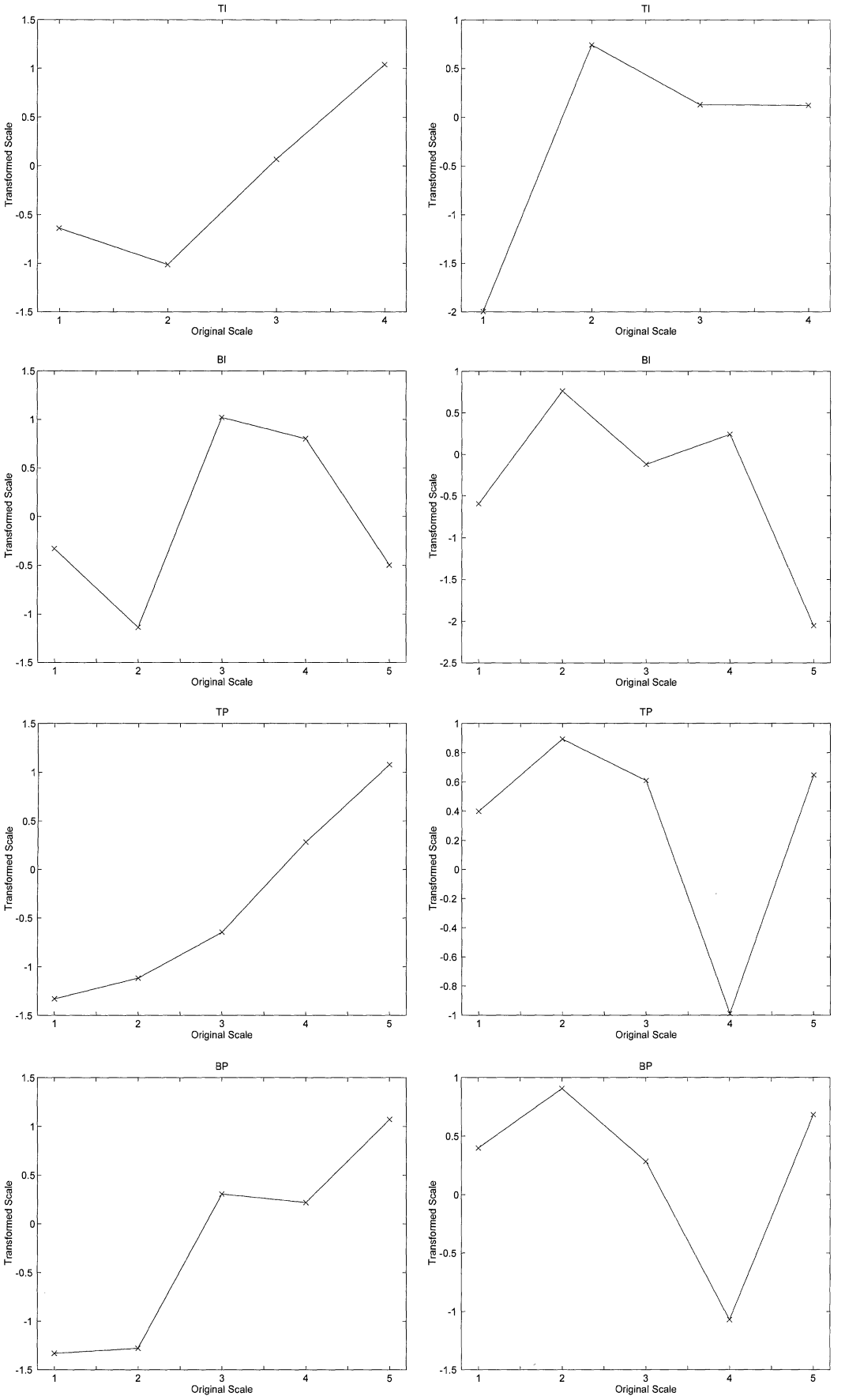

FIG. 9. *Optimal transformations of some of the variables in the mammals example: (left) dimension 1; (right) dimension 2.*

minants but also close to the bats, thus suggesting that its dentition is adapted to its food habits (the opossum eats almost anything, including insects, small mammals, fruits, berries and cultivated crops [27]).

Hartigan [52], using a tree-type clustering algorithm, found many similar groupings (e.g., beaver and squirrels, weasel and otters, deer and elk, various types of bats). However, it is worth noting that his clustering algorithm classified the hairy tail mole together with the opossum, instead of the bats, and the pack rat with the armadillo instead of the rodents (e.g., the squirrels). This is due to the nature of the particular algorithm used, which is quite sensitive to the order of the presentation of the objects and to the selected variables used at different levels of the tree. On the other hand, the Homals algorithm positions the objects by taking into consideration the similarity of the entire tooth profile and hence manages to avoid such misclassifications (see [52], page 171).

As mentioned above, a variable *discriminates* better to the extent that its category points are further apart in the derived map. The discrimination measures, shown in Figure 8, indicate that the variables TC, BC, TM and BM discriminate exclusively along the first dimension, while the remaining variables discriminate equally well on both dimensions. Reexamining Figure 5, we see that categories TC1, BC1, TM2 and BM2 are to the left of the origin, while categories TC2, BC2, TM1 and BM1 are to the right of the origin. This implies that this set of variables does a primary split of the objects into two groups (carnivores and other mammals), while the other set of variables does further splits especially along the second dimension (to a large extent distinguishing the rodents from the ruminants). It is also interesting to examine the plots of the original (i.e., $1, 2, 3, 4$) versus the transformed scales given in Figure 9. Obviously, such plots are totally uninteresting for binary variables, and therefore are omitted. However, for the remaining variables they reveal nonlinear patterns in both dimensions. In some cases, the patterns are monotone (e.g., variables TI in both dimensions, and BP in the first dimension) suggesting an implicit ordering in the original scale, while in others the pattern is not very clear.

2. OTHER ASPECTS OF HOMOGENEITY ANALYSIS

In this section we study several aspects of homogeneity analysis. More specifically, we provide some alternative ways of introducing the technique and study its connection to an eigenvalue problem. We briefly review how to handle missing data in this framework and more elaborate coding schemes of the data. Finally, we discuss how homogeneity analysis is related to other techniques proposed in the literature that deal with categorical variables.

2.1 Some Alternative Introductions to Homogeneity Analysis

In the previous section, homogeneity analysis was motivated and introduced in pure graphical language. The basic premise was that complicated multivariate data can be made more accessible by displaying their main regularities and patterns in plots. What the technique accomplished was to scale the N objects (map them into a low-dimensional Euclidean space) in such a way that objects with similar profiles were close together, while objects with different profiles were relatively apart. However, the technique can be introduced from a different starting point.

Another possibility for introducing homogeneity analysis is through linearization of the regressions. Consider a column of the object scores $X(\cdot, s)$ as N data values on the x-axis. Let the category quantifications in the same sth dimension of these N data points correspond to the y-axis values. The regression line of x on y has slope equal to 1. The reason is that the category quantifications y are averages of x-values within that category (follows from (1.4)). Remarkably enough, the regression of y on x is also perfectly linear with slope given by the eigenvalue γ_s. This is because the object scores x are proportional to the averages of the category quantifications applying to an object (follows from (1.5)). Therefore, the Homals solution could be defined on the basis of this property; it is the solution where object scores are proportional to category means and where category quantifications are proportional to object means.

A third possible interpretation of the Homals solution is in terms of a principal components analysis of the quantified data matrix. It can be shown [22] that the sum of squared correlations between the optimal quantified variables $G_j Y(\cdot, s)$ and the vector $X(\cdot, s)$ is maximized.

2.2 Homogeneity Analysis as an Eigenvalue and a Singular Value Decomposition Problem

One of the reasons why squared edge length is appealing as a criterion is that it makes the resulting minimization problem an eigenvalue problem. To see this, substitute the optimal $\hat{Y}_j = D_j^{-1} G_j' X$

for given X in the loss function (1.1), to get

$$\begin{aligned}\sigma(X;\star) &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}\bigl(X-G_jD_j^{-1}G_j'X\bigr)'\\ &\qquad\cdot\bigl(X-G_jD_j^{-1}G_j'X\bigr)\\ &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}\bigl(X'X-X'G_jD_j^{-1}G_j'X\bigr),\end{aligned}\tag{2.1}$$

where the star has replaced the argument over which the loss function is minimized. Let $P_j = G_jD_j^{-1}G_j'$ denote the orthogonal projector on the subspace spanned by the columns of the indicator matrix G_j. Let $P_\star = J^{-1}\sum_{j=1}^{J}P_j$ be the average of the J projectors. Equation (2.1) can be rewritten as

$$\begin{aligned}\sigma(X;\star) &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}\bigl(X-P_jX\bigr)'\bigl(X-P_jX\bigr)\\ &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}\bigl(X'X-X'P_jX\bigr).\end{aligned}\tag{2.2}$$

This together with the normalization constraints (1.2) and (1.3) gives that minimizing (2.2) is equivalent to maximizing $\operatorname{tr}(X'\mathscr{L}P_\star\mathscr{L}X)$, where $\mathscr{L} = I_n - u_Nu_N'/u_N'u_N$ is a centering operator that leaves $\mathscr{L}X$ in deviations from its column means. The optimal X corresponds to the largest p eigenvectors of the matrix $\mathscr{L}P_\star\mathscr{L}$. We can then write the minimum loss as follows:

$$\sigma(\star;\star) = N\Bigl(p-\sum_{s=1}^{p}\lambda_s\Bigr),\tag{2.3}$$

where λ_s, $s=1,\dots,p$, are the largest p eigenvalues of $P_\star$. Therefore, the minimum loss of homogeneity analysis is a function of the p largest eigenvalues of the average projector $P_\star$. Another derivation starts by combining the J indicator matrices G_j into a superindicator matrix $G = [G_1|\cdots|G_J]$ (the symbol $|$ stands for concatenating matrices horizontally, so that $[A_1|A_2] = [A_1A_2]$, provided that they have the same number of rows) and the marginal frequencies into $D = \bigoplus_{j=1}^{J}D_j$. (The symbol $\bigoplus$ stands for direct sum. This operation is defined for two matrices A_1 and A_2 as

$$A_1\bigoplus A_2 = \begin{bmatrix}A_1 & 0\\ 0 & A_2\end{bmatrix}$$

and similarly for any number of matrices.) The solution for the optimal X can then be obtained (see [33]) by the singular value decomposition of

$$J^{-1/2}\mathscr{L}GD^{-1/2} = U\Lambda V,\tag{2.4}$$

where the left-hand side is the superindicator matrix in deviations from column means and weighted by the marginal frequencies. The optimal X corresponds to the first p columns of the matrix U (the first p left-singular vectors). Notice that the complete eigenvalue and singular value solutions have $q = \sum_{j=1}^{J}\ell_j - J$ dimensions. The advantage of employing the ALS algorithm is that it only computes the first $p \ll q$ dimensions of the solution, thus increasing the computational efficiency and decreasing the computer memory requirements.

2.3 Missing Data

The present loss function makes the treatment of missing data a fairly easy exercise. Missing data can occur for a variety of reasons: blank responses, coding errors and so on. Let M_j, $j\in\mathbf{J}$, denote the $N\times N$ binary diagonal matrix with entries $M_j(ii) = 1$ if observation i is present for variable j and 0 otherwise. Define $M_* = \sum_{j=1}^{J}M_j$. Notice that since G_j is an incomplete indicator matrix (has rows with just zeros), we have that $M_jG_j = G_j$, $j\in\mathbf{J}$. The loss function then becomes

$$\begin{aligned}&\sigma(X;Y_1,\dots,Y_J)\\ &\quad= J^{-1}\sum_{j=1}^{J}\operatorname{tr}\bigl(X-G_jY_j\bigr)'M_j\bigl(X-G_jY_j\bigr),\end{aligned}\tag{2.5}$$

subject to the normalization restrictions $X'M_*X = JNI_p$ and $u_N'M_*X = 0$. The $\hat{Y}_j$'s are given by (1.4), whereas the object scores are given by $\hat{X} = M_*^{-1}\sum_{j=1}^{J}G_jY_j$. In the presence of missing data, it is no longer the case that $u_{l_j}'D_jY_j = 0$ (the category quantifications are not centered), because in the weighted summation with respect to the row scores of X, some of the scores are skipped. This option is known in the literature [33] as *missing data passive* or *missing data deleted*, because it leaves the indicator matrix G_j incomplete. There are two other possibilities: (i) *missing data single category* and (ii) *missing data multiple categories*. In the first case, all missing observations for a particular variable are treated as a new category, thus augmenting the indicator matrix of the variable with an additional column. In the second case, every missing observation for a particular variable is treated as a new category, so that the number of columns added to the indicator matrix of the variable corresponds the number of missing observations. The missing data passive option essentially ignores the missing observations, while the other two options make specific strong assumptions regarding the pattern of the missing data.

2.4 Alternative Coding Schemes

The coding scheme considered so far is the so-called *crisp coding* of the indicator matrix. The main

advantages it presents are as follows: it is simple and computationally efficient (due to the sparseness of the indicator matrix); it allows for nonlinear transformation of the variables; it is very robust even when coding noisy data; and the number of parameters (categories) per variable is generally small. Its disadvantages are that in many data analytic situations the determination of the categories is arbitrary and that when coding interval data there is uncertainty about the allocation of values near the category boundaries. Many alternatives have been suggested in the literature (for a thorough account see [102]), but the most commonly used alternative coding scheme is called *fuzzy coding*, a generalization of the strict logical coding of the indicator matrix. Instead of having a single 1 indicating a specific category, with zeros everywhere else, a whole set of nonnegative values adding up to 1 can be assigned to each object. In some cases these values can even be considered probabilities that the object lies in the respective categories. The main advantage of this scheme is that when a value lies near the boundary between categories, it may be allocated to both categories in appropriate amounts. The main disadvantage of any more general coding scheme is the computational burden it introduces to the ALS procedure.

