# CONSTRAINED HOMOGENEITY ANALYSIS WITH APPLICATIONS TO HIERARCHICAL DATA

GEORGE MICHAILIDIS AND JAN DE LEEUW

ABSTRACT. In this paper we extend the techniques of homogeneity analysis and nonlinear principal components analysis to a multilevel sampling design framework. We also propose some models that take advantage of the multilevel nature of the sampling design, and allow us to make within-groups and between-groups comparisons. Furthermore, it is shown that several models proposed in the literature for panel and event history data, can be casted naturally into our framework. A data set from the National Educational Longitudinal Study (NELS:88) is used to illustrate the techniques introduced in the paper.

## 1. **Introduction to Homogeneity Analysis**

The basic technique studied in this paper is known under many different names. For example, we have *principal components of scale analysis* [19, 20], *factorial analysis of qualitative data* [7], *second method of quantification* [21], *multiple correspondence analysis* [2, 17, 27] and *homogeneity analysis* [10, 15]. The technique has been derived from various data analytic points of view, starting with ideas from principal component analysis, from multidimensional scaling and from scale analysis. It combines the aspect of maximal correlation between variables with that of optimal scaling. Pearson in 1907 discovered some of the basic facts connected with the technique while studying variation of the correlation coefficient of two variables under different choices of quantification of the variable, and Fisher [14] was the first one to use it in data analysis. On the other hand, the first optimal scaling techniques were introduced by Shepard and Kruskal (e.g. [24]) in the early sixties. Kruskal, Roskam and Lingoes wrote families of computer programs that combined different linear and nonlinear models with the idea of optimal scaling. De Leeuw, Young and Takane continued along a similar path in the seventies (for a review see [42]). More recently Breiman and Friedman [4] developed similar methods using conditional expectations and exploring the structure of the $L_2$ space.

We next give a brief introduction to homogeneity analysis in graphical language. In order to make things concrete, we motivate the technique using an example from education. The data come from the National Education Longitudinal Study of 1988 (NELS:88). A brief discussion of the sampling design of NELS:88, along with a description of the variables used in the examples is given in the Appendix.

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 [36] states that by the year 2000 ”every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning.” Because in many situations learning is constrained in an atmosphere of fear and disorderliness, student behavior influences school atmosphere and the climate for learning (whether it takes the form of violence and risk taking activities such as bringing weapons to school or using alcohol and drugs) or a low commitment to academic effort (such as poor attendance, lack of discipline or study habits) [8]. These student behaviors also play a key role in determining student success in school and beyond (see [22] 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, 37]. 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 data in this example come from a public urban school in the West and the sample size is 37. Due to the categorical nature of the variables many studies (e.g. [22]) look at the data and draw

conclusions about the overall school climate by examining only the univariate frequencies of the variables (shown in Figure 1.1). Such pictures tell only part of the story contained in the data, since they ignore possible interactions between the various variables, and reveal nothing about possible patterns of student responses. Nevertheless, graphical displays are very useful in uncovering the main regularities of complicated multivariate data. In what follows, we describe a way of making such a multivariate picture for the data at hand, and discuss its properties.

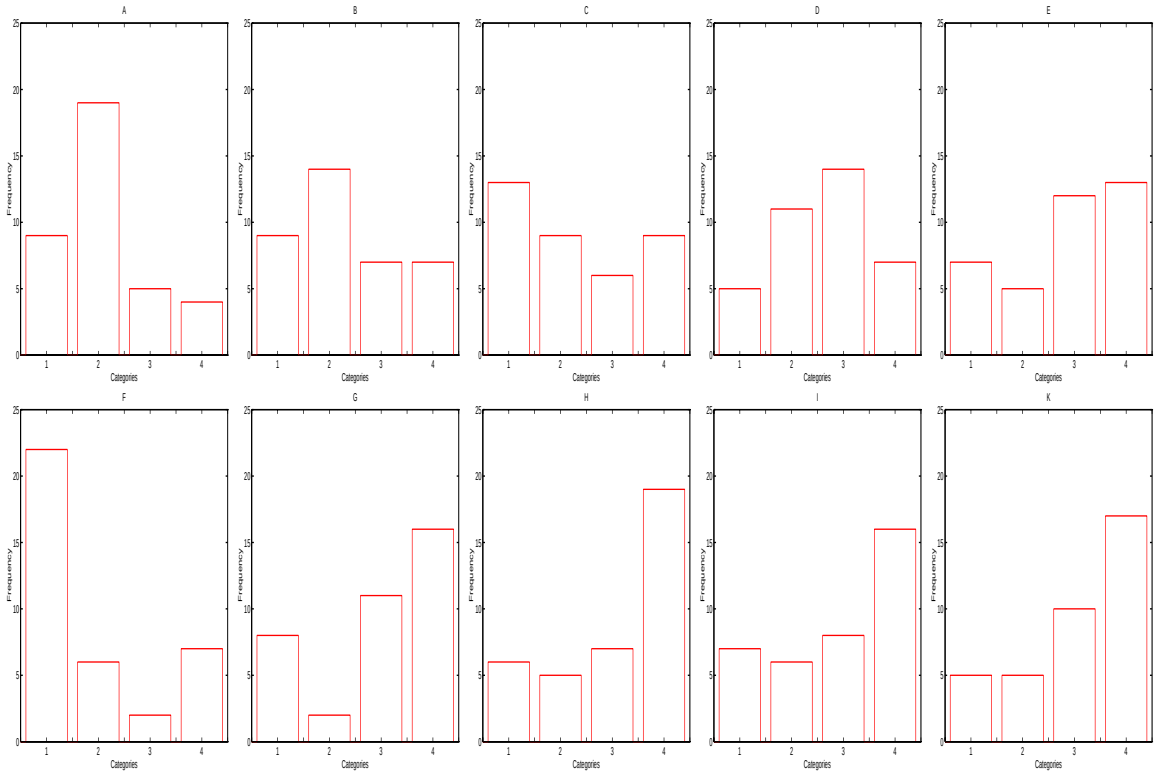


FIGURE 1.1. Frequencies of the variables

In general, the data consist of $N$ objects and $J$ categorical variables (37 students and 10 variables in the example), with variable $j \in \mathbf{J} = \{1, 2, \ldots, J\}$ taking $\ell_j$ possible values (4 in this case). Let $L = \sum_{j \in \mathbf{J}} \ell_j$ denote the total number of categories over all variables. The data are coded as $J$ *indicator* matrices $G_j$, where $G_j(i,t) = 1,\ i = 1, \ldots, N,\ t = 1, \ldots, \ell_j$, if object $i$ belongs to category $t$ for variable $j$ and $G_j(i,t) = 0$ if it belongs to some other category. The $N \times L$ matrix $G = [G_1, G_2, \ldots, G_J]$ is called the *super-indicator* matrix.

One can represent all information in the data by a bipartite graph with $N + L$ vertices (nodes) and $NJ$ edges [26]. The first set of vertices represents the objects, while the second the categories. Each edge connects an object and a category. Therefore, each object is connected to $J$ categories (one edge for each variable). On the other hand, each category is connected to the objects they belong to it (hence, the number of edges varies). The super-indicator matrix $G$ corresponds to the *adjacency* matrix of the graph [26]. Drawing such a graph would not yield useful insights into the

data, except in cases with a very small number of objects and variables. An alternative approach is to find a way to *embed* the bipartite graph into a low dimensional Euclidean space ($\mathbb{R}^p$) or in other words draw a joint map of objects and categories in $\mathbb{R}^p$. The choice of low dimensionality (usually $p \leq 3$) is because the map can be plotted and the choice of Euclidean space stems from its nice properties (projections, triangle inequality) and our overall familiarity with Euclidean geometry. In order to achieve our goal, we need to *scale* (assign numerical values to) the objects and the categories. 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 $\ell_j$ vertices of variable $j$. We call $X$ the *object scores* matrix and the $Y_j$'s the *category quantifications* matrices. By assigning random values to $X$ and the $Y_j$'s and then plot the $N + L$ vertices and the corresponding edges we will typically get a picture similar to the one shown in the left panel of Figure 1.3. It can be seen that very little has been gained by this 2-dimensional representation. A more aesthetically pleasing and 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. Thus, our goal becomes of making a *graph plot* that minimizes the average *squared* length of the edges. This criterion is chosen because it leads to an eigenvalue problem, and thus connects nicely to many classical multivariate analytic techniques (e.g. principal components analysis, canonical correlation analysis etc).

To formalize our criterion in a convenient way, we make use of the $X$, $Y_j$ and $G_j$ matrices. The squared length of the $N$ edges for variable $j$ is given by

$$\sigma_j(X, Y_j) = \text{SSQ}\left(X - G_j Y_j\right),\ j \in \mathbf{J}, \tag{1.1}$$

where, $\text{SSQ}(H)$ denotes the sum of squares of the elements of the matrix $H$. The corresponding graph drawing with $N + \ell_j$ vertices and $N$ edges is known as the *star plot* for variable $j$. The graph plot is then the overlay of the $J$ star plots. Hence, the average squared edge length over all variables is given by

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\left(X - G_j Y_j\right), \tag{1.2}$$

and this is the function we want to minimize simultaneously over $X$ and $Y_j$. In order to avoid the trivial solution corresponding to $X = 0$, and $Y_j = 0$ for every $j \in \mathbf{J}$, we require in addition

$$X'X = NI_p, \tag{1.3}$$

$$u'X = 0, \tag{1.4}$$

where $u$ is a vector of appropriate dimensions comprised of all ones. The second normalization restriction basically requires the graph plot to be centered around the origin of the $p$-dimensional Euclidean space, while the first restriction standardizes the squared length of the object scores (to be equal to $N$), and in two or higher dimensions requires the columns of $X$ to be in addition orthogonal.

A close examination of (1.2) provides a different interpretation closer to ideas from principal component analysis. Suppose that the matrices $Y_j$ contain weights that transform the original categorical variables $G_j$ into new scales $G_jY_j$, and suppose that $X$ is considered to be a common hypothetical variable. Then, (1.2) measures (in an appropriate metric) the differences between the original $J$ variables and the hypothetical variable $X$ that serves as a new overall scale for the objects. In case the weights gave rise to *linear* transformations of the categorical variables, the solution to the above minimization problem would correspond to performing linear principal components analysis. It is worth noting that the variables enter symmetrically in the criterion (1.2), and no special role is assigned to any of them (like in regression models where one of the variables serves as the dependent/outcome variable). The solution to the minimization problem given by (1.2), (1.3) and (1.4) is found by employing an *Alternating Least Squares* (ALS) algorithm. In the first step of the algorithm, (1.2) is minimized with respect to $Y_j$ for fixed $X$. The set of normal equations is given by

$$D_jY_j = G_j'X,\ j \in \mathbf{J}, \tag{1.5}$$

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

$$\hat{Y}_j = D_j^{-1}G_j'X,\ j \in \mathbf{J}. \tag{1.6}$$

In the second step of the algorithm, (1.2) is minimized with respect to $X$ for fixed $Y_j$'s. The normal equation is given by

$$JX = \sum_{j=1}^{J} G_jY_j, \tag{1.7}$$

so that

$$\hat{X} = J^{-1}\sum_{j=1}^{J} G_jY_j. \tag{1.8}$$

In the third step of the algorithm the $X$ matrix is column centered by setting $W = \hat{X} - u(u'\hat{X}/N)$, and then orthonormalized by the modified Gram-Schmidt procedure [15] $X = \sqrt{N}\text{GRAM}(W)$, so that the normalization restrictions (1.3) and (1.4) are satisfied. These three steps are repeated until the algorithm finds the global minimum of (1.2) (see chapter 3 in [15]). This solution is known in the literature ([10, 11, 15]) as the Homals solution (homogeneity analysis by means of alternating least squares).

**Remark 1.1.** *Rotational Invariance.* It is worth mentioning the *rotational invariance* property of the Homals solution. To see this, suppose we select a different basis for the column space of the matrix $X$; that is, let $X^\sharp = X \times R$, where $R$ is a rotation matrix satisfying $R'R = RR' = I$. We then get from (1.6) that $Y_j^\sharp = D_j^{-1}G_j'X^\sharp = \hat{Y}_jR$. Thus, any rotation of the object scores and of the category quantifications corresponds to a solution to the problem given in (1.2).

Equation (1.6) shows that a category point is in the centroid of the object scores that belong to it (see Figure 1.2), while equation (1.8) shows that an object point is the average of the quantifications of the categories it belongs to. Hence, the Homals solution accomplishes the goal set forth of producing a graph plot with objects close to the categories they fall in and categories close to the objects belonging in them.

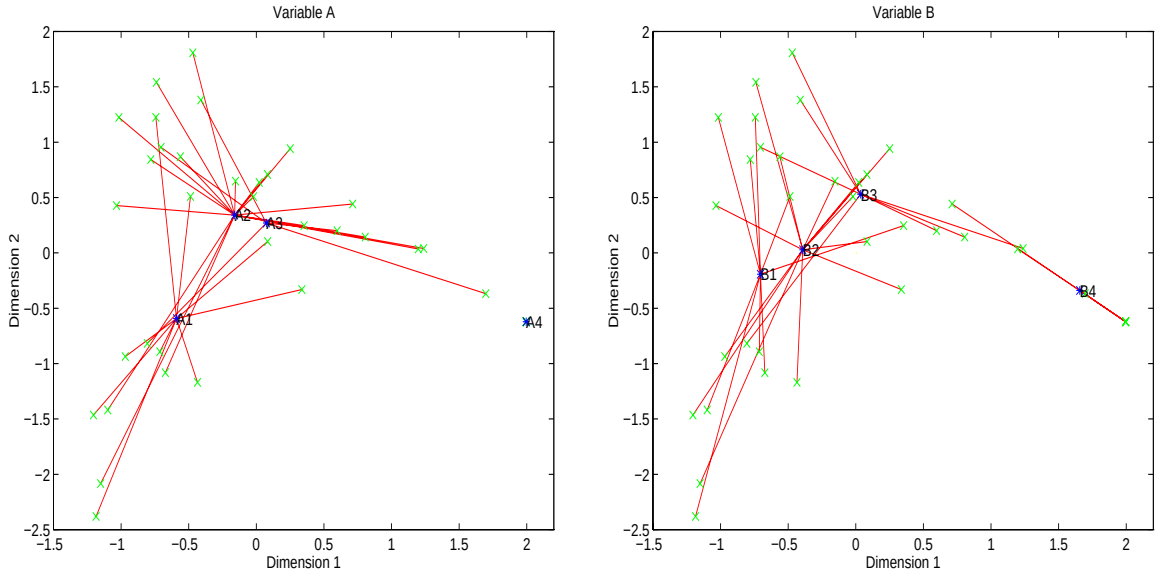


FIGURE 1.2. Star Graphs for Variables A and B

We introduce next some quantities that allow us to evaluate the fit of the derived solution. Once the ALS algorithm has converged, by using the fact that $Y_j'D_jY_j = Y_j'G_j'X$, we can write (1.2) as

$$J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(X-G_jY_j\big)'\big(X-G_jY_j\big)=J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(X'X+Y_j'G_j'G_jY_j-2Y_j'G_j'X\big)= \tag{1.9}$$

$$J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(X'X-Y_j'D_jY_j\big)=J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(NI_p-Y_j'D_jY_j\big)=Np-J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(Y_j'D_jY_j\big).$$

We can define the *discrimination measures* given by

$$\eta_{js}^2\equiv Y_j'(.,s)D_jY_j(.,s)/N,\ j\in\mathbf{J},\ s=1,\ldots,p, \tag{1.10}$$

where $Y_j(.,s)$ denotes the $s^{th}$ column of the matrix $Y_j$ and represents the quantification for variable $j$ in the $s^{th}$ dimension of the solution. Geometrically the discrimination measures give the average squared distance (weighted by the marginal frequencies) of 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 $G_jY_j(.,s)$ and the corresponding column of object scores $X(.,s)$ (see chapter 3 in [15]). Hence,

(1.2) can also be expressed as

$$N\big(p - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2\big). \tag{1.11}$$

The quantities $\gamma_s = J^{-1}\sum_{j=1}^{J}\eta_{js}^2,\ s = 1, \ldots, p$ are called the *eigenvalues* (see Remark 1.2) and correspond to the average of the discrimination measures. The eigenvalues give an overall measure of fit of the derived map in each of the $p$ dimensions.

