Title
Interactive Correspondence Analysis in a Dynamic Object-Oriented Environment

Authors
Jason Bond
George Michailides

Publication Date
2011-10-13

INTERACTIVE CORRESPONDENCE ANALYSIS IN A DYNAMIC OBJECT-ORIENTED ENVIRONMENT

JASON BOND AND GEORGE MICHAILIDIS

ABSTRACT. A highly interactive, user-friendly object-oriented software package written in Lisp-Stat is introduced that performs simple and multiple correspondence analysis, and profile analysis. These three techniques are integrated into a single environment driven by a user-friendly graphical interface that takes advantage of Lisp-Stat's advanced graphical capabilities. Techniques that assess the stability of the solution are also introduced. Some of the features of the package include colored graphics, incremental graph zooming capabilities, manual point separation to determine identities of overlapping points, and stability and fit measures. The features of the package are used to show some interesting trends in a large educational dataset.

1. **Introduction**

Exploratory data analytic techniques have become increasingly popular over the last decade. One of the main reasons for their popularity is that they are primarily intended to reveal features in the data, by producing low dimensional maps in order to summarize the data, rather than to test hypotheses about the underlying mechanism that generated the data. This practice is particularly suitable for various fields in the social and biological sciences, where data practitioners are confronted by large data sets, especially in terms of the number of variables involved, and therefore a specific model is hard to postulate. However, most implementations of these techniques have followed the example of other classical statistical methods, with lots of printed output, a few low quality static graphs, and a batch processing mode. Obviously such programs are unsuitable for these techniques that by nature require a high degree of interaction between the analyst and the data , and also heavily depend on high quality graphical displays. Recent advances in computer technology (dynamic real time graphics, menu driven programs, etc) have made possible a shift in the development of statistical software towards truly interactive and dynamic environments. In this paper we integrate three exploratory data analytic methods suitable for categorical data, namely correspondence analysis of contingency tables, multiple correspondence analysis and analysis of profiles into a program written in the Lisp-Stat language [27] that offers the user a high degree of interaction with the data, high quality dynamic graphics, and the capability of assessing the stability of the derived maps. The latter is usually an integral part of exploratory data analysis, since the data analyst has to examine whether the discovered patterns are real or merely due to chance.

Date: June 17, 1997.

2. **A Brief Account of Correspondence Analysis**

Correspondence analysis (CA) is an exploratory multivariate technique that converts frequency table data into graphical displays in which the rows and the columns of the table are depicted as points. Mathematically, CA decomposes the χ^2-measure of association of the table into components in a manner similar to that of principal component analysis for continuous data.

In CA no model is introduced, and no assumptions about the underlying stochastic mechanism that generated the data at hand are made, contrary to the approach taken in logliner analysis [2], one of the most frequently used alternatives for the analysis of multivariate categorical data. The primary interest in CA is in the presentation of the structure of the observed data. This rationale has been developed into an official principle by Benzécri and his co-workers [1]. CA can be traced back in the work of Hirschfeld [14], although some of the basic ideas can be found in the work of Pearson and his debate with Yule [5]. It has been rediscovered in various forms and in different contexts in the work of Fisher [7], Guttman [12], Hayashi [13] and especially Benzécri who paid special attention to the geometric form of the method. Extensive accounts of the history of the technique and its similarities and differences with other methods such as dual scaling, simultaneous linear regressions, and canonical correlation are provided in the books by Nishisato [19] and Greenacre [9].

We can distinguish between *simple* CA (CA of contingency tables) and *multiple* CA, a generalization of CA designed to handle more than two categorical variables.

2.1. **Simple Correspondence Analysis.**

Let F be an $I \times J$ contingency table, whose entries $F(i,j)$[1] give the frequency 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, where u is the unit vector. Let $D_r = \text{diag}\,(r)$ denote the diagonal matrix containing the elements of vector r and $D_c = \text{diag}\,(c)$ the diagonal matrix containing the elements of vector c.

Correspondence analysis is a technique with which it is possible to find a multidimensional representation of the dependencies between the rows and the columns of F. We can calculate the so-called χ^2-distances between rows, as well as between columns. The χ^2-distance between rows i and i' of table F is given by

$$\delta^2(i,i') = N \sum_{j=1}^{J} \frac{\left(F(i,j)/r(i) - F(i',j)/r(i')\right)^2}{c(j)}. \tag{2.1}$$

[1]The (i,j) element of matrix A is denoted by $A(i,j)$, the i^{th} row by $A(i,.)$ and the j^{th} column by $A(.,j)$. Similarly, the i^{th} element of a vector a is denoted by $a(i)$.

Formula (2.1) shows that $\delta^2(i,i')$ is a measure for the difference between the profiles of rows i and i'. Whenever rows i and i' have the same profile $\delta^2(i,i')=0$. The difference between profiles i and i' for column j is divided by $c(j)$, thus giving less influence to points for column categories that have large marginals. The configuration of I row points is located in a Euclidean space of dimension $I-1$. In that space, coordinates X can be found so that $\delta^2(i,i')$ would be the same as the squared Euclidean distance between rows i and i' of X. The profile of column marginals $c(j)$, being the mean row profile, is the weighted average of the row points, where the row marginals are used as weights, and is located in the origin of the space. The χ^2-distance concept can be used in interpreting the configuration of points. It tells us that when two rows are close together, their profiles must be similar and moreover they should be related in a similar manner to the columns. On the other hand, whenever two rows are far apart, they are related in different ways to the columns. When a row point is near the center of the X space, its profile is similar to the profile of column marginals $c(j)$. Finally, when two row points are in opposite directions from the center, they deviate in opposite ways from the profile of column marginals [11].

We would like to associate the configuration X with the matrix F. Define $E=rc'/N$. Note that the elements of E have the form $E(i,j)=r(i)c(j)/N$. We consider the singular value decomposition (SVD) of the matrix

$$D_r^{-1/2}\big(F-E\big)D_c^{-1/2}=K\Lambda L', \tag{2.2}$$

with $K'K=L'L=I$, and Λ the diagonal matrix containing the singular values. The dimensionality of the solution equals $\min(I-1,J-1)$. Matrix K contains scores corresponding to the row categories. The scores are normalized to give

$$X=N^{1/2}D_r^{-1/2}K, \tag{2.3}$$

so that, $X'D_rX=NI$ and $u'D_rX=0$. Since, CA is symmetric, we can also look at the column categories, which after a suitable normalization are given by

$$Y=N^{1/2}D_c^{-1/2}L, \tag{2.4}$$

so that, $Y'D_cY=NI$ and $u'D_cY=0$. Hence, in each dimension the row and the column scores have a weighted variance of one and a weighted average of zero.

Given, the above solution for the row points, we can compute the column point configuration as follows

$$\tilde{Y}=N^{1/2}D_c^{-1/2}L\Lambda=Y\Lambda, \tag{2.5}$$

with the effect that $\tilde{Y}'D_c\tilde{Y}=N\Lambda^2$. Similarly, for the row points we have that

$$\tilde{X}=N^{1/2}D_r^{-1/2}K\Lambda=X\Lambda, \tag{2.6}$$

so that $\tilde{X}'D_r\tilde{X} = N\Lambda^2$. Moreover, some algebra shows that

$$D_r^{-1/2}\big(F-E\big)D_c^{-1/2} = K\Lambda L' \Leftrightarrow \qquad D_r^{-1/2}\big(F-E\big)D_c^{-1/2}L = K\Lambda \Leftrightarrow$$
$$D_r^{-1/2}\big(F-E\big)Y/\sqrt{N} = K\Lambda \Leftrightarrow D_r^{-1}\big(F-E\big)Y = \sqrt{N}D_r^{-1/2}K\Lambda = \tilde{X} \Leftrightarrow$$
$$D_r^{-1}FY = \tilde{X} \tag{2.7}$$

where, the second relation follows from the fact that $L'L = I$, the third from (2.4), and the last one from the fact that $D_r^{-1}EY = D_r^{-1}(D_r uu'D_c)Y = uu'D_cY = 0$. Similarly, we get that

$$\tilde{Y} = D_c^{-1}F'X. \tag{2.8}$$

Relations (2.7) and (2.8) are known as the *transition formulae* and can be used to interpret distances between row and column points. When a row profile is equal to the average row profile, the first relation shows that the row point will be at the weighted average of the columns, i.e. the origin of the X space, and similarly for the column points. When for some column j the row profile value $F(i,j)/r(i)$ is larger than the average $c(j)$, the column will attract the row point in its direction.

Regarding plotting the results, there are the following choices.

(I) Plot the pair $(\tilde{X}, Y)$, which shows that the row points are in the center of gravity of column points.
(II) Plot the pair $(X, \tilde{Y})$, which shows that the column points are in the center of gravity of row points.
(III) Plot the pair $(\tilde{X}, \tilde{Y})$.
(IV) Plot the pair $(X\Lambda^{1/2}, Y\Lambda^{1/2})$.

The last two options abandon the centroid principle present in the first two options. However, using $\tilde{X}$ as row scores (or $\tilde{Y}$ as column scores), distances between row points are equivalent to χ^2-distances (similarly for column points). For this reason the third option that treats rows and columns symmetrically is used most frequently in the French literature [11]. These options are illustrated using Fisher's eye and hair color example [7] (for a description of this data set see Appendix A). It is worth observing that options (III) and (IV) produce identical arrangements of the points, and they are rescaled versions of each other, as expected.

Using the fact that CA decomposes the matrix $D_r^{-1/2}(F-E)D_c^{-1/2} = K\Lambda L'$, and using relations (2.3) and (2.4) we have

$$D_r^{-1/2}\big(F-E\big)D_c^{-1/2} = K\Lambda L' \Leftrightarrow D_r^{-1}\big(F-E\big)D_c^{-1} = \frac{1}{N}\big(N^{1/2}D_r^{-1/2}K\Lambda L'D_c^{-1}N^{1/2}\big) \Leftrightarrow$$
$$D_r^{-1}\big(F-E\big)D_c^{-1} = \frac{1}{N}X\Lambda Y' \Leftrightarrow F = E + \frac{1}{N}D_rX\Lambda Y'D_c = \frac{1}{N}D_r\big(rc' + X\Lambda Y'\big)D_c, \tag{2.9}$$