2.5 Comparison to Other Techniques

In this section we attempt to relate homogeneity analysis to other widely used multivariate techniques, such as correspondence analysis, multidimensional scaling and so on.

2.5.1 *Relationship to correspondence analysis.* A special case of homogeneity analysis is the analysis of a crosstable that represents the association of two categorical variables. In this case the rows and columns of the table correspond to the categories of the two variables. Fisher's [31] eye color and hair color data set represents a prototypical example. Fisher described his objective as finding systems of row scores and column scores that maximize the correlation coefficient of the two variables, and also provided other interpretations of these scores in terms of analysis of variance, discriminant analysis and canonical correlation analysis. However, if one switches from a one-dimensional solution to a higher-dimensional solution, it is then possible to regard the systems of row and column scores as coordinates in a certain space and an elegant geometric interpretation of the solution can be given. The French approach to correspondence analysis is mainly characterized by the emphasis on geometry [3, 10, 48, 71]. In the French literature the analysis of a crosstable is called correspondence analysis ("analyse des correspondences") and the analysis of a collection of indicator matrices, which is equivalent to homogeneity analysis, is called multiple correspondence analysis ("analyse des correspondences multiple").

Let F be an $I \times J$ contingency table, whose entries f_{ij} give the frequencies with which row category i occurs together with column category j. Let $r = Fu_J$ denote the vector of row marginals, $c = F'u_I$ the vector of column marginals and $n = u'c = u'r$ the total number of observations. Finally, let $D_r = \text{diag}(r)$ be the diagonal matrix containing on its diagonal the elements of the vector r, and let $D_c = \text{diag}(c)$ be the diagonal matrix containing the elements of the vector c. The χ^2-distances between rows i and i' of table F are given by

$$\delta^2(i, i') = n \sum_{j=1}^{J} \frac{(f_{ij}/r_i - f_{i'j}/r_{i'})^2}{c_j}. \tag{2.6}$$

Formula (2.6) shows that $\delta^2(i, i')$ is a measure for the difference between the rows i and i'. It also shows that since the entries of the table are corrected for the row marginals, proportional rows yield zero distances. In addition, remaining squared differences between entries are weighted heavily if the corresponding column marginals are small, while these differences do not contribute much to the χ^2-distances if the column marginals are large. Finally, due to the role of the column marginals, the distances between the rows change when new observations are added to the crosstable. In a similar manner, χ^2-distances can be defined between columns of the crosstable.

The objective of correspondence analysis is to approximate the χ^2-distances by Euclidean distances in some low-dimensional space. In order to derive the coordinates X of the row categories of table F in the new Euclidean space, we consider the singular value decomposition of the matrix of the observed frequencies minus the expected frequencies corrected for row and column marginals

$$D_r^{-1/2}(F - E)D_c^{-1/2} = U\Lambda V', \tag{2.7}$$

where $E = rc'/n$. The optimal scores X are then given (after normalization) by

$$X = n^{1/2} D_r^{-1/2} U, \tag{2.8}$$

so that $X'D_rX = nI$ and $u_I'D_rX = 0$. It is worth noting that this solution places the origin of the space in the weighted centroid (since $u_I'D_rX = 0$)

and represents the row scores in weighted principal axes positions (since $X'D_rX = nI$).

Consider now the superindicator matrix $G = [G_1|\cdots|G_J]$ of J indicator matrices G_j. Define now the matrices D_r and D_c as follows: let $D_r = JI_N$, $D_c = D \equiv \bigoplus_{j=1}^{J} D_j$ (see Section 2.2) and $n = NJ$. Correspondence analysis of the superindicator matrix G corresponds to performing a singular value decomposition of the matrix

$$J^{-1/2}\left(G - \frac{J}{JN}Gu_N u_N'\right)D^{-1/2} = J^{-1/2}\mathscr{L}GD^{-1/2} = U\Lambda V, \tag{2.9}$$

which is identical to (2.4). This result shows that homogeneity analysis could also be viewed as approximating the χ^2-distances between the rows of the superindicator matrix. This special situation is due to the fact that the row marginals of the superindicator matrix are all equal to J. Subsequently the characteristic row weights in correspondence analysis are eliminated, and hence we deal with an unweighted origin ($u'X = 0$) and unweighted principal axes ($X'X = NI_p$). Obviously this does not hold in the presence of missing values that are coded as zeros, thus rendering unequal row marginals. Finally, it is worth noting that the special relationship between homogeneity and correspondence analysis holds only in case rows of the superindicator matrix G are analyzed, despite the fact that correspondence analysis is a symmetric technique in terms of the treatment of rows and columns. The problem arises, when considering column differences of the superindicator matrix, from the fact that the unequal column marginals enter into the picture.

2.5.2 *Relationship to multidimensional scaling.* In multidimensional scaling (MDS) the objective is to approximate given measures of association $\Delta = \{\delta_{ij}\}$, often called *dissimilarities* or *proximities*, between a set of objects by distances $D(X) = \{\delta_{ij}(X)\}$ between a set of points in some low-dimensional space. In multivariate analysis the object of the analysis is a multivariate data matrix Z and the distance approach chooses the association between the rows of the data matrix as the prime target of the analysis. This implies that each row of Z is regarded as a *profile* and the dissimilarities $\Delta(Z)$ are derived among the profiles. It is easy then to see that any multivariate analytic technique can be regarded as an MDS method by correctly specifying the kind of dissimilarity measure involved in it. In MDS, since the dissimilarities are approximated in a low-dimensional space, a loss function is used to measure the difference between Δ and the low-dimensional $D(X)$. In practice, squared dissimilarities Δ^2 are used, because of the additivity implied by Pythagoras' theorem. A typical loss function in MDS is given by

$$\sigma(X) = \operatorname{tr}\mathscr{L}\left(\Delta^2 - D^2(X)\right)' \cdot \mathscr{L}\left(\Delta^2 - D^2(X)\right)\mathscr{L}, \tag{2.10}$$

which shows that squared dissimilarities are approximated by squared distances. A moment of reflection shows that if we consider as squared dissimilarity measures the χ^2-distances of the rows of the superindicator matrix G (see previous subsection), then homogeneity analysis can be regarded as an MDS technique. The primary difference between homogeneity analysis and a general MDS technique is that the homogeneity analysis solution is obtained at the expense of stronger normalization conditions and a metric interpretation of the data (i.e., the absolute magnitude of the dissimilarities is taken into account by the technique, as opposed to a nonmetric MDS analysis where only the order of the dissimilarities matters).

2.5.3 *Relationship to cluster analysis.* As we have seen in the previous subsection homogeneity analysis provides us with an approximation of the χ^2-distances of the rows of the superindicator matrix G by low-dimensional Euclidean distances. The χ^2-distances are a measure of the dissimilarities between objects, based on the classification given in the data. The data indicate which categories are shared by objects, and also how many objects belong to each category. These two pieces of information are contained in G and in D (the matrix of univariate marginals for all variables). In homogeneity analysis, classification follows from interpreting the configuration of object points in the p-dimensional space. To put it differently, we are looking to identify clouds (clusters) of object scores and characterize them. In that sense, homogeneity analysis resembles a cluster technique.

Van Buuren and Heiser [95] have developed a technique called Groupals that simultaneously allocates the object points to only one of K groups and optimally scales the variables. Hence, the clustering and transformation problems are treated simultaneously. An alternating least squares algorithm is proposed to solve this problem. Groupals can be regarded as a forced classification method with optimal scaling features. A problem that often arises in practice is that the algorithm converges to local minima.

2.5.4 *Relationship to discriminant analysis and analysis of variance.* Homogeneity analysis can be stated in discriminant analysis and analysis of variance terms. Suppose for the time being that the matrix of object scores X is known. Each categorical variable $j \in \mathbf{J}$ defines a partitioning of these object scores. This means that we can decompose the total variance T of X in a between B and a within (group) W component. We now wish to scale the objects (find the optimal X) in such a way that W will be as small as possible, while keeping T equal to a constant (e.g., the identity matrix).

This leads to a trivial solution: all objects in the first category of the variable get the same score, all objects in the second category get another score and so on. The location of the points X is arbitrary but they satisfy $W = 0$ and $B = T = I$. However, in the presence of more than one variable, a trivial solution for one is not a trivial solution for another variable. Hence, we have to seek a compromise solution to the problem. For given X let us define T_*, B_* and W_*, which are averages over all J variables. Clearly for all variables the total variance of X is the same. The objective becomes to find a configuration X so that W_* becomes minimum, subject to the constraint $T_* = T = I$. This is another way of defining homogeneity analysis.

In homogeneity analysis terminology we have the total variance given by $T \equiv X'X = NI_p$, the variance between categories of X given by $X'P_jX$ for variable j (with $P_j = G_jD_j^{-1}G_j'$), and the variance within categories of X given by $X'(I_N - P_j)X$ for variable j. Thus, homogeneity analysis maximizes the average between categories variance, while keeping the total variance fixed. Consequently, the main difference between discriminant analysis and homogeneity analysis is that in the former we have a single categorical variable and X must be of the form UV, with U known and weights V unknown. In homogeneity analysis the number of variables J is greater than one and X is completely unknown (or $U = I$).

2.6 Other Approaches to Analyzing Categorical Data

As we have seen, homogeneity analysis is primarily a data descriptive technique of (primarily) categorical data, and its origins can be traced back to the work of Hirschfeld [57], Fisher [31] and especially Guttman [50], although some ideas go further back to Pearson (see the discussion in [17]). The main objective is to scale (assign real values to) the categories so that a particular criterion is optimized, for example, the edge length loss function (1.1). The technique has been rediscovered many times and is also known as (multiple) correspondence analysis [3], dual scaling [79, 81], quantification theory [55] and also simultaneous linear regression, centroid scaling, optimal scoring and biplot, each name emphasizing some particular aspect of the technique. For example, the French group around Benzécri paid particular attention to contingency tables and emphasized the geometrical aspects of the technique, while Nishisato's derivation stems from analysis of variance considerations (see Section 2.5.3), and Guttman was trying to apply principal component analysis to categorical data. However, in spite of the fact that the various approaches have a common starting point, most of them have passed the stage of basic formulation and moved toward their own unique advancement. Hence, we have Nishisato's efforts to apply dual scaling techniques to a wider variety of data such as multiway data matrices, paired comparisons, rank order, successive categories and sorting [80]. On the other hand, a lot of work has been done by the French group on extending and generalizing correspondence analysis beyond simply examining the interaction of row and column variables, by assuming stronger underlying mechanisms that generated the data [28–30, 101]. The Gifi group by considering generalizations of the loss function (1.1), and by placing restrictions on the category quantifications, attempts to incorporate other popular multivariate techniques into the system, while retaining the focus on the graphical representations of the data and the exploratory nature of the techniques (for more details see Sections 3 and 4). Thus, we see the various groups and approaches branching out, diverging from their starting point and exploring new directions. However, a common point that all of them retain is that the methods and techniques are usually not introduced by way of an estimation problem based on a model involving parameters and error terms. Rather, one directly poses an optimization problem for some type of loss function, while statistical inference takes a back seat [9]. Nevertheless, there have been many attempts to transform correspondence analysis of contingency tables into a model-based approach appropriate for formal inference. In this line of research we have association models and correlation models [15, 34, 35, 41–45] and their extensions to handle ordinal data [36, 83, 87]. On another line we have the development of latent structure models for analyzing a single or a set of multidimensional contingency tables [13, 14, 40, 51, 69]. Finally, it is worth mentioning that the ideas of optimal scaling of the variables can be found in the ACE methodology [8], in the

ALSOS system [106] and in recent developments in discriminant analysis [53, 54].

3. NONLINEAR PRINCIPAL COMPONENT ANALYSIS

In the Gifi system nonlinear PCA is derived as homogeneity analysis with restrictions [22]. The starting point for this derivation is the loss function given in (1.1). However, *rank-1 restrictions* of the form

$$Y_j = q_j\beta_j', \quad j \in \mathbf{J}, \tag{3.1}$$

are imposed on the multiple category quantifications, with q_j being an ℓ_j-column vector of *single* category quantifications for variable j, and β_j a p-column vector of weights (component loadings). Thus, each quantification matrix Y_j is restricted to be of rank-1, which implies that the quantifications in p-dimensional space become proportional to each other. The introduction of the rank-1 restrictions allows the existence of multidimensional solutions for object scores with a single quantification (optimal scaling) for the categories of the variables, and also makes it possible to incorporate the measurement level of the variables (ordinal, numerical) into the analysis. This is impossible in the multiple quantification framework (homogeneity analysis) presented in Section 1. First, consider a multiple treatment of numerical variables. In this case, the quantification of the categories must be the same as the standardized a priori quantification. This implies that multiple numerical quantifications contain incompatible requirements. Second, consider a multiple treatment of ordinal variables. This is not contradictory in itself; however, the different quantifications must have the same order as the prior quantifications, thus resulting in being highly intercorrelated. It follows that such an option does not have much to offer.