**Remark 1.2.** *Homogeneity Analysis as an Eigenvalue and Singular Value Decomposition Problem.* One of the reasons that squared edge lengths are appealing is that they give rise to an eigenvalue or a singular value decomposition problem [10, 31]. It can be shown that the object scores $X$ are the eigenvectors corresponding to the $p$ largest eigenvalues of the matrix $\mathcal{L}P_\star\mathcal{L}$, where $P_\star$ is the average of the $J$ projectors spanned by the columns of the indicator matrices $G_j$ and $\mathcal{L}$ a centering operator of the form $I - uu'/u'u$, and also to the first $p$ left singular vectors of the matrix $J^{-1/2}(I - uu'/N)GD^{-1/2}$, which is the superindicator matrix $G$ in deviation from column means and corrected for marginal frequencies respectively.

We summarize next some basic properties of the Homals solution.

- ◇ Category quantifications and object scores are represented as points in a joint space.
- ◇ A category point is the centroid of objects belonging to that category. This is a direct consequence of (1.6).
- ◇ Objects with the same response pattern (identical profiles) receive identical object scores (follows from (1.8)). 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 (1.10)).
- ◇ If a category applies uniquely to only one 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 (1.6)).
- ◇ 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 normalization of the object scores, since $u'D_jY_j = u'D_jD_j^{-1}G_j'X = u'G_jX = u'X = 0$.

Let us now turn our attention to what we can accomplish by applying homogeneity analysis to the data of that particular school from NELS:88. The graph plot of a 2-dimensional solution is given in

the right panel of Figure 1.3, a considerable improvement over the random representation depicted in the left panel. The solution exhibits a satisfactory fit with eigenvalues of .64 and .35 for the first two dimensions respectively. The graph plot shows several interesting patterns. In order to examine them more closely we study the arrangement of only the category points on the 2-dimensional map (see Figure 1.4). It can be seen that they form 3 distinct groups. In the right part of the graph we find the 'not a problem' categories of all the 10 variables. Thus, students located in this area of the map (see Figure 1.5) are associated with these categories, which means that this set of $8^{th}$ graders believe that their school does not have any problem areas, or in other words the overall school climate can be characterized as positive. On the other hand, in the lower left quadrant of Figure 1.4 we find the 'serious problem' categories of all the variables indicating that the students in that part of the graph view the school climate as negative. Finally, in the upper left quadrant we find the 'moderate' and 'minor' category points. It is interesting to note that the 'clustering' of the students is done according to the same category levels, which implies that there are groups of students responding primarily with 4's (not a problem) to all the questions, other groups mixing 2's and 3's and other students using only 1's (serious problems) for all of their answers. However, the solution indicates that there are very few students mixing 1's and 4's. The fact that the majority of the students are located in the left part of Figure 1.5 indicates that most of the students think that violence (physical conflicts, theft and robbery, abuse of teachers), absenteeism, use of alcohol and drugs are to some degree a problem in their school. Moreover, due to the fact that category points E1, G1, H1, I1, K1, A4, B4, C4 and F4 are further away from the origin implies that few students chose those categories in their responses, which is consistent with the univariate response patterns shown in Figure 1.1. To a large extent the analysis cleanly separates the small group of students that thinks their school is problem free, from the majority of students that believes that the overall climate in their school is not positive. The discrimination measures, shown in Figure 1.6, indicate that variables B, C, D, E, F separate the category points particularly well along the first dimension, while variables G, H and I along the second dimension. Thus, the 'scale' expressed by the first dimension summarizes information about tardiness, absenteeism, cutting class, physical conflict between students, robbery and theft, and the 'scale' expressed by the second dimension information about illegal use of drugs, alcohol and possession of weapons. The derived joint map of students and categories revealed many interesting things about the school and its students that could not be found by only examining the univariate frequencies.

**Remark 1.3.** *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 $M_j(i,i) = 1$ 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

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \mathrm{tr}\big(X - G_j Y_j\big)' M_j \big(X - G_j Y_j\big), \tag{1.12}$$

subject to the normalization restrictions $X' M_* X = JNI_p$ and $u' M_* X = 0$. The $\hat{Y}_j$'s are given by (1.6), 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

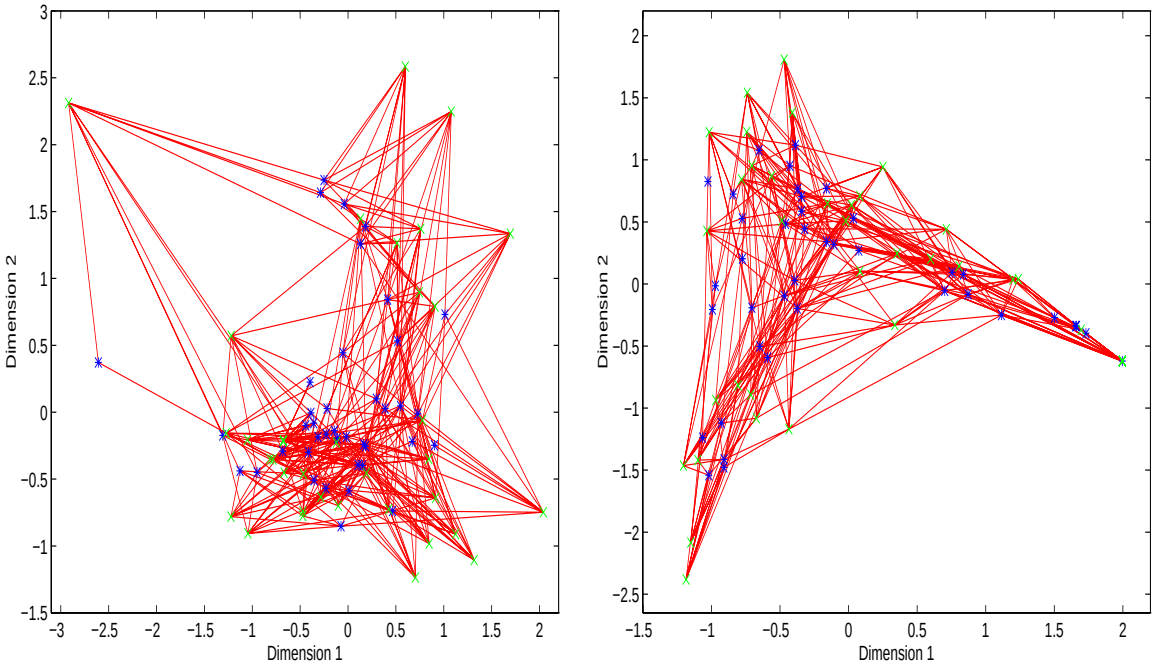


FIGURE 1.3. Graph Plot of the School Data; Left: Random Graph, Right: Homals Solution (*=category points, x=object points)

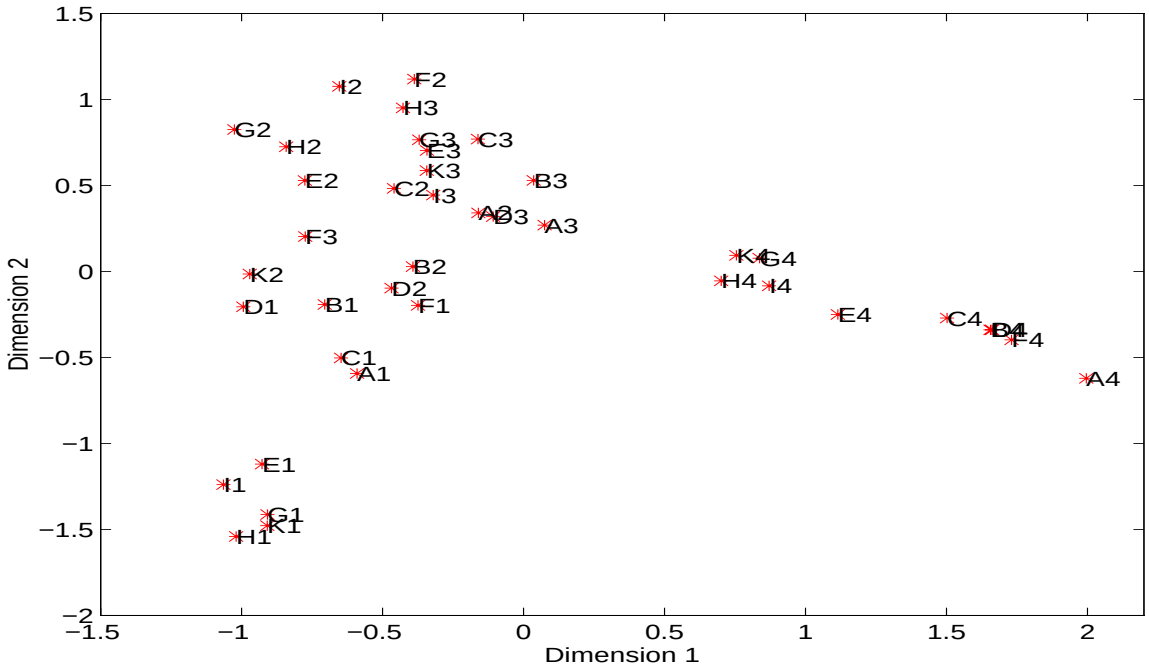


FIGURE 1.4. Category Quantifications

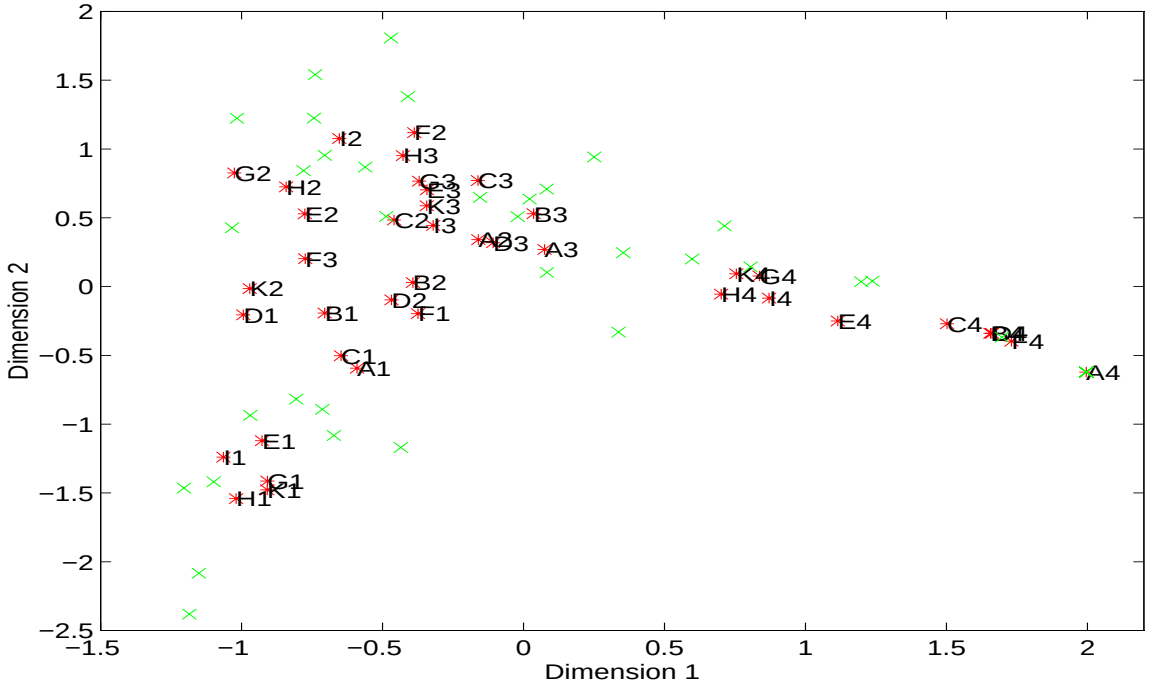


FIGURE 1.5. Category Quantifications and Object Scores (*=category points, x= object points)

weighted summation with respect to the row scores of $X$, some of the scores are skipped. This option is known in the literature [15] 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. On the other hand, the missing data multiple categories option amounts to minimizing (1.12) subject to $X'X = NI_p$ and $u'X = 0$, which tends to make observations with many missing data dominate the solution.

.

**Remark 1.4.** *Rank-one Restrictions and Nonlinear Principal Components Analysis.* The Homals solution leaves the category quantifications $Y_j$ free, thus treating the variables as *nominal*, that is, only the fact that the objects fall in some category plays a role. However, in many data analytic situations we are dealing with ordinal type of variables (e.g. variables measured on a Likert type of scale) or numerical (continuous) type of variables (e.g. age, income). Homogeneity analysis simply ignores this information. The reason is that it can not readily be incorporated in the multiple quantification framework outlined in the previous section. In homogeneity analysis, categories might be in a certain order on the first dimension, in a different order on the second dimension

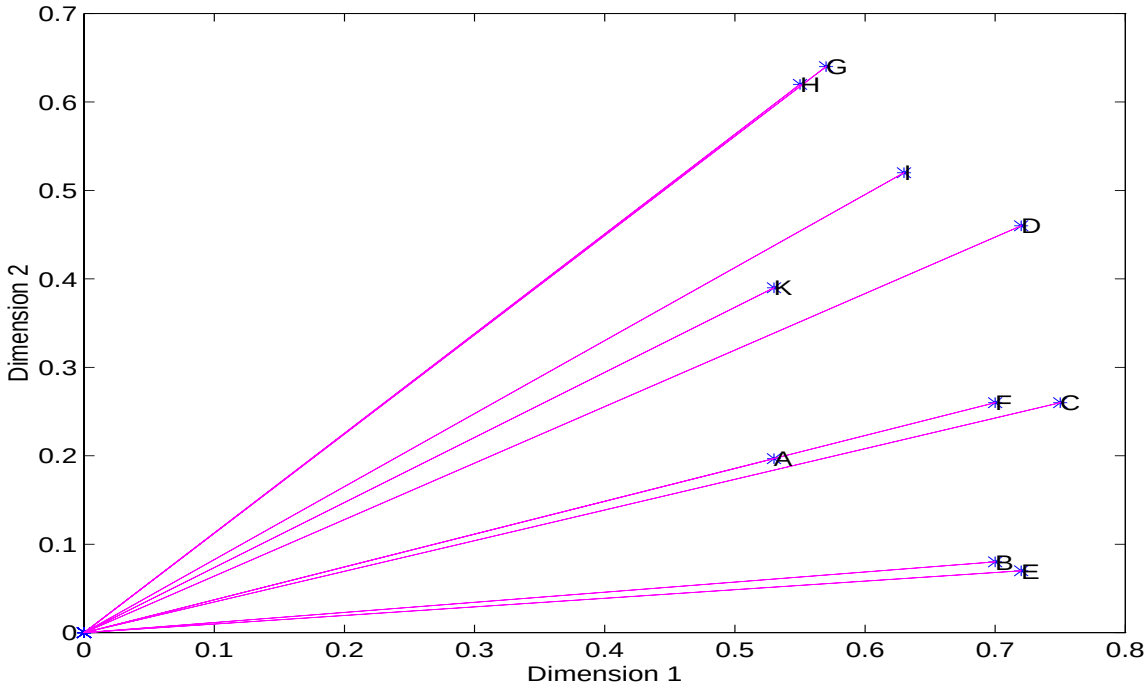


FIGURE 1.6. Discrimination Measures

and yet in a different one on the third dimension, and so on. If on the other hand we require the categories to be in a prespecified order in all dimensions (as should be the case for ordinal data), this would result in the various quantifications being highly intercorrelated. In order to overcome these difficulties, it is also required that the quantifications satisfy a *rank-one* restriction (see chapter 4 in [15]); that is,

$$Y_j = q_j \beta_j', \ j \in \mathbf{J}, \tag{1.13}$$

where, $q_j$ is a $\ell_j$-column vector containing the *single category quantifications* and $\beta_j$ a $p$-column vector of *component loadings*. In this case the quantifications in $p$ dimensions become proportional to each other. So, in this case the problem becomes to minimize (1.2) subject not only to the normalization conditions (1.3) and (1.4), but also the restriction (1.13). The solution to this problem is known in the literature as the Princals solution (principal components analysis by means of alternating least squares) [15, 31], since in the presence of numerical variables it gives rise to the classical principal components analysis of the correlation matrix. The Princals model allows the data analyst to treat each variable differently, some as nominal, some as ordinal and some as numerical. In that sense, Princals generalizes the Homals model.