Fisher's Eye and Hair Color Example

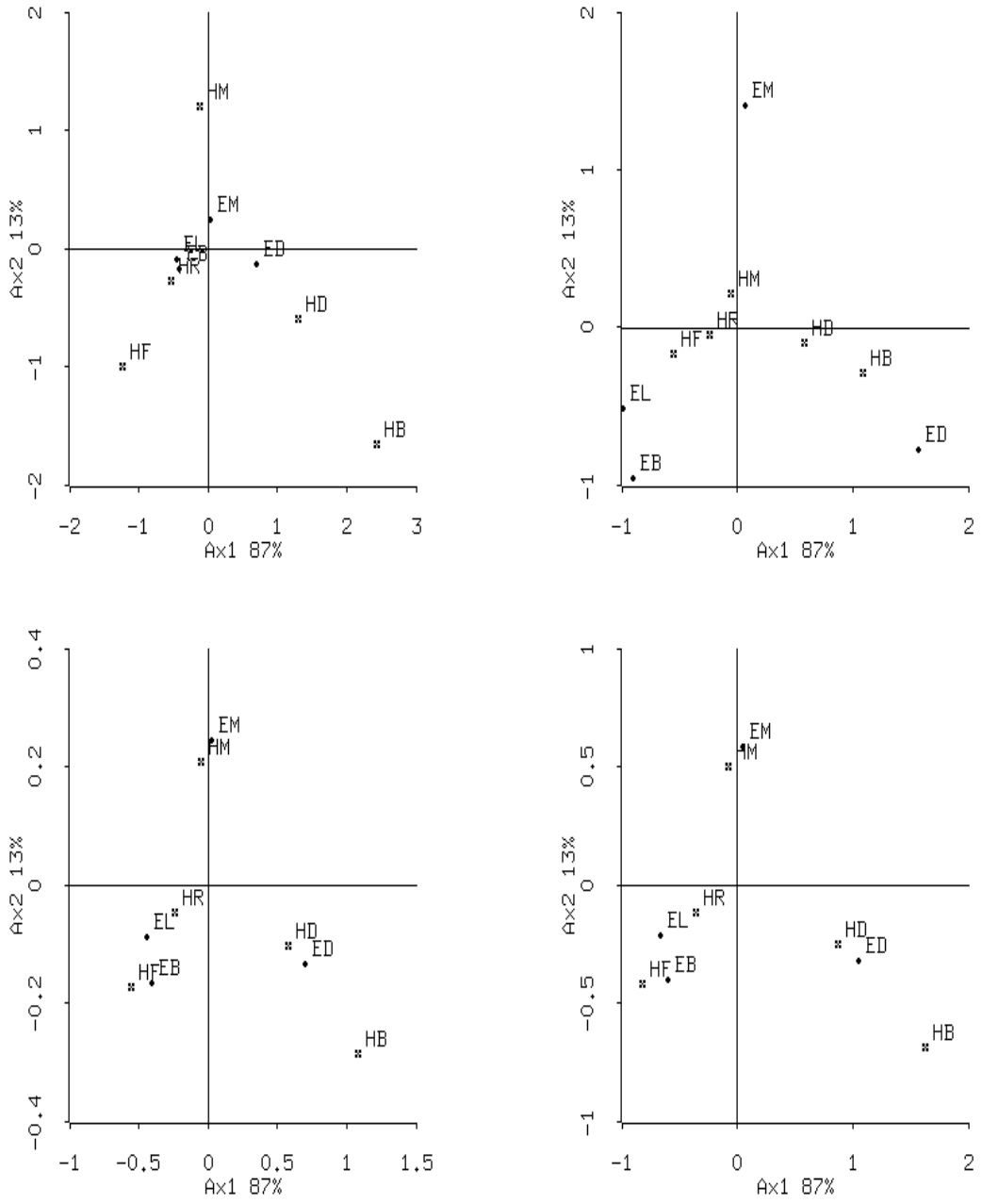

FIGURE 2.1. Upper Left: Normalization I, Upper Right: Normalization II, Bottom Left: Normalization III, Bottom Right: Normalization IV (EL, EB, EM, ED: light, blue, medium, dark eye color respectively; HF, HR, HM, HD, HB: fair, red, medium, dark, black hair color respectively)

which shows that CA decomposes the departure from independence in the matrix F. The usual test for independence of a contingency table is given by Pearson χ^2-statistic that is related to CA by

$$\text{tr}\Lambda^2 = \chi^2/N, \tag{2.10}$$

which in the French literature is known as the *total inertia* [8]. Equation (2.10) shows that CA also decomposes the χ^2-statistic for testing independence of a contingency table. In closing, CA performs a decomposition of the residuals of a contingency table in the absence of independence, and the resulting picture allows a close inspection of the interactions between its rows and columns.

Remark 2.1. *Passive rows and columns.* In some data analytic situations, one may want to exclude certain rows or columns from the initial stage of the analysis, while still being able to inspect the projections of such points onto the eventual CA maps. CA can easily accommodate such a situation and a more complete account is provided in section 4.

2.2. **Multiple Correspondence Analysis.** In the presence of more than two categorical variables we can proceed as follows. Suppose we have collected data on N objects (individuals, etc.) and J variables, with k_j categories per variable. Let G_j be a $N \times k_j$ matrix with entries $G(i,t) = 1,\ i = 1,\ldots,N,\ t = 1,\ldots,k_j$, if object i belongs to category t of variable j, and $G(i,t) = 0$ if it belongs to some other category. We denote by $G = [G_1|G_2|\ldots|G_J]$ the *super*-indicator matrix of all variables, and by $C = G'G$ the symmetric matrix known as the Burt table [8]. The Burt table contains all the category marginals on the main diagonal and all possible cross-tables of the J variables in the off-diagonal. MCA corresponds to performing simple CA on the Burt table C, so the solution is given by

$$J^{-1}D^{-1/2}\big(C - Duu'D/N\big)D^{-1/2} = L\Lambda^2 L', \tag{2.11}$$

where, $D = \text{diag}\,(C)$. In this case the category points are given by

$$Y = \sqrt{N}D^{-1/2}L\Lambda. \tag{2.12}$$

MCA can be thought of as the joint analysis of all the two-way tables composing the Burt table. Hence, it uses the information contained in both the diagonal and off-diagonal blocks of the Burt table. However, the diagonal blocks contain just the univariate marginals of each variable, and do not contribute any information regarding associations of the variables. Each of these blocks of perfect association has the highest inertia possible in a frequency matrix and corresponds geometrically to the profiles coinciding with the vertices, since the profiles of a diagonal matrix are unit vectors. The most apparent symptom of this problem is that the total inertia in an MCA is generally high while the percentages of inertia along principal axes are invariably low, thus suggesting a bad representation of the data. Possible alternatives to MCA of the Burt table are joint correspondence analysis [10], a technique that only takes into consideration the off-diagonal blocks of the Burt table, and *homogeneity analysis* which jointly analyzes objects and variables, a version of which is presented next.

2.3. **Analysis of Profile Frequencies.** The setup for the technique known in the literature as ANAPROF [8] is the following. Consider the superindicator matrix G defined in the previous

section. In case the number of objects N is much larger than the number of profiles (response patterns) that occur in the data matrix, it is convenient to express the superindicator matrix as $G = G_p S$, where S is a $q \times \sum_j k_j$ binary matrix with $S(h,l) = 1$ if category l belongs to the h^{th} profile, and 0 otherwise, and G_p a $N \times q$ profile indicator matrix, with entries $G_p(i,h) = 1$ if the i^{th} object has the h^{th} profile in S, and 0 otherwise. Define $\tilde{F} = G_p' G_p S$, which is a $q \times \sum_j k_j$ matrix, with $G_p' G_p$ being a diagonal matrix containing the marginal frequencies of the profiles. Matrix $\tilde{F}$ has row marginals $D_r = G_p' G_p$ and column marginals $D_c = D = \text{diag}\,(G'G)$. We can now apply simple CA to the matrix $\tilde{F}$, which is similar to homogeneity analysis [8]. However, the advantage of this technique is that CA is performed on a small matrix (q presumably is much smaller than N) and by using an explicit SVD decomposition we can look at the full solution, instead of the first p dimensions that homogeneity analysis by means of an *alternating least squares* algorithm permits. The solution is contained in the following SVD given by

$$D_r^{-1/2}\left(\tilde{F} - D_r u u' D_c\right) D_r^{-1/2} = K^\star \Lambda L', \tag{2.13}$$

where $K^\star = (G_p' G_p)^{-1/2} G_p' K$, with K given by $J^{-1/2} G D^{-1/2} = K \Lambda L'$. The solution for the variables is given by

$$Y = \sqrt{N} D^{-1/2} L \Lambda, \tag{2.14}$$

while the solution for the objects by

$$X = G_p (G_p' G_p)^{-1/2} K^\star. \tag{2.15}$$

However, since we are only interested in plotting unique profiles, we can set $X = I_p (G_p' G_p)^{-1/2} K^\star$. This technique computes object coordinates, thus allowing the user to examine interactions between specific profiles that might be of special interest to him.

Remark 2.2. *The Common SVD.* It is seen that at the heart of each of these three techniques lies the solution of a SVD problem, which implies that for their implementation a single computational routine is needed. The rest of the program contains input-output routines, such as reading the data and creating the appropriate data matrices, plotting and formatting the results.

3. **Stability Issues**

All three techniques discussed in this paper (CA, MCA and ANAPROF) are primarily data analytic techniques. However, when examining the graphical displays produced by the program the user is usually confronted with the following question: "Are the patterns in the plots real, or merely chance effects?" This question leads directly to the issue of "stability" of the results and their "significance" in some statistical sense. The question of stability seems to be particularly relevant when the data arise from some well defined random sampling scheme, or in other words when it can be safely assessed that the data are a representative "image" of an underlying population. However, this ideal situation, on which most conventional statistical inference is based, occurs rather infrequently and many data sets are collected in a deliberate nonrandom fashion.

The previous observation leads to the notions of *external* and *internal* stability. 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 the techniques discussed here are externally stable in case any other sample from the same population produces roughly the same results (e.g. singular values, row and column profiles, etc.). Internal stability deals with the specific data set at hand. An internally stable solution implies that the derived results give a good summary of that specific data set. In this case, 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 observations or categories that have a large influence on the results. Internal stability can be thought of as a form of robustness. An extensive discussion of these two notions and their implications in data analysis can be found in Michailidis and de Leeuw [17] and in the numerous references cited there.

In order to assess the stability of the techniques we resort to the nonparametric approach of bootstrapping, that is suited for both external and internal stability. The bootstrap relies on a "new" fictitious perturbed sample created by resampling with replacement from the data set (sample) at hand. So, we attempt to assess stability by examining what would have happened if a truly "new" sample was drawn from the underlying population. In the case of internal stability, bootstrapping can be thought of as a form of data based perturbation analysis. In the remainder of this section we present the appropriate method of bootstrapping for each of the three techniques. We also present some analytical results for the singular values from simple CA (based on perturbation results for eigenvalues [15]).

3.1. **CA.** In this case, the information contained in the original data set has been collapsed to the observed contingency table. Thus, a moment of reflection shows that bootstrapping in this setting is equivalent to simulating data from a multinomial with sample size N and cell probabilities given by the observed proportions (the elements of the matrix F/N). The algorithm employed in the program can be found in [6].

When distributing N throughout the contingency table, the possibility arises that the sum of an entire row or column is 0. This is likely to occur when the original contingency table has rows or columns with fairly low marginals. To avoid problems arising when computing D_r or D_c, generalized inverses are used. The result is that the mass assigned to a particular row or column with zero marginal is zero. Counts of the number of times each row or column has zero marginals during the bootstrap iterations is provided as output and rows or columns that frequently are entirely zero are likely to have rather unstable solutions.