To minimize (1.1) under restriction (3.1), we start by computing the $\hat{Y}_j$'s as in (1.4). We then partition the Gifi loss function as follows:

$$\begin{aligned}
&\sum_{j=1}^{J} \operatorname{tr}\bigl(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\bigr)' \\
&\quad \cdot \bigl(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\bigr) \\
&\quad = \sum_{j=1}^{J} \operatorname{tr}\bigl(X - G_j\hat{Y}_j\bigr)'\bigl(X - G_j\hat{Y}_j\bigr) \\
&\qquad + \sum_{j=1}^{J} \operatorname{tr}\bigl(Y_j - \hat{Y}_j\bigr)' D_j \bigl(Y_j - \hat{Y}_j\bigr).
\end{aligned} \tag{3.2}$$

We impose the rank-1 restrictions on the Y_j's and it remains to minimize

$$\sum_{j=1}^{J} \operatorname{tr}\bigl(q_j\beta_j' - \hat{Y}_j\bigr)' D_j \bigl(q_j\beta_j' - \hat{Y}_j\bigr), \tag{3.3}$$

with respect to q_j and β_j. We do this by going to another ALS loop (alternate over q_j and β_j), which gives, for fixed q_j's,

$$\hat{\beta}_j = \bigl(\hat{Y}_j' D_j q_j\bigr)/\bigl(q_j' D_j q_j\bigr), \quad j \in \mathbf{J}, \tag{3.4}$$

and, for fixed β_j's,

$$\hat{q}_j = \hat{Y}_j\beta_j/\bigl(\beta_j'\beta_j\bigr), \quad j \in \mathbf{J}. \tag{3.5}$$

At this point we need to take into consideration the restrictions imposed by the measurement level of the variables. This means that we have to project the estimated vector $\hat{q}_j$ on some cone $\mathscr{C}_j$. In the case of ordinal data the relevant cone C_j is the cone of monotone transformations given by $C_j = \{q_j | q_j(1) \leq q_j(2) \leq \cdots q_j(l_j)\}$. The projection to this cone is solved by a weighted *monotone* regression in the metric D_j (the weights) (see [16] and references therein). In the case of numerical data the corresponding cone is a ray given by $C_j = \{q_j | q_j = \gamma_j + \delta_j s_j\}$, where s_j is a given vector; for example, the original variable quantifications. The projection to this cone amounts to a linear regression problem. However, it can be seen that there is no freedom for choosing q_j different than s_j, and so $\hat{q}_j$ becomes irrelevant. Finally, in the case of nominal data the cone is the $\mathbb{R}^{l_j}$ space and the projection is done by simply setting $q_j = \hat{q}_j$. We then set $\hat{Y}_j = \hat{q}_j\hat{\beta}_j'$ and proceed to compute the object scores. This solution that takes into consideration the measurement level of the variables is referred in the literature [33, 22] as the *Princals solution* (principal component analysis by means of alternating least squares). It can be shown that if all variables are treated as single numerical, the Princals solution corresponds to an ordinary principal component analysis on the s_j variables appropriately standardized (e.g., $u_{l_j}' D_j s_j = 0$ and $s_j' D_j s_j = 1$) [22] (see also [33], Chapter 4). Hence, we have a technique that is invariant under all nonlinear transformations of the variables, and in the special case in which we allow for linear transformations only we get back to ordinary principal components analysis. The Princals model allows the data analyst to treat each variable differently; some may be treated as multiple nominal and some others as single nominal, ordinal or numerical. Moreover, with some additional effort (for details see [76]) one can also incorporate into the analysis categorical variables of mixed measurement level, that is, variables with some categories

measured on an ordinal scale (e.g., Likert scale) and some on a nominal scale (e.g., categories in survey questionnaires corresponding to the answer "not applicable/don't know"). In that sense, Princals generalizes the Homals model.

Therefore, the complete Princals algorithm is given by the following steps:

0. Initialize X, so that $u_N'X = 0$ and $X'X = NI_p$.
1. Estimate the multiple category quantifications by $\hat{Y}_j = D_j^{-1}G_j'X,\ j \in \mathbf{J}$.
2. Estimate the component loadings by $\hat{\beta}_j = (\hat{Y}_j'D_jq_j)/(q_j'D_jq_j'),\ j \in \mathbf{J}$.
3. Estimate the single category quantifications by $\hat{q}_j = \hat{Y}_j\beta_j/(\beta_j'\beta_j),\ j \in \mathbf{J}$.
4. Account for the measurement level of the jth variable by performing a monotone or linear regression.
5. Update the multiple category quantifications by setting $\hat{Y}_j = \hat{q}_j\hat{\beta}_j',\ j \in \mathbf{J}$.
6. Estimate the object scores by $\hat{X} = J^{-1}\sum_{j=1}^{J} G_jY_j$.
7. Column center and orthonormalize the matrix of the object scores.
8. Check the convergence criterion.

In principle, to obtain the minimum of (3.3) over all monotone or linear transformations of the q_j's, steps 2–5 should be repeated until convergence is reached. However, since the value of the loss function will be smaller after a single iteration of the inner ALS loop, a single pass through steps 2–5 is used in practice. The above algorithm is implemented in the Princals program in SPSS [90].

REMARK 3.1. *On the single options.* The most common options in treating variables in Princals are single ordinal and single numerical. The single nominal treatment of a variable makes little sense. A nominal treatment of a variable implies that the data analyst has no a priori idea of how categories should be quantified. If that is the case, then there is no reason in requiring the same quantification on p dimensions. If the data analyst has some prior knowledge, she will be better off by employing one of the other two single options.

We proceed to define the notions of *multiple* and *single loss*. The Gifi loss function can be partitioned into two parts, as follows:

$$\begin{aligned}(3.6)\qquad &\sum_{j=1}^{J}\operatorname{tr}(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j)\\ &\quad + \sum_{j=1}^{J}\operatorname{tr}(\hat{q}_j\hat{\beta}_j' - \hat{Y}_j)'D_j(\hat{q}_j\hat{\beta}_j' - \hat{Y}_j).\end{aligned}$$

Using (1.8), the first term in (3.6) can be also written as $N(p - \sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2)$, which is called the *multiple loss*. The discrimination measure η_{js}^2 is called the *multiple fit* of variable j in dimension s. Imposing the normalization restriction $q_j'D_jq_j = N$, and using the fact that $\hat{Y}_j'D_jq_j\beta_j' = N\beta_j\beta_j'$ (from (3.4)), the second part of (3.6) can be written as

$$\begin{aligned}(3.7)\qquad &\sum_{j=1}^{J}\operatorname{tr}(\hat{Y}_j'D_j\hat{Y}_j - N\beta_j\beta_j')\\ &\quad = N\Bigl(\sum_{j=1}^{J}\sum_{s=1}^{p}(\eta_{js}^2 - \beta_{js}^2)\Bigr),\end{aligned}$$

which is called the *single loss*, since it corresponds to the additional loss incurred by imposing the rank-1 restriction (3.1). The quantities β_{js}^2, $s = 1, \ldots, p$, are called *single fit* and correspond to squared component loadings (see [33], Chapter 4).

From (3.6) it can be seen that if a variable is treated as multiple nominal, it does not contribute anything to the single loss component. Furthermore, two components are incorporated into the single loss part: first, the rank-1 restriction, that is, the fact that single category quantifications must lie on a straight line in the joint space; second, the measurement level restriction, that is, the fact that single quantifications may have to be rearranged to be either in the right order (ordinal variables) or equally spaced (numerical variables). For the mammals' dentition data set the latter would imply that the plots containing the transformed scales (see Figure 9) would only show straight lines with the categories arranged in an increasing or decreasing order for ordinal variables and additionally being equally spaced for numerical variables. Of course, one can immediately see that for binary variables these distinctions are of no essence.

REMARK 3.2. *Nonlinear principal components analysis and eigenvalue problems.* In Section 2.2 we showed that homogeneity analysis under the loss function (1.1) corresponds to an eigenvalue problem, and an ALS algorithm was primarily used for computational efficiency purposes. For the problem at hand an ALS procedure becomes a necessity, because except for the special case where all the variables are treated as single numerical, the problem does not admit an eigenvalue (or a singular value) decomposition. The latter fact also implies that in some cases the ALS algorithm might converge to a local minimum (see [33], Chapter 4).

REMARK 3.3. *Missing data.* In the presence of missing data (3.2) becomes

$$\begin{aligned}
&\sum_{j=1}^{J} \operatorname{tr}(X - G_j Y_j)' M_j (X - G_j Y_j) \\
&\quad = \sum_{j=1}^{J} \operatorname{tr}(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j)))' \\
&\qquad \cdot M_j (X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))) \\
&\quad = \sum_{j=1}^{J} \operatorname{tr}(X - G_j \hat{Y}_j)' M_j (X - G_j \hat{Y}_j) \\
&\qquad + \sum_{j=1}^{J} \operatorname{tr}(Y_j - \hat{Y}_j)' D_j (Y_j - \hat{Y}_j).
\end{aligned} \tag{3.8}$$

This shows that missing data do not affect the inner ALS iteration loop where the single category quantifications and the component loadings are estimated.

3.1 An Example: Crime Rates of U.S. Cities

The data in this example give crime rates per 100,000 people in seven areas—murder, rape, robbery, assault, burglary, larceny, motor vehicle theft—for 1994 for each of the largest 72 cities in the United States. The data and their categorical coding are given in Appendix B. In principal, we could have used homogeneity analysis to analyze and summarize the patterns in this data. However, we would like to incorporate into the analysis the underlying monotone structure in the data (higher crime rates are worse for a city) and thus have treated all the variables as ordinal in a nonlinear principal components analysis. In Figure 10 the component loadings of the seven variables of a two-dimensional solution are shown. In case the loadings are of (almost) unit length, then the angle between any two of them reflects the value of the correlation coefficient between the two corresponding quantified variables. It can be seen that the first dimension (component) is a measure of overall crime rate, since all variables exhibit high loadings on it. On the other hand, the second component has high positive loadings on rape and larceny and negative ones on murder, robbery and auto theft. Thus, the second component will distinguish cities with large numbers of incidents involving larceny and rape from cities with high rates of auto thefts, murders and robberies. Moreover, it can be seen that murder, robbery and auto theft are highly correlated, as are larceny and rape. The assault variable is also correlated, although to a lesser degree, with the first set of three variables and also with burglary. It is interesting to note that not all aspects of violent crimes are highly correlated (i.e., murder, rape, robbery and assault) and the same holds for property crimes (burglary, larceny and auto thefts).

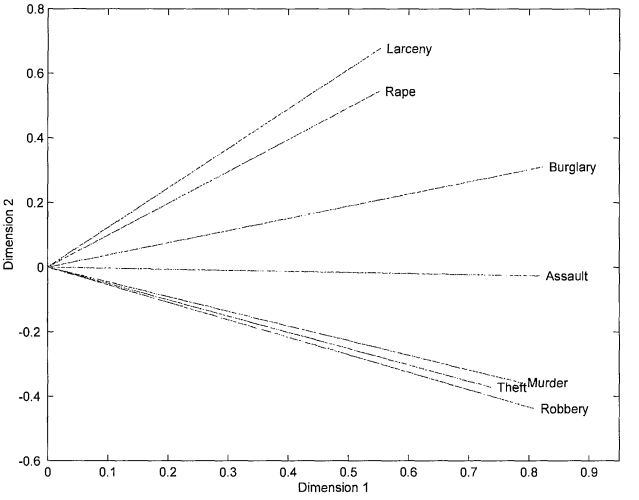

FIG. 10. *Component loadings of the seven crime variables.*

In Figure 11 some of the variable transformations are shown. It can be seen that some variables such as murder and robbery (not shown here) receive linear transformations, while some others (e.g., assault and larceny) distinctly nonlinear ones. It can be seen that the middle categories for some of the variables (e.g., assault and larceny) receive small and approximately equal weights in the optimal solution, thus indicating that they do not contribute much in the direction of maximum variability. Finally, in Figure 12 the variable quantifications along with the object scores are depicted. Notice that all the quantifications lie on straight lines passing through the origin, a result of the imposed rank-1 restriction (3.1). On the right of the graph we find the cities with high crime rates on all seven areas (Tampa, Atlanta, Saint Louis, Miami), and on the left cities with few crime incidents (Virginia Beach, Honolulu, San Jose, El Paso, Raleigh, Mesa, Anaheim). In the lower part of the graph and somewhat to the left there is a cluster of cities that have few rapes and larcenies, but are somewhere in the middle with respect to the other crime areas (New York, Philadelphia, Los Angeles, Long Beach, Houston, San Francisco, Jersey City) and in the lower right cities with many murder, robbery and auto theft incidents (Detroit, Newark, Washington, DC, Oakland, New Orleans, Chicago, Fresno). On the other hand, cities in the upper right part of the graph are characterized by large numbers of rapes, larcenies and bur-