## 2. **Multilevel Modeling**

In this Section we introduce the main theme of this paper, which is to extend the technique of homogeneity analysis to a multilevel sampling design framework. In many situations individual objects (level-1 units) can be naturally grouped (*clustered*) into groups (clusters, level-2 units). This would be the case if we wanted to examine more than one schools in the NELS:88 example. Other examples of a multilevel nature can be found in educational research with students grouped by class or school or school district, in sociological research with individuals grouped by socioeconomic status, in marketing research with consumers clustered in geographical regions, while in longitudinal studies we have repeated measurements on individuals. In the first example clusters correspond to classes/schools/ school districts, in the second to various a priori defined levels of socioeconomic status, in the third to regions (such as counties, states or even the northeast, the southwest etc), and in the fourth example time points are clustered within individuals, thus individuals correspond to the level-2 units. Formally, we collect data on $N$ objects grouped naturally in $K$ clusters, with $n_k$ objects per cluster, $k \in \mathbf{K} = \{1, \ldots, K\} (\sum_{k=1}^{K} n_k = N)$. Once again, we want to examine $J$ categorical variables, with $\ell_j,\ j \in \mathbf{J}$ categories each. The purpose of this study is twofold. Our first goal is to extend homogeneity analysis to the multilevel sampling framework. In many cases however, this approach is either not very meaningful, or not feasible. For example in the National Education Longitudinal Study of 1988 (NELS:88) data set there are approximately 24,500 students grouped in over 1,000 schools. It is easy to see that examining the category quantifications for each cluster separately is not a particularly informative or useful task. This leads us to our second goal which is to build models that take advantage of the clustering of the objects. More specifically, we shall desire models which can simultaneously express how one variable is connected to another variable across all clusters, and also how one cluster varies (differs) from another.

Very little has be done on applying homogeneity analysis techniques to multilevel data. De Leeuw, van der Heijden and Kreft [12] and van der Heijden and de Leeuw [40] have used these techniques to examine panel and event history data. In that case, data are collected on $n_k = n$ objects for $K$ time periods. The authors introduce three way indicator matrices with objects in the rows, categories of variables in the columns, and time points in the layers to code the data, and apply homogeneity analysis to the collection of such matrices. Their approach is not applicable to other types of multilevel data (such as students clustered within schools). We propose an alternative approach. Let $G_{jk},\ j \in \mathbf{J},\ k \in \mathbf{K}$ denote the $n_k \times \ell_j$ indicator matrix of variable $j$ for cluster $k$. Let $X_k,\ k \in \mathbf{K}$ be the $n_k \times p$ matrix of object scores of cluster $k$, and let $X = [X_1', \ldots, X_K']'$. Similarly, let $Y_{jk}$ be the $\ell_j \times p$ matrix of multiple category quantifications of the $j^{th}$ variable for the $k^{th}$ cluster, and let $Y_j = [Y_{j1}', \ldots, Y_{jK}']'$. We collect the $K$ indicator matrices of variable $j$ in the superindicator matrix

$$G_j = \begin{pmatrix} G_{j1} & 0 & \ldots & \ldots & 0 \\ 0 & G_{j2} & \ldots & \ldots & 0 \\ 0 & \ldots & G_{jk} & \ldots & 0 \\ 0 & \ldots & 0 & \ldots & G_{jK} \end{pmatrix},$$

or in more compact notation $G_j = \bigoplus_{k=1}^{K} G_{jk}$, which is called the *design* matrix. In the remainder of this study we hold the design matrix fixed, since its versatile and general form proves extremely convenient. However, by imposing restrictions on the category quantifications, we are able to generate interesting and useful models and incorporate prior knowledge. It is also shown that the approach taken in [12] and [40] can be derived from our framework. In the case of multiple clusters the average squared edge length function becomes

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\left(X - G_j Y_j\right) = J^{-1} \sum_{j=1}^{J} \sum_{k=1}^{K} \text{SSQ}\left(X_k - G_{jk} Y_{jk}\right). \tag{2.1}$$

In order to avoid the trivial solution we impose the following normalization restriction:

$$X_k' X_k = n_k I_p, \; u' X_k = 0, \text{ for every } k \in \mathbf{K}. \tag{2.2}$$

The other possibility $u'X = 0$ and $X'X = NI_p$ is briefly discussed later on in Remark 2.3.

The problem in (2.1) is identical to the one presented in (1.2); only the normalization differs. Thus, its solution is given by

$$\hat{Y}_j = D_j^{-1} G_j' X, \; j \in \mathbf{J}, \tag{2.3}$$

where $D_j = G_j' G_j = \bigoplus_{k=1}^{K} (G_{jk}' G_{jk}) = \bigoplus_{k=1}^{K} D_{jk}$ is the $K\ell_j \times K\ell_j$ diagonal matrix containing the univariate marginals of variable $j$ for all $K$ clusters. This implies that $\hat{Y}_{jk} = D_{jk}^{-1} G_{jk}' X_k, \; j \in \mathbf{J}, \; k \in \mathbf{K}$. The object scores are given by

$$\hat{X} = \frac{1}{J} \sum_{j=1}^{J} G_j Y_j, \tag{2.4}$$

which gives that $\hat{X}_k = J^{-1} \sum_{j=1}^{J} G_{jk} Y_{jk}$, for every $k \in \mathbf{K}$. Equations (2.3) and (2.4) show that the basic geometric properties of the Homals solution, namely that category points are the centroids of the object scores they belong to them, and object scores are the average of the quantifications of the categories they belong to, continue to hold for every cluster. We define next the *cluster discrimination measures*

$$\eta_{jks}^2 \equiv Y_{jk}'(., s) D_{jk} Y_{jk}(., s)/n_k, \; j \in \mathbf{J}, \; k \in \mathbf{K}, \; s = 1, \ldots, p, \tag{2.5}$$

where $Y_{jk}(., s)$ contains the elements of the $s^{th}$ column of the category quantification matrix $Y_{jk}$. Since, the category quantifications have a weighted sum equal to zero, they are interpreted the usual way; the larger the $\eta_{jks}^2$, the better the categories of that variable in that cluster discriminate between level-1 units. The cluster discrimination measures allow the data analyst to examine variations in the discriminatory power of the variables across the clusters. It is also useful to define the *total discrimination measures* for each variable as

$$\eta_{js}^2 \equiv Y_j'(., s) D_j Y_j(., s)/N, \; j \in \mathbf{J}, \; s = 1, \ldots, p, \tag{2.6}$$

where $Y_j(., s)$ contains the elements of the $s^{th}$ column of $Y_j$. These quantities represent an overall measure of the discriminatory power of each variable. We examine next the relationship between

the total and the cluster discrimination measures. Some algebra shows that

$$\eta_{js}^2 \equiv \frac{1}{N} Y_j'(.,s) D_j Y_j(.,s) = \frac{1}{N} \sum_{k=1}^{K} Y_{jk}'(.,s) D_{jk} Y_{jk}(.,s), \tag{2.7}$$

so that is easy to see that

$$\eta_{js}^2 = \frac{1}{N} \sum_{k=1}^{K} n_k \eta_{jks}^2, \ j \in \mathbf{J}, \ s = 1, \ldots, p. \tag{2.8}$$

Thus, the total discrimination measures of variable $j$ can be expressed as a weighted average of the discrimination measures of the clusters for variable $j$, with the weights given by $n_k/N$ and representing the contribution of cluster $k$ to the total. Thus, larger clusters are weighted more in the total. Finally, we can define *cluster eigenvalues* given by $\gamma_{ks} = J^{-1} \sum_{j=1}^{J} \eta_{jks}^2$, and *total eigenvalues* given by $\gamma_s = J^{-1} \sum_{j=1}^{J} \eta_{js}^2$. The cluster and the total eigenvalues are related by $\gamma_s^2 = N^{-1} \sum_{k=1}^{K} n_k \gamma_{ks}^2$, similarly to the discrimination measures. It can be seen that we have a proportional to size representation of the clusters to the overall fit of the solution.

**Remark 2.1.** *Model Equivalences.* It is worth noting that under normalization (2.2) this model is equivalent to applying the ordinary Homals algorithm (see Section 2) to each of the $K$ clusters separately.

**Remark 2.2.** *Comparing Clusters.* As we have seen in Remark 1.1 the Homals solution is rotationally invariant. The latter combined with the fact that the multilevel Homals solution amounts to calculating $K$ separate solutions (see Remark 2.1), introduces the problem of making meaningful comparisons between clusters, since different clusters may have different orientations of the axes. We would like to make the clusters as similar as possible by rotating their axes to a target solution. Any of the $K$ solutions can be used as the target one. This amounts to solving a *Procrustes orthogonal rotation problem* [16]. Let $X_k(t)$ be the matrix of object scores of the cluster designated as the target solution, and $X_k$ the object scores of some other cluster. We then have to minimize $\text{tr}(X_k(t) - X_k R)'(X_k(t) - X_k R)$ with respect to the $p \times p$ rotation matrix $R$ (i.e. $R'R = RR' = I_p$). The solution is given by first calculating the singular value decomposition of $X_k' X_k(t) = U \Lambda V'$, and then setting $R = UV'$, the orthogonal polar factor of $X_k' X_k(t)$.

**Remark 2.3.** *On another possible normalization.* Instead of normalizing the object scores locally (within every cluster $k \in \mathbf{K}$), we might require a global scaling given by $u'X = \sum_{k=1}^{K} u'X_k = 0$ and $X'X = \sum_{k=1}^{K} X_k' X_k = N I_p$. Some algebra shows that under this normalization the multilevel Homals model is equivalent to a single cluster Homals model with interactive coding; that is, we introduce $K \times \ell_j$ categories for each variable, so that each cluster has its own set of categories. In this case, the clusters are pulled together through the global scaling of the object scores. However, this option allows the Homals algorithm to focus on the cluster differences, thus producing trivial solutions.

**Remark 2.4.** Analogously, to Homals we can extend the Princals model to the multilevel framework. We require the category quantifications of each cluster to satisfy a rank-one restriction of the form $Y_{jk} = q_{jk} \beta_{jk}', \ j \in \mathbf{J}, \ k \in \mathbf{K}$. This means that $y_{jks}(t) = \beta_{jks} q_{jk}(t), \ t = 1, \ell_j$, which

implies that if we plot $y_{jks}(t)$ against $q_{jk}(t)$ for different values of $s \in \{1, \ldots, p\}$ we see parallel straight lines. Details about the multilevel Princals model can be found in [32].

2.1. **NELS:88 Example.** For this example we selected 12 schools with 35 or more students in each one, resulting in a total sample size of 498 students and an average school size of $41.5$ students. The reason for selecting these particular schools was, that due to their relatively large size, it was expected that each category of every variable would contain some responses. The school we examined in the first Section corresponds to school #1. Some background characteristics of the schools are presented in the following Table.

| School # | # of Students | Type | Region |
|---|---|---|---|
| 1 | 37 | Public Urban | West |
| 2 | 44 | Public Urban | West |
| 3 | 36 | Public Urban | West |
| 4 | 40 | Public Suburban | West |
| 5 | 38 | Public Suburban | West |
| 6 | 35 | Public Suburban | West |
| 7 | 36 | Public Rural | West |
| 8 | 38 | Public Rural | North Central |
| 9 | 38 | Public Rural | North Central |
| 10 | 56 | Private Urban | North Central |
| 11 | 54 | Private Suburban | South |
| 12 | 46 | Private Urban | South |

TABLE 2.1. Background Characteristics of the 12 Schools

Clearly, this sample of 12 schools is not a representative sample of the school population, since a large number of rural schools is present and no schools from the Northeast are included in the sample. The latter fact indicates that the sample at hand is not suitable for drawing inferences for the country's school and student populations. However, this sample is suitable for addressing the following question. Suppose that a student (parent, teacher) is interested in attending (sending to, working) in one of these 12 schools. Knowledge regarding the basic structure of these variables and an overall idea of the school climate is essential to the student (parent, teacher) for those 12 schools. Information about other schools is marginally interesting to them. The techniques previously developed are used primarily for descriptive and not for inferential purposes [13, 33] in this example. As observed in [6], statistical inference in the classical sense takes a back seat in this approach; however, we return to a form of inference suitable for this framework called "replication stability" (see Section 12.3 in [15]) in the next Section.

A two-dimensional Homals analysis was performed on the school data set. The fit of the third dimension was a rather poor one (total eigenvalue .18). The fit of the solution (eigenvalues) for each school separately and for the overall sample is given in Table 2.2. The overall fit can be

| School # | Dimension 1 | Dimension 2 |
|---|---|---|
| 1 | .642 | .352 |
| 2 | .724 | .429 |
| 3 | .643 | .422 |
| 4 | .668 | .438 |
| 5 | .559 | .335 |
| 6 | .620 | .320 |
| 7 | .711 | .506 |
| 8 | .479 | .381 |
| 9 | .618 | .410 |
| 10 | .636 | .505 |
| 11 | .575 | .373 |
| 12 | .442 | .337 |
| Overall | .608 | .403 |

TABLE 2.2. School Eigenvalues

characterized as satisfactory. Some schools exhibit a very good fit in both dimensions (e.g. schools 2, 7, 10), while some others a rather poor one in both dimensions (e.g. schools 8, 12). Some schools have a good fit in the first dimension and a satisfactory one in the second (e.g. schools 1, 6, 11). Overall the schools present enough variation in terms of fit. This can also be seen by examining the school and total discrimination measures for each variable that are shown in Figure 2.1 (the school discrimination measures are connected by a dashed line in order to emphasize their variation). It is worth noting that the discrimination measures of the schools exhibiting a good fit (2, 7, 10) are in general larger than the total measures for all the variables, while those with a poor fit (8, 12) have discrimination measures smaller than the total ones for all the variables. This is consistent with the definition of the eigenvalues (both cluster and total) and the fact that there are no large differences between the clusters in terms of sample sizes. The remaining schools have discrimination measures larger than the total ones for some of the variables, and smaller than the total measures for the rest of the variables. Finally, some schools (e.g. 8, 9, and to a certain extent 11) have smaller discrimination measures than the total for the majority of the variables; however, for a couple of variables the cluster measures are much larger than the total ones, thus indicating the possible presence of outliers. Figure 2.2 displays the total discrimination measures of the ten variables. All variables discriminate (the category points are further apart) equally well in both dimensions. Hence, it is difficult to associate a particular dimension with a certain subset of the variables. However, variables C (students cutting class), E (robbery or theft), F (vandalism of school property), G and H (student use of alcohol and illegal drugs) discriminate best among students in both dimensions. The optimal category transformations for the variables are given in Figure 2.3. The optimal transformations of all variables in the first dimension are monotone increasing functions of the original categories, while they are quadratic functions of the original categories in the second dimension. The almost linearity of the optimal transformations in the first dimension suggests that the variables are originally measured on an interval scale (i.e. Likert scale); hence, the Homals solution can be interpreted as a test of that assumption. The second

dimension contrasts the two mid-categories (’minor’ and ’moderate’) with the two extreme ones (’not a problem’ and ’serious’). Going from the lower left corner upwards to the middle, and from there to the right lower corner, one finds categories ordered from ’not a problem’ towards ’serious problem.’ It is interesting to observe that the two mid-categories receive the same quantifications for a number of variables (E, F, G and H).