3.2. **MCA and ANAPROF.** We briefly outline the method in a general context (for a comprehensive account see also [28]). Suppose we have J categorical variables. Each variable takes values in a set S_j (the range of the variable [8]) of cardinality ℓ_j (number of categories of variable j). Define $S = S_1 \times \ldots S_J$ to be the *profile space*, that has cardinality $\ell = \prod_{j=1}^{J} \ell_j$. That is the space $S = \{(s_1, \ldots, s_j),\ s_j \in S_j,\ j \in \mathbf{J}\}$ 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 h^{th} profile contains category t, and 0 otherwise; that is S maps the space of profiles 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 t^{th} object (individual etc) has the h^{th} profile in S, and $G_S(t,h) = 0$ otherwise. The superindicator matrix $G = [G_1|\ldots|G_J]$ can now be written as $G = G_S S$, which immediately shows that there is a one-to-one correspondence between G and S.

Consider a probability distribution P on S. Since the space S is finite, 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 of the techniques 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 the present context, 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 mild regularity conditions on the $\phi(.)$ 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 $P(\phi(p) \leq z|p)$ [24], where P_* denotes the conditional probability given p_m. The previous discussion indicates that the appropriate way to bootstrap in MCA and ANAPROF is to sample objects with replacement, or in other words, sample rows of the data matrix.

However, in many occasions this approach may lead to the following problem. If the frequency of a profile is low in the original data set, then there is the possibility of not appearing in the bootstrap indicator matrix G^m. In this case some categories will be absent from the m^{th} bootstrap replication. In MCA, the problem of categories with zero marginals is treated identically as it is in simple CA. Generalized inverses are used in the computation of D, the diagonal matrix of column marginals of the super indicator matrix G. The solution is computed for all categories with nonzero marginals and once again, counts are provided for the number of times a particular category has zero marginals during the bootstrap iterations. The problem of empty profiles is alleviated in ANAPROF by first filling out the diagonal of the m^{th} bootstrap resampled matrix $(G^m)'(G^m)$ with ones. This ensures that all profiles show up at least once in each bootstrap iteration. The remaining $N - q$ bootstrapped observations are distributed to the diagonal of $(G^m)'(G^m)$ according to a multinomial distribution with parameters $N - q$ and diag $(G'G)/N$ (the sample matrix). The underlying assumption of $q << N$ makes this a reasonable approach.

3.3. **Analytical Results for Singular Values.** In a series of papers O'Neill [20, 21, 22, 23] has derived the asymptotic distribution of the singular values (canonical correlations) of contingency tables. We give a brief description of his main result relevant to the present work.

The starting point is the reconstitution formula (2.9). Dividing both sides by N we get

$$F/N = E/N + (D_r/N)X\Lambda Y'(D_c/N), \tag{3.1}$$

which can be written as

$$p_{ij} \approx p_{i\cdot}p_{\cdot j}\big(1+\sum_{h=1}^{H^\star}\Lambda(h,h)X(i,h)Y(j,h)\big),\ i=1,\ldots,I,\ j=1,\ldots,J, \tag{3.2}$$

where $p_{ij}=F(i,j)/N$ denotes the sample proportion of cell (i,j), $p_{i\cdot}$ $(p_{\cdot j})$ the marginal proportion of row i (column j), and $H^\star \geq 1$ the number of nonzero singular values $\Lambda(h,h)$. O'Neill shows that the variables $\sqrt{N}(\hat{\Lambda}(h,h)-\Lambda(h,h)),\ h=1,\ldots,H^\star$, where $\hat{\Lambda}$ denotes the sample values and Λ the population ones, are asymptotically normally distributed with zero means and second order moments depending on the canonical correlations and on third and fourth order moments of the elements of X and Y. Specifically,

$$\begin{aligned}\sigma_\ell^2 = (1+\frac{1}{2}\Lambda^2(\ell,\ell))\big[1+\sum_{h=1}^{H^\star}\Lambda(h,h)E(X(.,h)X^2(.,\ell))E(Y(.,h)Y^2(.,\ell))\big] \\ -\frac{3}{4}\Lambda^2(\ell,\ell)\big[E(X^4(.,\ell))+E(Y^4(.,\ell))\big],\ \ell=1,\ldots,H^\star,\end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\sigma_{\ell\nu} =& \frac{1}{2}\Lambda(\ell,\ell)\Lambda(\nu,\nu)-\frac{3}{4}\Lambda(\ell,\ell)\Lambda(\nu,\nu)\big[E(X^2(.,\ell)X^2(.,\nu))+E(Y^2(.,\ell)Y^2(.,\nu))\big] \\ &+\sum_{h=1}^{H^\star}\Lambda(h,h)\big[E(X(.,h)X(.,\ell)X(.,\nu))E(Y(.,h)Y(.,\ell)Y(.,\nu)) \\ &+\frac{1}{4}\Lambda(\ell,\ell)\Lambda(\nu,\nu)\big\{E(X(.,h)X^2(.,\ell))E(Y(.,h)Y^2(.,\nu)) \\ &+E(X(.,h)X^2(.,\nu))E(Y(.,h)Y^2(.,\ell))\big\}\big],\ \ \ell,\nu=1,\ldots,H^\star.\end{aligned} \tag{3.4}$$

An example of the notation and how to calculate the expectations in (3.3) and (3.4) is

$$E\big(X.,h)X^2(.,\ell)\big) = \sum_{i=1}^{I}\sum_{j=1}^{J}X(i,h)X^2(i,\ell)p_{ij} \tag{3.5}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}X(i,h)X^2(i,\ell)p_{i\cdot}p_{\cdot j}\big(1+\sum_{h=1}^{H^\star}\Lambda(h,h)X(i,h)Y(j,h)\big). \tag{3.6}$$

and similarly for the other moments.

4. **Implementation**

The program is implemented in the Lisp-Stat [27] language. Its main features are discussed below.

4.1. **Computation.** As mentioned in Remark 2.2 all three techniques are based on an explicit singular value decomposition. Such a routine is readily available in the *Lisp-Stat* [27] language. However, there are several simplifications that can be made before such a routine should be used, and are outlined next.

4.1.1. **Correspondence Analysis.** In CA, the matrix

$$D_r^{-\frac{1}{2}}(F - rc')D_c^{-\frac{1}{2}} = K\Lambda L' \tag{4.1}$$

is formed first. However, $F,\ r,\ c,\ D_r,\ D_c$ are computed only from the submatrix of the contingency table that contains the subset of active rows and columns (the rows and columns that the user has requested; see also Remark 2.1). At this point it must be checked whether any zero marginals for rows or columns is produced from this active subset of the contingency table. Although for small to medium sizes r and c (e.g. no more than 10 categories) the formation and multiplication of the diagonal matrices D_r and D_c does not constitute a large computational burden, the process occurs many more times than just in the formation of (4.1). Diagonal matrix multiplication is therefore carried out by vectorizing the multiplication operation, or breaking apart the matrix to be multiplied and performing separate vector multiplications on each row (or column). For example, to form $D_r^{-1/2}(F - rc')$ one would break the matrix $F - rc'$ into its rows and multiply these rows by the diagonal elements of $D_r^{-1/2}$. Passive row and column coordinates are computed from the row and column solutions obtained from using the active rows and columns points. These points are found by projecting the passive row and column profiles onto the respective row and column solution space.

Let Sq$[\cdot]$ denote the elementwise squaring of the elements of the matrix argument and diag $(\cdot)$ a diagonal matrix with the vector argument on the diagonal with zeros elsewhere. Let K_i be the i^{th} row of K, and L_i the i^{th} row of L. The following statistics are also printed.

1. *Inertias.* The inertia due to the i^{th} principal axis is $I_i = \Lambda^2(i,i)$, the i^{th} squared singular value in the decomposition (4.1).
2. *Partial Inertias.* The partial inertias due to the i^{th} row (column) point for each of the p dimensions of the solution is given by the vector I_{r_i} = Sq$[K_i]$ (or Sq$[L_i]$ for the columns). For a given principal axis and point, the partial inertia contribution is defined to be the squared length of the projection of the point onto the principal axis. These are defined only for the active points.

3. *Squared Cosines.* For the active points, the squared cosine of the i^{th} row (column) point for each of the p dimensions of the solution is given by the vector diag $(Sq(D_r^{-1/2}K_i\Lambda_p))^{-1}Sq(D_r^{-1/2}K_i\Lambda_p)$. For a given principal axis and point, the squared cosine is the proportion of the distance from the point to the centroid of the cloud taken up by the length of the squared projection of the point onto the axis. These are computed for the active as well the passive row and column points.

4.1.2. **ANAPROF**. Aside from the computation of the appropriate SVD, the main computational burden of ANAPROF is in the reading in of the data and the simultaneous formation of the matrices S and $G_p'G_p$. This is done when the data file is first specified. If some of the variables in the data set are to be treated as passive, the data set must be re-read from the data file. This was found to be more efficient than routines to reduce (expand) the profile matrix and profile counts based on the columns to be treated as passive (not passive). Once the data has been read, the following singular value decomposition can be performed

$$(\text{diag } (JG_p'G_p))^{-1/2}FD^{-1/2} = K^*\Lambda L' \tag{4.2}$$

where D is the diagonal matrix of the column marginals of $F = G_p'G_pS$. Again, diagonal matrices are stored as lists and products of full matrices with diagonal matrices are performed by breaking the full matrix into its rows (or columns) and multiplying these rows (or columns) by the appropriate diagonal elements of the diagonal matrix. For plotting, we are only interested in unique profiles; hence, we set $X = I_p(G_p'G_p)^{-1/2}K^*$.

4.1.3. **MCA.** Computational considerations for correspondence analysis on the Burt matrix are similar to those for simple CA. Diagonal matrix multiplication is treated as in simple CA and ANAPROF. As each variable in the Burt matrix comprises several columns (or rows), the specification of passive variables actually removes a block of the rows and columns from the Burt matrix.

