Title
The Gifi System for Nonlinear Multivariate Analysis

Permalink
https://escholarship.org/uc/item/0789f7d3

Authors
Michailides, George
de Leeuw, Jan

Publication Date
1998

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. A brief account of stability issues and areas of applications of the techniques is also given.

1

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 details 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 $j \in \mathbf{J} = \{1, 2, \ldots, J\}$ has ℓ_j categories (possible values). 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 1.1 the bipartite graph of a toy example corresponding to a 4×3 contingency table with 7 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
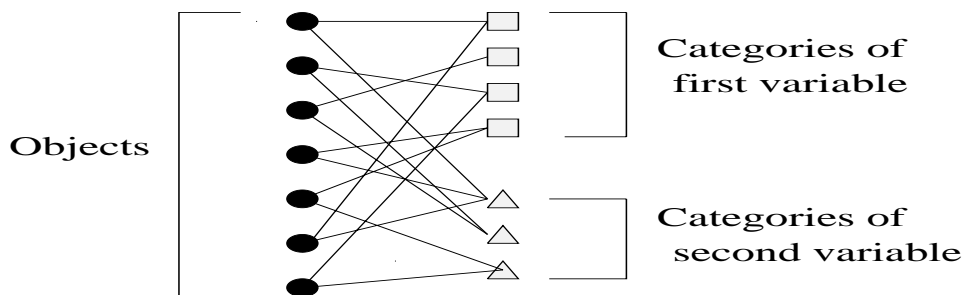

FIGURE 1.1. A Toy Example

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 in a manner that they are easy to understand and present has attracted a lot of attention in

the computer science literature [25]. There are different approaches of 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 Y_j, $j \in \mathbf{J}$ 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 1.2 for the mammals dentition data set that is analyzed later on in subsection 1.1. It can be seen that very little has been gained by this 2-dimensional
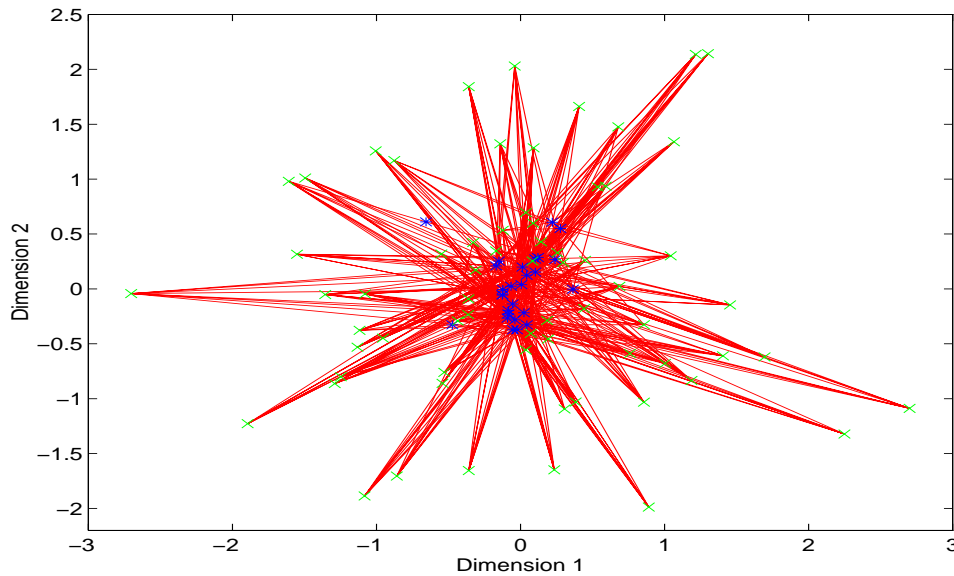

FIGURE 1.2. An arbitrary two dimensional graph plot of the mammals data set (x=objects, *=categories)

representation. The picture has too much 'ink' and no interesting patterns have emerged. A more informative picture would emerge if the edges are short, or in other words if objects are close to the categories they fall in, and categories are close to the objects belonging in them [24]. Hence, our goal becomes 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 rlated to many classical multivariate analytic techniques.

The data are coded by using *indicator* matrices G_j, with entries $G_j(i, t) = 1$, $i = 1, \ldots, N$, $t = 1, \ldots, \ell_j$[1] if object i belongs to category t, and $G_j(i, t) = 0$ if it belongs to some other category.

---

[1]In the paper we employ the following notational conventions. Upper case letters are used for matrices (e.g. A), and lower case letters for vectors (e.g. a). The $(s, t)^{th}$ element of a matrix is denoted by $A(s, t)$, the sth row by $A(s, .)$ and the tth column by $A(., t)$. Analogously, the sth element of a vector is denoted by $a(s)$. Finally, u denotes a vector comprised of only ones, and I_p the identity matrix of order p.

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

$$(1.1) \qquad \sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\left(X - G_j Y_j\right) = J^{-1} \text{tr}\left(X - G_j Y_j\right)'\left(X - G_j Y_j\right),$$

where 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 [32], 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

$$(1.2) \qquad\qquad\qquad\qquad\qquad X'X = NI_p,$$

$$(1.3) \qquad\qquad\qquad\qquad\qquad u'X = 0.$$

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 in addition orthogonal. Although this is computationally convenient, in many respects is not completely satisfactory, a fact already noted by Guttman [49].

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. In the first step, (1.1) is minimized with respect to Y_j for fixed X. The set of normal equations is given by

$$(1.4) \qquad\qquad\qquad\qquad D_j Y_j = G_j'X, \; j \in \mathbf{J},$$

where $D_j = G_j'G_j$ is the $\ell_j \times \ell_j$ diagonal matrix containing the univariate marginals of variable j. Hence, we get that

$$(1.5) \qquad\qquad\qquad\qquad \hat{Y}_j = D_j^{-1}G_j'X, \; j \in \mathbf{J}.$$

In the second step of the algorithm, (1.1) is minimized with respect to X for fixed Y_j's. The normal equation is given by

$$(1.6) \qquad\qquad\qquad\qquad JX = \sum_{j=1}^{J} G_j Y_j,$$

so that

$$(1.7) \qquad\qquad\qquad\qquad \hat{X} = J^{-1} \sum_{j=1}^{J} G_j Y_j.$$

In the third step of the algorithm the object scores X are column centered by setting $W = \hat{X} - u(u'\hat{X}/N)$, and then orthonormalized by the modified Gram-Schmidt procedure [38] $X =$

$\sqrt{N}$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 (for more details see Section 2.2).

Equation (1.5) expresses the so-called first centroid principle [3] (a category quantification is in the centroid of the object scores they belong to it), while equation (1.7) 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 [30] and Guttman [49]). This solution is known in the literature ([32, 19, 17]) 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 [88], 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

$$(1.8) \quad J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(\hat{X} - G_j \hat{Y}_j\big)'\big(\hat{X} - G_j \hat{Y}_j\big) = J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(\hat{X}'\hat{X} + \hat{Y}_j' G_j' G_j \hat{Y}_j - 2\hat{Y}_j' G_j' \hat{X}\big) =$$

$$J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(\hat{X}'\hat{X} - \hat{Y}_j' D_j \hat{Y}_j\big) = J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(N I_p - \hat{Y}_j' D_j \hat{Y}_j\big) = Np - J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(\hat{Y}_j' D_j \hat{Y}_j\big).$$

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

$$(1.9) \qquad\qquad \eta_{js}^2 \equiv \hat{Y}_j'(.,s) D_j \hat{Y}_j(.,s)/N, \; j \in \mathbf{J}, \; s = 1, \ldots, p.$$

Geometrically, 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. 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 $G_j \hat{Y}_j(.,s)$ in dimension s, and the corresponding column of object scores $\hat{X}(.,s)$ (see chapter 3 in [32]). Hence, the loss function can also be expressed as

$$(1.10) \qquad\qquad N\big(p - \frac{1}{J} \sum_{j=1}^{J} \sum_{s=1}^{p} \eta_{js}^2\big) = N\big(p - \sum_{s=1}^{p} \gamma_s\big),$$

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

We summarize next some basic properties of the Homals solution.

◇ Category quantifications and object scores are represented in a joint space (see Figure 1.4 for the mammals dentition data set).

◇ A category point is the centroid of objects belonging to that category, a direct consequence of (1.5) (see Figure 1.3 for two variables from the mammals dentition data set).

◇ Objects with the same response pattern (identical profiles) receive identical object scores (follows from (1.7)) (see Figure 1.7). 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 (1.9)).

◇ If a category applies uniquely to only a single object, then the object point and that category point will coincide.

◇ Category points with low marginal frequencies will be located further away from the origin of the joint space, whereas categories with high marginal frequencies will be located closer to the origin (follows from (1.5)).

◇ 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 normalization of the object scores, since $u'D_j\hat{Y}_j = u'D_jD_j^{-1}G_j'\hat{X} = u'G_j'\hat{X} = u'\hat{X} = 0$.

◇ 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 also the maximum eigenvalue subject to the constraint that $X(.,2)$ is uncorrelated to $X(.,1)$, and so forth.

◇ The solutions for the object scores are uncorrelated (follows from (1.3)). However, the solutions for the 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 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 $X^\sharp = X \times R$, where R is a rotation matrix satisfying $R'R = RR' = I_p$. We then get from (1.5) that $Y_j^\sharp = D_j^{-1}G_j'X^\sharp = \hat{Y}_jR$. Thus, the axes of the joint space can not be uniquely identified.

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. The complete data together with the coding of the variables are given in Appendix A. The main question is whether the technique managed to produce a fairly clean picture and uncover some interesting features of this data set.
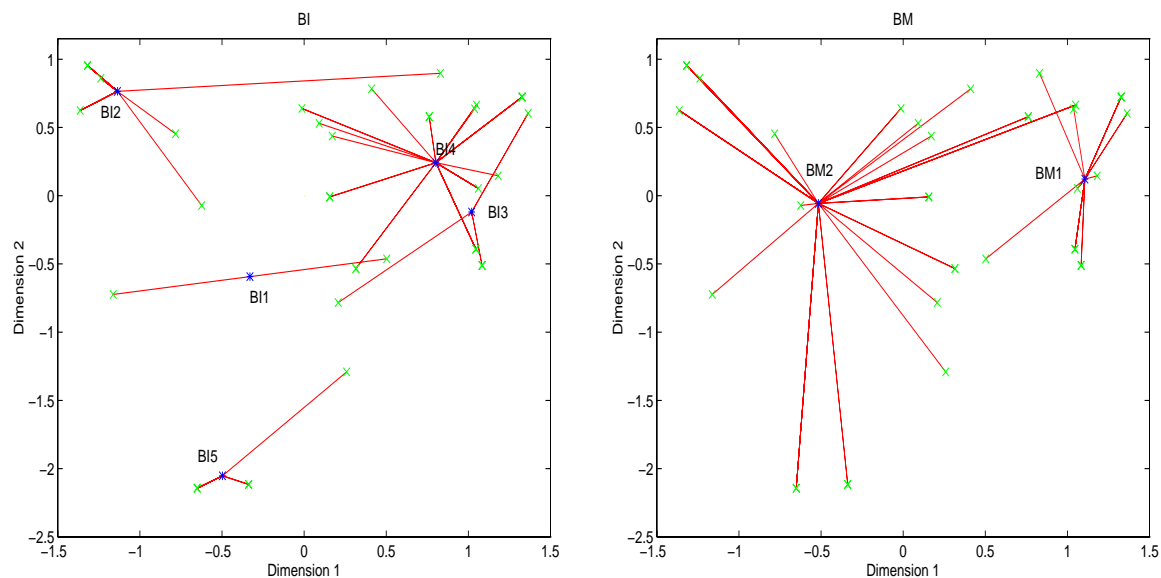
FIGURE 1.3. Star plots of variables bottom incisors (BI) and bottom molars (BM) from the mammals dentition example