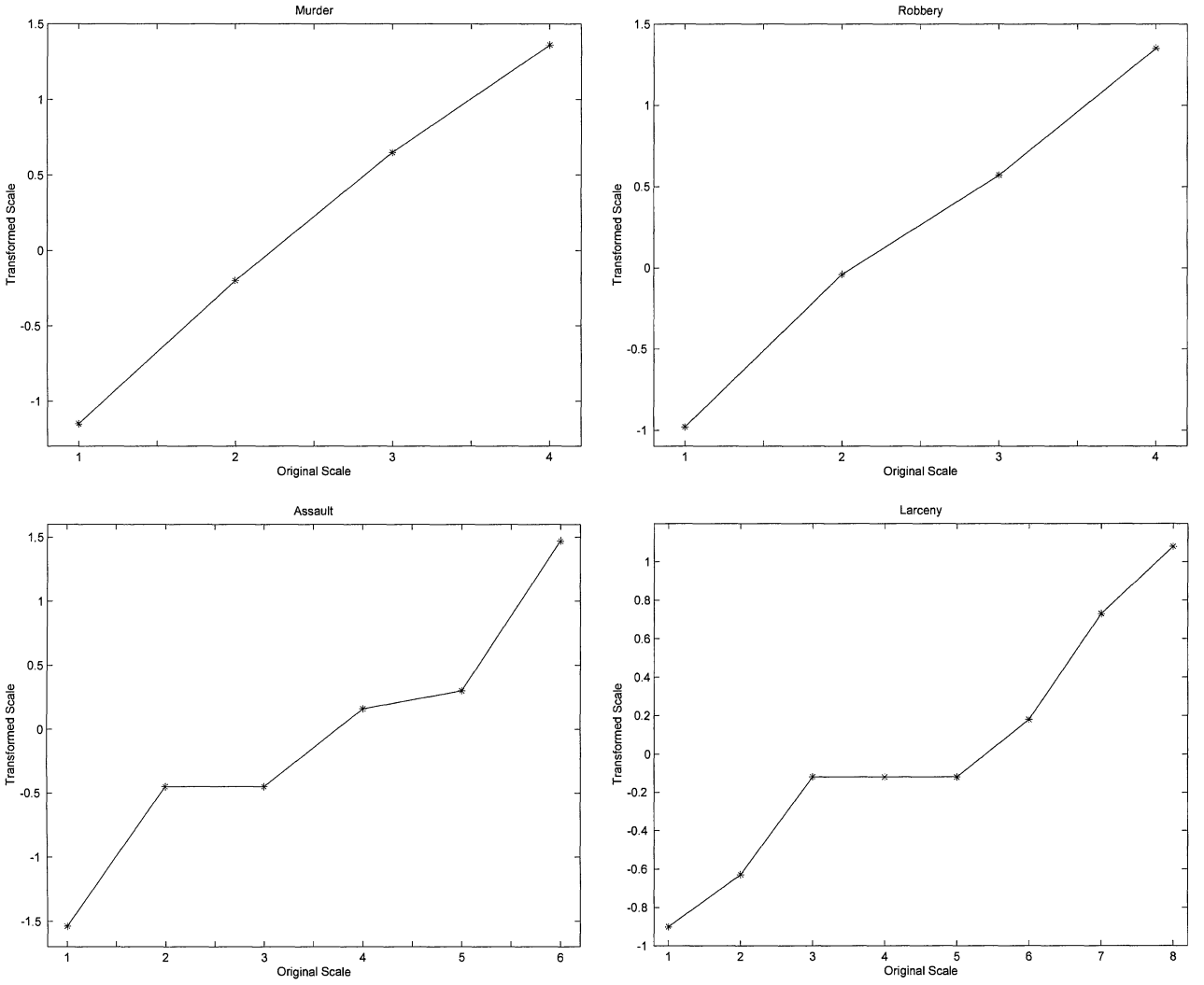

FIG. 11. *Optimal transformations of some of the crime variables.*

glaries (Oklahoma City, Minneapolis, Baton Rouge, Kansas City, Birmingham), while cities in the center are somewhere in the middle with respect to crime (e.g., Denver, Dallas, Las Vegas, Phoenix, Boston, Cleveland and Seattle to name a few). Finally, in the upper left we find a cluster of cities that have somewhat higher numbers of larceny and rape incidents, but few other types of crime incidents (Colorado Springs, Lexington, Anchorage, San Antonio, Akron, Aurora). It should be mentioned that the nature of the original data (numerical variables) makes it possible to run an ordinary principal components analysis (equivalent to treating all the variables as single numerical in the present framework), and many of the patterns discussed above would be present. The use of nonlinear transformations sharpened some of the findings and reduced the effect of some outlier observations.

4. EXTENSION TO MULTIPLE SETS OF VARIABLES

Hotelling's [61] prime goal was to generalize multiple regression to a procedure in which the criterion set contained more than one variable. He proposed a replacement of a set of criterion variables by a new composite criterion that could be predicted optimally by the set of predictor variables. His objective, formulated asymmetrically, was to maximize the proportion of variance in the composite criterion that was attributable to the predictor set. In a subsequent paper Hotelling [62] retained the idea of partitioning the variables into sets, but formulated a more symmetric account of the technique. More specifically, he wanted to study the relationship between two sets of variables after having removed linear dependencies of the variables within each of

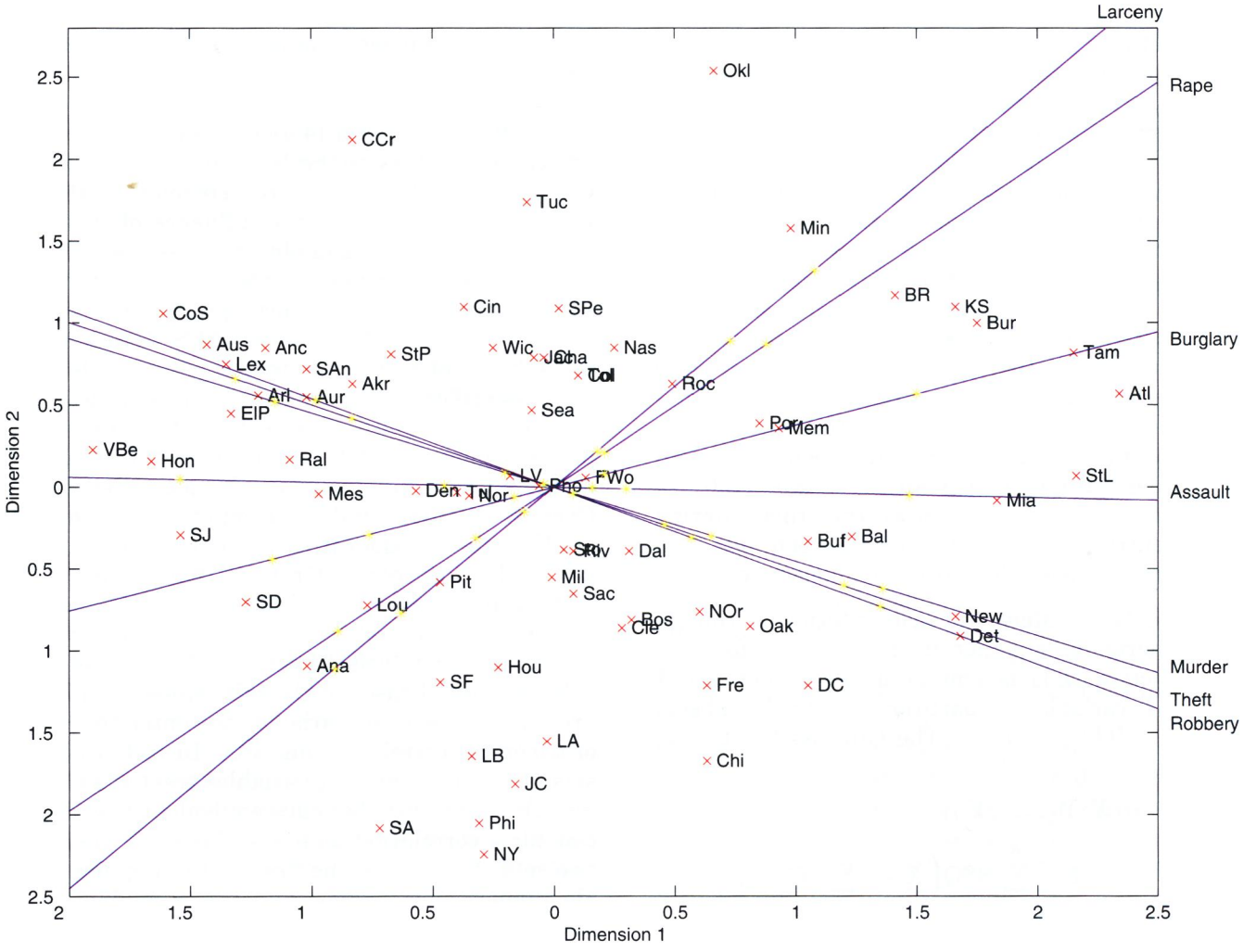

FIG. 12. *Variable quantifications (yellow points, not labeled) and U.S. cities: (solid lines) projections of the transformed variables; the placement of the variable labels indicates the high crime rate direction (e.g., for the assault variable category 1 is on the left side of the graph, while category 6 is on the right side).*

these two sets. Hence, any variable may contribute to the analysis in as much as it provides independent information with respect to the other variables within its own set and to the extent that it is linearly dependent with the variables in the other set. The relationship between the two sets was channeled through a maximum correlation, labeled the canonical correlation, between a linear combination (called canonical variables) of the variables in the first set and a linear combination of the variables in the second one. If the data analyst is interested in more than a single solution, a second pair of canonical variables orthogonal to the first one is to be found, exhibiting the second largest correlation, and the procedure is repeated until a p-dimensional solution is determined. Hotelling's procedure is known in the literature as canonical correlation analysis [37].

Starting with Steel [91], various attempts have been made and approaches suggested to generalize Hotelling's procedure to K sets of variables. In a K-problem there exist $K(K-1)/2$ canonical correlations between the K canonical variables that can be collected in a $K \times K$ correlation matrix $\mathscr{R}$. The generalizations deal with different criteria that can be formulated as functions of the matrix $\mathscr{R}$. The most common ones are as follows:

1. Minimize the determinant of $\mathscr{R}$, or equivalently minimize the product of the eigenvalues of $\mathscr{R}$, proposed by Steel [91].
2. Maximize the sum of the correlations in $\mathscr{R}$, proposed by Horst [59, 60].
3. Maximize the variance of the first principal component of the set of canonical variables, which turns out to be equivalent to maximizing the

largest eigenvalue of $\mathscr{R}$, also proposed by Horst [59, 60].
4. Maximize the sum of squares of the eigenvalues of $\mathscr{R}$, proposed by Kettenring [66].
5. Minimize the smallest eigenvalue of $\mathscr{R}$, also proposed by Kettenring [66].
6. Maximize the sum of squares of the correlations between each canonical variable and an unknown coordinate vector x, proposed by Carroll [12] and discussed by Kettenring [66], who showed that it is equivalent to maximizing the largest eigenvalue of $\mathscr{R}$. The introduction of the comparison vector x brings this criterion close to formulations of homogeneity analysis presented above. This criterion is also discussed in the works of Saporta [86], de Leeuw [18] and van der Burg, de Leeuw and Verdegaal [100]. In van de Geer [96] it is shown that this criterion concentrates on correlations between sets and ignores the possible within-sets structure.

In the Gifi system the last criterion is considered; therefore, a generalization of the familiar loss function (1.1) is employed. The index set $\mathbf{J}$ of the J variables is partitioned into K subsets $J(1), \ldots, J(k), \ldots, J(K)$. The Gifi loss function is given by

$$\sigma(X; Y_1, \ldots, Y_J) = K^{-1} \sum_{k=1}^{K} \mathrm{SSQ}\Big(X - \sum_{j \in J(k)} G_j Y_j\Big), \tag{4.1}$$

subject to the constraints $X'X = NI_p$ and $u_N'X = 0$, where SSQ stands for sum of squares. Equation (4.1) implies that all variables within each set $J(k)$, $k = 1, \ldots, K$, are treated as *additive* (ignoring interactions between variables within the same set). However, the optimal transformations of a variable j within a set $J(k)$ still depend on the optimal transformations of the remaining variables of set $J(k)$. This calls for a correction for the contribution of the other variables and is reflected in the following ALS algorithm.

STEP 1. For given X the optimal Y_j is given by

$$\hat{Y}_j = D_j^{-1} G_j'(X - V_{kj}), \quad j \in \mathbf{J}, \tag{4.2}$$

where $V_{kj} = \sum_{j \in J(k)} G_j Y_j - G_j Y_j$, $k = 1, \ldots, K$, $j \in \mathbf{J}$.

STEP 2. For given Y_j's, the optimal X is given by

$$\hat{X} = K^{-1} \sum_{k=1}^{K} \sum_{j \in J(k)} G_j Y_j. \tag{4.3}$$

STEP 3. The object scores are column centered and orthonormalized in order to satisfy the normalization constraints.

Equations (4.2) and (4.3) illustrate the centroid principle, which is at the heart of the Gifi system. Category quantifications are centroids of the object scores corrected for the influence of the other variables in the set, and object scores are averages of quantified variables. In the presence of rank-1 restrictions for the category quantifications (i.e., $Y_j = q_j \beta_j'$, $j \in \mathbf{J}$) an inner ALS iteration loop must be employed for estimating the single category quantifications q_j and the component loadings β_j (see Section 3). The restricted minimization problem given by (4.1) is known in the literature as the Overals problem, the ALS algorithm as the Overals algorithm and the computer program that implements the algorithm as the Overals program [90, 103]. The reason for these particular names is that if we consider a single variable per set, (4.1) reduces to (1.1), the ordinary loss function for homogeneity analysis. Therefore, Homals and Princals are special cases of Overals. Moreover, if there are only two sets of variables we enter the realm of canonical correlation analysis. In fact, with two sets of variables and all variables treated as single numerical, Overals becomes equivalent to ordinary canonical correlation analysis. Finally, if there are two sets of variables, the first containing many single numerical variables, and the second a single categorical variable, Overals can be used to perform canonical discriminant analysis.

REMARK 4.1. *Overals as an eigenvalue problem.* Following steps analogous to those considered in Section 2.2, it can be shown that the minimum of the loss function is given by

$$\sigma(\star, \star) = NK\Big(p - \sum_{s=1}^{p} \lambda_s\Big), \tag{4.4}$$

where λ_s, $s = 1, \ldots, p$ are the eigenvalues of the matrix $\mathscr{L} P_\star \mathscr{L}$, with $P_\star = K^{-1} \sum_{k=1}^{K} P_k$ and $P_k = \sum_{j \in J(k)} G_j D_j^{-1} G_j'$. Therefore, the minimum loss is a function of the p largest eigenvalues of the average projector of the K subspaces spanned by the columns of the matrices $\sum_{j \in J(k)} G_j$.