Figure 2.4 displays the category quantifications of the variables for each school. The points in the graph represent the centers of gravity of the object points (students) associated with each category. Several different patterns can be observed between the variable categories. For example for some schools (1, 4, 6, 10, 11 and 12) the following pattern emerges. In the lower left quadrant of the graph we find the ’serious problem’ categories for cutting class, physical conflicts, robbery and vandalism, use of alcohol and drugs, possession of weapons and physical and verbal abuse of teachers. However, the ’serious problem’ category for student tardiness and absenteeism (variables A and B) was located at different places in different schools. Thus, students in this area of the map are associated with these categories, which implies that they consider their school to be seriously affected by these problems. In the upper half of the graph, we find the ’minor/moderate problem’ categories for almost all the variables. Students associated with these categories believe that these problem areas are present only to a certain degree in their schools. Finally, in the lower right quadrant of the graph we find the ’not a problem’ categories for all the variables; hence, students in that area of the graph think that there are no problem areas in their schools. It is interesting to observe that the ’clustering’ of the students is done according to the same category levels. Thus, students consider all the areas representing either a serious, or a minor/moderate or not a problem in their school. In principle, in this set of schools we do not have students that indicate some areas as being a serious problem and some other areas as not a problem. Hence, to a large extent the analysis cleanly separates the students that think there exist serious problems in their schools, from the ones that think their schools are problem free (as far as the areas identified in the data set are concerned). Moreover, the analysis reveals distinctly nonlinear student response patterns; that is, variable categories are not linear with the dimensions of the space. For some other schools (7, 9) the solution separates students that indicated that all the areas examined represent a ’serious’ problem in their schools, from the rest of the students that indicate ’moderate/minor’ to ’not’ a problem. It is worth noting that the presence of outliers (students with a unique response profile different than that of other students) in school 9 distorts the picture and might affect the interpretation. For some schools (2, 3, 5, 8) the students that said ’not’ a problem are separated from the rest of the students. In this set of schools, unlike the first two, we observe mixed response patterns. There are students that consider some of the areas being a ’serious’ problem in their schools, while some other only a ’moderate’ and in a few cases a ’minor’ problem. Some other interesting points arising from examining the category quantifications plots are (i) the fact that use of alcohol is a ’serious’ problem in the rural schools 8 and 9 (but not in 7) (see the position of G1 in the respective graphs), and (ii) the fact that student tardiness and to a certain degree absenteeism are ’serious’ problems in the private schools (position of A1 and B1, especially in school 12). In general, these 12 schools exhibit a wide range of student response patterns. By closely examining the optimal category quantification plots we have identified three ’main’ groups of schools: those where the majority of the students believe there are problems, those where most of the students believe there are no

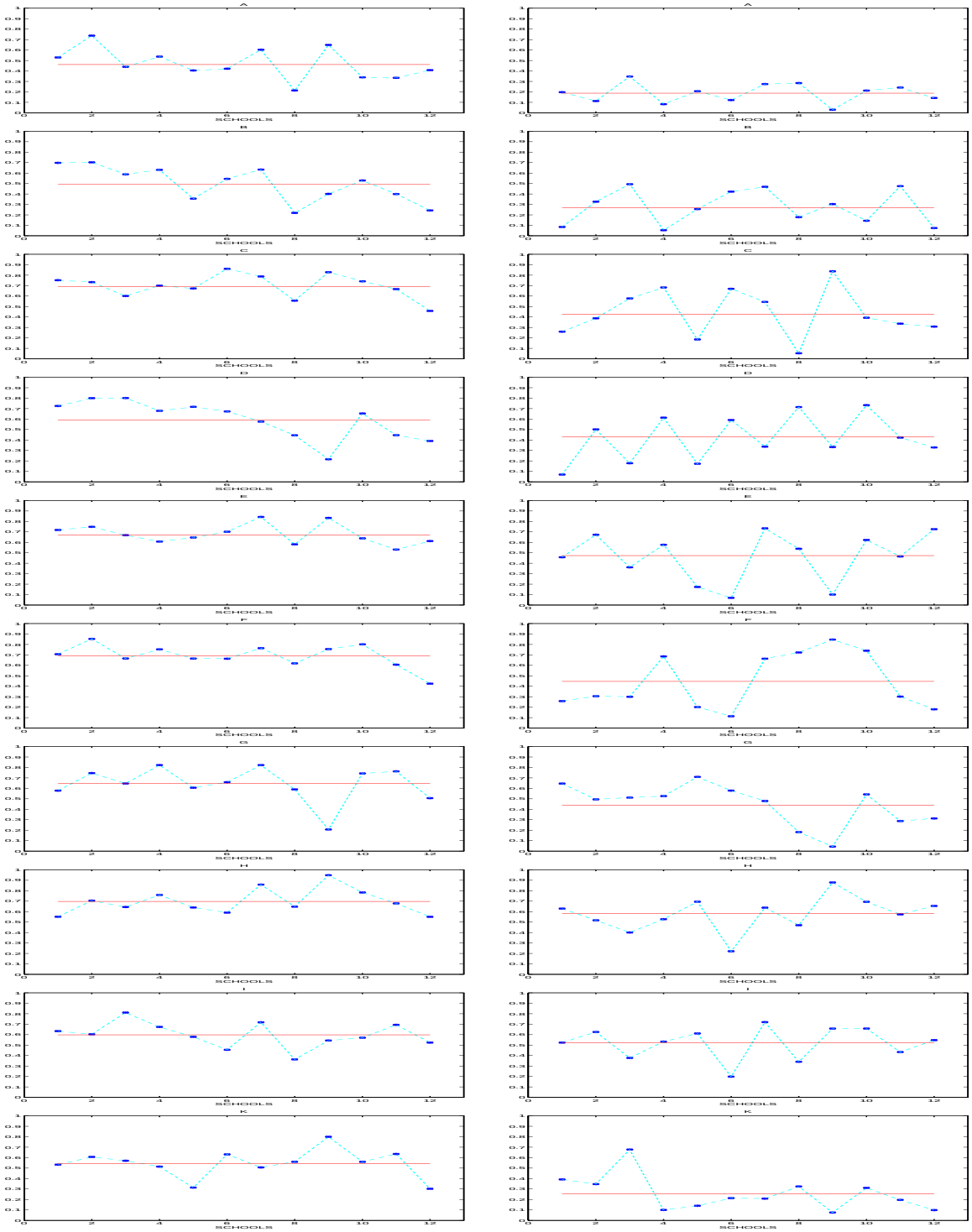

FIGURE 2.1. School (*) and Total (lines) Discrimination Measures of the Variables; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12; the solid line represents the variable's overall discrimination measure (Left: dimension 1, Right: dimension 2).

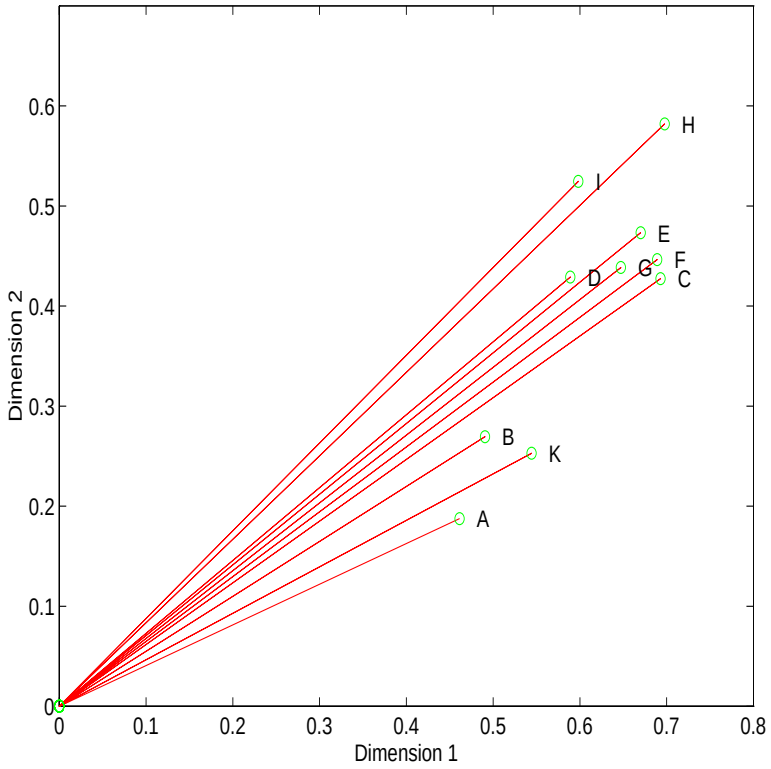


FIGURE 2.2. Total Discrimination Measures

problems, and those where the students are equally distributed among 'serious', 'moderate/minor' and 'not a problem' subgroups. However, even within these three groups there exists variation in the response patterns. This can be more clearly seen from the plots of object scores shown in Figure 2.5 (all graphs have the same scale). The distance between two student points is related to the homogeneity of their profiles, or more generally, their response patterns (see also Section 1). These plots reveal the presence of outliers in the group of rural schools (7, 8 and 9). They also show differences between schools within the same group of response patterns identified after examination of the category quantifications. For example, although schools 1, 4, 10, and 12 have similar quantification profiles, their object scores exhibit differences; those of schools 1 and 12 are evenly distributed in the space, while those of schools 1 and 10 tend to cluster into two groups: the 'serious problem' and the rest. Similar variations can be observed within the other two groups of schools.

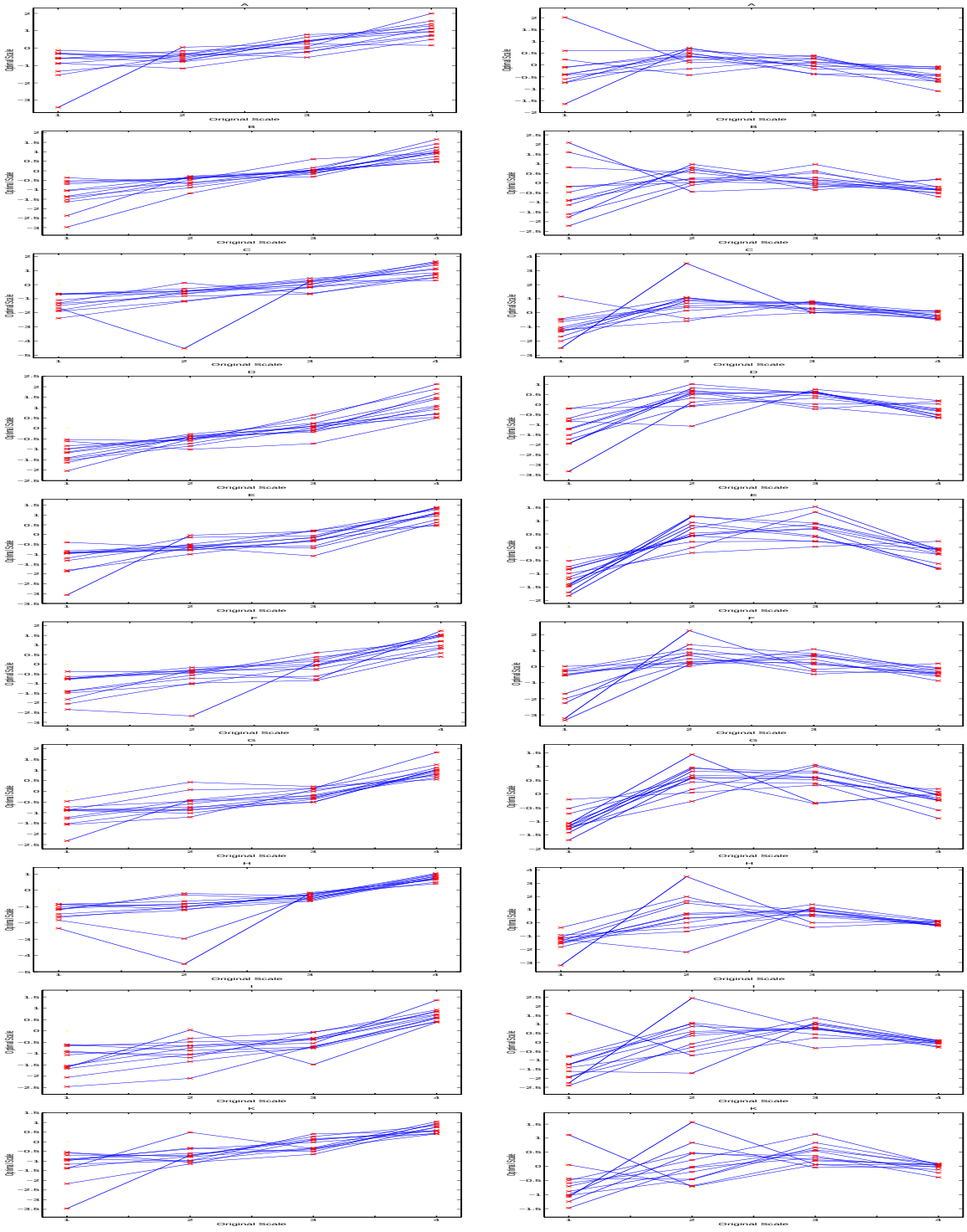


FIGURE 2.3. Optimal Transformations of the Variables (Left: dimension 1, Right: dimension 2)