4.2. **Object Structure.** Figure (4.1) shows the inheritance tree for the program. The structure of many of these prototypes are similar to those used in ([4]). As the types of operations (numerical and data manipulation) performed, the types of plots available, the similarity of the desired type of interactiveness of the plots, and the types of output are somewhat similar for all three analyses, it was decided that one program that encompasses all three would be more efficient than three separate non-interacting programs. To implement this idea, a single parent prototype *anacor-proto*, was created which holds all information involving the data and the computed solution for all three techniques. This parent prototype also controls the computational aspects of the analysis. The three major groups of prototypes used in the accompanying program are the *anacor-proto* parent prototype, the dialog prototypes, and the plotting prototypes. Dialog prototypes are efficient due to their ability to move between types of analyses and to reduce the amount of code produced by consolidating similarities in output functions, types of plots available, and similarities in dialog

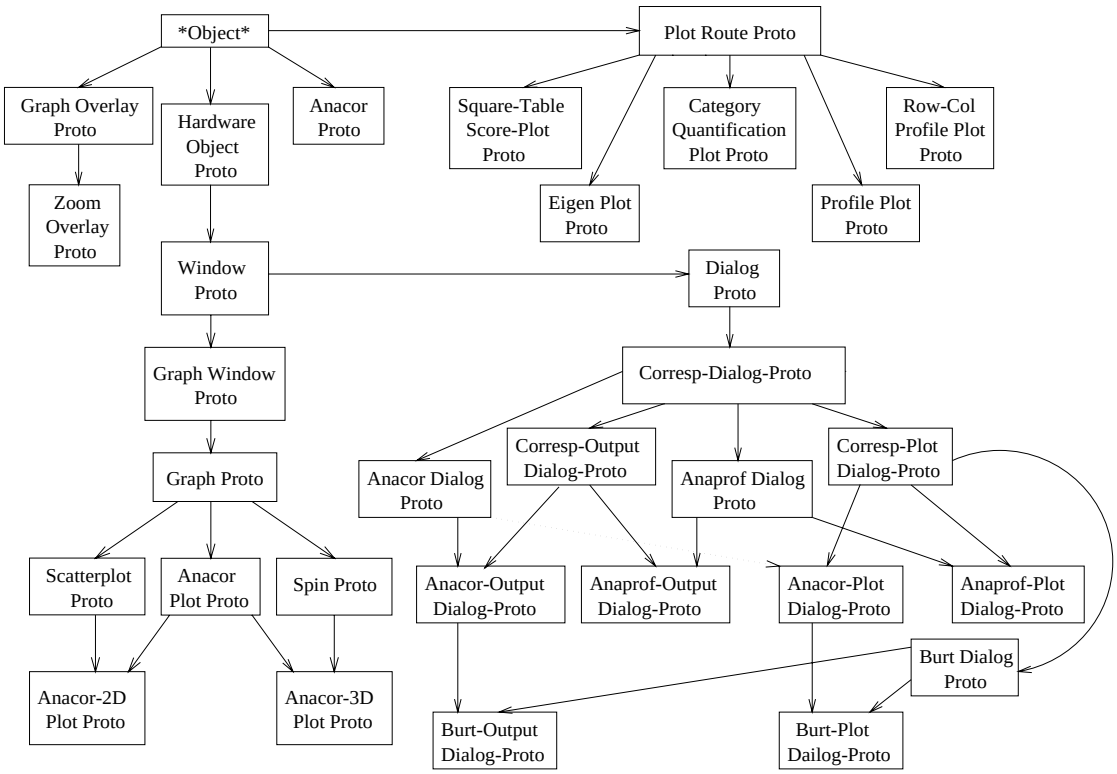

FIGURE 4.1. Inheritance Tree

functions such as reading a new file or requesting a plot. The plotting prototypes take advantage of Lisp-Stat's interactive environment, especially the availability of mouse modes for manipulating the contents of a plot.

Plotting is controlled and routed by the *plot-route-proto* prototype. Each specific type of plot is managed by its own prototype which initializes, fills, and stores the plot. Creating a separate type of plot only requires the creation of a new managing prototype. The methods required are *:isnew*, *:make-points*, *:make-point-labels*, *:init-selected-points*, and *:make-lines*. Each plot is created as an instance of either the *anacor-2d-plot-proto* or the *anacor-3d-plot-proto* depending on the requested dimensionality of the solution.

The true interactive nature of the program can be found in the dialog and zooming prototypes. Due to the fact that the techniques require different types of data input, additional options, and types of output, different dialogs were implemented. It is possible, however, to switch between different types of analyses by pressing the *New Data File* button in each main data dialog.

Zooming is performed through the *zoom-proto* prototype. This prototype is a descendant of the *graph-overlay-proto* prototype and controls the mouse interaction with the plot. Three mouse

modes are available within the *zoom-proto* prototype and these mouse modes are accessed by mouse clicks in the appropriate locations in the plots margin. The mouse mode *'newselecting* is used to override the standard *'selecting* mouse mode in order to capture mouse clicks that fall outside of the plot margin. This mode is accessed by clicking inside the box marked *Selecting* in the plots margin. Selection of points may be performed at any state of zooming. To keep points that are currently selected, whether they are showing or not, the shift key should be held down while drawing a box around the desired points.

The ability to zoom in on points has been found to be very useful, not only for examining the solution of an ANAPROF or CA analysis, but for any plot that contains a cloud of points and where it of interest to distinguish these points from each other. A version of the type of zooming implemented here, but with fewer features, is also used in the companion paper ([4]). Due to its usefulness, it is available as a separate module which can be used on any Lisp-Stat plot by a simple *:add-overlay* call. The mouse mode *'zoom* may be selected by clicking in the square next to the symbol "+" in the plots margin. Zooming may be performed any number of times, and the process of "stepping out" of the zoom may be carried out by clicking in the square next to the symbol "-" in the plots margin. "Stepping Out" of a zoom refers to returning to the previous zoomed state. For example, one may select a set of points to be zoomed in on and then select a subset of these zoomed points to be zoomed in on again. If the box next to the "-" symbol is pressed the plot is returned from the "zoomed-zoomed" state to the "zoomed" state. Zooming out completely is accomplished by clicking in the square next to *Out* in the plots margin.

A common problem with Lisp-Stat plots is that points that are plotted directly on top of each other are not distinguishable; their labels overlap, thus making them unreadable. This can occur in an ANACOR analysis when two rows or columns have exactly the same profile. This problem is solved by the mouse mode *sep*. When the square next to *Expand* is selected, a box may be drawn around overlapping points. When the mouse is released, these points are centered in the plot and are expanded (contracted) radially outward from each other by clicking on the up (down) arrow symbols in the plots margin.

4.3. **Using the Package.** The flow and use of the program is very similar to [4]. Data needs to be stored in a white space (space or tab) delimited file. For simple, the data needs to be in the form of a contingency table, for ANAPROF and MCA the data matrix need be stored as itself. Missing values are not allowed in any of the analyses but are planned for future upgrades of the program. As an example, consider the Fisher eye/hair color data set. The initial process of loading the data set can be seen in Figure 4.2.

Once the data is loaded, the dialog in Figure 4.3 appears. At this point, filenames for row status and column status files describing the active or passive state of rows or columns may be provided. These files need to be white space delimited files containing 0's and 1's. The length of the row status file should be in accord with the number of rows in the data set, and analogously for the

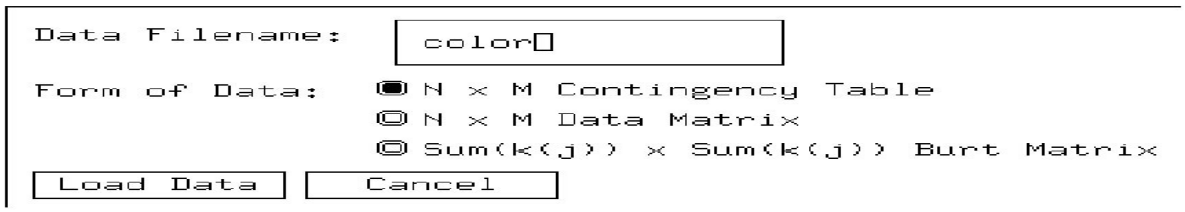

FIGURE 4.2. Data Loading Dialog

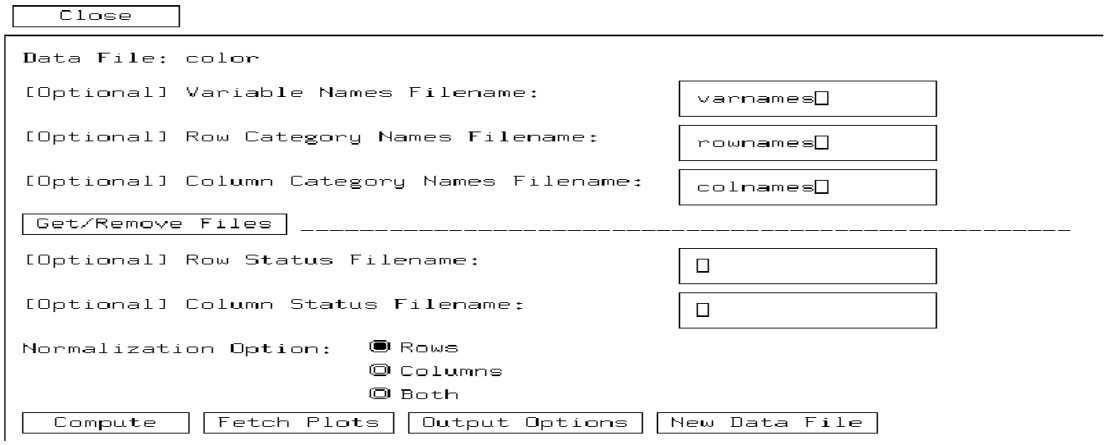

FIGURE 4.3. ANACOR Dialog Screen

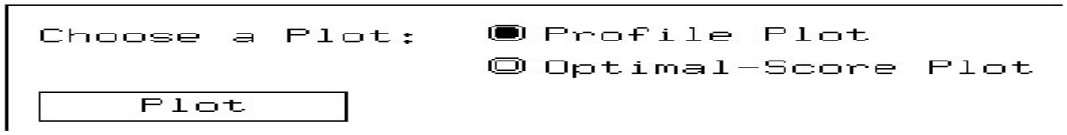

FIGURE 4.4. CA Plots

columns status file. A 1 corresponds to a row/column being treated as active and a 0 corresponds to a row/column being treated as passive. MCA of the Burt Matrix only requires a column status file with the number of entries equal to the number of columns in the data set, since the symmetric Burt matrix is analyzed. ANAPROF also only requires a column status file. These status files are optional and if not provided, the program will treat all rows and columns as active.

At this point, either just the solution may be computed or bootstrapping may be performed. If bootstrapping is chosen, the number of bootstrap iterations is requested in a dialog. Once the solution is computed, the user may plot various aspects of the solution or request printed output of the solution. For simple CA, the plots available can be seen in Figure 4.4 and the output options in Figure 4.5.

An example of the ANAPROF dialog can be seen in Figure 4.6. Again variable labels may be provided as well as a variable status file. Also displayed is the ratio q/N and the value of N, where q is the number of unique profiles and N is the number of rows in the data file. Ratios closer to zero indicate fewer unique profiles. At this stage the solution can be computed or bootstrapping may be performed. Available plots and output options can be seen in Figures 4.7, and 4.8, respectively. MCA dialogs are identical to the ANAPROF dialogs, while plots and output options are identical to the simple CA ones.

4.4. **Normalization of Bootstrap Samples.** Using (2.7), it can be easily seen that under normalizations (I) and (II) the solution of CA is rotational invariant. For example, if R is a rotation matrix satisfying $R'R = RR' = I$, and we set $Y^\sharp = YR$, we get $\tilde{X}^\sharp = D_r^{-1}FY^\sharp = D_r^{-1}FYR = \tilde{X}R$. It is a similar case for the ANAPROF solution. Therefore, bootstrap replications of normalizations (I) and (II) of CA, and also of ANAPROF suffer from the same problem, thus making it impossible to compare the bootstrapped solutions to the original ones. In order to make them comparable, we need to rotate the solution of each bootstrap sample accordingly.