A two dimensional analysis gives an adequate fit, with eigenvalues .73 and .38 for the two dimensions respectively. The graph plot of this solution is given in Figure 1.4. 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. In order to study those patterns more closely we turn our attention to the arrangement of only the category vertices on the map (see Figure 1.5). It can be seen that they form 4 groups. In the upper right corner we can find the 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. In the upper left corner we find the 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 the categories TP4 and BP4 are close together, thus suggesting that objects with 3 top premolars, usually have 3 bottom premolars as well; similarly, for the subgroup TC1, BC1, TM1 and BM1, we have that animals with 0 top and bottom canines have more than 3 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 1.6) along with their respective frequencies (see Figure 1.7). 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 three separate clusters located in the upper right, upper left and lower left areas of the picture. For example in the lower left area we find the following animals: elk, deer, moose, reindeer, antelope, bison, mountain goat, muskox, and mountain sheep. Similarly, in the upper right area we find among others various types of squirrels (ground, gray, fox), and rats (kangaroo, pack, field mouse, black, etc), and in the upper right animals such as jaguar, cougar, lynx, 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 bytes. Hartigan [51] 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 the clustering algorithm classified the hairy tail mole together with the opposum, instead of the bats, and the pack rat with the armadillo instead of the squirrels. This is due to the nature of the particular algorithm, 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.
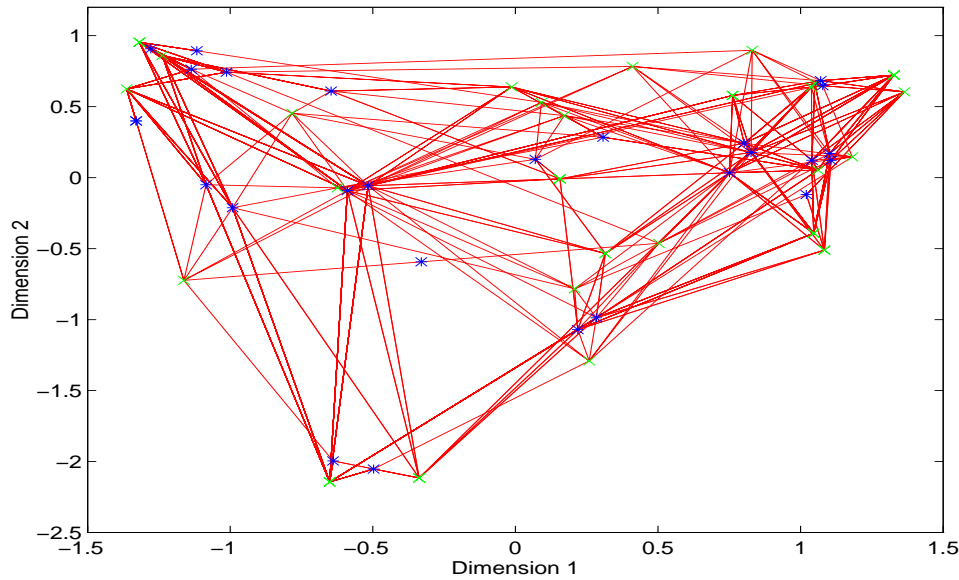

FIGURE 1.4. A two dimensional graph plot of the mammals data set produced by Homals (x=objects, *=categories)

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 1.8, 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 1.5, 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 splitting of the objects into two groups, while the other set of variables does further splittings especially along the second dimension. It is also interesting to examine the plots of the original versus the transformed scales given in Figure 1.9. Obviously, such plots are totally uninteresting for binary variables, and therefore are ommitted. 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.
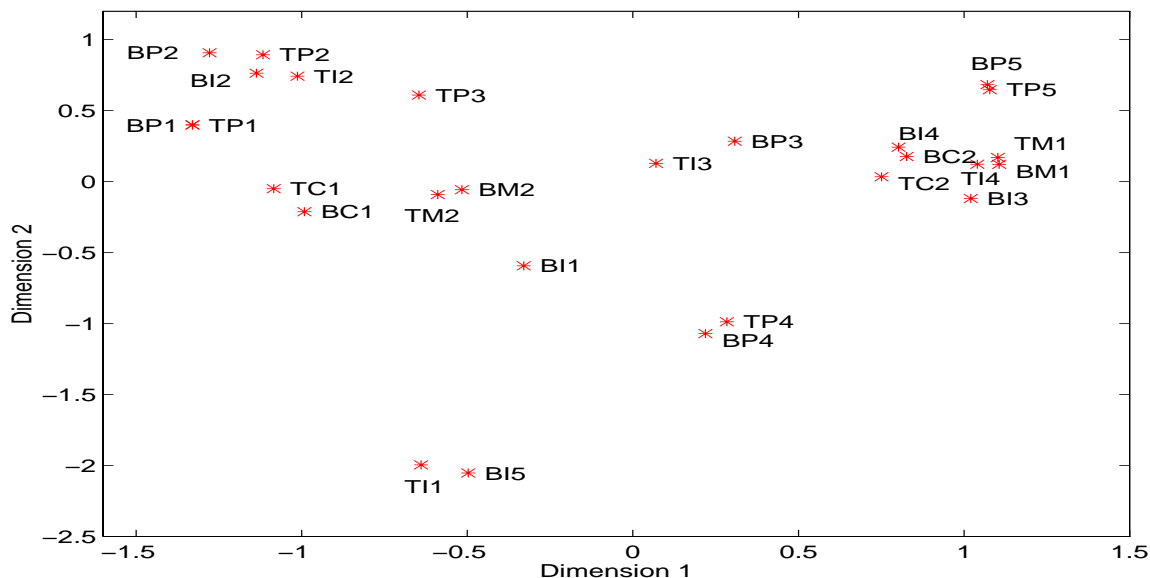
FIGURE 1.5. Category Quantifications

2. **Other Aspects of Homogeneity Analysis**

In this section we study various 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 of Homogeneity Analysis.** Homogeneity analysis has been motivated and introduced in the previous section 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 to introduce homogeneity analysis is through linearization of the regressions. Consider a column of the object scores $X(., s)$ as N data values on the x-axis. Let the category quantifications in the same s^{th} 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 (1.5)). Remarkably enough,
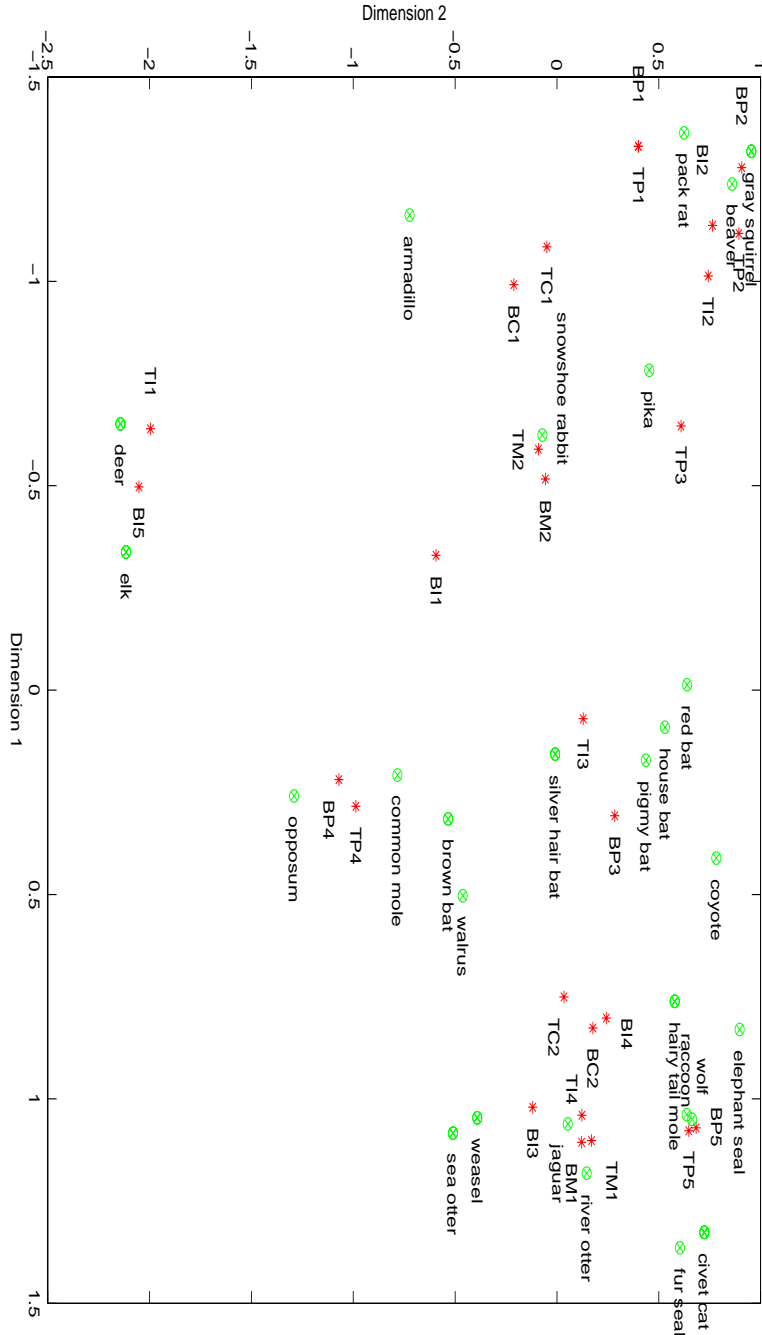
FIGURE 1.6. Category Quantifications and Object Scores (ommitting mammals which have identical dentition to the one shown on the graph)