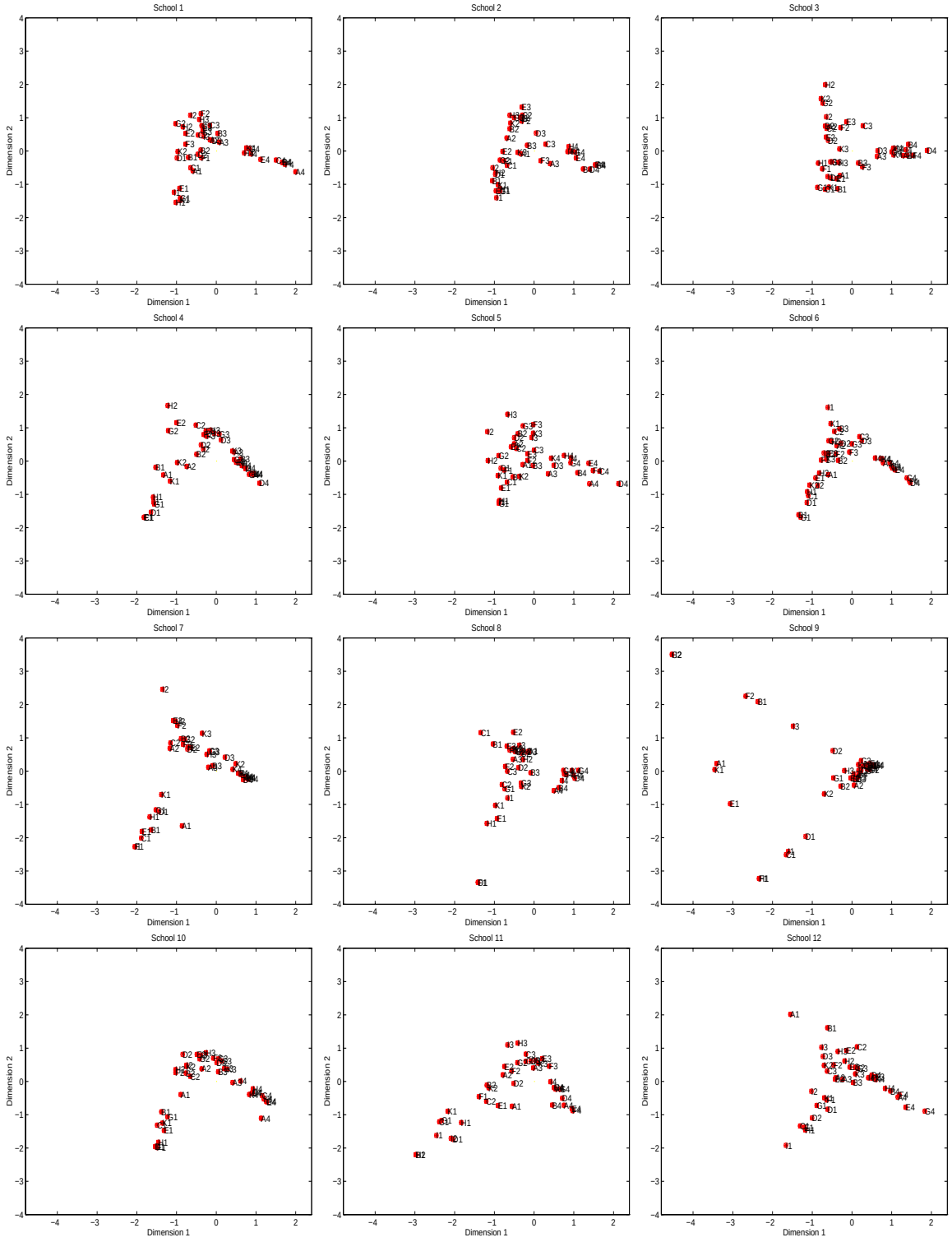


FIGURE 2.4. Optimal Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

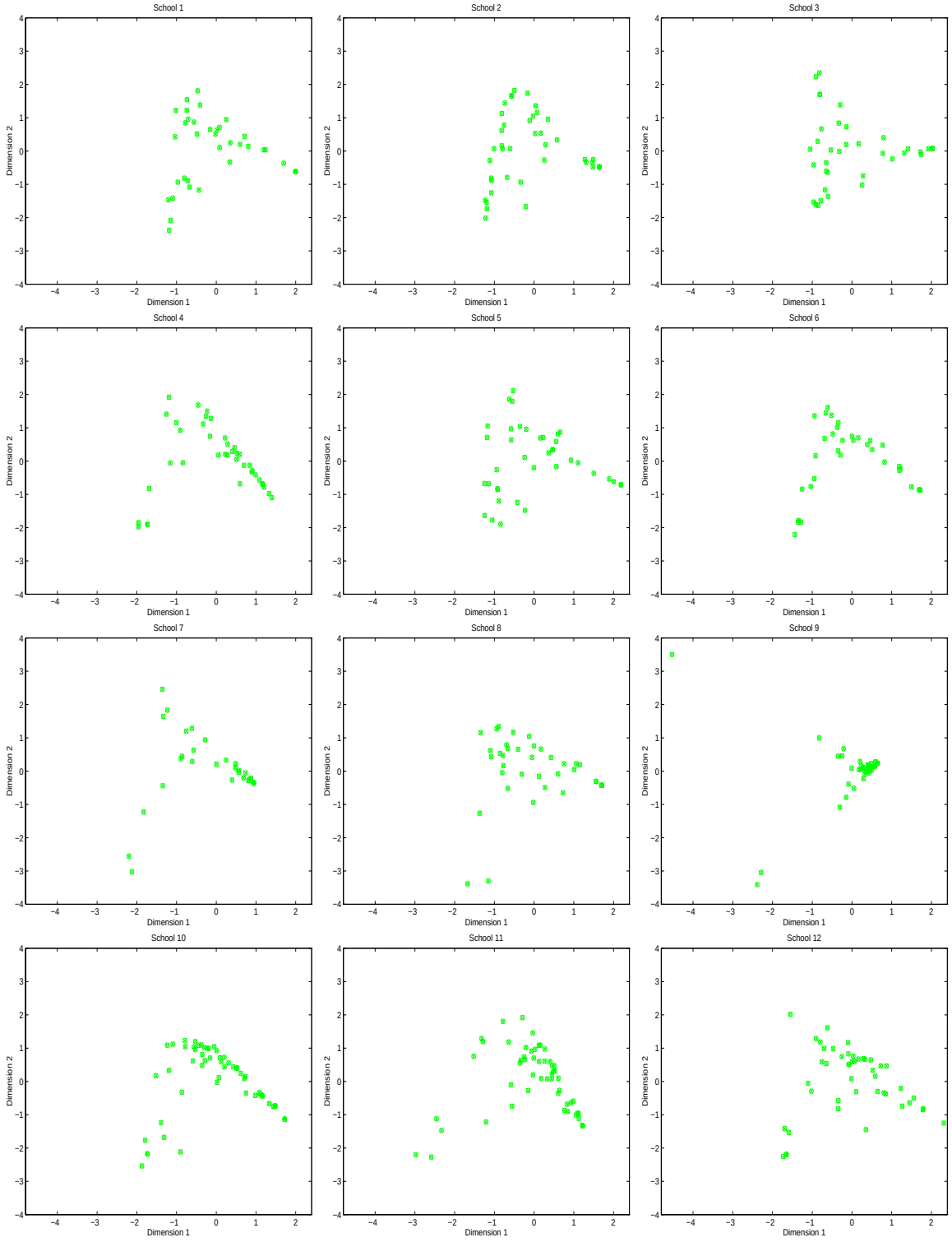


FIGURE 2.5. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

## 3. **Equality Restrictions on the Category Quantifications**

The multilevel modeling framework introduced in Section 2, allowed us to examine the student response patterns in 12 schools, to identify common features among the schools and to see their differences. However, the above analysis suffers from the following shortcomings: (i) the number of parameters to be estimated is large (e.g. $40 = 4 \times 10$ category quantifications per school) (ii) for some clusters (e.g. schools 8, 9) the solution is unstable (a direct consequence of (i)), and (iii) the analysis ignores the multilevel structure in the data. Moreover, in case we were interested in examining a large number of schools, say over 50, the previous exercise becomes prohibitive, since looking at the category quantifications and object scores for each school separately is not a particularly informative or useful task.

In this Section we examine imposing equality restrictions on the category quantifications between clusters. Such restrictions reduce the number of parameters that need to be estimated, thus improving the stability of the solution. Moreover, they allow the data analyst to incorporate prior knowledge into the analysis. This model is midway between the totally restricted model of Section 1 (single cluster case) and the totally unrestricted one of Section 2.

Let $\Gamma^j_{\mathbf{K}}$ denote a partition of the clusters $k \in \mathbf{K}$ for variable $j$; that is, $\Gamma^j_{\mathbf{K}} = \{\mathcal{K} \subseteq \mathbf{K} : \cup_{\mathcal{K}\in\mathbf{K}}\mathcal{K} = \mathbf{K},\ \mathcal{K} \cap \mathcal{K}' = \emptyset,\ \forall\ \mathcal{K}, \mathcal{K}' \subseteq \mathbf{K}\}$. We then require $\tilde{Y}_{jk} = u\alpha'_{jk} + Z_{j\mathcal{K}},\ j \in \mathcal{J},\ k \in \mathcal{K},\ \mathcal{K} \in \Gamma^j_{\mathbf{K}}$, where $Z^{\mathcal{K}}_j$ is the $\ell_j \times p$ matrix of *restricted* category quantifications for cluster $k \in \mathcal{K}$, and $\alpha_{jk}$ a $p$ column vector of intercepts. The parameters $\alpha_{jk}(s),\ s = 1, \ldots, p$ are used to ensure that the category quantifications $\tilde{Y}_{jk}(t, s)$ have a weighted sum over $t$ equal to zero for all combinations of $(j,\ k,\ s)$. This is a useful restriction in cases where we examine the same set of variables in different contexts or at different time points [11].

We begin by introducing the *constraint* matrix $C_j,\ j \in \mathbf{J}$ that maps $\mathbf{K} \to \Gamma^j_{\mathbf{K}}$. It has entries $C_j(k, r) = 1,\ k = 1, \ldots, K,\ r = 1, \ldots, R_j$ ($R_j$ denoting the cardinality of the set $\Gamma^j_{\mathbf{K}}$) if cluster $k \in \mathbf{K}$ belongs to the collection of clusters $\mathcal{K} \in \Gamma^j_{\mathbf{K}}$ and $0$ otherwise. Some examples of constraint matrices are given next:

$$C_j = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad C_j = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad C_j = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In the first example the four clusters are collapsed to a single 'super-cluster', which implies that the category quantifications of variable $j$ should be equal for all four clusters. In the second example the first two clusters would correspond to the first super-cluster and the last two to the second one; hence, we require equality of the category quantifications of variable $j$ for the first two clusters and also for the last two ones. Finally, in the last example equality of the category quantifications is imposed only on the first two clusters, while the last two are left unrestricted. It is worth noting

the similarity between the $C_j$ matrices and the $G_j$ matrices. Let $H_j = C_j \bigotimes I_{l_j},\ j \in \mathbf{J}$, that is

$$H_j = \begin{pmatrix} C_j(1,1)I_{l_j} & \dots & C_j(1,R_j)I_{l_j} \\ C_j(2,1)I_{l_j} & \dots & C_j(2,R_j)I_{l_j} \\ \dots & \dots & \dots \\ C_j(K,1)I_{l_j} & \dots & C_j(K,R_j)I_{l_j} \end{pmatrix}$$

We can then write the average squared edge function as

$$\sigma\big(X; Y_1, \dots, Y_J\big) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\ \big(X - G_j H_j Z_j\big), \tag{3.1}$$

where $Z_j = [Z_{j1}', \dots, Z_{jR_j}']'$. We employ the ALS algorithm to minimize (3.1) with respect to $Z_j$ and $X$. Fixing first $X$ we get

$$\hat{Z}_j = \big(H_j' D_j H_j\big)^{-1} H_j' G_j' X,\ j \in \mathbf{J}. \tag{3.2}$$

Some algebra shows that $Z_{j\mathcal{K}} = (\sum_{k\in\mathcal{K}} D_{jk})^{-1} \sum_{k\in\mathcal{K}} G_{jk}' X_k,\ \mathcal{K} \in \Gamma_{\mathbf{K}}^j$. Minimizing (3.1) with respect to $X$ we get

$$\hat{X} = J^{-1} \sum_{j=1}^{J} G_j H_j Z_j. \tag{3.3}$$

Note than in case $D_j^{-1}$ exists, we can also write (3.2) as

$$\hat{Z}_j = \big(H_j' D_j H_j\big)^{-1} H_j' D_j D_j^{-1} G_j' X = \big(H_j' D_j H_j\big)^{-1} H_j' D_j Y_j,\ j \in \mathbf{J}. \tag{3.4}$$

Therefore, the restricted category quantifications can be expressed as a weighted combination of the unrestricted category quantifications, with the weights given by $(H_j' D_j H_j)^{-1} H_j' D_j$; or to put it differently, the element $Y_{jk}(t,s),\ k \in \mathcal{K}$ participates with a weight $D_{jk}(t,t)/(\sum_{k\in\mathcal{K}} D_{jk}(t,t))$ in the calculation of element $Z_{j\mathcal{K}}(t,s)$. Finally, we also have that $Y_j^\star = H_j Z_j$, where $Y_{jk}^\star = Z_j^{\mathcal{K}}$ for every $k \in \mathcal{K}$, thus making (3.3) equivalent to (2.4). In the presence of equality constraints on the category quantifications the ALS algorithm becomes: (i) estimate the restricted category quantifications using (3.2), (ii) calculate $Y_j^\star = H_j \hat{Z}_j$, (iii) estimate the object scores using (3.3), and (iv) orthonormalize the $X_k,\ k \in \mathbf{K}$ matrices.

Notice that the restricted category quantifications have a weighted sum over categories equal to zero for the collection of clusters $\mathcal{K}$ and not for the individual clusters $k$; that is, $u'(\sum_{k\in\mathcal{K}} D_{jk}) Z_{j\mathcal{K}} = 0$. However, we want the category quantifications centered for every cluster $k \in \mathbf{K}$, in order to ease the presentation and interpretation of the category quantification plot. Using the intercept parameters, we set $\tilde{Y}_{jk} = u\hat{\alpha}_{jk}' + Y_{jk}^\star$, where

$$\hat{\alpha}_{jk}' = -\big(u' D_{jk} Y_{jk}^\star\big)/n_k,\ j \in \mathbf{J},\ k \in \mathbf{K}. \tag{3.5}$$

Obviously, for the variables that we do not impose restrictions, we have $\hat{\alpha}_{jk} = 0$ for all $k \in \mathbf{K}$. Thus, once the ALS algorithm has converged, we center the category quantifications and calculate the object scores using $X = J^{-1} \sum_{j=1}^{J} \sum_{k}^{K} G_{jk} \tilde{Y}_{jk}$, so that the category quantification points are the centroid of objects belonging to that category.

At this point, it is rather hard for us to think of practical situations where one might want to impose a different set of equality restrictions between clusters for each variable. However, the framework is developed and it does not introduce major computational difficulties, thanks to the sparseness of the constraint matrices.

We examine next what happens to the fit of the solution when we impose equality restrictions on the category quantification between clusters. We will assume for ease of presentation that $D_j^{-1}$ exists, and thus use the relationship given in (3.4). Some algebra shows that the total discrimination measures of the unrestricted solution can also be written as

$$\eta_{js}^2 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s),\ j\in\mathbf{J},\ s=1,\ldots,p, \tag{3.6}$$

while the cluster discrimination measures as

$$\eta_{jks}^2 = \frac{1}{n_k}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s),\ j\in\mathbf{J},\ k\in\mathbf{K},\ s=1,\ldots,p. \tag{3.7}$$

The total discrimination measure of the restricted solution for variable $j\in\mathbf{J}$ is given by

$$\begin{aligned}\tilde{\eta}_{js}^2 &= \frac{1}{N}\mathrm{tr}\big(\tilde{Y}_j' D_j \tilde{Y}_j\big) = \frac{1}{N}\sum_{k=1}^{K}\mathrm{tr}\big(\tilde{Y}_{jk}' D_{jk}\tilde{Y}_{jk}\big) = \frac{1}{N}\sum_{k=1}^{K}\mathrm{tr}\big(\tilde{Y}_{jk}' D_{jk}(u\hat{\alpha}_{jk}' + Y_{jk}^{\star})\big) = \\ &\frac{1}{N}\sum_{k=1}^{K}\mathrm{tr}\big(\tilde{Y}_{jk}' D_{jk}u\hat{\alpha}_{jk}' + \tilde{Y}_{jk}' D_{jk}Y_{jk}^{\star}\big) = \frac{1}{N}\sum_{k=1}^{K}\mathrm{tr}\big(Y_{jk}^{\star'} D_{jk}Y_{jk}^{\star} + \hat{\alpha}_{jk}u' D_{jk}Y_{jk}^{\star}\big) = \\ &\frac{1}{N}\sum_{k=1}^{K}\mathrm{tr}\big(Y_{jk}^{\star'} D_{jk}Y_{jk}^{\star} - n_k\hat{\alpha}_{jk}\hat{\alpha}_{jk}'\big) = \frac{1}{N}\mathrm{tr}\big(Y_j^{\star'} D_j Y_j^{\star}\big) - \frac{1}{N}\sum_{k=1}^{K} n_k\mathrm{tr}\big(\hat{\alpha}_{jk}\hat{\alpha}_{jk}'\big).\end{aligned} \tag{3.8}$$

But we also have that (using (3.4))

$$\begin{aligned}\frac{1}{N}\mathrm{tr}\big(Y_j^{\star'} D_j Y_j^{\star}\big) &= \frac{1}{N}\mathrm{tr}\big(Z_j' H_j' D_j H_j Z_j\big) = \frac{1}{N}\mathrm{tr}\big(Y_j' D_j H_j \big(H_j' D_j H_j\big)^{-1} H_j' D_j Y_j\big) = \\ &\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)\Big(\sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}(i,s)\Big)^2.\end{aligned} \tag{3.9}$$

An application of Cauchy's inequality gives that

$$\Big(\sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}(i,s)\Big)^2 \le \sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}^2(i,s). \tag{3.10}$$

Combining relations (3.8), (3.9) and (3.10) and after some algebra we get that

$$\tilde{\eta}_{js}^2 \le \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s) = \eta_{js}^2. \tag{3.11}$$

Hence, we also get

$$\tilde{\gamma}_s = \frac{1}{J}\sum_{j=1}^{J}\tilde{\eta}_{js}^2 \leq \frac{1}{J}\sum_{j=1}^{J}\eta_{js}^2 = \gamma_s. \tag{3.12}$$

Therefore, restrictions have a negative effect both on the overall fit of the solution and on the variables' discrimination measures. The magnitude of this effect depends primarily on the distribution of the objects over the categories (see (3.10)). However, nothing can be said on the effect of the restrictions on the cluster discrimination measures and fit.