For CA under normalization (I), suppose X is the row solution for the original sample. Let $X(m)$ denote the solution for the row coordinates for the m^{th} bootstrap sample. The problem of rotation in the presence of orthogonal constraints can be stated as

$$Min_R \mathrm{tr}[(X - X(m)R)' D_r(m)(X - X(m)R)] \tag{4.3}$$

over $R'X'(m)D_r(m)X(m)R = I$. Since from the definition of the normalization of the row solution we have $X'(m)D_r(m)X(m) = I$, the constraint reduces to $R'R = I$. This problem is known as an orthogonal Procrustes rotation problem and the solution is given by $R = UV'$ where $X'D_r(m)X(m) = U\Lambda V'$. The rotated solution for normalization (II) in CA is analogous. For ANAPROF, we solve equation (4.3) for R, using the identity matrix in the inner product instead of D_r. The "other portion" of the solution, meaning the column scores in normalizations (I) and the row scores in normalization (II) of ANACOR, and the category quantifications in ANAPROF are rotated using the same orthogonal rotation matrix R.

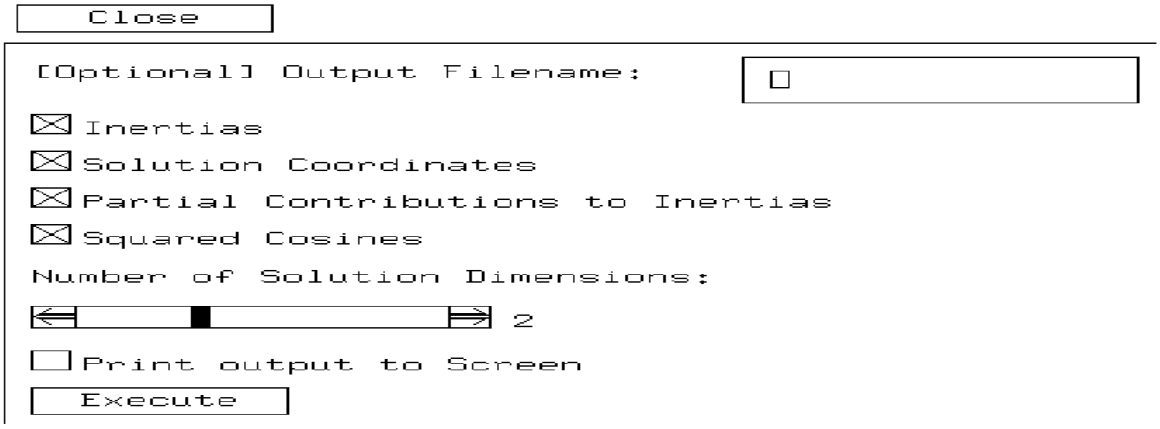

FIGURE 4.5. CA Output

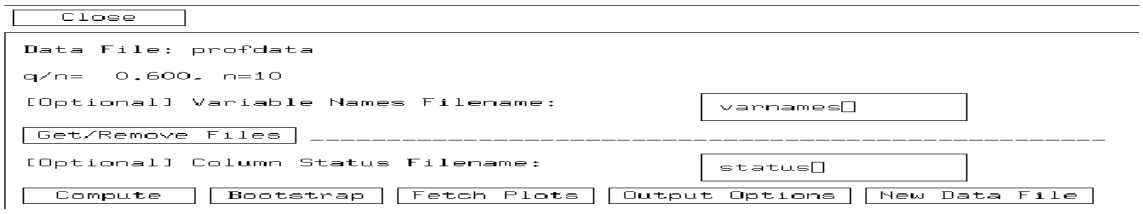

FIGURE 4.6. ANAPROF Dialog

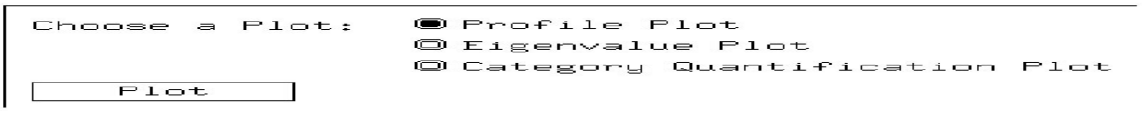

FIGURE 4.7. ANAPROF Plots

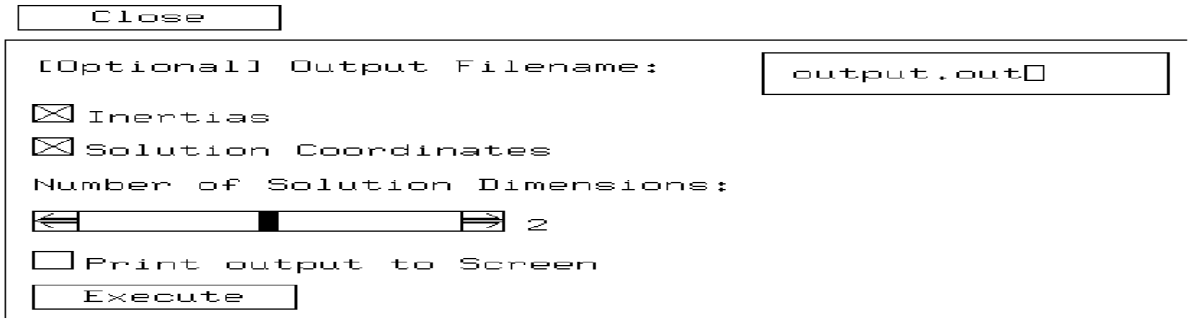

FIGURE 4.8. ANAPROF Output

It can be seen why normalization (III) and (IV) in CA and MCA on the Burt table do not suffer from rotational invariance, by noticing that no such orthogonal R exists that simultaneously satisfies the required normalization on the solution.

5. **Comparisons to Other Programs**

Most commercial software packages contain a procedure that performs CA and MCA -PROC CORRESP in SAS, program CA in BMDP, program ANACOR in SPSS [25, 3, 26]. Our program is very close to the commercial ones in terms of the output produced and the options offered (active and passive categories, partial inertias, squared cosines). The main advantage of the commercial programs is that they come with all of the data manipulation functions that are part of a general statistical package. The main advantage of this program is that it is menu driven, offers high quality dynamic graphic capabilities (rotation of the plots, zoom-in-zoom-out options, selection of points), and performs stability analysis. In summary, it utilizes all the recent advances in computer technology and is written taking into consideration the modern practice of exploratory data analysis. Finally, it is an open platform, so that users can add modules suitable to their particular needs.

6. **Applications**

6.1. **CA Application.** The data in this example comes from the NELS:88 data set (for other applications that used the NELS:88 data set see [4, 18, 16]. A brief description of the variables is given in Appendix B.

Rows and columns are treated symmetrically through the use of normalization III, so that distances between rows and distances between columns are approximately χ^2 distributed. Figure 6.1 shows the first two dimensions of the solution. The rows (F1S48A variable; how far the father wants the student to go in school) and columns (F1S53B variable; type of occupation the student expects to have at age 30) seem to exhibit the Guttman effect [9], falling in a horseshoe like pattern.

The first axis accounts for approximately 81% of the total inertia. The remaining eigenvalues die off slowly to zero, as can be seen in Figure 6.3. Projecting the rows onto the first axis one can see that there is an ordering by education., from $< HS$ to HS to $2-YR$, etc. Note that *NA* - Not Applicable, *DC* - Don't Care, and *DK* - Don't Know, fall into the middle range of the projections onto the first axis. The far right of the plot is rather cluttered and makes it difficult to distinguish point labels. Figure 6.2 shows the zoomed in portion of the cluster of points on the far right in Figure 6.1. As suspected, desired education is ordered in this cluster as well, from $4-YR$ - some Four Year college to *CGRAD* - College Graduate to *PGRAD* - Post Graduate School.