4.1 An Example: School Climate

The data for this large example (23,248 students) come from the National Education Longitudinal Study of 1988 (NELS:88). A description of the 11 variables used, their coding and their univariate marginals are given in Appendix C. We provide

next some motivation on the issues related to this particular data set. Recently, there has been a lot of interest among researchers and policy makers on the importance of the school learning environment and the influence of individual and peer behavior on student performance. For example goal 6 of the National Education Goals Panel [78] states that by the year 2000 "every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning." Because in many situations learning is constrained in an atmosphere of fear and disorderliness, student behavior influences school atmosphere and the climate for learning (whether it takes the form of violence and risk-taking activities such as bringing weapons to school or using alcohol and drugs) or a low commitment to academic effort (such as poor attendance, lack of discipline or study habits) [11]. These student behaviors also play a key role in determining student success in school and beyond (see [64] and references therein), as well as the way students, teachers and administrators act, relate to one another and form their expectations and to a certain extent beliefs and values [1, 82]. Thus, this particular set of variables from NELS:88 addresses issues directly related to the school culture and climate, as seen from the students' point of view.

The variables were divided into three sets $J(1) = (A - C)$, $J(2) = (D - I)$, $J(3) = (J, K)$ and treated as single ordinal, since lower values indicate that the students perceive the particular problem as more serious in their school. The first set characterizes attendance patterns of students, the second set deals with issues that affect the overall school environment and the third deals with attitudes of students toward their teachers. The eigenvalues (measure of fit of the solution) for the two-dimensional solution are 0.65 and 0.36, respectively, and the multiple correlations of the three sets with each mean canonical variable (dimension) are 0.67, 0.86 and 0.75 for the first dimension and 0.45, 0.69 and 0.58 for the second one. The multiple correlations are a measure of the amount of variance of the canonical variable explained by the particular set. It can be seen that the second set does the best job, followed by the third and the first set in both dimensions.

The component loadings of the variables are shown in Figure 13. The loadings of all variables are pretty high on the first dimension (canonical variable), with the exception of variable B (student absenteeism). On the other hand, variables E (robbery), H (drugs), I (weapons) and J (physical abuse of teachers) load positively on the second canonical variable, while variables A (tardiness), D (physical

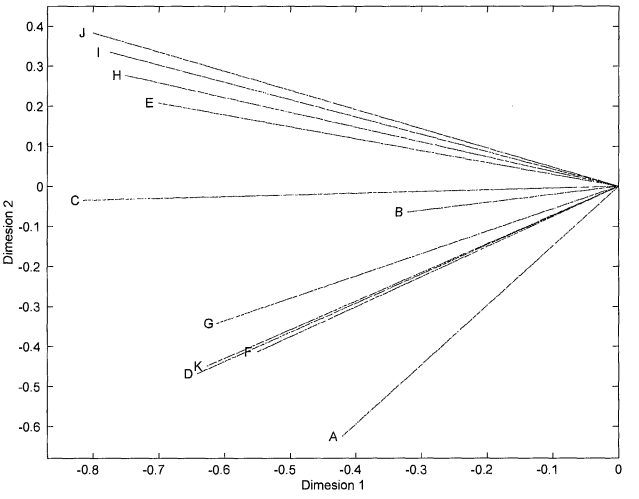

FIG. 13. *Component loadings of the variables in the school climate example.*

conflicts), F (vandalism), G (alcohol) and K (verbal abuse of teachers) load negatively on the second canonical variable. It is interesting to observe that the robbery, drugs and weapons variables are fairly highly correlated, while the variables on physical and verbal abuse of the teachers are uncorrelated. In general, the first canonical variable can be interpreted as an overall measure of school climate, while the second one distinguishes between students (and consequently schools) that experience a rough and potentially dangerous school environment, with those that experience a simply rough environment. Finally, it is worth noting that from the first set only the tardiness variable discriminates along the second dimension, a fact reflected on the low multiple correlation coefficient of the first set of variables and the second canonical variable.

By examining the category points plot (see Figure 14) we expect to find the students attending schools where the overall climate is perceived to be poor in the right part of the picture and, in particular, in the lower right, those that believe the climate at their school is good in the lower left, those going to schools with some sort of problems somewhere in the middle and those going to schools with problems described by variables A, D, F, G and K in the upper middle part of the picture. In Figure 15 the object scores are plotted, together with their frequencies and the category points (positioned along the red lines). The object point with the very large spike on the left end of the graph corresponds to the profile consisting of only 4's; that is, these students (approximately 11% of the sample) indicated that none of the areas covered by the 11 variables is a problem in their school.

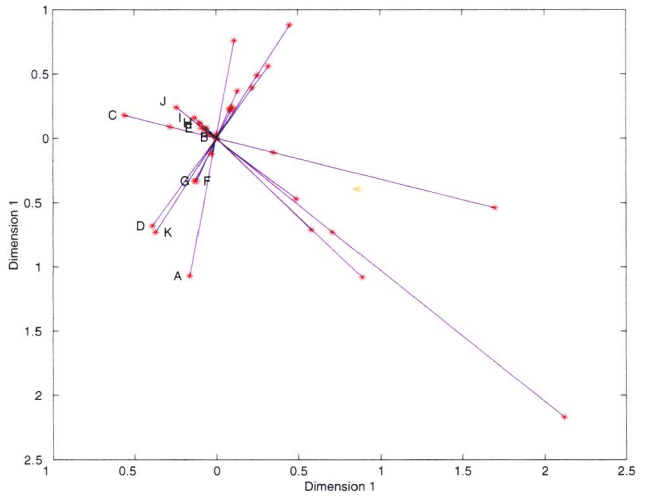

FIG. 14. *Category points of the school climate data set (variable label points toward the highest, i.e., "not a problem," category).*

On the other hand, the big spike on the right end of the picture corresponds to the other extreme profile of only 1's; hence, close to 2% of the students attend pretty rough schools. It can also be seen that approximately 40% of the object scores are concentrated on the lower left part, indicating that these areas can be characterized as at most a minor problem in the respective schools. Overall, the solution reveals that very few students mixed "not a problem" responses with "serious problem" ones; such students can be found in the center-lower part of Figure 13. On the other hand, the majority of the students used "neighbor" responses (e.g., "not a problem" and "minor problem" or "moderate" and "serious" problem), thus resulting in the observed clustering of students attending problem-free schools in the lower left of the graph, those attending somewhat rough schools in the upper left and finally those attending really rough schools in the lower right of the graph. The object scores can be subsequently used as input in other types of analysis; for example, in regression analysis they can be related to outcome variables such as test scores and used to examine relevant hypotheses.

5. STABILITY ISSUES

The techniques presented so far aim at the uncovering and representation of the structure of categorical multivariate data. However, there has been no reference to any probabilistic mechanism that generated the data under consideration. The

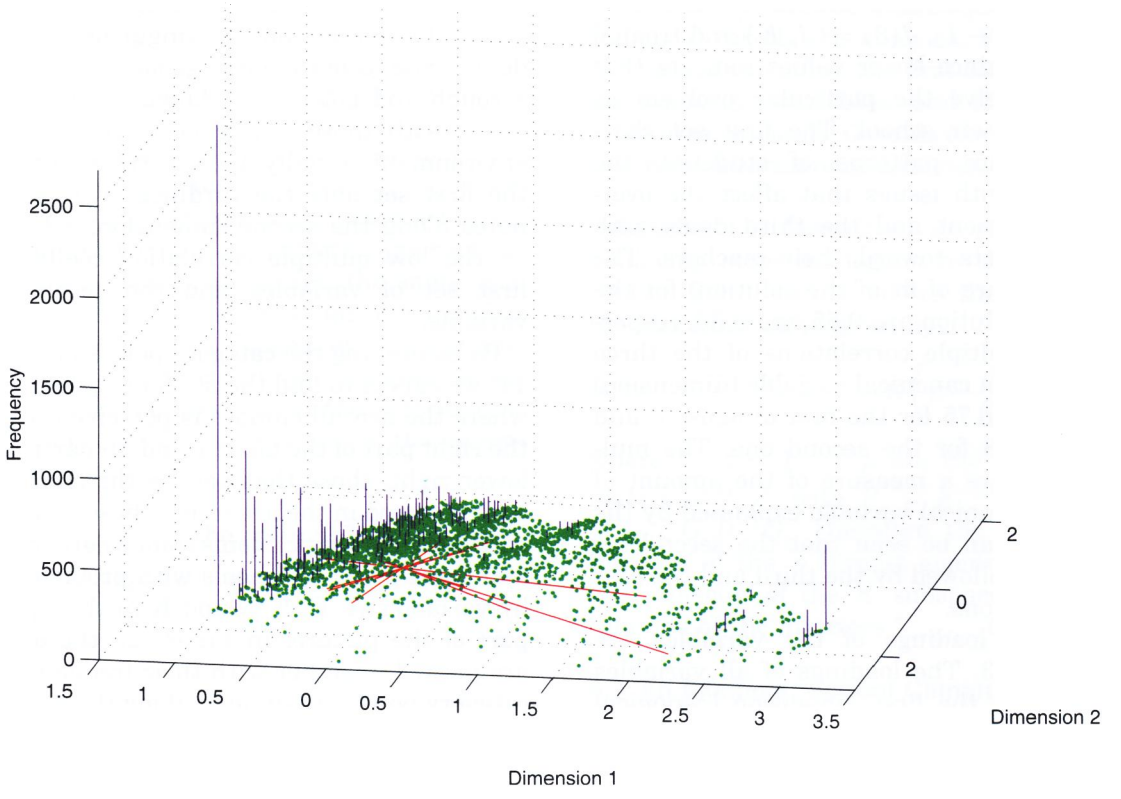

FIG. 15. *Object scores of the school climate data set and their respective frequencies: (green points) frequencies less than or equal to 5; (red lines) location of the category points on the map.*

main focus of these techniques is on providing a low-dimensional representation of the original high-dimensional space (where presumably the dependencies and interdependencies in the data are much easier to describe and characterize). As Kendall points out "many of the practical situations which confront us are not probabilistic in the ordinary sense.... It is a mistake to try and force the treatment of such data into a classical statistical mould, even though some subjective judgment in treatment and interpretation may be involved in the analysis" (see [65], page 4). Nevertheless, one may always pose the question of whether the patterns in the various plots are real or mere "chance" effects. Thus, the goal in this section is to give a brief overview of the question of *stability* of the representation. The concept of stability is central in the Gifi system and is used in the following sense: data analysis results are stable when small and/or unimportant changes of *input* lead to small and unimportant changes in the results (*output*) (see [33], page 36). By reversing this definition, we get that in the Gifi system a result can be characterized as *unstable* in those instances where small and unimportant changes in the input result in significant changes in the output. In our analysis, we consider as input the data set at hand (objects and variables), the coding of the variables, the dimensionality of the solution, the measurement level of the variables, the type of technique employed (Homals, Princals), and as output category quantifications, object scores, discrimination measures, eigenvalues, component loadings and so on. It should be noted that, while in other types of data analysis the output usually consists of a small number of point estimates and their standard errors (e.g., regression analysis), for the techniques under consideration there exists a whole series of outputs.

The most important forms of stability relevant to the techniques previously presented are the following:

(a) *Replication stability*—if a new data set is sampled and we apply the same technique to this new set, then the results do not change dramatically.
(b) *Stability under data selection*—variations in the data are considered (omitting either objects from the data set or variables from subsequent analysis) and the stability of the analysis is considered.
(c) *Stability under model selection*—small changes in the model result in small changes in the results obtained.
(d) *Numerical stability*—rounding errors and computation with limited precision do not greatly influence the results given by the techniques.
(e) *Analytic and algebraic stability*—refers to the derivation of analytical results for the output (e.g., error bounds on the parameters of interest) by considering "perturbations" of the input.
(f) *Stability under selection of technique*—application of a number of different techniques to the same data set, aiming at answering the same question, results in approximately the same information.

In this section, we focus primarily on stability under data selection. However, we also look briefly into analytic and algebraic stability. It should be noted that issues of numerical stability have been addressed during the presentation of the various models (e.g., normalization issues, etc.).

The distinction between *internal* and *external* stability may provide a better understanding of the concept of stability [48] as used in the Gifi system. External stability refers to the conventional notions of statistical significance and confidence. In the conventional statistical framework, the aim of the analysis is to get a picture of the empirical world and the question is to what extent the results do indeed reflect the real population values. In other words, the results of any of our models are externally stable in case any other sample from the same population produces roughly the same results (output). Consequently, the confidence regions of the output parameters are relatively small. Internal stability deals with the specific data set at hand. The models produce a simple representation of the data and reveal associations between the variables. An internally stable solution implies that the derived results give a good summary of that specific data set. We are not interested in population values, because we might not know either the population from which the data set was drawn or the sampling mechanism; in the latter case, we might be dealing with a sample of convenience. Possible sources of instability in a particular data set are outlying objects or categories that have a large influence on the results. Internal stability can be thought of as a form of robustness.

Both external and internal stability play a role in the practice of data analysis. It is often the case that a data analyst wants to get insight into the structure of the population, particularly whenever the data set has been collected by a well-defined sampling mechanism. In such cases, external stability of the results allows the practitioner to draw firmer conclusions about the structure of the un-

derlying population. On the other hand, when a data analyst is only interested in the structure of the specific data set, internal stability ensures the invariance of the sample solution. The notions of external and internal stability are directly linked with the notions of descriptive and inferential use of the models introduced in the previous sections. The main distinction between these two notions is whether the models are (i) exclusively used to reduce the complexity of the data and uncover their basic structure or (ii) used to draw conclusions and generalize them from the sample to the population (see the debate between de Leeuw [20] and Molenaar [77] and references therein; see also Leamer [70] for some provocative thoughts on the subject from the Bayesian viewpoint). When dealing with external stability a "new" sample should be drawn and the technique applied to it. The word "new" may mean (i) a truly new sample from the population, (ii) a fictitious "new" sample, common in classical statistical analysis of stability [22], or (iii) a "new" perturbed sample created by resampling with replacement from the sample at hand. In case of a fictitious "new" sample, the data analyst attempts to assess stability of the technique by examining what would have happened if a truly "new" sample was drawn from the underlying population. When dealing with internal stability, only the third possibility is available.