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 (1.7)). Therefore, the Homals solution could be defined on the basis of
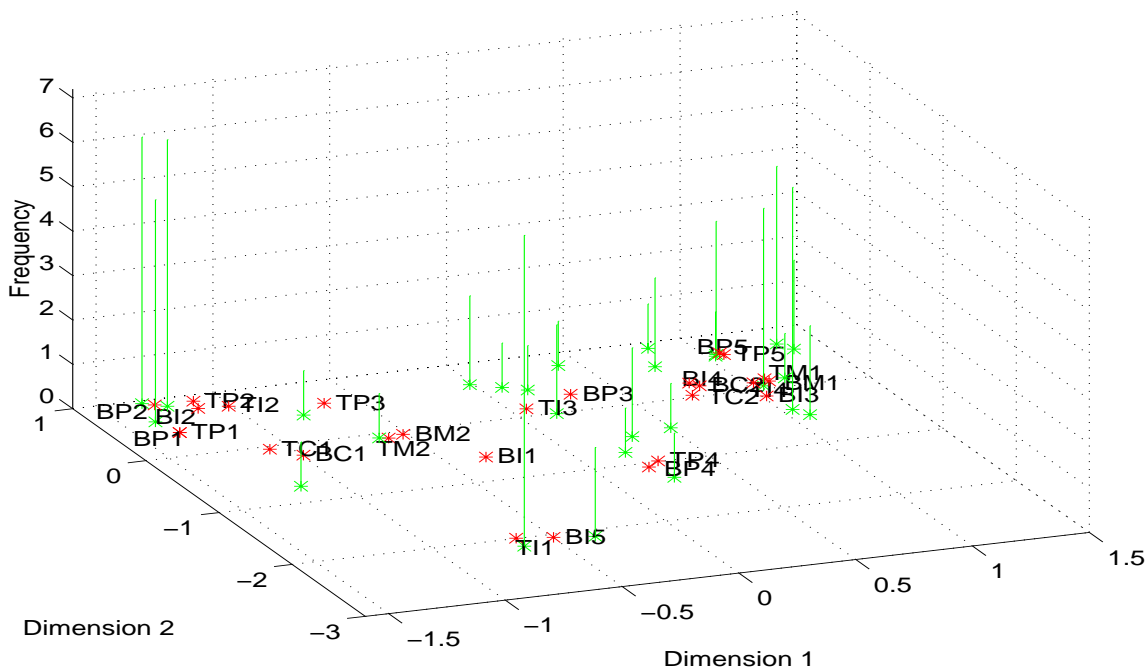
FIGURE 1.7. Category Quantifications and Object Scores (height of the object scores represents their respective frequencies)
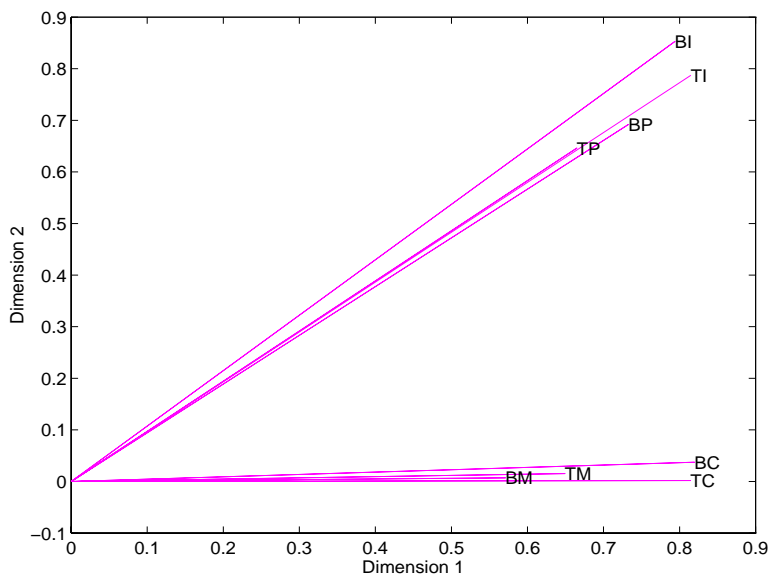
FIGURE 1.8. Discrimination Measures

this property; it is the solution where object scores are proportional to category means, and where category quantifications are proportional to object means.
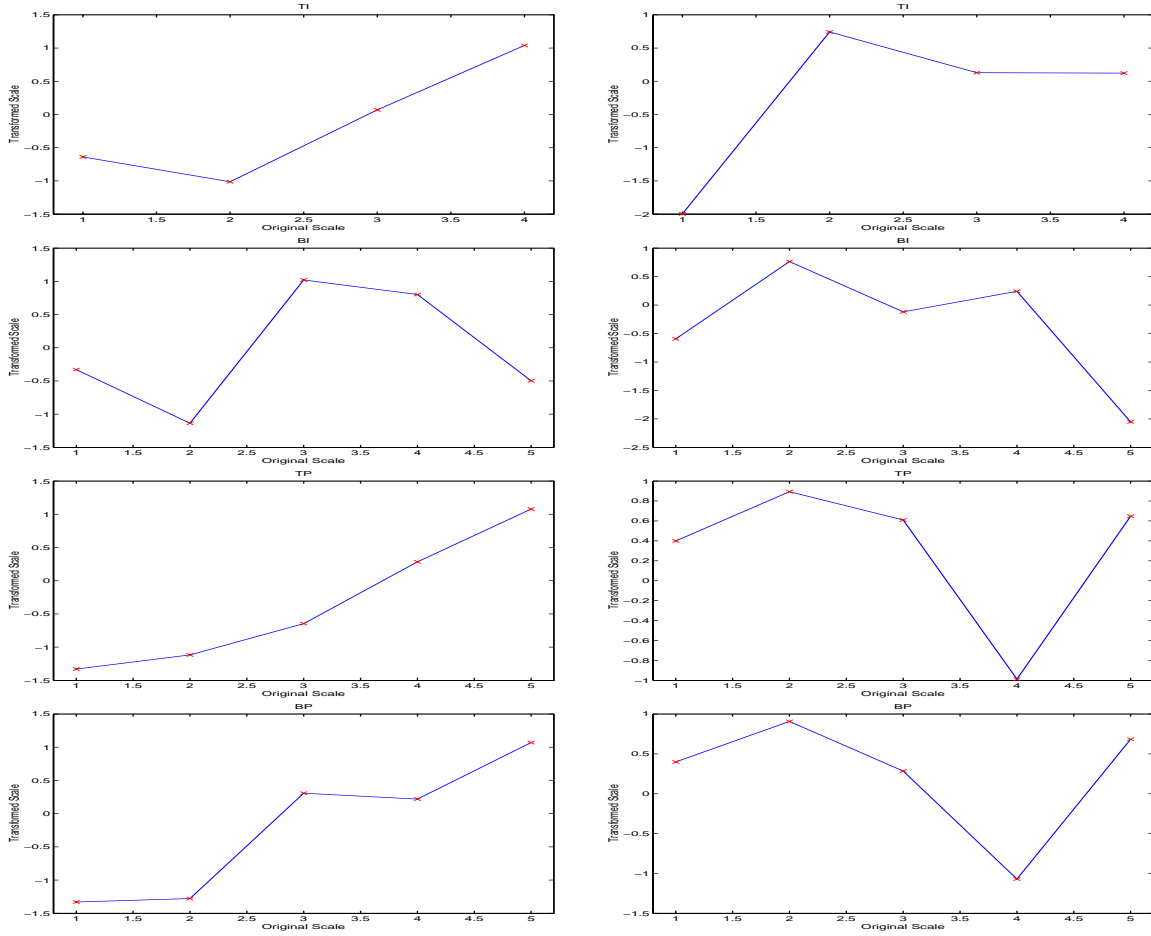
FIGURE 1.9. Optimal Transformations of some of the variables (Left: dimension 1, Right: dimension 2)

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 [17] that the sum of squared correlations between the optimal quantified variables $G_j Y(., s)$ and the vector $X(., s)$ 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 is that it makes the 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

$$(2.1) \qquad \sigma(X; \star) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}(X - G_j D_j^{-1} G_j' X)'(X - G_j D_j^{-1} G_j' X)$$

$$= J^{-1} \sum_{j=1}^{J} \operatorname{tr}(X'X - X'G_j D_j^{-1} G_j' X),$$

where the asterisk has replaced the argument over which the loss function is minimized. Let $P_j = G_j D_j^{-1} G_j'$ 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

$$(2.2) \qquad \sigma(X; \star) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}(X - P_j X)'(X - P_j X) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}(X'X - X'P_j X).$$

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

$$(2.3) \qquad \sigma(\star; \star) = N\left(p - \sum_{s=1}^{p} \lambda_s\right),$$

where λ_s, $s = 1, \ldots, p$ are the first 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 $G = [G_1| \ldots |G_J]$ and the marginal frequencies into $D = \bigoplus_{j=1}^{J} D_j$. The solution for the optimal X can then be obtained (see [32]) by the singular value decomposition of

$$(2.4) \qquad J^{-1/2} \mathcal{L} G D^{-1/2} = U \Lambda V,$$

where the left-hand side is the superindicator matrix in deviations from column means and corrected for 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 $p << q$ 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 etc. 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 $M_j G_j = G_j$, $j \in \mathbf{J}$. The loss function then becomes

$$(2.5) \qquad \sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}\big(X - G_j Y_j\big)' M_j \big(X - G_j Y_j\big),$$

subject to the normalization restrictions $X' M_* X = J N I_p$ and $u' M_* X = 0$. The $\hat{Y}_j$'s are given by (1.5), while the object scores by $\hat{X} = M_*^{-1} \sum_{j=1}^{J} G_j Y_j$. In the presence of missing data, it is no longer the case that $u' D_j Y_j = 0$ (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 [32] 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*, where the indicator matrix is completed with a single additional column for each variable with missing data, and (ii) *missing data multiple categories*, where each missing observation is treated as a new category. 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: 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: in many data analytic situations the determination of the categories is arbitrary, when coding interval data there is uncertainty about the allocation of values near the category boundaries, and for some data subsequent intervals are functionally related. Many alternatives have been suggested in the literature (for a thorough account see [99]), 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. Finally, various ways of handling missing data have been proposed (see [47]).

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 etc.

2.5.1. **Relationship with 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 [30] 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 can be given. The French approach to correspondence analysis is mainly characterized by the emphasis on geometry [3, 10, 70, 47]. 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, 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 $r = Fu$ denote the vector of row marginals, $c = F'u$ the vector of column marginals and $n = u'c = u'r$ the total number of observations. Finally, let $D_r = \operatorname{diag}(r)$ be the diagonal matrix containing the elements of vector r and $D_c = \operatorname{diag}(c)$ the diagonal matrix containing the elements of vector c. The χ^2-distances between rows i and i' of table F is given by

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

Formula (2.6) shows that $\delta^2(i, i')$ is a measure for the difference between the profiles of rows i and i'. It also shows that since the entries of the table are corrected for the row marginals, proportional rows obtain 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

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

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

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

so that, $X'D_rX = nI$ and $u'D_rX = 0$. It is worth noting that this solution places the origin of the space in the weighted centroid (since $u'D_rX = 0$) and represents the coordinates in weighted principal axes positions (since $X'D_rX = nI$).

Consider now the superindicator matrix $G = [G_1 | \ldots | G_J]$ of J indicator matrices G_j. Define also $D_r = JI$, $D_c = D \equiv \bigoplus_{j=1}^{J} D_j$ and $n = NJ$. Correspondence analysis of the superindicator matrix G resorts to the singular value decomposition of the matrix

$$(2.9) \qquad J^{-1/2}\big(G - \frac{J}{JN}Guu'\big)D^{-1/2} = J^{-1/2}\mathcal{L}GD^{-1/2} = U\Lambda V,$$

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 are analyzed, despite the fact that correspondence analysis is a symmetric technique regarding 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 with 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 a 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

$$(2.10) \qquad \sigma(X) = \mathrm{tr}\mathcal{L}\big(\Delta^2 - D^2(X)\big)'\mathcal{L}\big(\Delta^2 - D^2(X)\big)\mathcal{L},$$

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 a 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.