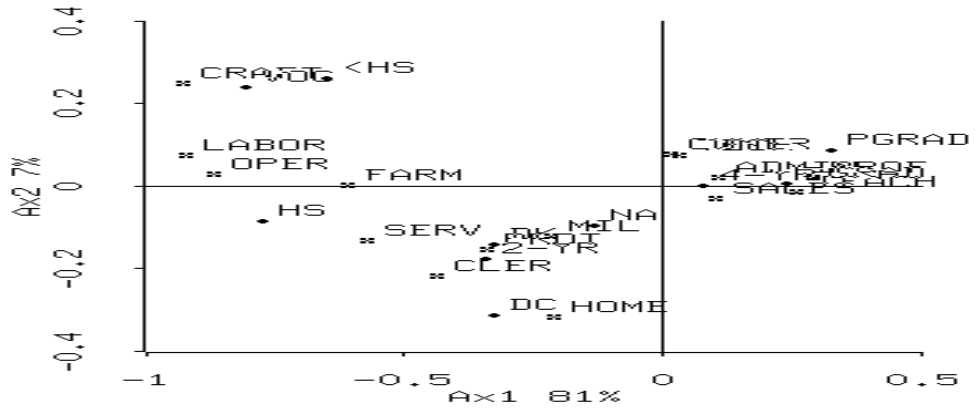

FIGURE 6.1. First 2 Dimensions

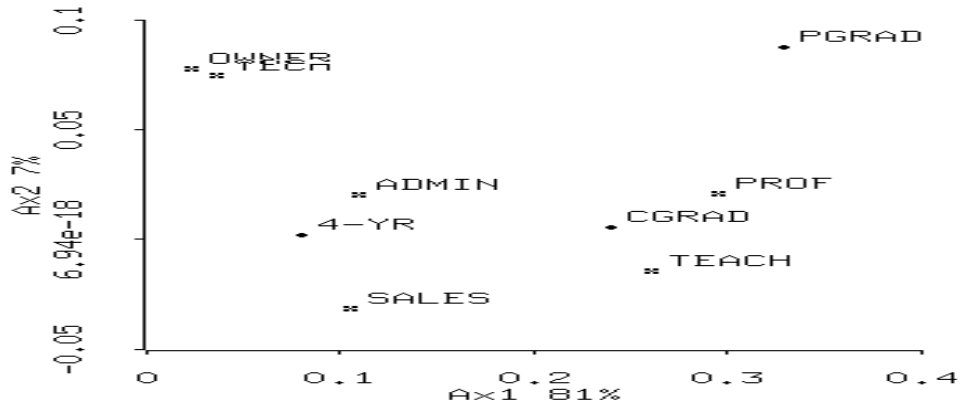

FIGURE 6.2. Zoom on first 2 Dimensions

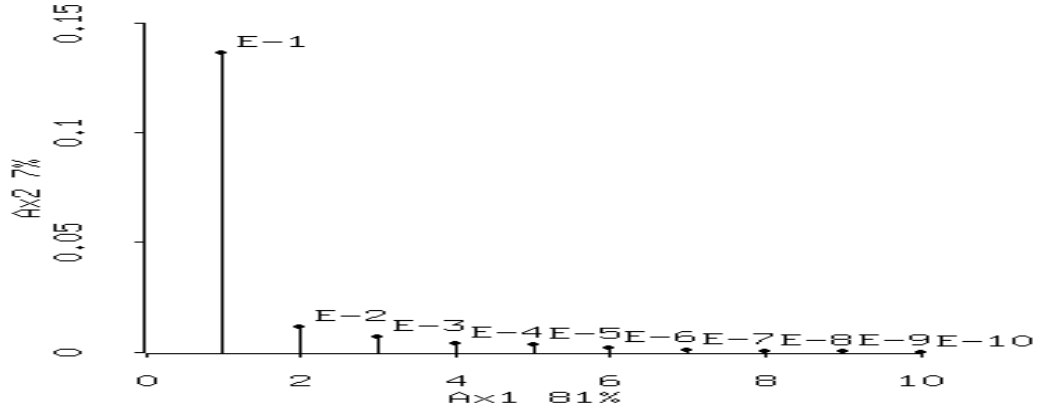

FIGURE 6.3. Eigenvalues

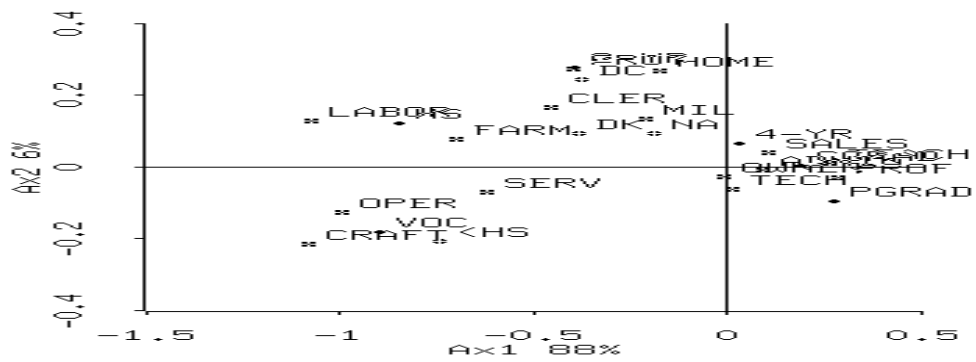

FIGURE 6.4. First 2 Dimensions With Passive Points

The projections of the columns on the first axis follow a similar pattern, in that jobs that require a lower level of education fall to the left while ones that require more education fall to the right. It is also evident that jobs that require special types of schooling, that is Professional $(PROF)$, Teaching $(TEACH)$, Craft Jobs $(CRAFT)$, etc fall close to the relevant education type.

In the interest of clearing up the picture, we treat *DC*, *DK*, and *NA* as passive, as well as $< HS$ due to its low row marginal. Figure 6.4 shows the first two dimensions of this solution. Again the first dimension comes out particularly strong, taking up 88% of the total inertia and 7% more than the previous solution. The second dimension only takes up around 6% of the total inertia. The horseshoe like pattern remains and the general pattern of clustering doesn't change. However, there are some distinct changes that can be seen. *OPER* - Machinery Operator and *SERV* - Service Type Jobs have moved closer to *VOC* - Specialty Vocational Schooling. Similarly, *LABOR* - Labor Type Occupations and *FARM* - Farming Jobs have moved towards *HS*. Both of these changes seem more reasonable than their previous positions and are due mainly to a larger relative proportion of these jobs falling into $< HS$ than in other profiles. The points $2-YR$ - 2 Year Schooling and *PROT* - Public Protection Jobs have moved to the top of the horseshoe while *DC*, *DK*, and *NA* remain relatively in the same positions. Again, the first axis seems to be ordering points by level of education or by jobs that require different levels of education.

The second axis, although rather weak, also seems to suggest an ordering. Notice that jobs/education levels that require/provide a higher level of specific/technical training are towards the bottom of the second axis. For education level, *VOC*, *PGRAD*, *CGRAD* are lower than *HS*, $2-YR$, and $4-YR$ in their second axis projection. For expected job at age 30, $CRAFT$, *OPER*, *SERV*, $TECH$, *PROF*, $TEACH$, and *OWNER* are lower in their second axis projection than *PROT*, *HOME*, *MIL*, *LABOR*, *CLER*, *SALES*, and *FARM*. This interpretation is somewhat debatable in that *CLER* and *MIL*, among others might be argued to be jobs that require a great deal of specialty training but again the second axis is somewhat weak in its contribution to the total inertia.

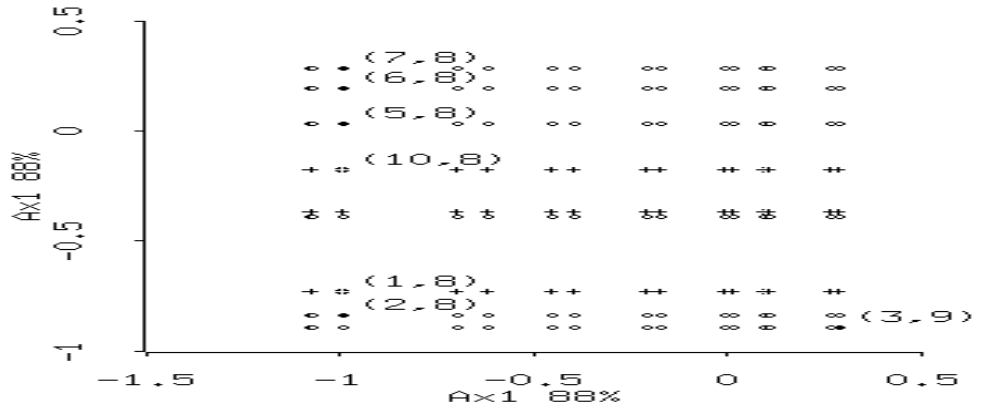

FIGURE 6.5. Optimal Quantifications, 1*st* Dimension

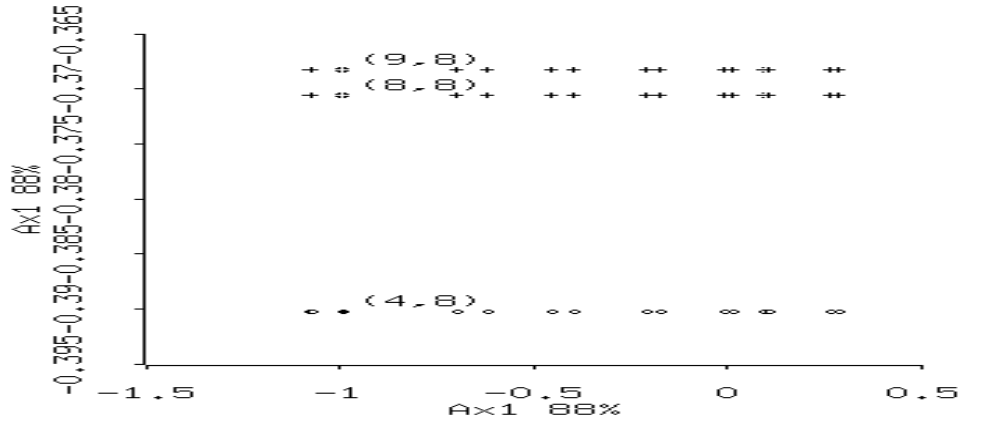

FIGURE 6.6. Zoom of Optimal Quantifications

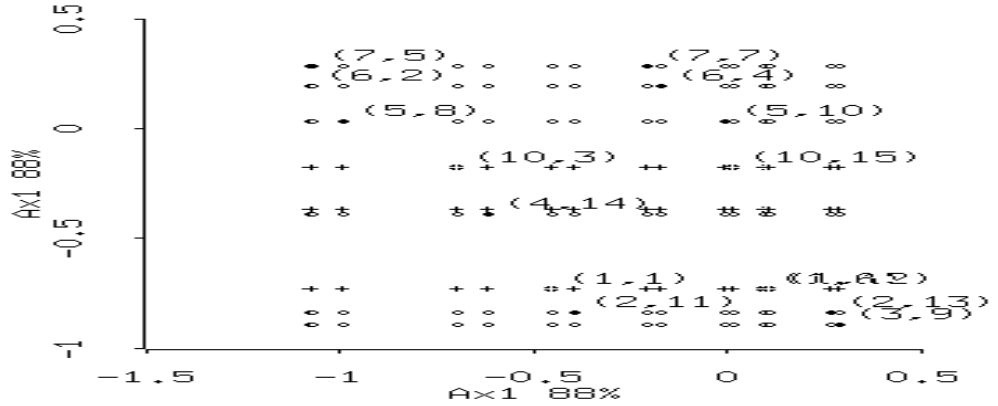

FIGURE 6.7. Optimal Quantifications, Columns Selected

To get an idea of the distances between rows and column for the first dimension of the solution, one can look at the coordinates assigned to the rows and columns (see Figure 6.5). Each point has an index corresponding to a $(row, column)$ pair. The numerical ordering can be seen in Table B1. A clustering of the rows and columns can be seen, with more distinct clusters formed in the columns. Some rows are quantified so close that they cannot be recognized by selecting them. A zoom of the middle rows can be seen in Figure 6.6. The columns are selected in Figure 6.7 and the columns are numbered according to the ordering in Table B1.

We turn our attention to the question of stability of our solution. The following Table shows a subset of the output for the first 2 dimensions of the original and the bootstrapped solution, for 20 replications. Judging from the bootstrap inertia means, the eigenvalues seem to be fairly stable.

```
Decomposition of total inertia along principal axes

AXES   INERTIA (eigenvalues)  %of INERTIA    Cum %
  1       .13658                 80.967       80.967
  2        1.2053E-2             7.1452       88.112

Total  .1686857456
```

```
Decomposition of total inertia along
principal axes for 20 Bootstrap Samples

AXES  INERTIA BS MEANS   INERTIA BS SDs    %of INERTIA  Cum %
  1    .1374782007       6.3043487563E-3    75.648      75.6
  2    1.5977882653E-2   2.4649213898E-3    8.7919      84.4

Total  .18173
```

A nice graphical display of the bootstrap inertia points, along with the original inertia points can be seen in Figure 6.8.

The bootstrapped solution points can be simultaneously plotted along with the original solution. This allows the inspection of the degree to which individual solution points are stable. This becomes impossible for a moderate number of categories and a moderate number of bootstrap replications as the points in the plots become totally indistinguishable (see Figure 6.8).