**Remark 3.1.** *Analysis of Panel Data.* Van der Heijden and de Leeuw [40] used correspondence analysis techniques to examine panel data. If in our approach we take $n_k = n$ for every $k \in \mathbf{K}$, $\Gamma_{\mathbf{K}}^j \equiv \{1\}$ for every $j \in \mathbf{J}$, then the above analysis corresponds to the analysis of their LONG indicator matrix. This type of analysis provides only a single set of category quantifications for the objects, but $K$ different sets of object scores, one for each time point. A possible drawback of such an analysis as pointed out in [40] is that the restricted category quantifications might be distinguishing the different time points rather than the different objects. This will happen, if the distributions of the categories of each variable differ considerably over time points. In our approach, by allowing to impose equality restrictions only on a subset of the variables, we might be able to avoid this rather uninteresting solution.

**Remark 3.2.** *Equality Restrictions and Clustering.* We briefly examine the relationship between imposing equality restrictions on all the variables of all the clusters, and treating the data set as a single cluster. In the first case the category quantifications are given by (3.2), while in the second by (1.6). Notice that in general we have $G_j^{'\mathcal{K}} X^{\mathcal{K}} \neq \sum_{k\in\mathcal{K}} G_{jk}' X_k$, where $X^{\mathcal{K}}$ and $G_j^{'\mathcal{K}}$ are the single cluster object scores and indicator matrix, respectively. The latter implies that we get different results. When we impose equality restrictions we attempt to gain strength by pulling information from all the clusters involved, while preserving the 'local' (within cluster) scaling of the object scores. On the other hand, combining all the clusters to a single cluster introduces a different type of 'local' scaling for the object scores.

**Remark 3.3.** *3-Stage Modeling.* The present setup allows us to also look at collections of clusters, thus introducing a second stage of clustering. For example, we can examine students grouped in classes, which are naturally grouped in schools. However, in the present approach we are only interested in incorporating some prior information to the analysis (by using equality restrictions at a subset of variables, say at the school level or at the school district level), and not modeling explicitly the second stage of clustering. The latter would require an extension of our sampling framework and certain modifications to the structure of the design matrices $D_j,\ j \in \mathbf{J}$. In the present framework the focus remains on the clusters (e.g. schools, classes), but by incorporating some prior information we can improve both on the stability of the solution and the conclusions derived from the analysis.

**Remark 3.4.** *Equality Constraints in the Princals Model* In order to incorporate equality constraints between clusters in the Princals model we impose a rank-one type of constraint on the constrained category quantification matrices $Z_{j\mathcal{K}}$. Thus, we require

$$Z_{j\mathcal{K}} = v_{j\mathcal{K}}\theta_{j\mathcal{K}}',\ j \in \mathbf{J},\ \mathcal{K} \in \Gamma_{\mathbf{K}}^j, \tag{3.13}$$

where $v_{j\mathcal{K}}$ is a $\ell_j$ column vector of single constrained category quantifications and $\theta_{\mathcal{K}}$ a $p$ column vector of component loadings (for details see [32]).

3.1. **NELS:88 Example (continued).** We continue with the example discussed in Section 2. The unrestricted Homals solution revealed some common patterns among the schools, but also quite a few differences; however, the presence of low frequency student profiles (outliers) compromised the stability of the solution and resulted in distorting the pictures. In order to overcome these shortcomings, we are going to impose constraints on the category quantifications of some of the variables. In particular, variables A, B and C will be constrained across the following four school groups (public urban, public suburban, public rural and private), while variables H, I and K will be constrained across public and private schools. The reason for choosing these two particular clusterings of the variables is that various studies suggest that public urban schools are in many aspects (e.g. resources, socioeconomic status of the parents) different than public suburban and public rural schools and also that public schools are different than private schools [29]. Therefore, we incorporate prior knowledge in our analysis. The remaining variables were left unrestricted. Thus, the corresponding constraint matrices are given by

$$C_a = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad C_b = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

for variables $a = A, B, C$ and $b = H, I, K$ respectively. The first set of variables was selected after a close examination of the variable transformation plot (Figure 2.3) and the category quantification plot (Figure 2.4). These variables presented enough variation between the schools. Moreover, it is reasonable to assume that absenteeism, tardiness and cutting class are not school-specific problems, but are affected by the school environment (location etc), and schools adopt similar policies to eliminate such problems. On the other hand, variables H and I are mainly responsible for the presence of outliers (particularly in the rural public schools) and moreover it is assumed that possession of weapons and verbal and physical abuse of teachers will be regarded differently in public and in private schools. Thus, by imposing constraints we attempt to enhance the stability of the Homals solution and at the same time incorporate prior information.

A two-dimensional solution produced a satisfactory fit, with total eigenvalues .573 and .360 respectively (see Table 3.1). There is only some small loss in the fit (as expected from (3.12)). It

is worth noting that the rural schools experienced the largest decreases in the fit, because the unrestricted solution placed a lot of weight to the outlying observations. Some schools showed small improvements in their fit (e.g. school 12). The school discrimination measures of the constrained variables exhibit smaller variation around the total discrimination measures (see Figure 3.1). As before, most variables discriminate equally well on both dimensions (Figure 3.2), although variables A and B can now be primarily associated with the first dimension.

The transformation plots of all the variables are given in Figure 3.3. The solution produces clear monotonic transformations for the constrained variables in the first dimension. In the second dimension, a quadratic pattern seems to be emerging (thus distinguishing the two middle categories from the two extreme ones), although things are not that clear.

| School # | Dimension 1 | Dimension 2 |
|---|---|---|
| 1 | .607 | .318 |
| 2 | .683 | .365 |
| 3 | .624 | .359 |
| 4 | .584 | .410 |
| 5 | .607 | .342 |
| 6 | .590 | .278 |
| 7 | .634 | .386 |
| 8 | .496 | .376 |
| 9 | .489 | .304 |
| 10 | .588 | .472 |
| 11 | .493 | .308 |
| 12 | .509 | .352 |
| Overall | .573 | .360 |

TABLE 3.1. School Eigenvalues

The main advantage of the constrained solution can be seen in Figures 3.4 and 3.5 (both Figures use the same scale). First, observe that all the graphs are nicely centered. Second, when examining the object scores, we immediately notice that the outliers in schools 7, 8 and 9 have disappeared. Moreover, several of the other schools have cleaner pictures, with the public suburban and the private schools exhibiting a quadratic pattern along the second dimension, the public rural schools being concentrated to the right of the graph (minor and not a problem categories) and the public urban schools being primarily distributed around the serious and moderate categories. This observation is supported by the optimal category quantifications graphs. More specifically, we see that in the public urban schools the students that indicated 'no problem' form a separate cluster (especially in school 1). It is also worth noting the similar patterns exhibited in the public suburban and private schools. In most of them (with the possible exception of school 5), the 'serious problem' students are cleanly separated from the rest of the respondents. Finally, despite the pulling together of the public rural schools through the constraints, there seems to remain enough variation in their response patterns. Regarding variables $A, B$ and $G$ we have the following.

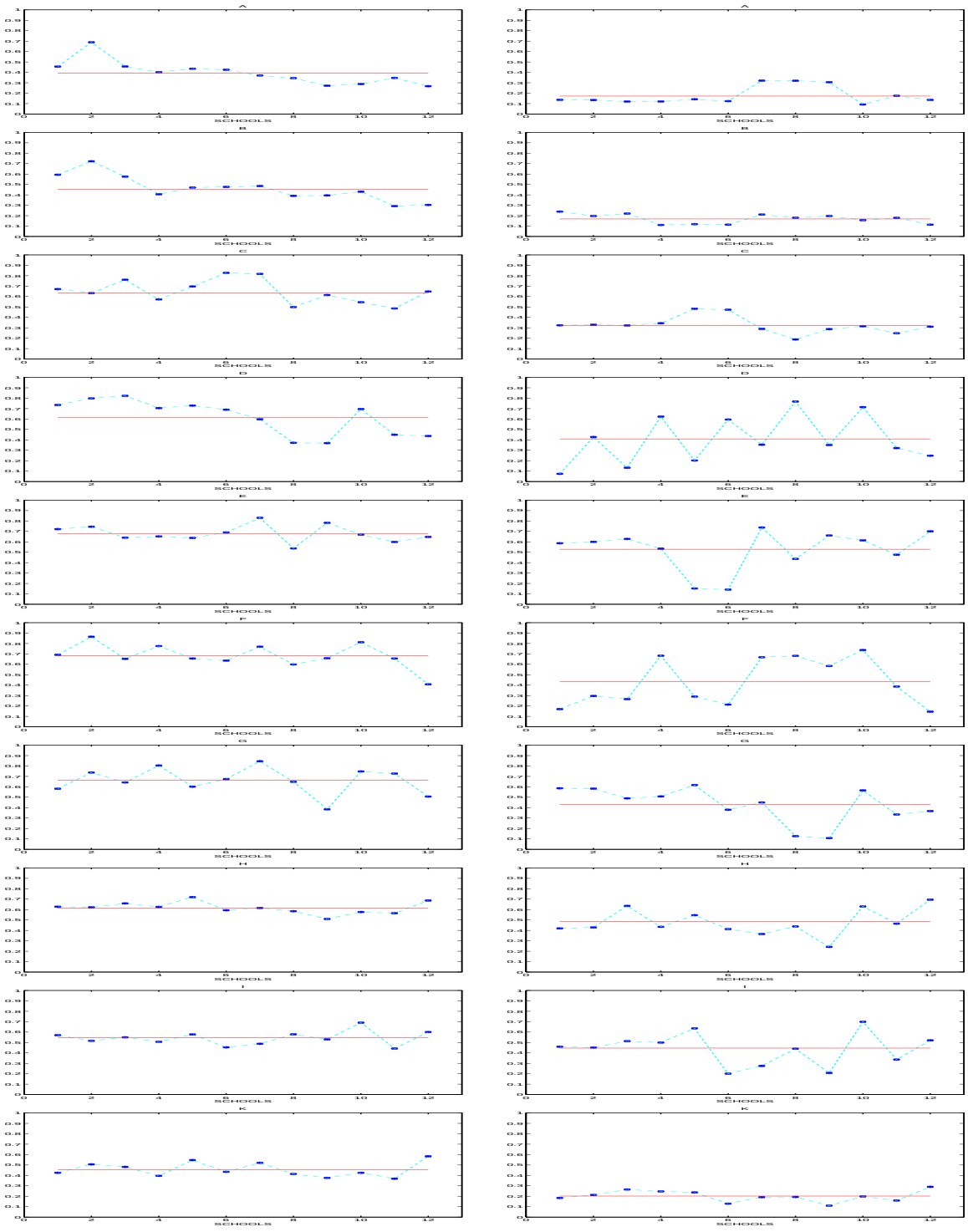

FIGURE 3.1. School (*) and Total (lines) Discrimination Measures of the Variables; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12; the solid line represents the variable's overall discrimination measure (Left: dimension 1, Right: dimension 2).

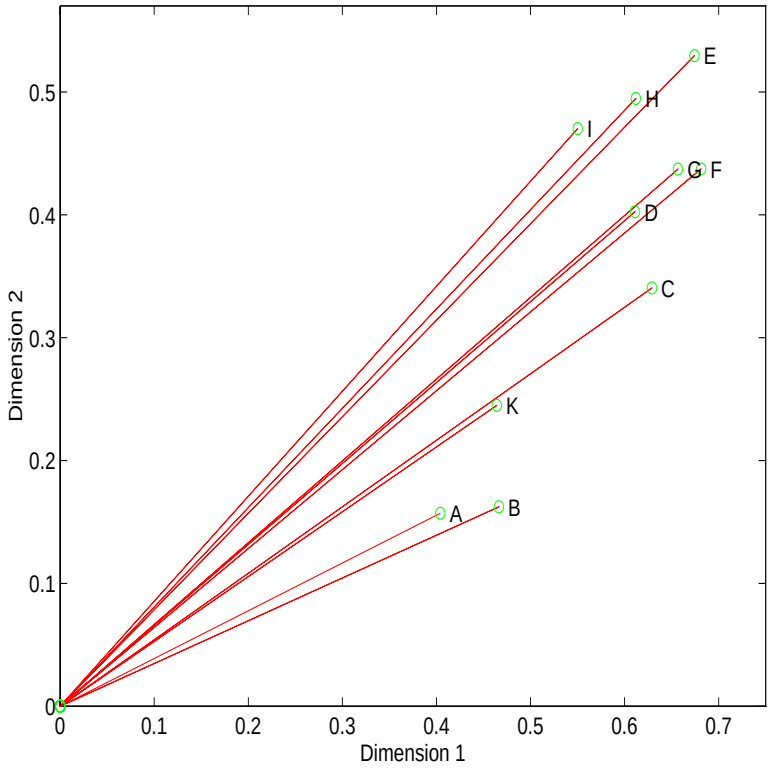


FIGURE 3.2. Total Discrimination Measures

Since variable $G$ was left unrestricted, the category quantifications graph continues to reveal that use of alcohol is a 'serious' problem in schools 8 and 9. On the other hand, student absenteeism was a 'serious' problem especially in school 12. This effect is permeated through the equality restrictions to the other two private schools.

The constraints managed to 'filter' most of the 'noise' present in the unconstrained Homals solution and strengthened the patterns that emerged there. It seems that the public urban schools in the sample can be characterized as 'rough' ones, while the public rural schools are, in principle, problem free. The public suburban and the private schools have a similar distribution of students across all categories, although most of them are leaning towards being problem free. However, student tardiness and absenteeism seem to be more noticeable problems in the private schools than in the public suburban ones. The constrained solution reaffirmed (even strengthened) our previous finding that the 'clustering' of the students is done according to the same category levels for all the variables. The implication for the student (teacher) who considers attending (working) at one of these 12 schools is, that it suffices to look at very few of these variables and classify the school. Moreover, the solution suggests that the main decision the student (parent, teacher) has to undertake is which group of schools (public urban, etc.) to attend, since the within group school differences appear to be small, or putting it differently, there exist different patterns between different clusters of schools and little variation in patterns within clusters of schools.