In the remainder of this section we will address stability issues related to merging categories, omitting a variable from the analysis, permutation tests and the bootstrap. The first two topics have been briefly addressed in [33], the third one in [21], while the last one has been examined in [19], [33], [73], [74] and [98].

5.1 Merging Categories

Merging categories can be formalized algebraically by introducing indicator matrices G_{C_j}, $j \in \mathbf{J}$, of dimension $\ell_j \times k_j$, with $k_j \leq \ell_j$, to replace G_j by $G_j G_{C_j}$. In the case $k_j = \ell_j$, we get that $G_{C_j} \equiv I_{\ell_j}$ and nothing changes. The orthogonal projector on the subspace spanned by the columns of $G_j G_{C_j}$ becomes $\tilde{P}_j = G_j G_{C_j} (G'_{C_j} D_j G_{C_j})^+ G'_{C_j} G'_j$, $j \in \mathbf{J}$, where A^+ denotes the Moore–Penrose (generalized) inverse of matrix A. By using perturbation results for eigenvalue problems (see, e.g., [63]), it can be shown that merging two categories with approximately the same quantifications hardly changes the eigenvalues (and hence the overall fit of the solution) when the number of variables is not too small. On the other hand, eliminating a very low frequency category (by merging it with some other category) improves, in general, the graphical representation of the solution.

5.2 Eliminating a Variable in Homogeneity Analysis

It can be shown that the importance of a variable for a dimension s can be expressed as $\eta^2_{js} - \gamma_s$, $s = 1, \ldots, p$, that is, the discrimination measure of that variable minus the average of the discrimination measures of all variables on dimension s (the eigenvalue). The latter implies that if a variable with a relatively small discrimination measure is eliminated from the analysis, the overall fit of the solution (eigenvalue) will not be affected much. Results for eigenvectors (object scores) are less complete, and it seems that a general pattern is hard to establish.

5.3 Permutation Methods

Although we emphasized the exploratory nature of the techniques described in this paper, nevertheless we would like to determine whether the structure observed in the data is "too pronounced to be easily explained away as some kind of fluke," to paraphrase Freedman and Lane [32]. Permutation tests can help to study the concept of "no structure at all." The idea behind using such tests is that they represent a nice way of formalizing the notion of no structure. The random variation is introduced *conditionally on the observed data*, which implies that we do not have to assume a particular model that generated the data, thus making them useful in nonstochastic settings as well [32]. Each new data set is generated by permuting the values the objects are assigned for each variable, resulting in destroying the original profiles of the objects. Then, the technique of interest is applied to the newly generated data and the eigenvalues of the solution computed. For small data sets in terms of both objects and variables (e.g., $J = 2$) it is possible to derive the permutation distribution of the eigenvalues by complete enumeration of the possible cases. However, for all other cases one has to resort to Monte Carlo methods [21].

We present next the results of such a test for the mammals' dentition example. The two panels of Figure 16 give the frequency distribution of the first and second eigenvalues of the homogeneity analysis solution for the mammals dentition example over 1000 replications. It can immediately be seen that the observed eigenvalues of 0.73 and 0.38 in the original data are way to the right, thus suggesting that the informal null hypothesis of absence of structure is false and hence the patterns in the data (e.g., various groupings of the mammals) are real.

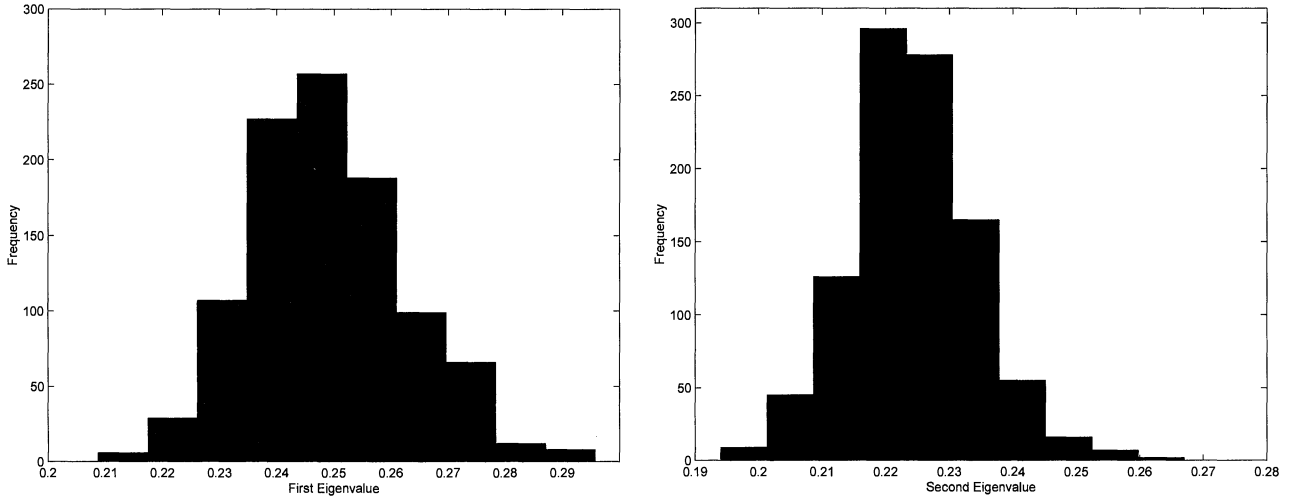

FIG. 16. *Histograms of the permutation distribution of the eigenvalues of the homogeneity analysis solution of the mammals' dentition example.*

5.4 Replication Stability and the Bootstrap

The previous subsections have provided some analytic and/or computational results on the stability of the eigenvalues of the Homals solution. However, very little is known (and analytic results seem very hard to get) for category quantifications, objects scores, component loadings and so on. On the other hand, recent advances in resampling techniques—particularly the bootstrap [26]—offer an interesting and useful alternative to study stability issues for other parameters of interest.

We develop the method in a general context (for a more comprehensive account see also [98]). Suppose we have J categorical variables. Each variable takes values in a set $\mathscr{S}_j$ (the range of the variable [33]) of cardinality ℓ_j (number of categories of variable j). Define $\mathscr{S} = \mathscr{S}_1 \times \cdots \mathscr{S}_J$ to be the *profile space*, which has cardinality $\ell = \prod_{j=1}^{J} \ell_j$. That is, the space $\mathscr{S} = \{(s_1, \ldots, s_j),\ s_j \in \mathscr{S}_j,\ j \in \mathbf{J}\}$ contains the J-tuples of profiles. Let S be an $\ell \times \sum_{j=1}^{J} \ell_j$ binary matrix, whose elements $S(h, t)$ are equal to 1 if the hth profile contains category t, and 0 otherwise; that is, S maps the space of profiles $\mathscr{S}$ to its individual components. Let also G_S be an $N \times \ell$ indicator matrix with elements $G_S(t, h) = 1$ if the tth object (individual etc.) has the hth profile in S, and $G_S(t, h) = 0$ otherwise. The superindicator matrix $G = [G_1|\cdots|G_J]$ can now be written as $G = G_S S$. Hence, there is a one-to-one correspondence between the ordinary indicator matrices G_j and the space of profiles $\mathscr{S}$.

Let $\mathscr{P}$ be a probability distribution on $\mathscr{S}$. Since the space $\mathscr{S}$ is finite, $\mathscr{P}$ corresponds to a vector of proportions $p = \{p_h\}$ with $\sum_{h=1}^{\ell} p_h = 1$. In the present framework, it is not difficult to see that each observed superindicator matrix G corresponds to a realization of the random variable π that has a multinomial distribution with parameters (N, p). The output (category quantifications, discrimination measures, object scores, component loadings etc.) of the techniques introduced in the previous sections can be thought of as functions $\phi(\pi)$.

From a specific data set of size N we can draw N^N sets also of size N, with replacement. In our case, each subset corresponds to a matrix G_S. The basic idea behind bootstrapping techniques is that we might as well have observed any matrix G_S of dimension $N \times \ell$ consisting of the same rows, but in different frequencies, than the one we observed in our original sample. So, we could have observed a superindicator matrix G^m, associated with a vector of proportions p_m, which is a perturbed version of π. The output of our techniques would then naturally be a function $\phi(p_m)$. Suppose that we have a sequence of p_m's and thus of functions $\phi(p_m)$. Then, under some regularity conditions on the $\phi(\cdot)$ (Hadamard or Fréchet differentiability [89]) it can be shown that $\phi(p_m)$ is a consistent estimator of $\phi(\pi)$ and that $P_*(\phi(p_m) \leq z|p_m)$ is a consistent estimator of $P(\phi(p) \leq z|p)$ [72, 88], where P_* denotes the conditional probability given p_m. The previous discussion indicates that the appropriate way to bootstrap in homogeneity analysis is to sample objects with replacement, or in other words, sample rows of the data matrix.

REMARK 5.1. *Bias correction and construction of confidence regions.* Two of the main issues in the theory of the bootstrap are (i) how to produce *unbi-*

ased bootstrap estimates and (ii) how to construct *confidence regions* with the correct coverage probability α [26]. The main problem in the present context is that by construction the parameters of the techniques (eigenvalues, category quantifications etc.) are multidimensional, and moreover the dimensions are correlated with each other. Regarding bias correction two possible solutions proposed by Markus [73] are (i) to adjust each bootstrap point by $\hat{\phi}^*_{c_b} = \hat{\phi}^*_b - 2(\hat{\phi}^* - \hat{\phi})$ and (ii) to adjust by $\hat{\phi}^*_{c_b} = 2\hat{\phi} - \hat{\phi}^*_b$, where $\hat{\phi}^*_b$ corresponds to the bth bootstrap point, $\hat{\phi}$ to the sample estimate and $\hat{\phi}^*$ to the mean of the B bootstrap points. The first one defines bias as a shift of the estimate with respect to the population value; the second, as a reflection with respect to the original sample value. Regarding the problem of constructing confidence regions, several approaches have been suggested in the literature. Weinberg, Carroll and Cohen [104] constructed ellipses based on the bootstrap variance–covariance matrix. They assumed that the sampling distribution is normal and the construction of confidence regions is based on F-values. A similar approach can be found in Takane and Shibayama [92]. Heiser and Meulman [56] suggested constructing ellipses by performing a singular value decomposition of the matrix of bootstrap points that are in deviations from their means. This procedure results in a spherical representation that determines the circle covering the $(1-\alpha)\times 100\%$ points with the shortest distance to the centroid. Subsequently, the circle is transformed into an ellipse. This construction avoids any link to the normal distribution. Markus [73] uses the convex hull of the scatter of the bootstrap points to construct the confidence regions. She then discards the $\alpha \times 100\%$ of the outer vertices of the hull, and the resulting hull is considered to be the desired confidence region (this algorithm is discussed in [46]; see also [47]). This method resembles the percentile method for estimating bootstrap confidence intervals [26].

5.5 Results of Previous Studies

There have been several studies that have used bootstrap methods to assess the stability of nonlinear multivariate techniques—homogeneity analysis, correspondence analysis, canonical correlation analysis [33, 98, 74]. The most comprehensive one is the monograph by Markus [73]. In this section we will attempt to summarize briefly the results of these studies:

1. The most important finding from a computational point of view is that to obtain valid results a large number of bootstrap replications is required (over 1000).
2. The bootstrap confidence regions give on average the right coverage probabilities. However, for categories with low marginal frequencies the coverage probabilities might be underestimated or overestimated.
3. Bias correction is beneficial to the coverage probabilities of eigenvalues but rather harmful to that of category quantifications, discrimination measures and component loadings. It seems that the translation method is the most appropriate for bias correction.
4. Marginal frequencies of about 8 seem to be the absolute minimum to ensure valid confidence regions. In light of this finding, merging categories appears to be not only beneficial overall, but necessary in many situations.
5. Both ellipses and peeled convex hulls produce valid confidence regions. However, this result heavily depends on a number of parameters, such as sample size, number of bootstrap replications, category marginal frequencies. In the case of small sample sizes, the behavior of confidence regions becomes erratic.
6. There are no results regarding the stability of patterns for ordinal and/or numerical variables (ordering of categories), and also in the presence of missing data.

In Figure 17 we present the bias corrected bootstrap means, along with ± 2 standard error bands, of some of the optimal transformations of the category quantifications of the mammals' dentition data set, based on 1000 bootstrap replications. The bootstrapped means and standard errors (in parentheses) of the eigenvalues in dimensions 1 and 2 are 0.738 (0.035) and 0.386 (0.027), respectively.

It can be seen that the fit of the solution in both dimensions is particularly stable, thus indicating that the patterns observed in the data set are real, thus confirming the results of the random permutation method. Regarding the category quantifications we see that the first dimension exhibits a far more stable behavior than the second. However, for the variables that discriminate along the second dimension (TI, BI) the results are satisfactory. Moreover, we see that categories with low marginal frequencies (e.g., BI1) exhibit more variation than categories with larger frequencies, thus confirming the results of previous studies.