2.5.3. **Relationship with Cluster Analysis.** As we have seen in the previous subsection homogeneity analysis provides 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 many 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. Or to put 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 [91] 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, a property inherited from the combinatorial nature of the problem.

2.5.4. **Relationship with 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 (the identity matrix for example).

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 variables, 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 [56], Fisher [30] and especially Guttman [49], although some ideas go further back to Pearson (see the discussion in de Leeuw [18]). The main objective is to scale (assign real values to) the categories so that a particular criterion is optimized (e.g. 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 [78, 79], quantification theory [54] and also simultaneous linear regression, centroid scaling, optimal scoring, 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 subsection 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 towards 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 comparison, 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 [27, 28, 29, 98]. 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 in 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 3-4). Thus, we see the various groups/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/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 [40, 41, 44, 33, 34, 42, 43, 15] and their extensions to handle ordinal data [85, 35, 82]. On another line we have the development of latent structure models for analyzing a single or a set of multidimensional contingency tables [68, 39, 50, 13, 14]. 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 [103] and in recent developments in discriminant analysis [52, 53].

3. **Nonlinear Principal Component Analysis**

In the Gifi system nonlinear PCA is derived as homogeneity analysis with restrictions [17]. The starting point for this derivation is the loss function given in (1.1). However, *rank-one restrictions* of the form

$$(3.1) \qquad Y_j = q_j \beta_j', \ j \in \mathbf{J},$$

are imposed on the multiple category quantifications, with q_j being a 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-one, which implies that the quantifications in p dimensional space become proportional to each other. The introduction of the rank-one 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 requirement 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 the restriction (3.1), we start by computing the $\hat{Y}_j$'s as in (1.5). We then partition the Gifi loss function as follows:

$$(3.2) \qquad \sum_{j=1}^{J} \text{tr}\big(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\big)' \big(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\big) =$$

$$\sum_{j=1}^{J} \text{tr}\big(X - G_j\hat{Y}_j\big)' \big(X - G_j\hat{Y}_j\big) + \sum_{j=1}^{J} \text{tr}\big(Y_j - \hat{Y}_j\big)' D_j \big(Y_j - \hat{Y}_j\big).$$

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

$$(3.3) \qquad \sum_{j=1}^{J} \text{tr}\big(q_j \beta_j' - \hat{Y}_j\big)' D_j \big(q_j \beta_j' - \hat{Y}_j\big),$$

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

$$(3.4) \qquad \hat{\beta}_j = \big(\hat{Y}_j' D_j q_j\big) / \big(q_j' D_j q_j\big), \ j \in \mathbf{J}.$$

and for fixed β_j's

$$(3.5) \qquad \hat{q}_j = \hat{Y}_j \beta_j / \big(\beta_j' \beta_j\big), \ j \in \mathbf{J}.$$

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 $\mathcal{C}_j$. In case of ordinal data the relevant cone C_j is the cone of monotone transformations given by $C_j = \{q_j | q_j(1) \leq q_j(2) \leq \ldots q_j(l_j)\}$. The projection to this cone is solved by a weighted *monotone* regression in the metric D_j (the weights) (see de Leeuw [16] and references therein). In 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 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 ([32], [17]) 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. $u'D_j s_j = 0$ and $s'_j D_j s_j = 1$), that is calculate the eigenvalues and the eigenvectors of the correlation matrix of the s_j variables [17] (see also chapter 4 in [32]). 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 [75]) one can also incorporate in 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:

Step 0: Initialize X, so that $u'X = 0$ and $X'X = NI_p$.

Step 1: Estimate the multiple category quantifications by $\hat{Y}_j = D_j^{-1} G'_j X$, $j \in \mathbf{J}$.

Step 2: Estimate the component loadings by $\hat{\beta}_j = (\hat{Y}'_j D_j q_j)/(q'_j D_j q'_j)$, $j \in \mathbf{J}$.

Step 3: Estimate the single category quantifications by $\hat{q}_j = \hat{Y}_j \beta_j/(\beta'_j \beta_j)$, $j \in \mathbf{J}$.

Step 4: Account for the measurement level of the jth variable by performing a monotone or linear regression.

Step 5: Update the multiple category quantifications by setting $\hat{Y}_j = \hat{q}_j \hat{\beta}'_j$, $j \in \mathbf{J}$.

Step 6: Estimate the object scores by $\hat{X} = J^{-1} \sum_{j=1}^{J} G_j Y_j$.

Step 7: Column center and orthonormalize the matrix of the object scores.

Step 8: Check the convergence criterion.

In principle, to obtain the minimum over all monotone or linear transformations of the q_j's, steps 2-5 should be repeated until convergence is reached for the criterion given by (3.3). However, since

the value of the loss function will be smaller after a single iteration of the inner ALS loop, inner iteration upon convergence is not necessary in practice. The above algorithm is implemented in the Fortran program from the University of Leiden called Princals and also in the corresponding program in SPSS [88].

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:

$$(3.6) \qquad \sum_{j=1}^{J} \text{tr}\big(X - G_j \hat{Y}_j\big)'\big(X - G_j \hat{Y}_j\big) + \sum_{j=1}^{J} \text{tr}\big(\hat{q}_j \hat{\beta}_j' - \hat{Y}_j\big)' D_j \big(\hat{q}_j \hat{\beta}_j' - \hat{Y}_j\big).$$

Using (1.10), the first term in (3.6) can be also written as $N\big(p - \sum_{j=1}^{J} \sum_{s=1}^{p} \eta_{js}^2\big)$, 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 $q_j' D_j q_j = N$, and using the fact that $\hat{Y}_j' D_j q_j \beta_j' = N \beta_j \beta_j'$ (from (3.4)), the second part of (3.6) can be written as

$$(3.7) \qquad \sum_{j=1}^{J} \text{tr}\big(\hat{Y}_j' D_j \hat{Y}_j - N \beta_j \beta_j'\big) = N \Big(\sum_{j=1}^{J} \sum_{s=1}^{p} (\eta_{js}^2 - \beta_{js}^2)\Big),$$

which is called the *single loss*. The quantities β_{js}^2, $s = 1, \ldots, p$ are called *single fit*, and correspond to squared component loadings (see chapter 4 in [32]).

From equation (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 in the single loss part: first the rank-one restriction, that is the fact that single category quantifications must lie on a straight line in the joint space, and 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 1.9) would only show straight lines with the categories arranged in an increasing or decreasing order for ordinal variables and additionally 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 from 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 chapter 4 in [32]).

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

(3.8)
$$\sum_{j=1}^{J} \mathrm{tr}\big(X - G_j Y_j\big)' M_j \big(X - G_j Y_j\big) =$$

$$\sum_{j=1}^{J} \mathrm{tr}\big(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))\big)' M_j \big(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))\big) =$$

$$\sum_{j=1}^{J} \mathrm{tr}\big(X - G_j \hat{Y}_j\big)' M_j \big(X - G_j \hat{Y}_j\big) + \sum_{j=1}^{J} \mathrm{tr}\big(Y_j - \hat{Y}_j\big)' D_j \big(Y_j - \hat{Y}_j\big).$$

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 is 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 treated all the variables as ordinal in a nonlinear principal components analysis. In Figure 3.1 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. 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.

In Figure 3.2 some of the variable transformations are shown. It can be seen that some variables such as murder and robbery (not shown here) received linear transformations, while some others (e.g. assault, larceny) distinctly nonlinear ones. Finally, in Figure 3.3 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-one 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
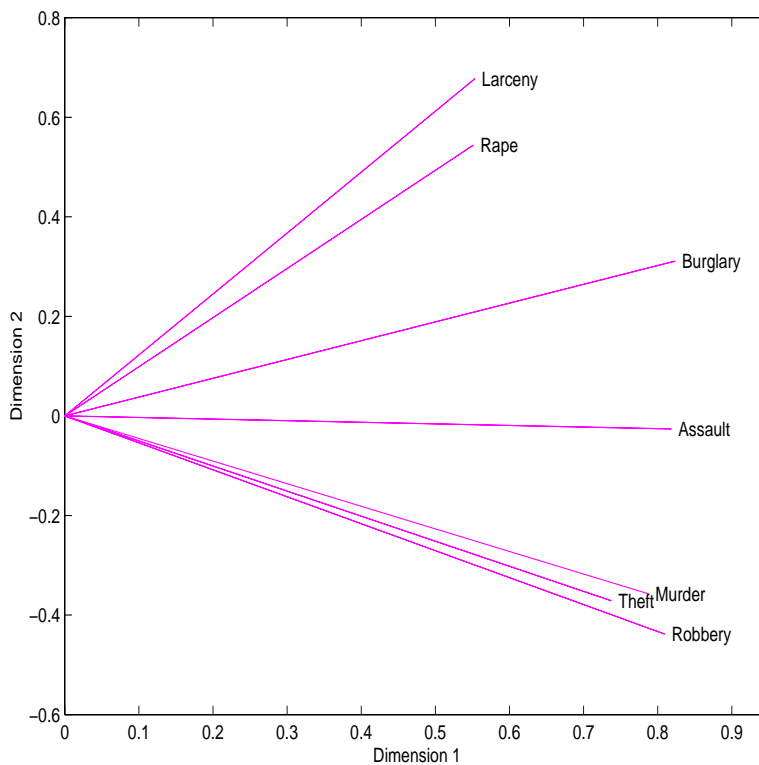
FIGURE 3.1. Component Loadings of the seven Crime Variables
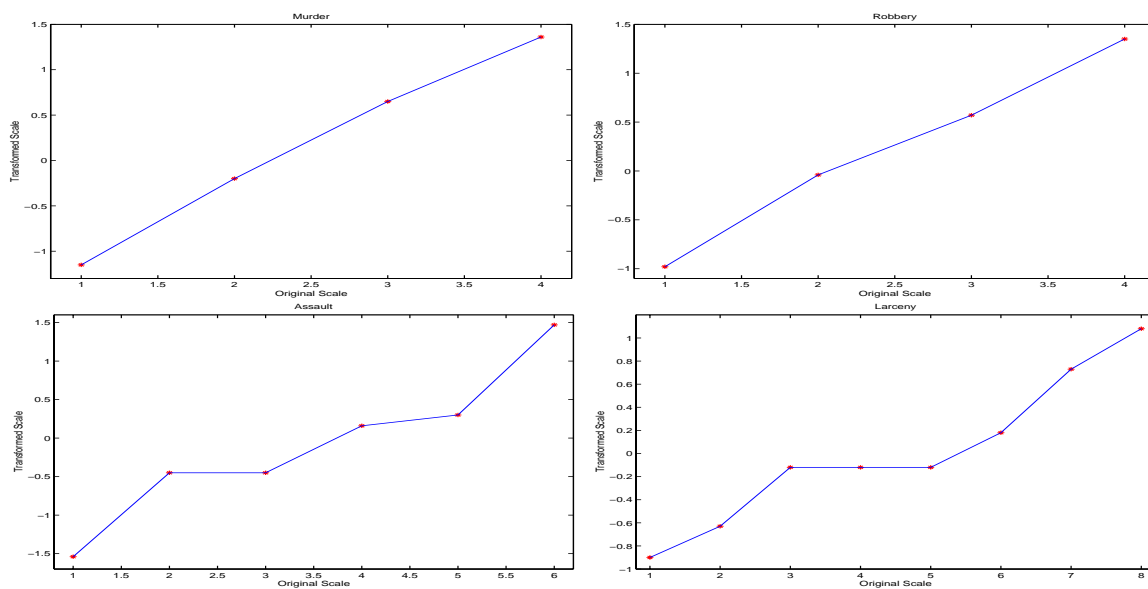

FIGURE 3.2. Variable Transformations

the other crime areas (New York, Philadelphia, Los Angeles, Long Beach, Houston, San Francisco, Jersey City) and in the lower right cities 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 burglaries (Oklahoma, Minneapolis, Baton Rouge, Kansas City, Burmingham), while cities in the center are somwhere in the middle with respect to crime (e.g. Denver, Dallas, Las Vegas, Phoenix, Boston, Cleveland, Seattle to name a few). Finally, in the upper left we find a cluster of cities that have higher numbers of larceny and rape 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 [60] 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 [61] 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 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 [36].