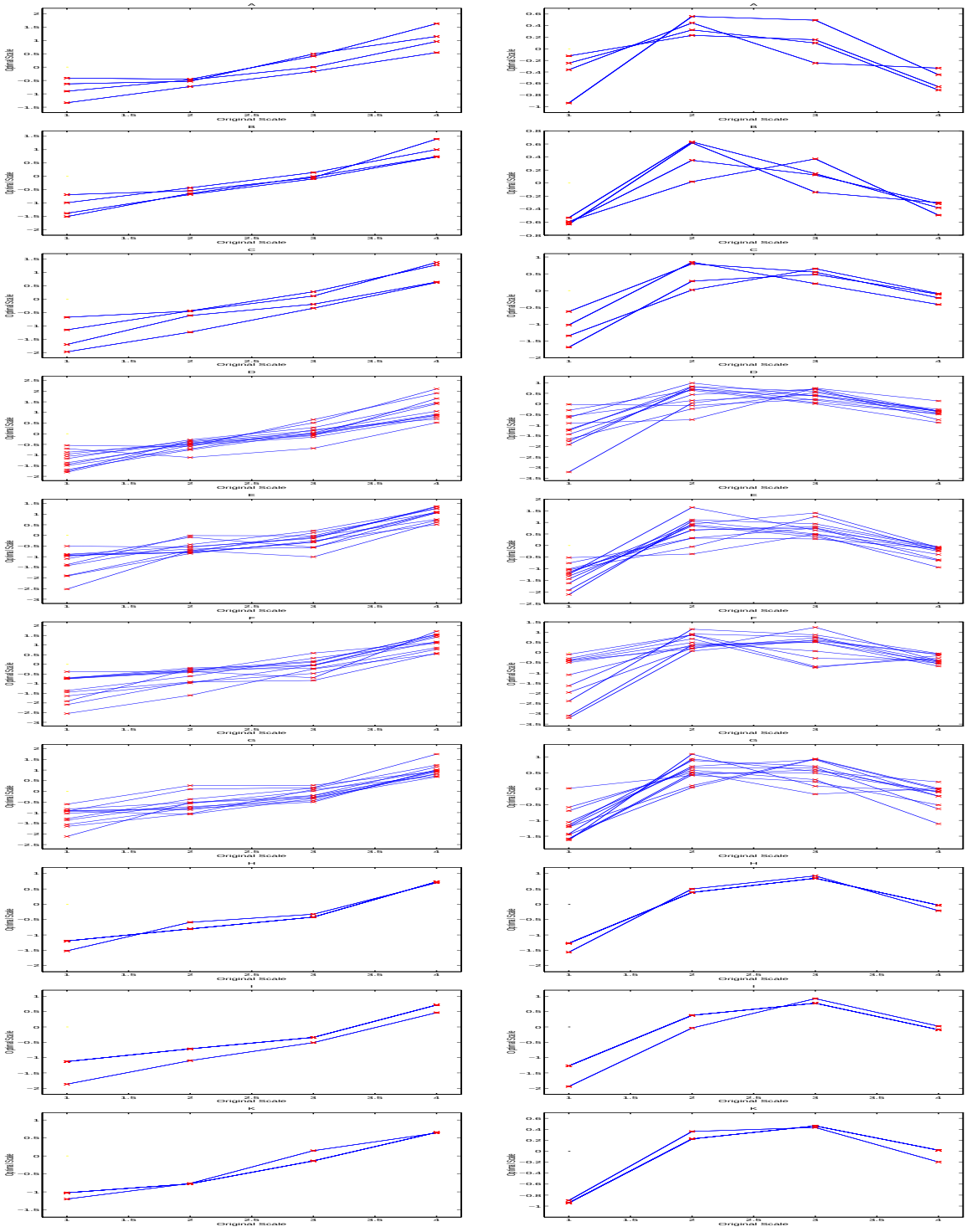


FIGURE 3.3. Optimal Transformations of the Variables (Left: dimension 1, Right: dimension 2)

The use of constraints allowed us to derive a more stable picture regarding the various patterns present in the data, by basically 'borrowing strength' from 'similar' schools; this idea permeates many other multilevel techniques such as hierarchical linear models etc. However, one might still pose the question, whether the patterns in the plots are real in the first place or merely chance effects. In order to provide an answer to this question we address briefly the question of stability of the solution.

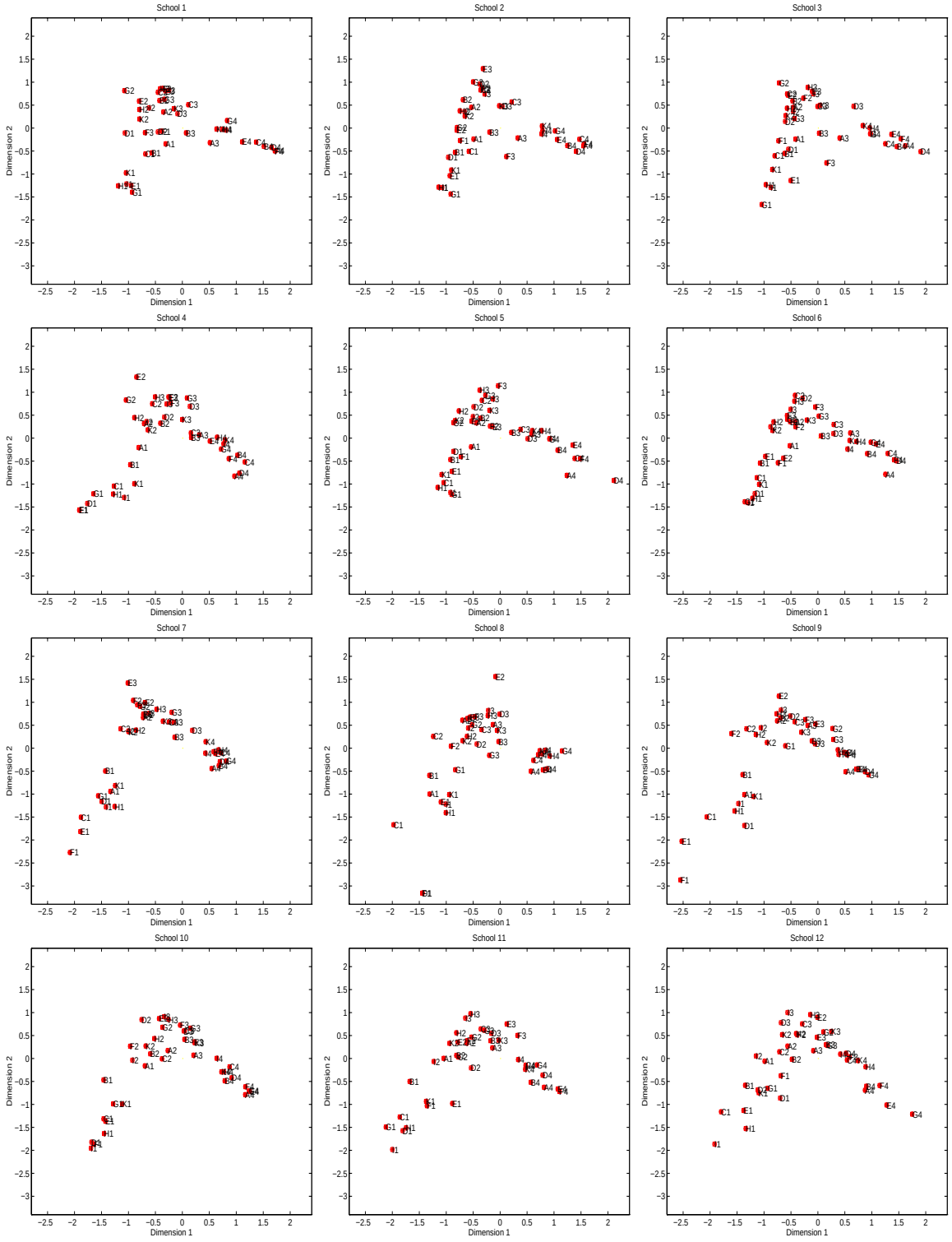


FIGURE 3.4. Optimal Constrained Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

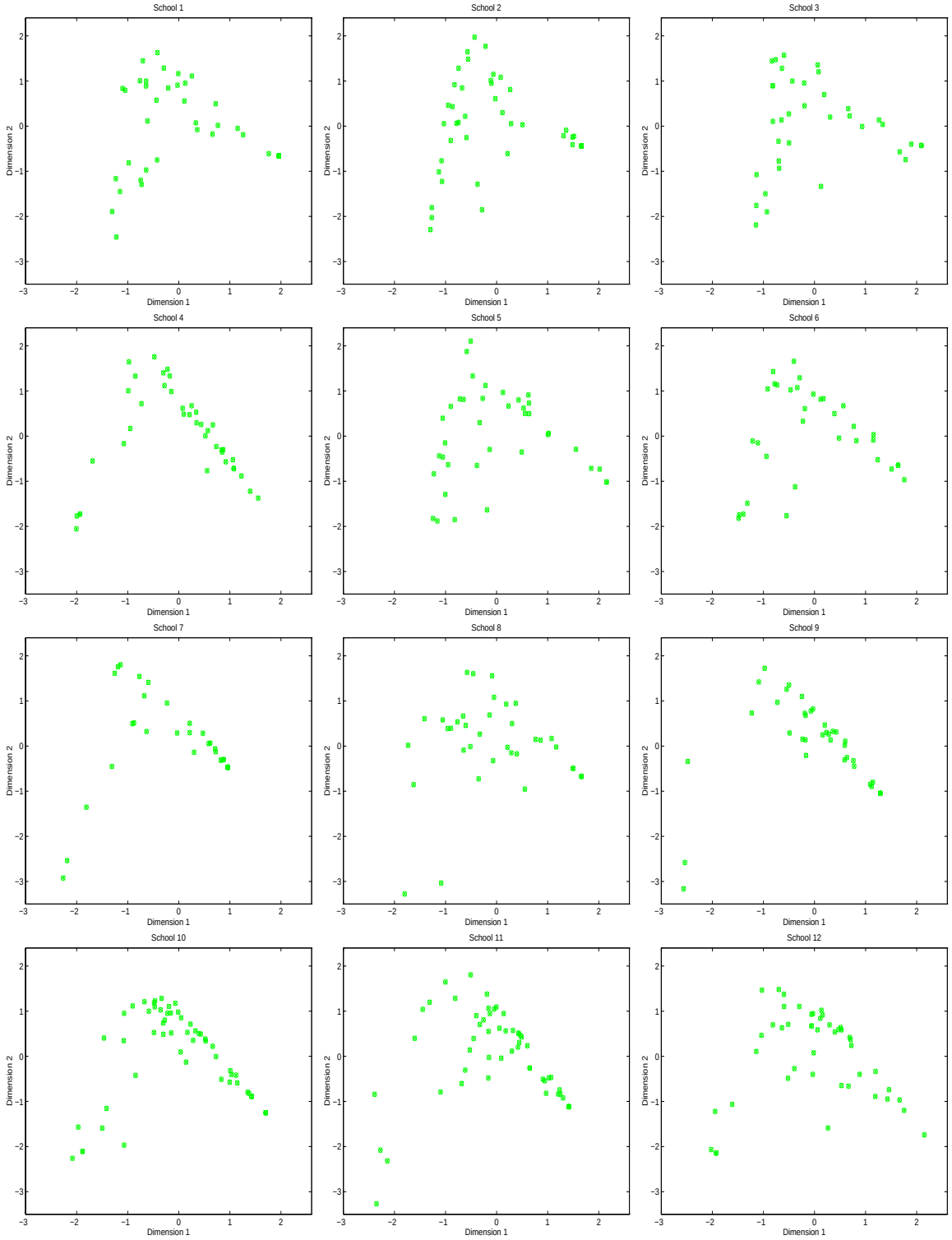


FIGURE 3.5. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

3.2. **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. As Kendall points out "many of the practical situations which confront us are not probabilistic in the ordinary sense ... It is a mistake to try and force the treatment of such data into a classical statistical mould, even though some subjective judgment in treatment and interpretation may be involved in the analysis" (see [23], p.4). Nevertheless, the question of stability of the chosen representation remains crucial; that is, are the various patterns observed in the plots real, or merely chance effects. In what follows we use the concept of stability in the following sense: data analytic results are stable when small and/or unimportant changes in the input lead to small and unimportant changes in the results (output), where as input we consider the data at hand (object and variables), the coding of the variables, the type of technique employed (unrestricted vs restricted multilevel Homals), the dimensionality of the solution, and as output category quantifications, object scores, total and cluster discrimination measures etc. At this point a distinction between *external* and *internal* stability is in order. External stability refers to the notions of statistical significance and confidence. We are interested in examining whether the results produced by a particular sample are similar (close) to the results produced by any other sample drawn from the same population. This allows the data practitioner to apply the conclusions from a particular sample to the entire population. On the other hand, internal stability deals with the specific data set at hand. An internally stable solution implies that the derived results provide a good summary of that specific data set. We are not interested in population values, because it may be difficult to determine the population from which the data set was drawn from or we may not know the sampling mechanism, or for the question we are dealing with the entire population might not be of particular interest. Internal stability can be thought of as a form of robustness [15]. The 12 schools from NELS:88 are a good example of the latter case.

The bootstrap technique allows us to examine the stability (external or internal) of the solution (for a detailed exposition see [31]). In the absence of any knowledge about the sampling mechanism that generated the data (e.g. stratified, clustered) the appropriate way to bootstrap in the present setting is to sample $n_k$ objects with replacement from each cluster, and then apply the technique under consideration (unrestricted or restricted Homals) to this new sample and repeat this exercise a large number of times. If on the other hand there exists additional information about the sampling mechanism, then more sophisticated ways of bootstrapping need to be applied in order to provide accurate results (see the discussion in chapter 6 of [30]). In Figure 3.6 some representative output of the bootstrapped unconstrained and constrained Homals solutions of our NELS:88 example is shown. Similar pictures are obtained for the remaining schools and variables. The eight plots on the left of Figure 3.6 show the original transformations (solid lines), the bias corrected (see chapter 6 in [30]) mean bootstrapped transformations (dashed dot lines), and the $\pm 2$ standard errors confidence bands of the transformations (dashed lines), for variables $C$ and $H$ for schools $3$ and $9$ in both dimensions of the unconstrained Homals solution. The eight plots on the right show the same things for the constrained Homals solution. School 3 was chosen because it exhibited fairly monotone transformations particularly in the first dimension, while the solution of school 9 was dominated by the presence of outliers, thus producing more erratic patterns in the variable transformations. Moreover, variable $C$ is an example of a variable constrained across 4

groups of clusters (public urban, public suburban, public rural and private schools), and variable $H$ a variable constrained across 2 groups of clusters (public and private schools). All results are based on 50 bootstrap replications of the solution. It is important to note that the vertical scales on the left and right plots differ greatly. A quick visual inspection of the plots of the unconstrained solution indicates that the first dimension is more stable than the second one (smaller standard errors), that the transformations of the variables of school 9 exhibit greater variability than those of school 3, and that the original transformations differ to some degree from the bias corrected mean transformations, especially along the second dimension. On the other hand, the plots of the constrained solution indicate that the imposed constraints reduce the variability substantially. This suggests that the patterns uncovered by the solution are real, and small perturbations in the data do not alter our findings. However, the first dimension continues to be more stable than the second one (smaller standard errors).