6. CONCLUDING REMARKS

In this paper a brief account of some varieties of multivariate analysis techniques for categorical

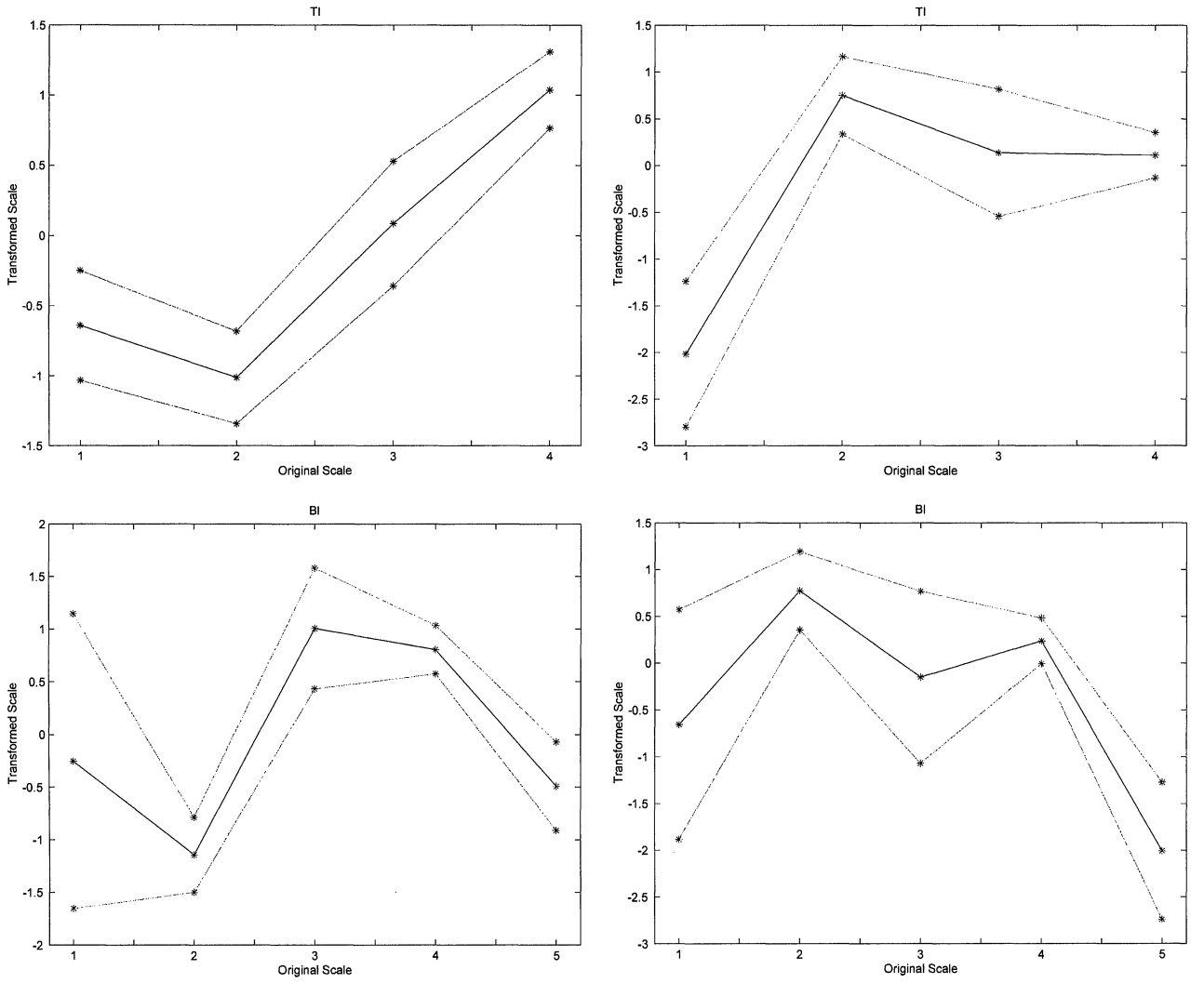

FIG. 17. *Bias corrected bootstrap means of the optimal transformations (solid lines) and respective error bands (broken lines) of some of the variables of the mammals' dentition example.*

data, known as the Gifi system, is given. The central themes of the system are the notion of optimal scaling of categorical data and its implementation through alternating least squares algorithms. The starting point of the system is homogeneity analysis, a particular form of optimal scaling. The use of various types of restrictions allows homogeneity analysis to be molded into other types of nonlinear multivariate techniques. These techniques have been extensively used in data analytic situations. In the Gifi book [33] all of Chapter 13 is devoted to applications covering the fields of education, sociology and psychology. Also, in their books Greenacre [48] and Benzécri [4] give a wide variety of applications of multiple correspondence analysis in the fields of genetics, social psychology, clinical research, education, criminology, linguistics, ecology, paleontology and meteorology. Other applications of the techniques include marketing [58], zoology [97], environmental studies [94], medicine [98] and food science [99]. However, the Gifi system has evolved beyond homogeneity analysis and its generalizations; hence, new techniques have been developed for path models [33], time series models [93], linear dynamical systems [5] and so on. In closing, it should be mentioned that the Gifi system is part of a still quite active research program.

APPENDIX A: DENTITION OF MAMMALS

The data (Table A1) for this example are taken from Hartigan [52]. Mammals' teeth are divided into

TABLE A1
Mammals' dentition data

Mammal		Mammal	
Opossum	45224422	Fox	44225512
Hairy tail mole	44225522	Bear	44225512
Common mole	43214422	Civet cat	44225511
Star nose mole	44225522	Raccoon	44225521
Brown bat	34224422	Marten	44225511
Silver hair bat	34223422	Fisher	44225511
Pigmy bat	34223322	Weasel	44224411
House bat	34222322	Mink	44224411
Red bat	24223322	Ferrer	44224411
Hoary bat	24223322	Wolverine	44225511
Lump nose bat	34223422	Badger	44224411
Armadillo	11111122	Skunk	44224411
Pika	32113322	River otter	44225411
Snowshoe rabbit	32114322	Sea otter	43224411
Beaver	22113222	Jaguar	44224311
Marmot	22113222	Ocelot	44224311
Groundhog	22113222	Cougar	44224311
Prairie dog	22113222	Lynx	44224311
Ground squirrel	22113222	Fur seal	43225511
Chipmunk	22113222	Sea lion	43225511
Gray squirrel	22112222	Walrus	21224411
Fox squirrel	22112222	Grey seal	43224411
Pocket gopher	22112222	Elephant seal	32225511
Kangaroo rat	22112222	Peccary	34224422
Pack rat	22111122	Elk	15214422
Field mouse	22111122	Deer	15114422
Muskrat	22111122	Moose	15114422
Black rat	22111122	Reindeer	15214422
House mouse	22111122	Antelope	15114422
Porcupine	22112222	Bison	15114422
Guinea pig	22112222	Mountain goat	15114422
Coyote	24225522	Musk-ox	15114422
Wolf	44225512	Mountain sheep	15114422

four groups: incisors, canines, premolars and molars. A description of the variables with their respective coding is given next:

- TI, top incisors—(1) zero incisors; (2) one incisor; (3) two incisors; (4) three or more incisors;
- BI, bottom incisors—(1) zero incisors; (2) one incisor; (3) two incisors; (4) three incisors; (5) four incisors;
- TC, top canine—(1) zero canines; (2) one canine;
- BC, bottom canine—(1) zero canines; (2) one canine;
- TP, top premolar—(1) zero premolars; (2) one premolar; (3) two premolars; (3) two premolars; (4) three premolars; (5) four premolars;
- BP, bottom premolar—(1) zero premolars; (2) one premolar; (3) two premolars; (3) two premolars; (4) three premolars; (5) four premolars;
- TM, top molar—(1) zero, one or two molars; (2) more than two molars;
- BM, bottom molar—(1) zero, one or two molars; (2) more than two molars.

Table A.2 gives the frequencies of the variables.

TABLE A2
Mammals' teeth profiles (in percent, $N = 66$)

	Categories				
Variable	1	2	3	4	5
TI	15.2	31.8	13.6	39.4	
BI	3.0	30.3	7.6	43.9	15.2
TC	40.9	59.1			
BC	45.5	54.5			
TP	9.1	10.6	18.2	39.4	22.7
BP	9.1	18.2	15.2	36.4	21.2
TM	34.8	65.2			
BM	31.8	68.2			

APPENDIX B: CRIME RATES IN U.S. CITIES IN 1994

The data for this example are taken from table No. 313 of the 1996 Statistical Abstract of the United States. The coding of the variables is given next:

- murder—(1) 0–10; (2) 11–20; (3) 21–40; (4) 40+;
- rape—(1) 0–40; (2) 41–60; (3) 61–80; (4) 81–100; (5) 100+;
- robbery—(1) 0–400; (2) 401–700; (3) 701–1000; (4) 1000+;
- assault—(1) 0–300; (2) 301–500; (3) 501–750; (4) 751–1000; (5) 1001–1250; (6) 1251+;
- burglary—(1) 0–1000; (2) 1001–1400; (3) 1401–1800; (4) 1801–2200; (5) 2200+;
- larceny—(1) 0–3000; (2) 3001–3500; (3) 3501–4000; (4) 4001–4500; (5) 4501–5000; (6) 5001–5500; (7) 5501–7000; (8) 7000+;
- motor vehicle theft—(1) 0–500; (2) 501–1000; (3) 1001–1500; (4) 1501–2000; (5) 2000+.

The data along with the city codes used in Figure 12 are given in Table B1.

APPENDIX C: SCHOOL CLIMATE

A description of the variables is given next:

- A—student tardiness is a problem at school;
- B—student absenteeism is a problem at school;
- C—students cutting class is a problem at school;
- D—physical conflicts among students is a problem at school;
- E—robbery or theft is a problem at school;
- F—vandalism of school property is a problem at school;
- G—student use of alcohol is a problem at school;
- H—student use of illegal drugs is a problem at school;
- I—student possession of weapons is a problem at school;
- J—physical abuse of teachers a problem at school;
- K—verbal abuse of teachers is a problem at school.

TABLE B1
Crime rates data

City	City code		City	City code	
New York (NY)	NY	3134213	Los Angeles (CA)	LA	3235223
Chicago (IL)	Chi	3*46343	Houston (TX)	Hou	3223323
Philadelphia (PA)	Phi	3232114	San Diego (CA)	SD	1113223
Phoenix (AZ)	Pho	3213464	Dallas (TX)	Dal	3424354
Detroit (MI)	Det	4546445	San Antonio (TX)	SAn	2211362
Honolulu (HI)	Hon	1111252	San Jose (CA)	SJ	1213112
Las Vegas (NV)	LV	2323333	San Francisco (CA)	SF	2133253
Baltimore (MD)	Bal	4445474	Jacksonville (FL)	Jac	2424462
Columbus (OH)	Col	2522453	Milwaukee (WI)	Mil	3322244
Memphis (TN)	Mem	3533534	Washington, DC	DC	4246363
El Paso (TX)	ElP	1213152	Boston (MA)	Bos	2435245
Seattle (WA)	Sea	2223373	Charlotte (NC)	Cha	2325462
Nashville (TN)	Nas	2425373	Austin (TX)	Aus	1211262
Denver (CO)	Den	2312323	Cleveland (OH)	Cle	3533314
New Orleans (LA)	NOr	4433444	Fort Worth (TX)	FWo	3423363
Portland (OR)	Por	2426375	Oklahoma City (OK)	Okl	2514583
Long Beach (CA)	LB	2133324	Tucson (AZ)	Tuc	1314384
Kansas City (MO)	KS	3536574	Virginia Beach (VA)	VBe	1111131
Atlanta (GA)	Atl	4546585	Saint Louis (MO)	StL	4346585
Sacramento (CA)	Sac	2223455	Fresno (CA)	Fre	3234455
Tulsa (OK)	Tul	2314323	Miami (FL)	Mia	3246585
Oakland (CA)	Oak	3445454	Minneapolis (MN)	Min	2534573
Pittsburgh (PA)	Pit	2322223	Cincinatti (OH)	Cin	2523351
Toledo (OH)	Tol	2522453	Buffalo (NY)	Buf	3445533
Wichita (KS)	Wic	2312463	Mesa (AZ)	Mes	1113353
Colorado Springs (CO)	Cos	1311151	Tampa (FL)	Tam	3546585
Santa Ana (CA)	SA	3122113	Arlington (VA)	Arl	1213242
Anaheim (CA)	Ana	1123223	Corpus Christi (TX)	CCr	1313371
Louisville (KY)	Lou	2222312	St. Paul (MN)	StP	2413332
Newark (NJ)	New	3346545	Birmingham (AL)	Bir	4536573
Norfolk (VA)	Nor	3322252	Anchorage (AK)	Anc	1313152
Aurora (CO)	Aur	1215252	St. Petersburg (FL)	SPe	1426462
Riverside (CA)	Riv	2225434	Lexington (KY)	Lex	1213241
Rochester (NY)	Roc	3332562	Jersey City (NJ)	JC	2134414
Raleigh (NC)	Ral	2113341	Baton Rouge (LO)	BRo	3326584
Akron (OH)	Akr	2413232	Stockton (CA)	Sto	2224454

NOTE: The asterisk for the Rape variable for Chicago denotes a missing observation.

The four possible answers to each of the variables are (1) serious, (2) moderate, (3) minor, and (4) not a problem.

Table C1 gives the frequencies of the variables.

TABLE C1
Student response frequencies

	Categories			
Variable	**1**	**2**	**3**	**4**
A	2708	6264	7685	6,591
B	2611	6356	7508	6,773
C	3435	4032	5862	9,919
D	3724	5774	7425	6,325
E	3178	3462	6918	9,690
F	3392	3536	6759	9,561
G	3508	3414	5057	11,269
H	3281	2412	4786	12,769
I	2642	2180	5335	13,091
J	1811	666	2294	18,377
K	2633	3258	6150	11,207

ACKNOWLEDGMENTS

We thank the past (Paul Switzer) and present (Leon Gleser) Executive Editors and three anonymous referees for many helpful suggestions and comments on earlier drafts of this paper that improved substantially the quality of the presentation and focus of this work.

REFERENCES

[1] ANDERSON, C. S. (1982). The search for school climate: a review of the research. *Review of Educational Research* **52** 368–420.