Starting with Steel [89] various attempts have been made and approaches suggested to generalize Hotelling's procedure to K sets of variables. In the two set problem, the canonical correlation serves as the single criterion. In a K set problem there exist $1/2K(K-1)$ canonical correlations between the optimal set of canonical variables that can be collected in a $K \times K$ correlation matrix $\mathcal{R}$. The generalizations deal with different criteria that can be formulated as functions of the matrix $\mathcal{R}$. The most common ones (discussed in the book by Gittins [36] that deals exclusively with this subject) are:

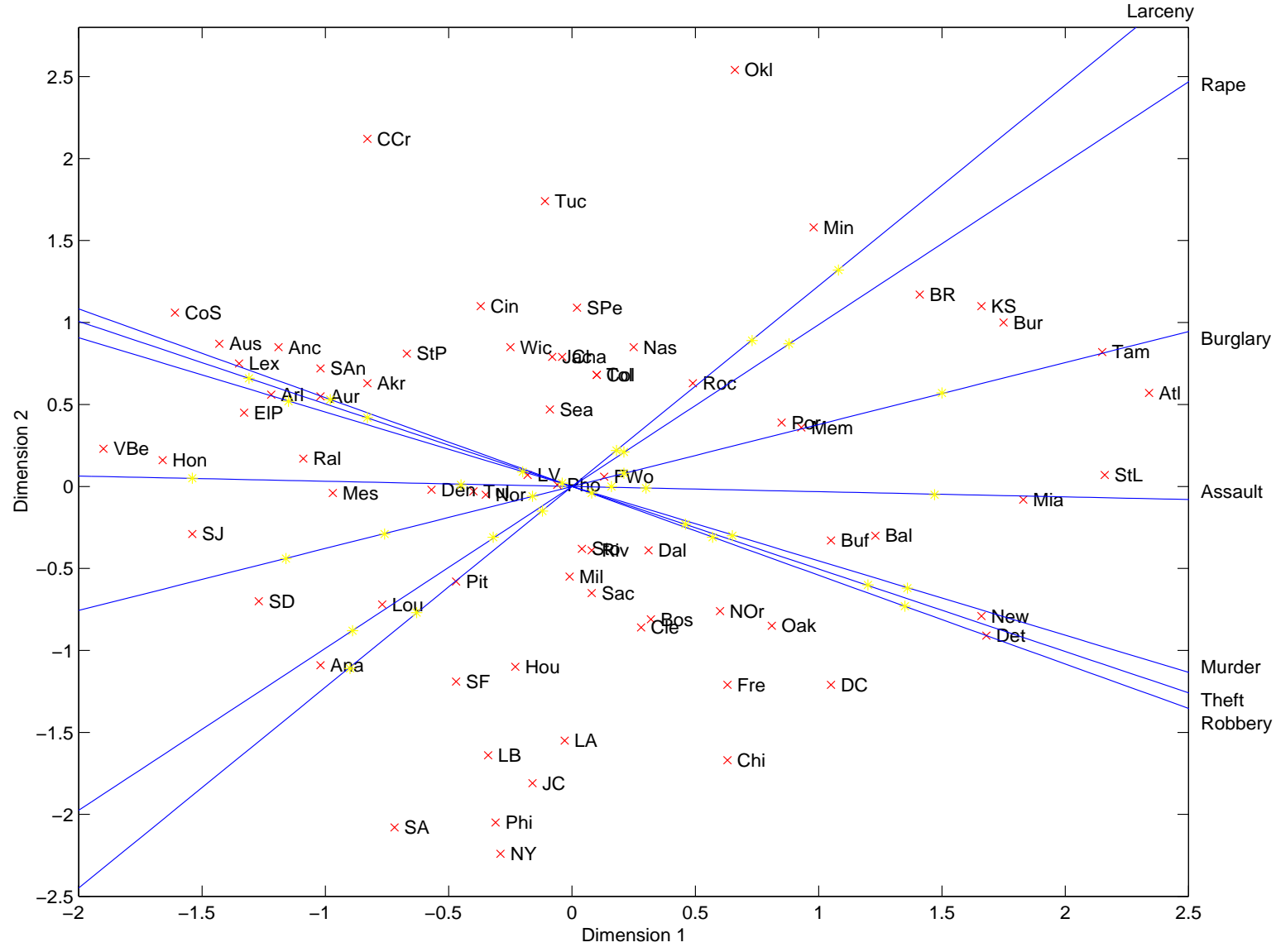THE GIFI SYSTEM OF DESCRIPTIVE MULTIVARIATE ANALYSIS

FIGURE 3.3. Variable Quantifications (not labeled) and U.S. Cities (The solid lines indicate the projections of the transformed variables. The placement of the variable labels indicates the high crime rate direction)

1. Minimize the determinant of $\mathcal{R}$, or equivalently minimize the product of the eigenvalues of $\mathcal{R}$, proposed by Steel [89].
2. Maximize the sum of the correlations in $\mathcal{R}$, proposed by Horst [58, 59].
3. Maximize the largest eigenvalue of $\mathcal{R}$, also proposed by Horst [58, 59].
4. Maximize the sum of squares of the eigenvalues of $\mathcal{R}$, proposed by Kettenring [65].
5. Minimize the smallest eigenvalue of $\mathcal{R}$, also proposed by Kettenring [65].
6. Maximize the largest eigenvalue of $\mathcal{R}$, proposed by Carroll [12].

The last criterion is equivalent to maximizing the sum of correlations between each canonical variable and an unknown coordinate vector x [32]. 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 [84], de Leeuw [19], and van der Burg et al. [96].

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

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

subject to the constraints $X'X = N I_p$ and $u'X = 0$. Equation (4.1) implies that all variables within each set $J(k)$, $k = 1, \ldots, K$ are treated as *additive*. The latter implies that optimal transformations of a variable j within a set $J(k)$ depends 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

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

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

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

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-one 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 Fortran computer program that implements the algorithm as the Overals program [100]. 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 analogous steps to those considered in section 2.2, it can be shown that the minimum loss function is given by

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

where λ_s, $s = 1, \ldots, p$ are the eigenvalues of the matrix $\mathcal{L}P_\star\mathcal{L}$, 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.**

4.1. **An Example: School Climate.** The data for this example come from the National Education Longitudinal Study of 1988 (NELS:88). 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 behaviors on student performance. For example goal six of the National Education Goals Panel [77] 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 anatmosphere 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 [63] 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, 81]. 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 $J(1) = (A - C), J(2) = (D - I), J(3) = (J, K)$. The first set characterizeds attendance patterns of students, the second set deals with issues that affect the overall school environment, and the third with attitudes of students towards their teachers. The generalized canonical correlations for the two dimensional solution are .65 and .36 respectively, and the mulitple correlations of the three sets with each canonical variable (dimension) are

.67, .86, .75 for the first dimension and .45, .69 and .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 4.1. 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 component, while variables A (tardiness), D (physical conflicts), F (vandalism), G (alcohol), and K (verbal abuse of teachers) load negatively on the second canonical variable. Hence, the first canonical variable can be interpreted as an overall measure of school climate, while the second one distinguishes between students that experience a rough and potentially dangerous school environment (upper half), with those that experience a simply rough environment.

By examining the category points plot (see Figure 4.1) 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 at 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 4.1 the object scores are ploted, together with their frequencies and the category points (lie on the red lines). The object point with the large spike on the left end 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. On the other hand, the big spike on the right end of the picture corresponds to the other extrem 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 are at most a minor problem in the respective schools.

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 focus of these techniques is in 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
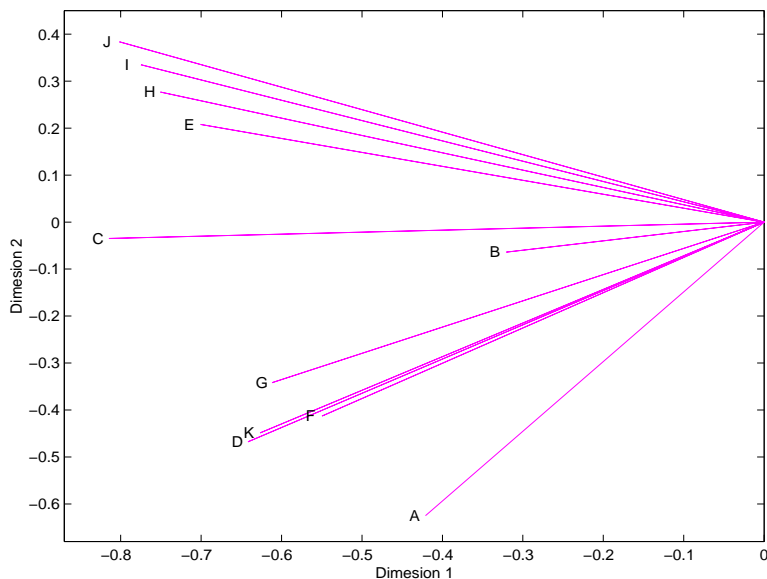
FIGURE 4.1. Component Loadings
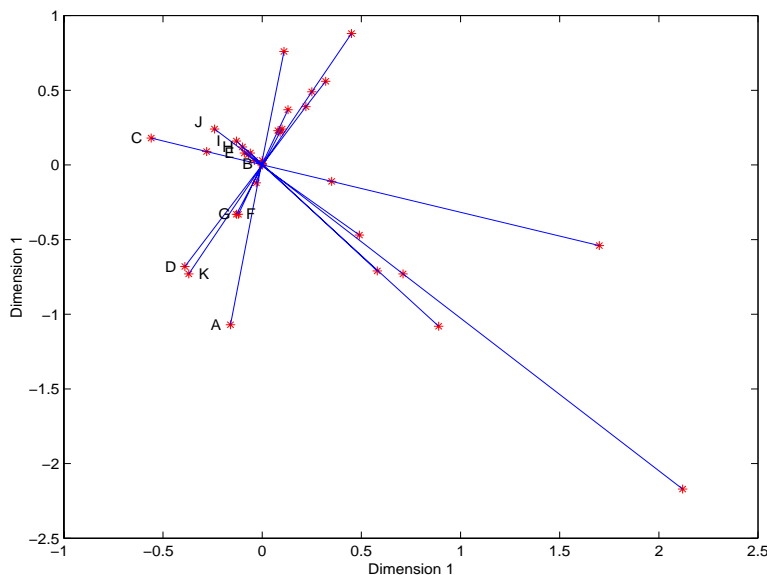
FIGURE 4.2. Category points (variable label points towards the highest -i.e. not a problem- category)

the analysis" (see [64], p. 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*)