A solution to this problem can be seen in the left panel of Figure 6.9. Using the provided dialog, individual variables may be selected. When a given category is selected, all of the bootstrap

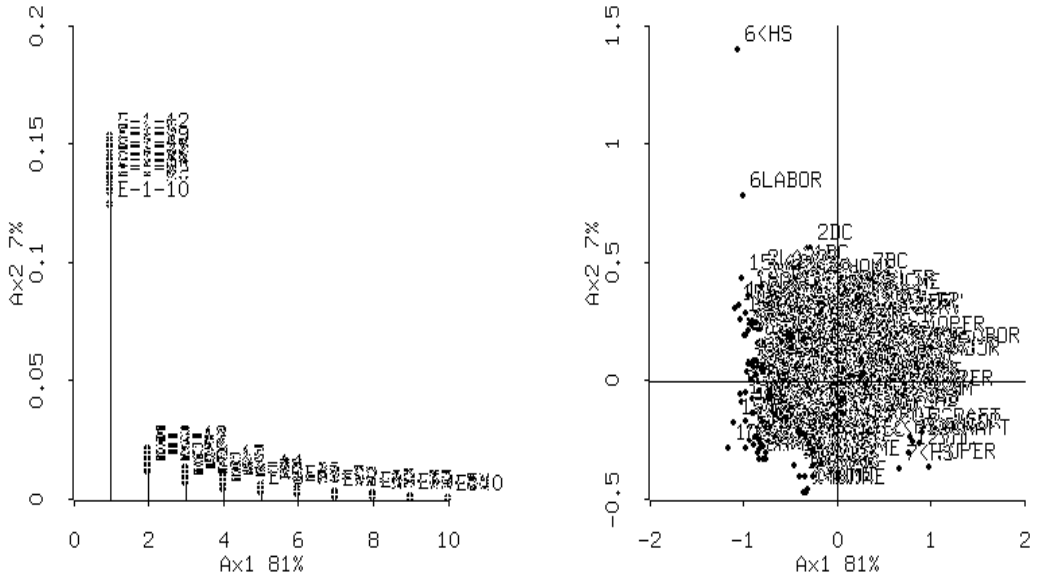

FIGURE 6.8. Bootstrapped Inertias and Bootstrapped Coordinates

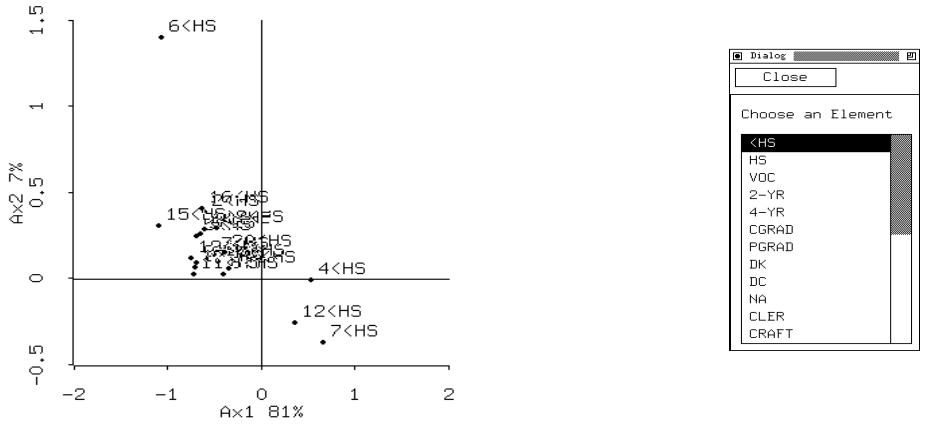

FIGURE 6.9. Bootstrap Solutions For $< HS$

solutions as well as the original solution for that category are shown. Other points are hidden from view. For example, we expect the bootstrapped solutions for the category $< HS$ to be fairly unstable, since its original marginal frequency is rather low. This is case as seen in the right panel of Figure 6.9. On the other hand, *CGRAD* taking up almost half of the total mass of the columns is expected to have a more stable bootstrap solution. This is also the case, as seen in Figure 6.10.

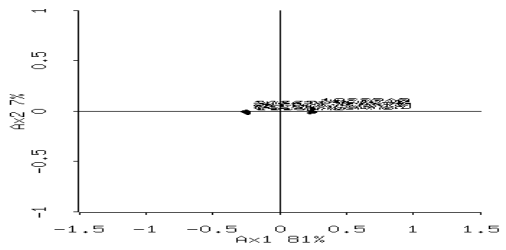

FIGURE 6.10. Bootstrap Solutions For *CGRAD*

To simplify the computations involved in the bootstrapping, each set of bootstrap iterations contains one replication of the original solution. Therefore, the original solution will be "covered up" by at least one bootstrap iteration. Point separation becomes a useful tool at this point. For a given category, one may wish to determine the position of the original solution in the cloud of its bootstrap solutions relative to the positions of the other category solution points. This may be done easily by expanding points that are overlapping until the original solution is found. On color monitors this becomes easier because the original solution points are colored and therefore easier to distinguish.

We compare the results from bootstrapping the singular values with those derived by the asymptotic expansion method. The following Table contains the estimated asymptotic covariance matrix of the singular values Λ ([23]). Tests of the hypotheses that the singular values are zero are strongly rejected, giving observed approximate Z values of $Z_1 = 36.72$ and $Z_2 = 9.06$ where $Z_k = N^{\frac{1}{2}}\Lambda(k,k)/\sigma_{k,k}$. Therefore, the results from the 20 bootstrapped samples agree with the asymptotic ones. It is worth noting that in order to conduct simultaneous hypothesis tests of the components of Λ, O'Neill also gives first and second order moments of the central Wishart matrix variate but those have not been implemented in the current version of this program.

Table 1
Decomposition of total inertia along principal axes

AXES	INERTIA	%of INERTIA	Cum %
1	.13658	80.967	80.967
2	1.2053E-2	7.1452	88.112
Total	.1686857456		

Table 2
Asymptotic Covariance Matrix of Singular Values

1.0235	0.2022
0.2022	1.4812

6.2. **MCA Application.** In this example we use a different set of variables from the NELS:88 data set. Excluding all observations with some information missing we end up with a sample size

Table 3

Decomposition of total inertia along principal axes for 20 Bootstrap Samples

AXES	INERTIA BS MEANS	INERTIA BS SDs	%of INERTIA	Cum %
1	.1376672426	5.9479779523E-3	75.278	75.3
2	1.5842090226E-2	2.20087987E-3	8.6627	83.9
Total	.18288			

of 21,562 observations (students). A description of the variables along with their coding is given in Appendix C.

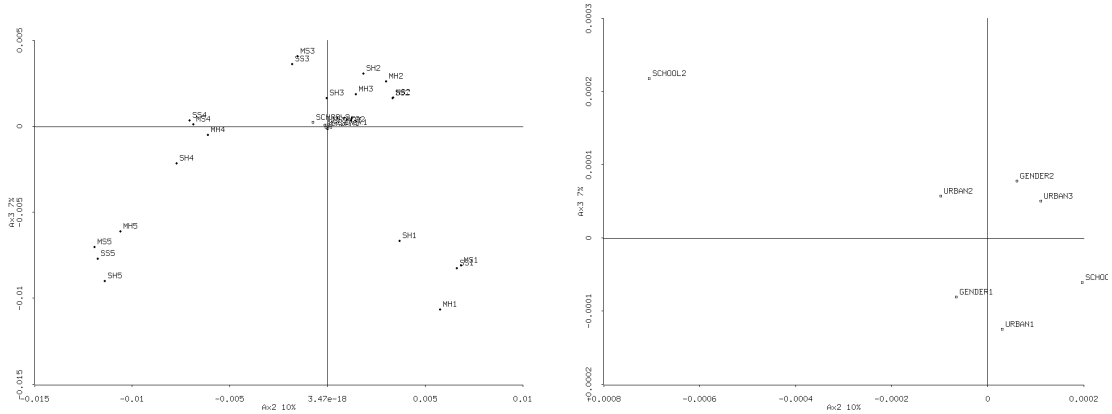

FIGURE 6.11. Left:MCA category points; Right:MCA category points of passive variables

For this data set the variables SCHOOL, URBAN and GENDER were treated as passive. A two dimensional solution accounts for 10% and 7% of the total inertia respectively. The category points of the solution are displayed in Figure 6.11. It can be seen that the amount of time spent on homework is associated with the scores received, for both subjects. More interestingly, there seems to be a clustering of students according to the same category levels. Thus, students that get high scores in mathematics and science tend to spend over 4 hours a week doing homework, while students that receive low scores tend not to allocate any time on homework. Similar findings hold for the other categories. The larger distance of points of categories 1 and 5 from the origin for all variables, as opposed to those of categories 2 to 4 is a result of their lower marginal frequencies (see Tables C1 and C2). Given the very large sample size, these results tend to confirm the stylized fact that scores are positively associated with the amount of time spent studying a subject. What about the effect of the background variables? Their category points are located around the origin, which implies that they do not exhibit any particular association with scores and time spent on homework. To get a better idea, a "zoom-in" display is shown in Figure 6.11. It seems that students attending private schools are more prone to studying and consequently receive higher scores. On the other hand gender and the degree of urbanicity seem to play no role, as expected. The problem with MCA is that it does not provide information about individual profiles. That's why we turn our attention to an analysis of profiles (ANAPROF) in order to get a better understanding of the association of time spent studying and scores.

6.3. **ANAPROF Application.** We continue with the analysis of the previous data set after dropping the background variables, since they contributed very little. The category points are shown in Figure 6.12, and they exhibit a very similar pattern (as expected) to the one derived from MCA. It is worth noting that in this case the first two axes account for 13% and 10% of the total inertia, respectively. The various profiles are shown in Figure 6.12 and in more detail in Figure 6.3. It can be seen that the profiles are arranged along a horseshoe (similar to the one exhibited by the category points), although the interior of the horseshoe is filled. This indicates that there are students with 'mixed' profiles, i.e. that spend a lot of time on homework and score poorly, and vice versa, students that score high and spend very little time studying. However, in the first two quadrants (top panel in Figure 6.14) the majority of the students spends some time studying and scores satisfactorily, while in the other two there are students with 'mixed' profiles along with students that study a lot and score high, or do not study at all and score low.

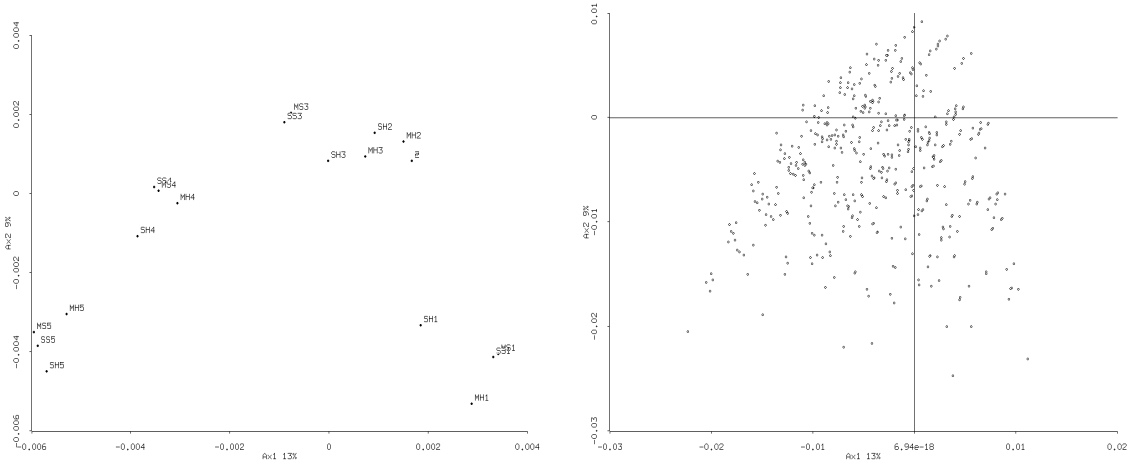

FIGURE 6.12. ANAPROF category points and ANAPROF profiles

7. **Appendix A - Fisher's Eye and Hair Color Example**

The following Table shows the 4×5 contingency table of 5387 school children from Caithness, Scotland, classified according to the two discrete variables, eye color and hair color.