[2] ANDERSON, T. W. (1984). *An Introduction to Multivariate Analysis Techniques*, 2nd ed. Wiley, New York.

[3] BENZÉCRI, J. P. (1973). *Analyse des Données*. Dunod, Paris.

[4] BENZÉCRI, J. P. (1992). *Handbook of Correspondence Analysis*. Dekker, New York.

[5] BIJLEVELD, C. C. J. H. (1989). *Exploratory Linear Dynamic Systems Analysis*. DSWO Press, Leiden.

[6] BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

[7] BOND, J. and MICHAILIDIS, G. (1996). Homogeneity analysis in Lisp-Stat. *J. Statistical Software* **1**.

[8] BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619.

[9] BUJA, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Ann. Statist.* **18** 1032–1069.

[10] CAILLIEZ, F. and PAGES, J. P. (1976). *Introduction à l'Analyse des Données*. SMASH, Paris.

[11] CARNEGIE FOUNDATION FOR THE ADVANCEMENT OF TEACHING (1988). *An Imperiled Generation: Saving Urban Schools*. Carnegie Foundation, Princeton, NJ.

[12] CARROLL, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th Convention of the American Psychological Association* **3** 227–228.

[13] CLOGG, C. C. (1981). New developments in latent structure analysis. In *Factor Analysis and Measurement in Sociological Research* (Jackson and Borgatta, eds.) 215–246. Sage, Beverly Hills, CA.

[14] CLOGG, C. C. (1984). Latent structure analysis of a set of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **79** 762–771.

[15] CLOGG, C. C. (1986). Statistical modeling versus singular value decomposition. *Internat. Statist. Rev.* **54** 284–288.

[16] DE LEEUW, J. (1977). Correctness of Kruskal's algorithms for monotone regression with ties. *Psychometrika* **42** 141–144.

[17] DE LEEUW, J. (1983). On the prehistory of correspondence analysis. *Statist. Neerlandica* **37** 161–164.

[18] DE LEEUW, J. (1984). The Gifi-system of nonlinear multivariate analysis. In *Data Analysis and Informatics III* (Diday et al., eds.) 415–424. North-Holland, Amsterdam.

[19] DE LEEUW, J. (1985). Jackknife and bootstrap methods in multinomial situations. Research Report 85-16, Dept. Data Theory, Leiden Univ.

[20] DE LEEUW, J. (1988). Models and techniques. *Statist. Neerlandica* **42** 91–98.

[21] DE LEEUW, J. and VAN DER BURG, E. (1984). The permutational limit distribution of generalized canonical correlations. In *Data Analysis and Informatics IV* (Diday et al., eds.) 509–521. North-Holland, Amsterdam.

[22] DE LEEUW, J. and VAN RIJCKEVORSEL, J. (1980). Homals and princals. Some generalizations of principal components analysis. In *Data Analysis and Informatics II* (Diday et al., eds.) 231–242. North-Holland, Amsterdam.

[23] EADES, P. and SUGIYAMA, K. (1990). How to draw a directed graph. *J. Inform. Process.* **13** 424–437.

[24] EADES, P. and WORMALD, N. C. (1994). Edge crossings in drawings of bipartite graphs. *Algorithmica* **11** 379–403.

[25] EADES, P., TAMASSIA, R., DI BATTISTA, G. and TOLLIS, I. (1994). Algorithms for drawing graphs: an annotated bibliography. *Comput. Geom.* **4** 235–282.

[26] EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[27] Encyclopedia Britannica.

[28] ESCOUFIER, Y. (1984). Analyse factorielle en référence à un modèle: application a l'analyse de tableaux d'échanges. *Rev. Statist. Appl.* **32** 25–36.

[29] ESCOUFIER, Y. (1985). L'analyse des correspondences: ses propriétés et ses extensions. *Bull. Internat. Statist. Inst.* **51** 1–16.

[30] ESCOUFIER, Y. (1988). Beyond correspondence analysis. In *Classification and Related Methods of Data Analysis* (Bock, ed.). North-Holland, Amsterdam.

[31] FISHER, R. A. (1938). The precision of discriminant functions. *Annals of Eugenics* **10** 422–429.

[32] FREEDMAN, D. A. and LANE, D. (1983). Significance testing in a nonstochastic setting. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 185–208. Wadsworth, Belmont, CA.

[33] GIFI, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

[34] GILULA, Z. (1986). Grouping and association in contingency tables: an exploratory canonical correlation approach. *J. Amer. Statist. Assoc.* **81** 773–779.

[35] GILULA, Z. and HABERMAN, S. J. (1986). Canonical analysis of two-way contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* **81** 780–788.

[36] GILULA, Z. and RITOV, Y. (1990). Inferential ordinal correspondence analysis: motivation, derivation and limitations. *Internat. Statist. Rev.* **58** 99–108.

[37] GITTINS, R. (1985). *Canonical Analysis: A Review with Applications in Ecology*. Springer, Berlin.

[38] GNANADESIKAN, R. and KETTENRING, J. R. (1984). A pragmatic review of multivariate methods in applications. In *Statistics: An Appraisal* (H. A. David and H. T. David, eds.) Iowa State Univ. Press.

[39] GOLUB, G. H. and VAN LOAN, C. F. (1989). *Matrix Computations*. Johns Hopkins Univ. Press.

[40] GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231.

[41] GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** 537–552.

[42] GOODMAN, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76** 320–334.

[43] GOODMAN, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13** 10–69.

[44] GOODMAN, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear approach in the analysis of contingency tables. *Internat. Statist. Rev.* **54** 243–309.

[45] GOODMAN, L. A. (1994). On quasi-independence and quasi-dependence in contingency tables, with special reference to ordinal triangular contingency tables. *J. Amer. Statist. Assoc.* **89** 1059–1063.

[46] GREEN, P. J. (1981). Peeling bivariate data. In *Interpreting Multivariate Data* (Barnett, ed.) Wiley, New York.

[47] GREEN, P. J. and SILVERMAN, B. W. (1979). Constructing the convex hull of a set of points in the plane. *Computer Journal* **22** 262–266.

[48] GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

[49] GREENACRE, M. J. and HASTIE, T. (1987). The geometric interpretation of correspondence analysis. *J. Amer. Statist. Assoc.* **82** 437–447.

[50] GUTTMAN, L. (1941). The quantification of a class of attributes: A theory and a method of scale construction. *The Prediction of Personal Adjustment* (Horst et al., eds.) Social Science Research Council, New York.

[51] HABERMAN, S. J. (1970). *Analysis of Qualitative Data. New Developments* **2**. Academic Press, New York.

[52] HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York.
[53] HASTIE, T., BUJA, A. and TIBSHIRANI, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23** 73–102.
[54] HASTIE, T., TIBSHIRANI, R. and BUJA, A. (1994). Flexible discriminant analysis with optimal scoring. *J. Amer. Statist. Assoc.* **89** 1255–1270.
[55] HAYASHI, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Statist. Math.* **5** 121–143.
[56] HEISER, W. J. and MEULMAN, J. J. (1983). Constrained multidimensional scaling. *Applied Psychological Measurement* **7** 381–404.
[57] HIRSCHFELD, H. O. (1935). A connection between correlation and contingency. *Proc. Cambridge Philos. Soc.* **31** 520–524.
[58] HOFFMAN, D. L. and DE LEEUW, J. (1992). Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing Letters* **3** 259–272.
[59] HORST, P. (1961). Relations among m sets of measures. *Psychometrika* **26** 129–149.
[60] HORST, P. (1961). Generalized canonical correlations and their application to experimental data. *Journal of Clinical Psychology* **17** 331–347.
[61] HOTELLING, H. (1935). The most predictable criterion. *Journal of Educational Psychology* **26** 139–142.
[62] HOTELLING, H. (1936). Relations between two sets of variables. *Biometrika* **28** 321–377.
[63] KATO, T. (1995). *Perturbation Theory of Linear Operators*. Springer, Berlin.
[64] KAUFMAN, P. and BRADBY, D. (1992). Characteristics of at risk students in NELS:88. Report 92-042, National Center for Education Statistics, Washington, DC.
[65] KENDALL, M. G. (1980). *Multivariate Analysis*, 2nd ed. Griffin, London.
[66] KETTENRING, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–460.
[67] KRUSKAL, J. B. and SHEPARD, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika* **39** 123–157.
[68] KSHIRSAGAR, A. N. (1978). *Multivariate Analysis*. Dekker, New York.
[69] LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.
[70] LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inferences from Nonexperimental Data*. Wiley, New York.
[71] LEBART, L., MORINEAU, A. and TABARD, N. (1977). *Technique de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*. Dunod, Paris.
[72] LIU, R. Y., SINGH, K. and LO, S. (1989). On a representation related to the bootstrap. *Sankhyā Ser. A* **51** 168–197.
[73] MARKUS, M. T. (1994). *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*. DSWO Press, Leiden.
[74] MEULMAN, J. J. (1984). Correspondence analysis and stability. Research Report 84-01, Dept. Data Theory, Leiden Univ.
[75] MICHAILIDIS, G. and DE LEEUW, J. (1995). Nonlinear multivariate analysis of NELS:88. UCLA Statistical Series Preprints 175. Univ. California, Los Angeles.
[76] MICHAILIDIS, G. and DE LEEUW, J. (1996). The Gifi system of nonlinear multivariate analysis. UCLA Statistical Series Preprints 204. Univ. California, Los Angeles.
[77] MOLENAAR, I. W. (1988). Formal statistics and informal data analysis, or why laziness should be discouraged. *Statist. Neerlandica* **42** 83–90.
[78] NATIONAL EDUCATION GOALS PANEL (1992). *The National Education Goals Report: Building a Nation of Learners*. Washington, DC.
[79] NISHISATO, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto Univ. Press.
[80] NISHISATO, S. (1994). *Elements of Dual Scaling. An Introduction to Practical Data Analysis*. Erlbaum, Hillsdale.
[81] NISHISATO, S. and NISHISATO, I. (1984). *An Introduction to Dual Scaling*. Microstats, Toronto.
[82] OAKES, J. (1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis* **11** 181–199.
[83] RITOV, Y. and GILULA, Z. (1993). Analysis of contingency tables by correspondence models subject to order constraints. *J. Amer. Statist. Assoc.* **88** 1380–1387.
[84] ROSKAM, E. E. (1968). *Metric Analysis of Ordinal Data in Psychology*. VAM, Voorschoten.
[85] RUTISHAUSER, H. (1969). Computational aspects of F. L. Bauers's simultaneous iteration method. *Numer. Math.* **13** 4–13.
[86] SAPORTA, G. (1975). *Liaisons entre Plusieurs Ensembles de Variables et Codage de Données Qualitatives*. Univ. Paris VI, Paris.
[87] SCHRIEVER, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. *Internat. Statist. Rev.* **51** 225–238.
[88] SHAO, J. (1992). Some results for differentiable statistical functionals. In *Nonparametric Statistics and Related Topics* (Saleh, ed.) 179–188. North-Holland, Amsterdam.
[89] SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
[90] SPSS INC. *SPSS Categories User's Manual.*
[91] STEEL, R. G. D. (1951). Minimum generalized variance for a set of linear functions. *Ann. Math. Statist.* **22** 456–460.
[92] TAKANE, Y. and SHIBAYAMA, T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika* **56** 97–120.
[93] VAN BUUREN, S. (1990). *Optimal Scaling of Time Series*. DSWO Press, Leiden.
[94] VAN BUUREN, S. (1994). Groupals analysis of abiotic measures from an environmental study in the archipelago of Hochelage. Research Report 94-14, Dept. Data Theory, Univ. Leiden.
[95] VAN BUUREN, S. and HEISER, W. J. (1989). Clustering N objects into K groups under optimal scaling of variables. *Psychometrika* **54** 699–706.
[96] VAN DE GEER, J. P. (1984). Linear relationships among k sets of variables. *Psychometrika* **49** 79–94.
[97] VAN DER BURG, E. (1985). Homals classification of whales, porpoises and dolphins. In *Data Analysis in Real Life Environment: Ins and Outs of Solving Problems* (J.-F. Marcotorchino, J.-M. Proth and J. Janssen, eds.) 25–35. North-Holland, Amsterdam.
[98] VAN DER BURG, E. and DE LEEUW, J. (1988). Use of the multinomial jackknife and bootstrap in generalized canonical correlation analysis. *Appl. Stochastic Models Data Anal.* **4** 154–172.
[99] VAN DER BURG, E., DE LEEUW, J. and DIJKSTERHUIS, G. (1994). Nonlinear canonical correlation with k sets of variables. *Comput. Statist. Data Anal.* **18** 141–163.
[100] VAN DER BURG, E., DE LEEUW, J. and VERDEGAAL, R. (1988). Homogeneity analysis with K sets of variables: an alternating least squares method with optimal scaling features. *Psychometrika* **53** 177–197.

[101] VAN DER HEIJDEN, P. G. M., DE FALGUEROLLES, A. and DE LEEUW, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *J. Roy. Statist. Soc. Ser. C* **38** 249–292.

[102] VAN RIJCKEVORSEL, J. L. A. (1987). *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. DSWO Press, Leiden.

[103] VERDEGAAL, R. (1986). OVERALS. Research Report UG-86-01. Dept. Data Theory, Leiden.

[104] WEINBERG, S. L., CARROLL, J. D. and COHEN, H. S. (1984). Confidence regions for INDSCAL using jackknife and bootstrap techniques. *Psychometrika* **49** 475–491.

[105] YOUNG, F. W., DE LEEUW, J. and TAKANE, Y. (1976). Regression with qualitative variables: an alternating least squares method with optimal scaling features. *Psychometrika* **41** 505–529.

[106] YOUNG, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika* **46** 357–388.