FIGURE 4.3. Object scores and their respective frequencies; green points corre-
spond to frequencies $leq5$ and the red lines indicate the location of the category
points on the map

(see [32] p. 36). By reversing this definition, we get that in the Gifi system a result can be charac-
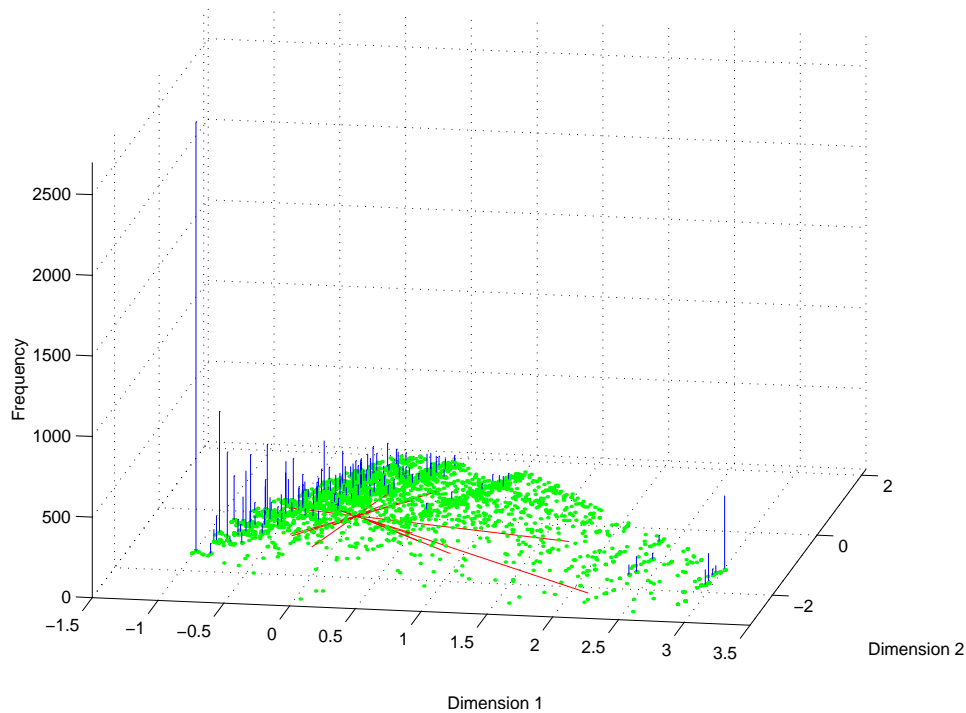terized as *unstable* in those instances where small and unimportant changes in the input result in
significant changes in the output. In our analysis, we consider as input the data set at hand -objects
and variables-, the coding of the variables, the dimensionality of the solution, the measurement
level of the variables, the type of technique employed -Homals, Princals-, and as output category
quantifications, object scores, discrimination measures, eigenvalues, component loadings etc. 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 output.

The most important forms of stability are:

(a) *Replication Stability.* If a new data set is sampled and apply the same technique to this new
    set, then the results should not change dramatically.
(b) *Statistical Stability.* It refers to the stability of the analysis whenever no new data set is
    formally sampled.

(c) *Stability under Data Selection.* Variations in the data are considered, by omitting either objects from the data set or variables from subsequent analysis. The former corresponds to rejection of outliers and *resampling* techniques. In this framework, resampling techniques can be thought of as a form of replication stability, but without formally sampling a new data set.

(d) *Stability under Model Selection.* Small changes in the model should result in small changes in the results obtained.

(e) *Numerical Stability.* It refers to the influence of rounding errors and of computation with limited precision on the results given by the techniques.

(f) *Analytic and Algebraic Stability.* If the data structures and possible representations have enough mathematical structure, then formal expressions of the input-output analysis can be drawn from considering perturbations of the input.

(g) *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 would focus primarily on stability under data selection. However, we are also going to 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 [47] 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 in 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 underlying 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. 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], (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.

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 [22] and Molenaar [76] and references therein, see also Leamer [69] for some provocative thoughts on the subject from the Bayesian viewpoint). The former approach is closer to the exploratory data analytic school, while the latter closer to the dictates of classical statistics. Moreover, it is easy to see that the inferential use of the models agrees with the concept of external stability, while the descriptive use with that of internal stability. We believe that both approaches are useful and informative whenever used in the appropriate context. For example, when dealing with a data set from a well designed study -correctly identified population, proper sampling mechanism, proper collection of the data set, so that another sample can be drawn in the future- it is reasonable to be interested not only in examining the structure of that particular data set, but also in generalizing the findings to the population. Stability of the results is crucial in this exercise, because if the solution is not stable -small changes in the input lead to large and/or important changes in the output (results)- then the second part of the exercise -generalization from the sample to the population- becomes irrelevant. On the other hand, sometimes it is not possible to specify with great accuracy the population from which the sample was drawn, or the sampling mechanism used to collect the sample, or we might be interested just in the structure of that particular data set. In this frame of reference, the descriptive approach is the appropriate or relevant one, and stability is associated with the concept of internal stability. The data set regarding multivariate analysis books in chapter 1 in the Gifi book [32], and samples of convenience are typical examples of this type.

In the remainder of this section we will address stability issues related to merging merging categories, omitting a variable from the analysis and the bootstrap. The first two topics have been briefly addressed in [32], while the third one has been examined in [32, 73, 20, 95, 72].

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 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 for example [62]), 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 $\eta_{js}^2 - \gamma_s,\ s = 1, \ldots, p$, 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 [31]. 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 [31]. 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 5.1 give the freqency distribution of the first and second eigenvalues of the homogeneity analysis solution over 1000 replications. It can immediately be seen that the observed eigenvalues of .73 and .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.

5.4. **Replication Stability and the Bootstrap.** The previous two subsections have provided some analytic 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 etc. On the other hand recent advances in resampling techniques -particularly
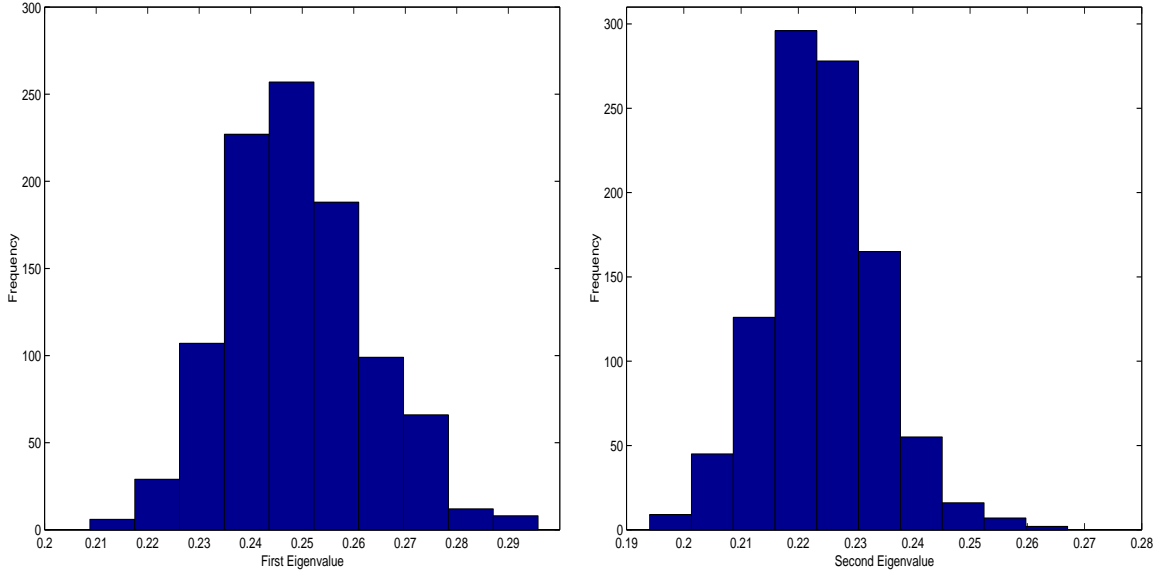
FIGURE 5.1. Permutation Frequency Distribution of the Eigenvalues of the Homogeneity Analysis Solution of the mammals dentition example

the bootstrap [26]- offers an interesting and useful alternative to study stability issues for the techniques presented in this paper.

We develop the method in a general context (for a comprehensive account see also [95]). Suppose we have J categorical variables. Each variable takes values in a set $\mathcal{S}_j$ (the range of the variable [32]) of cardinality ℓ_j (number of categories of variable j). Define $\mathcal{S} = \mathcal{S}_1 \times \ldots \mathcal{S}_J$ to be the *profile space*, that has cardinality $\ell = \prod_{j=1}^J \ell_j$. That is the space $\mathcal{S} = \{(s_1, \ldots, s_j), \ s_j \in \mathcal{S}_j, \ j \in \mathbf{J}\}$ contains the J-tuples of profiles. Let S be a $\ell \times \sum_{j=1}^J \ell_j$ 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 $\mathcal{S}$ to its individual components. Let also G_S be a $N \times \ell$ 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 $G = [G_1 | \ldots | G_J]$ 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 $\mathcal{S}$.

Let $\mathcal{P}$ be a probability distribution on $\mathcal{S}$. Since the space $\mathcal{S}$ is finite, $\mathcal{P}$ 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 $\phi(.)$ (Hadamard or Fréchet differentiability [87]) 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 $P(\phi(p) \leq z|p)$ [71, 86], 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 Cnfidence Regions* Two of the main issues in the theory of bootstrap are: (i) how to produce *unbiased* 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 [72] 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, while 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 et al. [101] 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 [90]. Heiser and Meulman [55] suggested to construct 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 [72] 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 [45]). 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- [32, 95, 73]. The most comprehensive one is the monograph by Markus [72]. In this section we will attempt to briefly summarize 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 1,000).