Table A1

	Hair Color					
Eye Color	Fair	Red	Medium	Dark	Black	Total
Light	688	116	584	188	4	1580
Blue	326	38	241	110	3	718
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

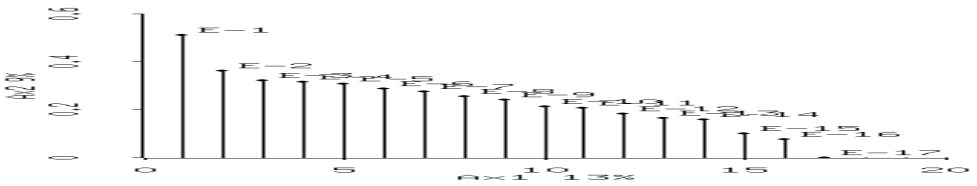

FIGURE 6.13. ANAPROF eigenvalues

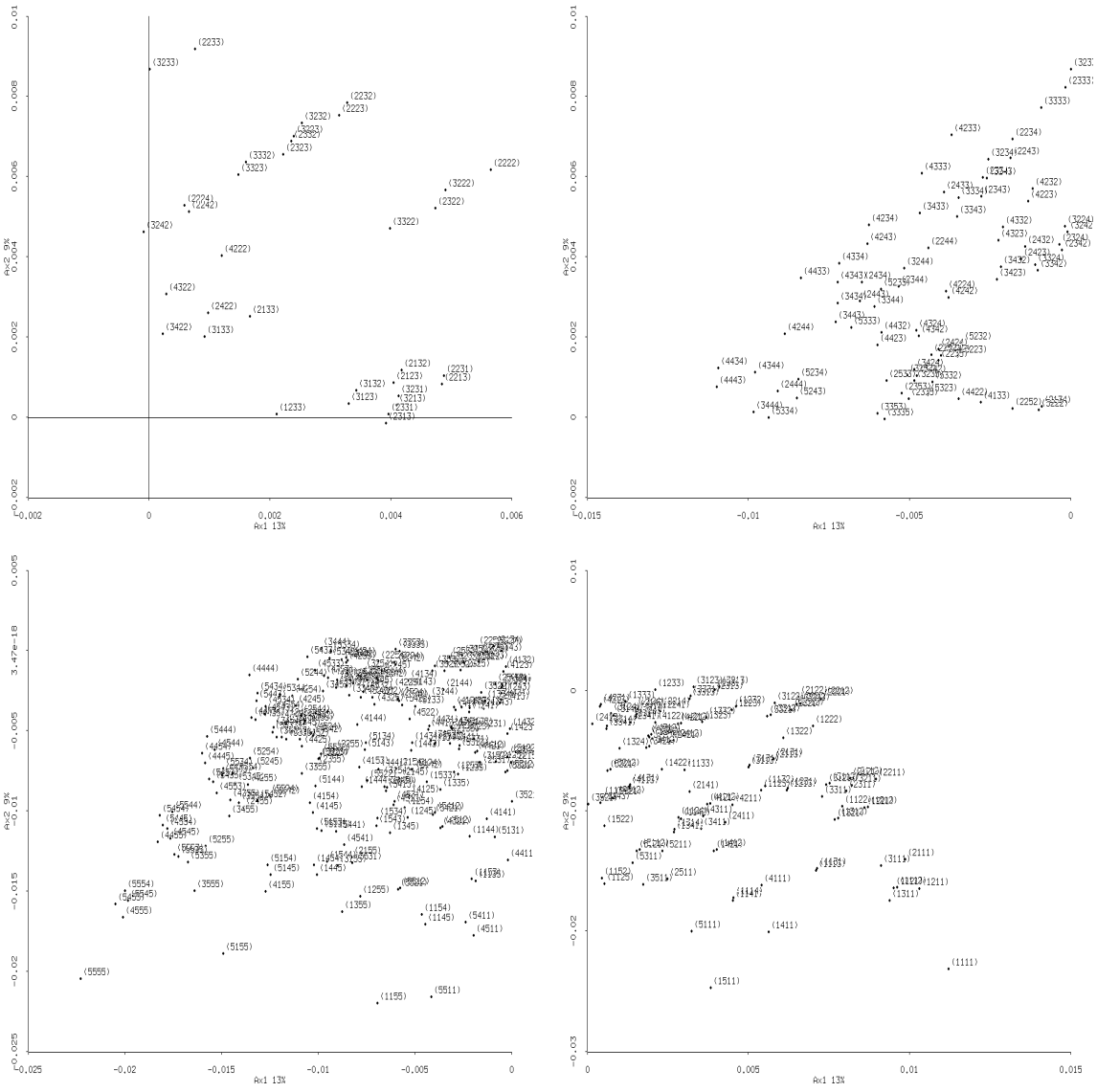

FIGURE 6.14. The four quadrants (counterclockwise) of ANAPROF profiles

8. **Appendix B - First NELS:88 Example**

The variables used are *F1S48A* - how far the father wants the student to go in school. and *F1S53B* - type of occupation student expects to have at age 30.

Table B1

	F1S48A										
F1S53B	$< HS$	HS	VOC	$2-YR$	$4-YR$	$CGRAD$	$PGRAD$	DK	DC	NA	Total
$CLER$	4	30	59	45	36	136	29	53	24	43	459
$CRAFT$	10	66	166	39	45	107	25	63	13	43	577
$FARM$	0	15	28	13	16	40	7	17	3	7	146
$HOME$	1	27	15	20	21	116	21	38	9	28	296
$LABOR$	5	13	20	10	7	14	4	8	2	12	95
$ADMIN$	4	29	51	41	86	354	144	37	13	75	834
MIL	4	30	40	32	48	157	46	50	10	35	452
$OPER$	4	21	33	9	6	28	11	22	7	10	151
$PROF$	14	64	114	131	365	1898	702	214	50	216	3768
$OWNER$	5	31	72	43	79	355	121	52	7	48	813
$PROT$	0	31	52	51	55	141	29	35	13	28	435
$SALES$	2	10	18	11	44	145	37	19	10	15	311
$TEACH$	5	15	23	17	77	385	90	36	16	39	703
$SERV$	2	23	41	15	13	64	21	41	11	26	257
$TECH$	2	15	80	42	84	361	112	58	14	40	808
Total	62	420	812	519	982	4301	1399	743	202	665	10105

9. **Appendix C - Second NELS:88 Example**

The set of variables examined in this example are (in parentheses the name of the variable in the Base Year Student Survey is given)

1. MH: Time spent on mathematics homework (BYS79A)
2. SH: Time spent on science homework (BYS79A)
3. MS: Mathematics standardized score (BY2XMSTD)
4. SS: Science standardized score (BY2XSSTD)
5. SCHOOL: Private or public (G8CTRL)

Some summary statistics on these variables are given in the following Tables.

Table C1 (%) N=21562

Variable	Categories				
	1	2	3	4	5
MH	8.3	41.6	22.9	18.2	9.1
SH	16.8	45.3	20.5	14.0	3.4

where the following *coding* is employed: 1=None, 2=Less than 1 hour, 3=1 hour, 4=2-3 hours, 5=more than 4 hours

Table C2 (%) N=21562

Variable	Mean	Std Dev	Min	Q1	Median	Q3	Max
MS	51.8	10.3	31.7	42.3	49.7	58.7	70.5
SS	50.8	10.2	32.6	42.9	49.6	57.8	80.1

Table C3 (%) N=21562

Variable	Categories				
	1	2	3	4	5
MS	15.9	35.3	24.5	19.5	4.8
SS	15.6	35.6	26.7	17.9	4.2

where the following *coding* is employed: 1=(0,40], 2=(40,50], 3=(50,58], 4=(58,65], 5=(65,100]

Finally, 78% of the students attended public schools and 22% private schools (including religious schools).

REFERENCES

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

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

[3] BMDP Manual, Version 7.0, University of California Press

[4] Bond, J. and Michailidis, G. (1996), "Homogeneity Analysis in Xlisp-Stat", *Journal of Statistical Software*, **1**, 2, 1-32

[5] De Leeuw, J. (1983), "On the Prehistory of Correspondence Analysis", *Statistica Neerelandica*, **37**, 161-164

[6] Devroye, L. (1986), *Nonuniform Random Variate Generation*, New York: Springer Verlag

[7] Fisher, R.A. (1940), "The Precision of Discriminant Functions", *The Annals of Eugenics*, **10**, 422-429

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

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

[10] Greenacre, M. (1988), "Correspondence Analysis of Multivariate Categorical Data by Weighted Least Squares," *Biometrika*, **75**, 457-467

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

[12] 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

[13] 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

[14] Hirschfeld, H.O. (1935), "A Connection between Correlations and Contingency", *Proceedings of the Cambridge Phylosophical Society*, **31**, 520-524

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

[16] Michailidis, G. and de Leeuw, J. (1995), "Nonlinear Multiavriate Analysis of NELS:88," UCLA Statistical Series, #176

[17] Michailidis, G. and de Leeuw, J. (1996), "The Gifi System for Nonlinear Multivariate Analysis," UCLA Statistical Series, #204

[18] Michailidis, G. and de Leeuw, J. (1996), "Constrained Homogeneity Analysis with Applications to Hierarchical Data," UCLA Statistical Series, #207
[19] Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications*, Toronot: Toronto University Press
[20] O'Neill, M.E. (1978), "Asymptotic Distributions of the Canonical Correlations from Contingency Tables," *Australian Journal of Statistics*, **20**, 75-82
[21] O'Neill, M.E. (1978), "Distributional Expansions for Canonical Correlations from Contingency Tables," *Journal of the Royal Statistical Socitey, B*, **40**, 303-312
[22] O'Neill, M.E. (1980), "The Distribution of Higher-Order Interactions in Contingency Tables," *Journal of the Royal Statistical Socitey, B*, **42**, 357-365
[23] O'Neill, M.E. (1981), "A Note on the Canonical Correlations from Contingency Tables," *Australian Journal of Statistics*, **23**, 58-66
[24] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag
[25] SAS/STAT User's Guide, Version 6, SAS Institute Inc.
[26] SPSS Categories User's Manual, SPSS Inc.
[27] Tierney, L. (1990)¡ *LISP-STAT*, New York: Wiley
[28] 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

PROGRAM IN STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, DEPARTMENT OF ENGINEERING-ECONOMIC SYSTEMS & OPERATIONS RESEARCH, STANFORD UNIVERSITY, STANFORD, CA 94305

E-mail address: jbond@stat.ucla.edu, gmichail@leland.stanford.edu