(2) The bootstrap confidence regions give on average the right coverage probabilities. However, for categories with low marginal frequencies the coverage probabilities might be under- 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 seems to be the absolute minimum to ensure valid confidence regions. In light of this finding, merging categories appears to be not only overall beneficial, 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 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 5.2 we present the bias corrected bootstrap means of some of the optimal transformations, along with ± 2 standard error bands of the category quantifications of the mammals dentition data set, based on 1,000 bootstrap replications. The bootstrapped means and standard errors (in parentheses) of the eigenvalues in dimensions 1 and 2 are .738 (.035) and .386 (.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. Regarding the category quantifications we see that the first dimension exhibits a far more stable behavior than the second one. 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, 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 various other types of nonlinear multivariate techniques. These techniques have been extensively used in data analytic situations. In the Gifi book [32] the entire chapter 13 is devoted to applications covering the fields of education, sociology and psychology. Also, in their books Greenacre [47] and Benzécri [4] give
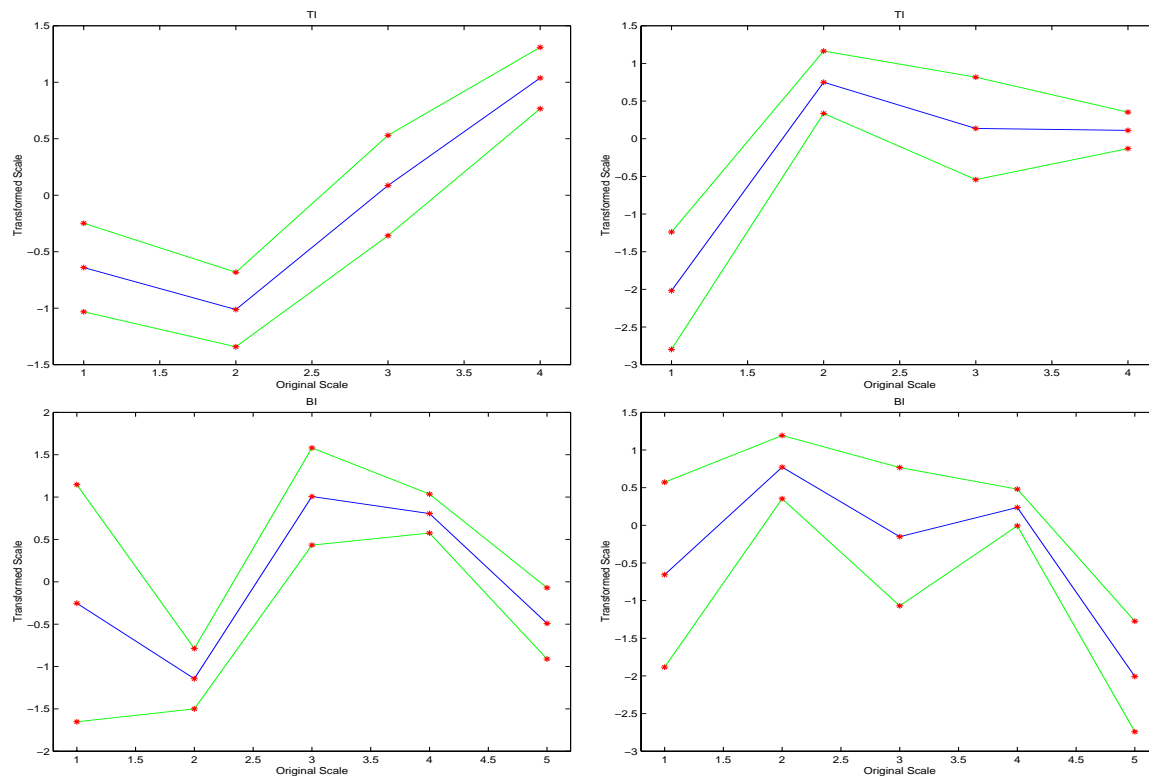
FIGURE 5.2. Bootstraps

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 [57], zoology [94], environmental studies [93], medicine [95] and food science [97]. However, the Gifi system has evolved beyond homogeneity analysis and its generalizations; hence, new techniques have been developed for path models [32], time series models [92], linear dynamical systems [5] etc. In closing, it should be mentioned that the Gifi system is part of a still quite active research program.

7. **Appendix A - Dentition of Mammals**

The data for this example are taken from Hartigan's book [51]. Mammals' teeth are divided into four groups: incisors, canines, premolars and molars. A description of the variables with their respective coding is given next.

TI: Top incisors; 1: 0 incisors, 2: 1 incisor, 3: 2 incisors, 4: 3 or more incisors
BI: Bottom incisors; 1 : 0 incisors, 2: 1 incisor, 3: 2 incisors, 4: 3 incisors, 5: 4 incisors
TC: Top canine; 1: 0 canines, 2: 1 canine
BC: Bottom canine; 1: 0 canines, 2: 1 canine
TP: Top premolar; 1: 0 premolars, 2: 1 premolar, 3: 2 premolars, 3: 2 premolars, 4: 3 premolars, 5: 4 premolars
BP: Bottom premolar; 1: 0 premolars, 2: 1 premolar, 3: 2 premolars, 3: 2 premolars, 4: 3 premolars, 5: 4 premolars
TM: Top molar; 1: 0-2 molars, 2: more than 2 molars
BM: Bottom molar; 1: 0-2 molars, 2: more than 2 molars

In the following table the frequencies of the variables are given.

Variable	Categories				
	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			

TABLE 7.1. Mammals teeth profiles (in %, N=66)

Mammals Dentition Data

Mammal		Mammal	
Opposum	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 Rabit	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	Muskox	15114422
Wolf	44225512	Mountain Sheep	15114422

8. Appendix B - Crime Rates in US 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+

Crime Rates Data

City	City Code		City	City Code	
New York (NY)	NY	2132212	Los Angeles (CA)	LA	2233222
Cicago (IL)	Chi	3*43332	Houston (TX)	Hou	2222322
Philadelphia (PA)	Phi	3231112	San Diego (CA)	SD	1112222
Phoenix (AZ)	Pho	2212442	Dallas (TX)	Dal	3422332
Detroit (MI)	Det	3543433	San Antonio (TX)	SAn	2211341
Honolulu (HI)	Hon	1111231	San Jose (CA)	SJ	1212111
Las Vegas (NV)	LV	1322322	San Francisco (CA)	SF	1132232
Baltimore (MD)	Bal	3443442	Jacksonville (FL)	Jac	2422441
Columbus (OH)	Col	2521432	Milwaukee (WI)	Mil	2321232
Memphis (TN)	Mem	3522522	Washnigton, DC	DC	3233342
El Paso (TX)	ElP	1212131	Boston (MA)	Bos	2423233
Seattle (WA)	Sea	1222352	Charlotte (NC)	Cha	2323441
Nashville (TN)	Nas	1423342	Austin (TX)	Aus	1211241
Denver (CO)	Den	2311322	Cleveland (OH)	Cle	3522312
New Orleans (LA)	NOr	3432432	Fort Worth (TX)	FWo	3422342
Portland (OR)	Por	1423353	Oklahoma City (OK)	Okl	1512552
Long Beach (CA)	LB	2122322	Tucson (AZ)	Tuc	1312352
Kansas City (MO)	KS	3533542	Virginia Beach (VA)	VBe	1111121
Atlanta (GA)	Atl	3544553	Saint Louis (MO)	StL	3344553
Sacramento (CA)	Sac	2222433	Fresno (CA)	Fre	2222433
Tulsa (OK)	Tul	1312322	Miami (FL)	Mia	3244553
Oakland (CA)	Oak	3433432	Minneapolis (MN)	Min	2532542
Pittsburgh (PA)	Pit	2321222	Cincinatti (OH)	Cin	1522331
Toledo (OH)	Tol	1521432	Buffalo (NY)	Buf	3433522
Wichita (KS)	Wic	1311442	Mesa (AZ)	Mes	1112332
Colorado Springs (CO)	Cos	1311131	Tampa (FL)	Tam	2534553
Santa Ana (CA)	SA	3121112	Arlington (VA)	Arl	1212231
Anaheim (CA)	Ana	1122222	Corpus Cristi (TX)	CCr	1312351
Louisville (KY)	Lou	2221311	St Paul (MN)	StP	1412321
Newark (NJ)	New	3344533	Birmingham (AL)	Bir	3524552
Norfolk (VA)	Nor	2321231	Anchorage (AK)	Anc	1312131
Aurora (CO)	Aur	1213231	St Petersburg (FL)	SPe	1424441
Riverside (CA)	Riv	2223422	Lexington (KY)	Lex	1212231
Rochester (NY)	Roc	3321541	Jersey City (NJ)	JC	2132412
Raleigh (NC)	Ral	1112331	Baton Rouge (LO)	BRo	3324552
Akron (OH)	Akr	1412221	Stockton (CA)	Sto	2222432

9. **Appendix C**

Description of variables:

A: Student tardiness a problem at school
B: Student absenteeism a problem at school
C: Students cutting class a problem at school
D: Physical conflicts among students a problem at school
E: Robbery or theft a problem at school
F: Vandalism of school property a problem at school
G: Student use of alcohol a problem at school
H: Student use of illegal drugs a problem at school
I: Student possession of of weapons a problem at school
J: Physical abuse of teachers a problem at school
K: Verbal abuse of teachers a problem at school

The four possible answers to each of the variables are:

1: Serious
2: Moderate
3: Minor
4: Not a problem

Variable	Categories			
	1	2	3	4
A	2708	6264	7685	6591
B	2611	6356	7508	6773
C	3435	4032	5862	9919
D	3724	5774	7425	6325
E	3178	3462	6918	9690
F	3392	3536	6759	9561
G	3508	3414	5057	11269
H	3281	2412	4786	12769
I	2642	2180	5335	13091
J	1811	666	2294	18377
K	2633	3258	6150	11207

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 edition, New York: Wiley

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

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

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

[6] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press

[7] Bond, J. and Michailidis, G. (1996), "Homogeneity Analysis in Lisp-Stat," *Journal of Statistical Software*, **1**, issue 2

[8] Breiman, L. and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, **80**, 580-619

[9] Buja, A. (1990), "Remarks on Functional Canonical Variates, Alternating Least Squares Methods and ACE," *The Annals of Statistics*, **18**, 1032-1069

[10] Cailliez, F. and Pagès, J.P. (1976), *Introduction a l'Analyse des Données*, Paris: SMASH

[11] Carnegie Foundation of the Advancement of Teaching (1988) *An Imperiled Generation: Saving Urban Schools*, Princeton: The Carnegie Foundation

[12] Carroll, J.D. (1968), "Generalization of Canonical Correlation Analysis to Three or More Sets of Variables," *Proceedings of the 76th Convention of the American Psychological Association*, **3**, 227-228

[13] Clogg, C.C. (1981), "New Developments in Latent Structure Analysis," *Factor Analysis and Measurement in Sociological Research*, Jackson and Borgatta (eds), Beverly Hills: Sage, 215-246

[14] Clogg, C.C. (1984), "Latent Structure Analysis of a Set of Multidimensional Contingency Tables," *Journal of the American Statistical Association*, **79**, 762-771

[15] Clogg, C.C. (1986), "Statistical Modeling Versus Singular Value Decomposition," *International Statistical Review*, **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., and van Rijckevorsel, J. (1980), "Homals and Princals. Some Generalizations of Principal Components Analysis," *Data Analysis and Informatics II*, Diday et al. (eds.), 231-242, Amsterdam: North Holland

[18] de Leeuw, J. (1983), "On the Prehistory of Correspondence Analysis," *Statistica Neerlandica*, **37**, 161-164

[19] de Leeuw, J. (1984), "The Gifi-system of Nonlinear Multivariate Analysis," *Data Analysis and Informatics III*, Diday et al. (eds.), 415-424, Amsterdam: North Holland

[20] de Leeuw, J. (1985), "Jackknife and Bootstrap Methods in Multinomial Situations," Research Report 85-16, Dept of Data Theory, Leiden University

[21] de Leeuw, J. and van der Burg, E. (1984), "The PermutationalLimit Distribution of Generalized Canonical Correlations," *Data Analysis and Informatics IV*, Diday et al. (eds), 509-521, Amsterdam: North Holland

[22] de Leeuw, J. (1988), "Models and Techniques," *Statistica Neerlandica*, **42**, 91-98

[23] Eades, P. and Sugiyama, K. (1990), "How to Draw a Directed Graph," *Journal of Information Processing*, **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. (1995), "Algorithms for Drawing Graphs: an Annotated Bibliography," *Computational Geometry: Theory and Applications*,

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

[27] Escofier, Y. (1984), "Analyse Factorielle en Référence a un Modele: Application a l'Analyse de Tableaux d'Echanges," *Revue Statistique Appliqué*, **32**, 25-36

[28] Escoufier, Y. (1985), "L'Analyse des Correspondances: Ses Propriétés et ses Extensions," *Bulletin of the International Statistical Institute*, **51**, 1-16

[29] Escoufier, Y. (1988), "Beyond Correspondence Analysis," *Classification and Related Methods of Data Analysis*, Bock (ed), Amsterdam: Elsevier

[30] Fisher, R. A. (1938), "The Precision of Discriminant Functions," *The Annals of Eugenics*, **10**, 422-429

[31] Freedman, D.A., and Lane, D. (1983), "Significance Testing in a Nonstochastic Setting," *A Festschrift for Erich L. Lehmann*, Bickel et al. (eds), 185-208, Belmont: Wadsworth

[32] Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Chichester: Wiley

[33] Gilula, Z. (1986), Grouping and Association in Contingency Tables: An Exploratory Canonical Correlation Approach," *Journal of the American Statistical Association*, **81**, 773-779

[34] Gilula, Z. and Haberman, S.J. (1986), "Canonical Analysis of Two-Way Contingency Tables by Maximum Likelihood," *Journal of the American Statistical Association*, **81**, 780-788

[35] Gilula, Z. and Ritov, Y. (1990), "Inferential Ordinal Correspondence Analysis: Motivation, Derivation and Limitations," *International Statistical Review*, **58**, 99-108

[36] Gittins, R. (1985), *Canonical Analysis; a Review with Applications in Ecology*, Berlin: Springer-Verlag

[37] Gnanadesikan, R. and Kettenring, J.R. (1984), "A Pragmatic Review of Multivariate Methods in Applications," *Statistics: an Appraisal*, in David and David (Eds.), Ames: The Iowa State University Press

[38] Golub, G.H. and van Loan C.F. (1989), *Matrix Computations*, Baltimore: Johns Hopkins University Press

[39] Goodman, L.A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, **61**, 215-231

[40] Goodman, L.A. (1979),"Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, **74**, 537-552

[41] Goodman, L.A. (1981), "Association Models and Canonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, **76**, 320-334

[42] Goodman, L.A. (1985), "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models and Assymetry Models for Contingency Tables with or without Missing Entries," *The Annals of Statistics*, **13**, 10-69

[43] 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," *International Statistical Review*, **54**, 243-309

[44] Goodman, L.A. (1994), "On Quasi-Independence and Quasi-Dependence in Contingency Tables, with Special Reference to Ordinal Triangular Contingency Tables," *Journal of the American Statistical Association*, **89**, 1059-1063

[45] 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

[46] Green, P.J. (1981), "Peeling Bivariate Data," *Interpreting Multivariate Data*, Barnett (ed.), New York: Wiley

[47] Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press

[48] Greenacre, M. and Hastie, T. (1987), "The Geometric Interpretation of Correspondence Analysis," *Journal of the American Statistical Association*, **82**, 437-447

[49] 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.), New York: Social Science Research Council

[50] Haberman, S.J. (1970), *Analysis of Qualitative Data, Vol. 2: New Developments*, New York: Academic Press

[51] Hartigan, J.A. (1975), *Clustering Algorithms*, New York: Wiley

[52] Hastie, T., Tibshirani, R. and Buja, A. (1994), "Flexible Discriminant Analysis with Optimal Scoring," *Journal of the American Statistical Association*, **89**, 1255-1270

[53] Hastie, T., Buja, A. and Tibshirani, R. (1995), "Penalized Discriminant Analysis," *The Annals of Statistics*, **23**, 73-102

[54] Hayashi, C. (1952), "On the Prediction of Phenomena From Qualitative Data and the Quantification of Qualitative Data from the Mathematico-statistical Point of View," *Annals of the Institute of Statistical Mathematics*, **5**, 121-143

[55] Heiser, W.J. and Meulman, J.J. (1983), "Constrained Multidimensional Scaling," *Applied Psychological Measurement*, **7**, 381-404

[56] Hirscfeld, H.O. (1935), "A Connection between Correlation and Contingency," *Cambridge Philosophical Society Proceedings*, **31**, 520-524

[57] Hoffman, D.L. and de Leeuw, J. (1992), "Interpreting Multiple Correspondence Analysis as a Multidimensional Scaling Method," *Marketing Letters*, **3**, 259-272

[58] Horst, P. (1961), "Relations among m Sets of Measures," *Psychometrika*, **26**, 129-149

[59] Horst, P. (1961), "Generalized Canonical Correlations and their Application to Experimental Data," *Journal of Clinical Psychology*, **17**, 331-347

[60] Hotelling, H. (1935), "The most Predictable Criterion," *Journal of Educational Psychology*, **26**, 139-142

[61] Hotelling, H. (1936), "Relations between two Sets of Variables," *Biometrika*, **28**, 321-377

[62] Kato, T. (1995), *Perturbation Theory of Linear Operators*, Berlin: Springer-Verlag

[63] Kaufman, P., Bradby, D. (1992), *Characteristics of at Risk Students in NELS:88*, Report 92-042, Washington: National Center for Education Statistics

[64] Kendall, M.G. (1980), *Multivariate Analysis*, 2nd edition, London: Griffin

[65] Kettenring, J.R. (1971), "Canonical Analysis of Several Sets of Variables," *Biometrika*, **58**, 433-460

[66] Kruskal, J.B. and Shepard, R.N. (1974), "A Nonmetric Variety of Linear Factor Analysis," *Psychometrika*, **39**, 123-157

[67] Kshirsagar, A.N. (1978), *Multivariate Analysis*, New York: Marcel Dekker

[68] Lazarsfeld, P.F. and Henry, N.W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin

[69] Leamer, E.E. (1978), *Specification Searches: Ad Hoc Inferences from Nonexperimental Data*, New York: Wiley

[70] Lebart, L., Morineau, A., Tabard, N. (1977), *Technique de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*, Paris: Dunod

[71] Liu, R.Y., Singh, K. and Lo, S. (1989), "On a Representation Related to the Bootstrap," *Sankhya, A*, **51**, 168-197

[72] Markus, M.T. (1994), *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*, Leiden: DSWO Press

[73] Meulman, J.J. (1984), "Correspondence Analysis and Stability," Research Report 84-01, Dept of Data Theory, Leiden University

[74] Michailidis, G. and de Leeuw, J. (1995), "Nonlinear Multivariate Analysis of NELS:88," UCLA Statistical Series Preprints, # 175

[75] Michailidis. G. and de Leeuw, J. (1996), "The Gifi System of Nonlinear Multivariate Analysism" UCLA Statistical Series Preprints, #204

[76] Molenaar, I.W. (1988), "Formal Statistics and Informal Data Analysis, or why Laziness Should Be Discouraged," *Statistica Neerlandica*, **42**, 83-90

[77] The National Education Goals Panel (1992), *The National Education Goals Report: Building a Nation of Learners*, Washington

[78] Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*, Toronto: Toronto University Press

[79] Nishisato, S. and Nishisato. I (1984), *An Introduction to Dual Scaling*, Toronto, Microstats

[80] Nishisato, S. (1994), *Elements of Dual Scaling. An Introduction to Practical Data Analysis*, Hillsdale: Lawrence Erlbaum Associates

[81] Oakes, J. (1989), "What Educational Indicators? The Case for Assessing the School Context," *Educational Evaluation and Policy Analysis*, **11**, 181-199

[82] Ritov, Y. and Gilula, Z. (1993), "Analysis of Contingency Tables by Correspondence Models Subject to Order Constraints," *Journal of the American Statistical Association*, **88**, 1380-87

[83] Roskam, E.E. (1968), *Metric Analysis of Ordinal Data in Psychology*, Voorschoten: VAM

[84] Saporta, G. (1975), "Liaisons entre plusieurs Ensembles de Variables et Codage de Données Qualitatives," Paris: Universite de Paris VI

[85] Schriever, B.F. (1983), "Scaling of Order Dependent Categorical Variables with Correspondence Analysis," *International Statistical Review*, **51**, 225-238

[86] Shao, J. (1992), "Some Results for Differentiable Statistical Functionals," *Nonparametric Statistics and Related Topics*, Saleh (ed), 179-188, Amsterdam: North Holland

[87] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag

[88] SPSS Categories User's Manual, SPSS Inc.

[89] Steel, R.G.D. (1951), "Minimum Generalized Variance for a Set of Linear Functions," *Annals of Mathematical Statistics*, **22**, 456-460

[90] Takane, Y. and Shibayama, T. (1991), "Principal Component Analysis with External Information on Both Subjects and Variables," *Psychometrika*, **56**, 97-120

[91] van Buuren, S. and Heiser, W.J. (1989), "Clustering N Objects into K Groups under Optimal Scaling of Variables," *Psychometrika*, **54**, 699-706

[92] van Buuren, S. (1990), *Optimal Scaling of Time Series*, Leiden: DSWO Press

[93] van Buuren, S. (1994), "Groupals Analysis of Abiotic Measures from an Environmental Study in the Archipel of Hochelage," Research Report 94-14, Dept of Data Theory, University of Leiden

[94] van der Burg, E. (1985), "Homals Classification of Whales, Porpoises and Dolphins," *Data Analysis in Real Life Environment: Ins and Outs of Solving Problems*, Proth et al. (Eds.), 25-35, Amsterdam: North Holland

[95] van der Burg, E. and de Leeuw, J. (1988), "Use of the Multinomial Jackknife and Bootstrap in generalized Canonical Correlation Analysis," *Applied Stochastic Models and Data Analysis*, **4**, 154-172

[96] 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

[97] van der Burg, E., de Leeuw, J. and Dijksterhuis, G. (1994), "Nonlinear Canonical Correlation with k Sets of Variables," *Computational Statistics & Data Analysis*, **18**, 141-163

[98] 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," *Applied Statistics*, **38**, 249-292

[99] van Rijckevorsel, J.L.A. (1987), *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*, Leiden: DSWO Press

[100] Verdegaal, R. (1986), "OVERALS," *Research Report UG-86-01*, Leiden: Department of Data Theory

[101] Weinberg, S.L., Caroll J.D. and Cohen, H.S. (1984), "Confidence Regions for INDSCAL using Jackknife and Bootstrap Techniques," *Psychometrika*, **49**, 475-491

[102] 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

[103] Young, F.W. (1981), "Quantitative Analysis of Qualitative Data," *Psychometrika*, **46**, 357-388

INTERDIVISIONAL PROGRAM IN STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, LOS ANGELES, CA 90095