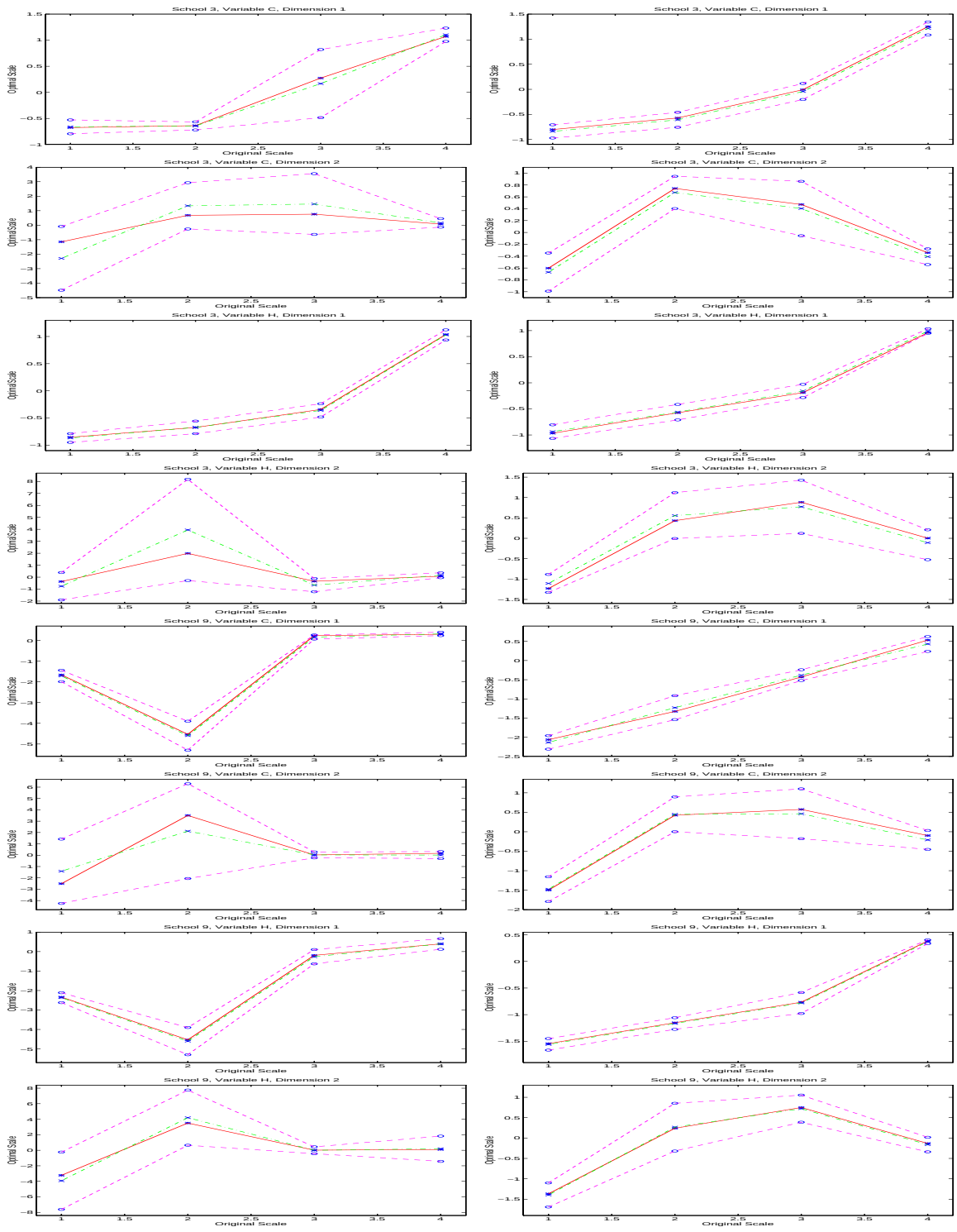


FIGURE 3.6. Stability of the Transformations; Left: Unconstrained Solution, Right: Constrained Solution (solid line: original transformation, dashed dot line: bias corrected mean bootstrapped transformation, dashed lines: confidence bands of the bootstrapped transformations)

## 4. **Fuzzy Coding Schemes for the Constraint Matrices**

The use of the constraint matrices $C_j$ (see Section 3) allowed us to cast the equality restrictions on the category quantifications in a very natural and computationally efficient mathematical framework. However, in many applications a partitioning of the clusters so that each cluster belongs to a single collection of clusters only, might not be particularly meaningful or even possible. For example, if we wanted to group the 12 schools in our example according to parents income, or socioeconomic status, or race, the "crisp" 0/1 coding would have presented problems, because schools are not 100% high income, or 100% white etc.

Fuzzy coding has been extensively used in multiple correspondence analysis [18, 41] to recode continuous data into ordered categories. We employ some ideas from fuzzy coding to enrich the framework for our constraint matrices. Instead of a 1 indicating a specific collection of clusters, with zeros elsewhere, we can assign a set of nonnegative values that add up to 1. These can even be considered probabilities that the cluster lies in the respective collection of clusters. For example, suppose we want to group the schools according to parents income, that is broken into three categories: high, middle, low. A possible constraint matrix might be

$$C_j = \begin{pmatrix} .7 & .2 & .1 \\ .3 & .4 & .3 \\ 0 & .45 & .55 \end{pmatrix}$$

It indicates that in the first school 70% of the parents belong to a high income bracket, 20% to a middle income bracket and 10% to a low one on average, while in the second school the respective percentages are 30%, 40% and 30%. Finally, in the third school there are no high income parents. This coding implies that the category quantifications of the first school for variable $j$ are given by $\tilde{Y}_{jk} = u\alpha'_{jk} + (.7Z_H + .2Z_M + .1Z_L)$, where $Z_H,\ Z_M$ and $Z_L$ are the category quantifications of the high, middle and low income groups of clusters, respectively. Hence, under the fuzzy coding scheme of the $C_j$'s the cluster category quantifications are restricted to be linear combinations of the group category quantifications.

The starting point for this general coding scheme is again the *combination* matrix $C_j,\ j \in \mathbf{J}$ that maps $\mathbf{K} \to \Gamma^j_{\mathbf{K}}$. Its entries satisfy the restriction

$$\sum_{r=1}^{R_j} C(k,r) = 1,\ k \in \mathbf{K}, \tag{4.1}$$

where $R_j$ denotes the cardinality of the set $\Gamma^j_{\mathbf{K}}$. The restriction (4.1) implies that the total mass of every cluster $k \in \mathbf{K}$ is distributed among the group of clusters defined by the columns of the combination matrix. Let $H_j = C_j \bigotimes I_{l_j},\ j \in \mathbf{J}$. Then, the squared edge length function can be

written as

$$\sigma\big(X; Z_1, \ldots, Z_J\big) = J^{-1} \sum_{j=1}^{Z} \text{SSQ}\,\big(X - G_j H_j Z_j\big), \tag{4.2}$$

where $Z_j = [Z'_{j1}, \ldots, Z'_{jR_j}]'$. We employ the ALS algorithm to minimize (4.2) with respect to $Z_j$ and $X$. Minimizing (4.2) for fixed $X$ we get

$$\hat{Z}_j = \big(H'_j D_j H_j\big)^{-1} H'_j G'_j X,\ j \in \mathbf{J}, \tag{4.3}$$

while minimizing (4.2) with respect to $X$ for fixed $Z_j$'s we get

$$\hat{X} = J^{-1} \sum_{j=1}^{J} G_j H_j Z_j. \tag{4.4}$$

When the $C_j$'s represented constraints matrices we had $H'_j D_j H_j = \sum_{k \in \mathcal{K}} D_{jk}$, a diagonal matrix, and therefore the inverse of the left-hand side always existed. The question is what happens in this case, where $C_j$ is a general matrix and not an indicator matrix. A well known result [16] gives that $(H'_j D_j H_j)^{-1}$ always exists (provided $D_j$ is of full rank), as long as $K \geq R_j$, that is, there are at least as many clusters as groups of clusters we intend to study. In our example, there were 3 clusters and 3 groups (high, middle, low income), so no problem existed. Otherwise, the use of a generalized inverse is required. The form of the $H'_j D_j H_j$ is also very interesting. We give the formula in case all entries $C_j(k, r) > 0,\ k \in \mathbf{K},\ r = 1, \ldots, R_j$

$$H'_j D_j H_j = \begin{pmatrix} \sum_{k=1}^{K} C_j^2(k,1) D_{jk} & \sum_{k=1}^{K} C_j(k,1) C_j(k,2) D_{jk} & \ldots & \sum_{k=1}^{K} C_j(k,1) C_j(k,R_j) D_{jk} \\ & \sum_{k=1}^{K} C_j^2(k,2) D_{jk} & \ldots & \sum_{k=1}^{K} C_j(k,2) C_j(k,R_j) D_{jk} \\ & & \ldots & \ldots \\ & & & \sum_{k=1}^{K} C_j^2(R_j R_j) D_{jk} \end{pmatrix}$$

So, $H'_j D_j H_j$ is a collection of diagonal matrices, and in case some entry of $C_j$ is zero then the respective block is also zero.

Finally, in case $D_j^{-1}$ exists we can express the $Z_j$'s as follows (similarly to (3.4))

$$\hat{Z}_j = \big(H'_j D_j H_j\big)^{-1} H'_j D_j D_j^{-1} G'_j X = \big(H'_j D_j H_j\big) H'_j D_j Y_j,\ j \in \mathbf{J}. \tag{4.5}$$

The latter implies that element $Y_{jk}(t, s),\ k \in \mathcal{K}$ participates with a weight $C(k, \mathcal{K}) D_{jk}(t, s) / (H'_j D_j H_j)^{-1}(t, t)$ in the calculation of element $Z_{j\mathcal{K}}(t, s)$. When dealing with combination matrices there is no simple expression for the inverse of the $H'_j D_j H_j$ matrix.

**Remark 4.1.** *On Constraint and Combination Matrices.* It can be seen that a constraint matrix is a special case of a combination matrix. The main reason for introducing them separately is to present the possibilities and the limits of each approach, and also avoid making the discussion too complicated.

It is worth noting that the current framework allows for mixing of constraint and combination matrices; an illustration of this possibility is given in the example that follows.

4.1. **NELS:88 Example (continued).** We continue with the NELS:88 example, by using both types of restrictions. Specifically, variables H, I and K continue to be constrained across public and private schools (crisp coding). For the remaining seven variables we consider constraints based on the family income variable, which has three categories -low (less than $20,000 yearly income), middle ($20,000-$35,000), and high (more than $35,000). The corresponding combination matrix (fuzzy coding) is given by

$$C_a = \begin{pmatrix} .258 & .228 & .514 \\ .212 & .430 & .358 \\ .360 & .334 & .306 \\ .125 & .500 & .375 \\ .454 & .334 & .212 \\ .030 & .412 & .558 \\ .640 & .300 & .060 \\ .486 & .514 & 0 \\ .472 & .306 & .222 \\ .036 & .607 & .357 \\ 0 & .130 & .870 \\ 0 & .109 & .891 \end{pmatrix}$$

for variables $a = A, B, C, \ldots, G$. It can be seen that the majority of the students in the public suburban (4,5,6) and the private (10,11,12) schools come from families with yearly incomes larger than $20,000, while those in the rural schools (7,8,9) from families with incomes less than $20,000. The students attending public urban schools (1,2,3) are more evenly distributed over the three income categories. Hence, the grouping of schools used in the previous section can be thought as a proxy for family income. The reason for applying restrictions to all seven variables is that by just restricting the first three (as in section 4.2), very similar results to the previous analysis were obtained.

A two-dimensional solution produced a satisfactory fit, with total eigenvalues $.529$ and $.315$ respectively. There is some loss of fit on both dimensions, as a consequence of applying restrictions on all ten variables. All schools experience a loss in fit, but schools exhibiting a good fit in the previous solution continue to do so (see Table 4.1).

The category quantifications and object scores plots are given in Figures 4.1 and 4.2 respectively. Almost all schools exhibit a quadratic pattern, with the 'serious problem' and the 'not a problem' categories forming separate clouds. This is expected since there are constraints (with differential weighting) imposed on the first seven variables across all schools. An examination of the object scores plot reveals that in the public rural schools only a small minority of the students responded using the 'serious problem' category. However, since for the majority of the schools use

| School # | Dimension 1 | Dimension 2 |
|---|---|---|
| 1 | .526 | .287 |
| 2 | .632 | .348 |
| 3 | .527 | .354 |
| 4 | .562 | .316 |
| 5 | .520 | .362 |
| 6 | .546 | .275 |
| 7 | .524 | .244 |
| 8 | .518 | .326 |
| 9 | .458 | .231 |
| 10 | .590 | .417 |
| 11 | .450 | .266 |
| 12 | .493 | .316 |
| Overall | .529 | .315 |

TABLE 4.1. School Eigenvalues

of alcohol does not represent a 'serious' problem (see Table 6.1), the constraints mask this particular feature for schools 8 and 9. On the other hand, they make the effect of student tardiness and absenteeism more prominent for the public rural schools. In general, the constraints 'filter' most of the 'noise' present in the unconstrained Homals solution. However, the more general coding scheme mixes the 'rougher' public urban schools (see section 4.2) with the relative 'problem free' private ones, thus producing a more uniform profile. Thus, some patterns specific to individual schools are weakened at best or eliminated at worst.

## 5. **Related Techniques and Concluding Remarks**

In this paper, the descriptive multivariate analytic techniques of homogeneity analysis and nonlinear principal components are extended to incorporate multilevel data structures. These techniques aim at the uncovering of the structure and representation of the interdependencies present in sets of categorical variables. Moreover, they are not introduced by way of an estimation problem based on a model involving parameters and error terms. Rather, one directly introduces a meaningful criterion (e.g. the average squared length function) which is subsequently optimized. All variables enter symmetrically in the criterion; no specific role is assumed for any of them, as is the case in regression models, where a particular variable (dependent) is singled out.

As it was pointed out in Remark 3.1 a similar technique of an exploratory nature, suitable for panel data, was introduced in [40]. Some interesting alternatives to our approach are given by latent structure models of multiway contingency tables [9], and multiple group item response theory

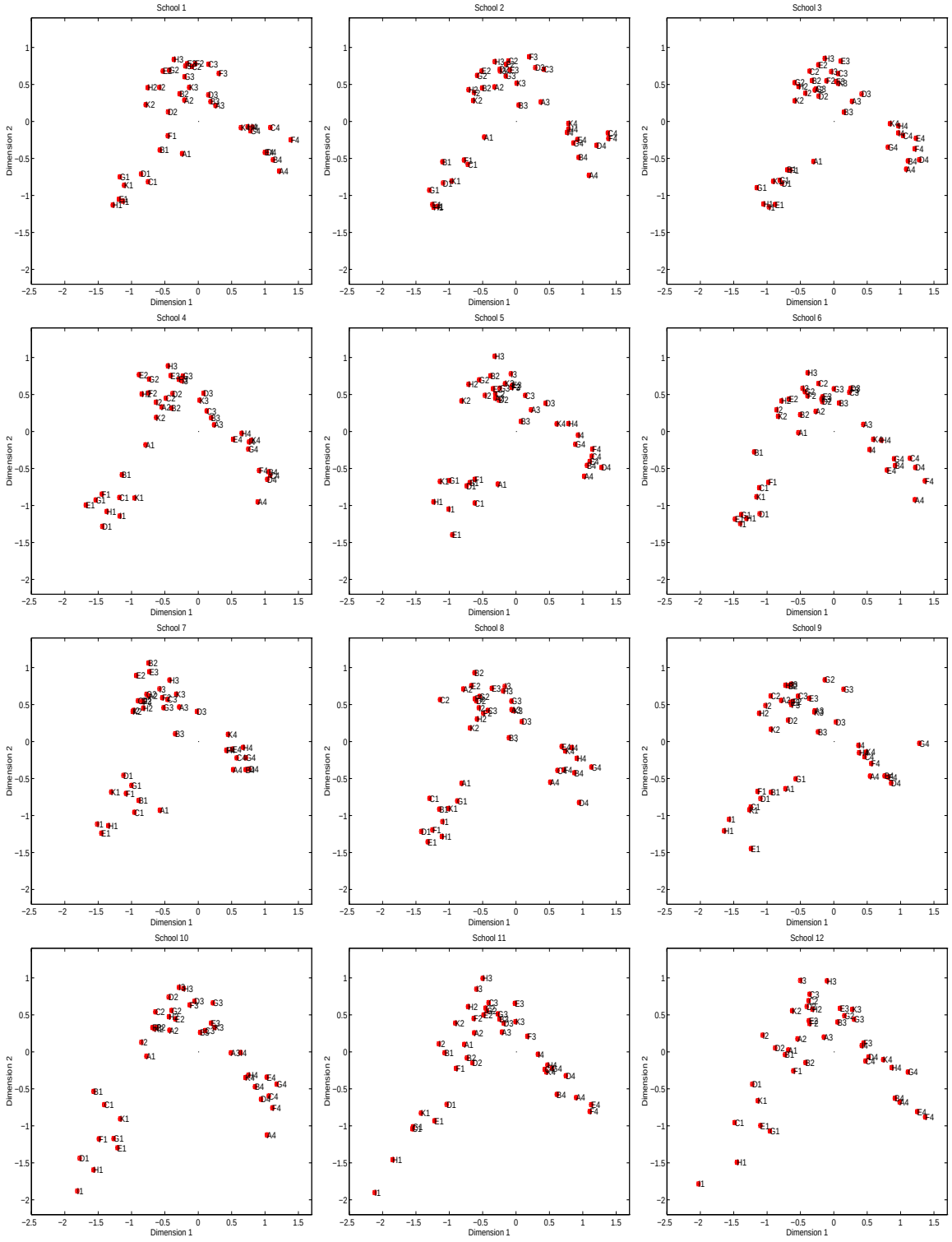


FIGURE 4.1. Optimal Constrained Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

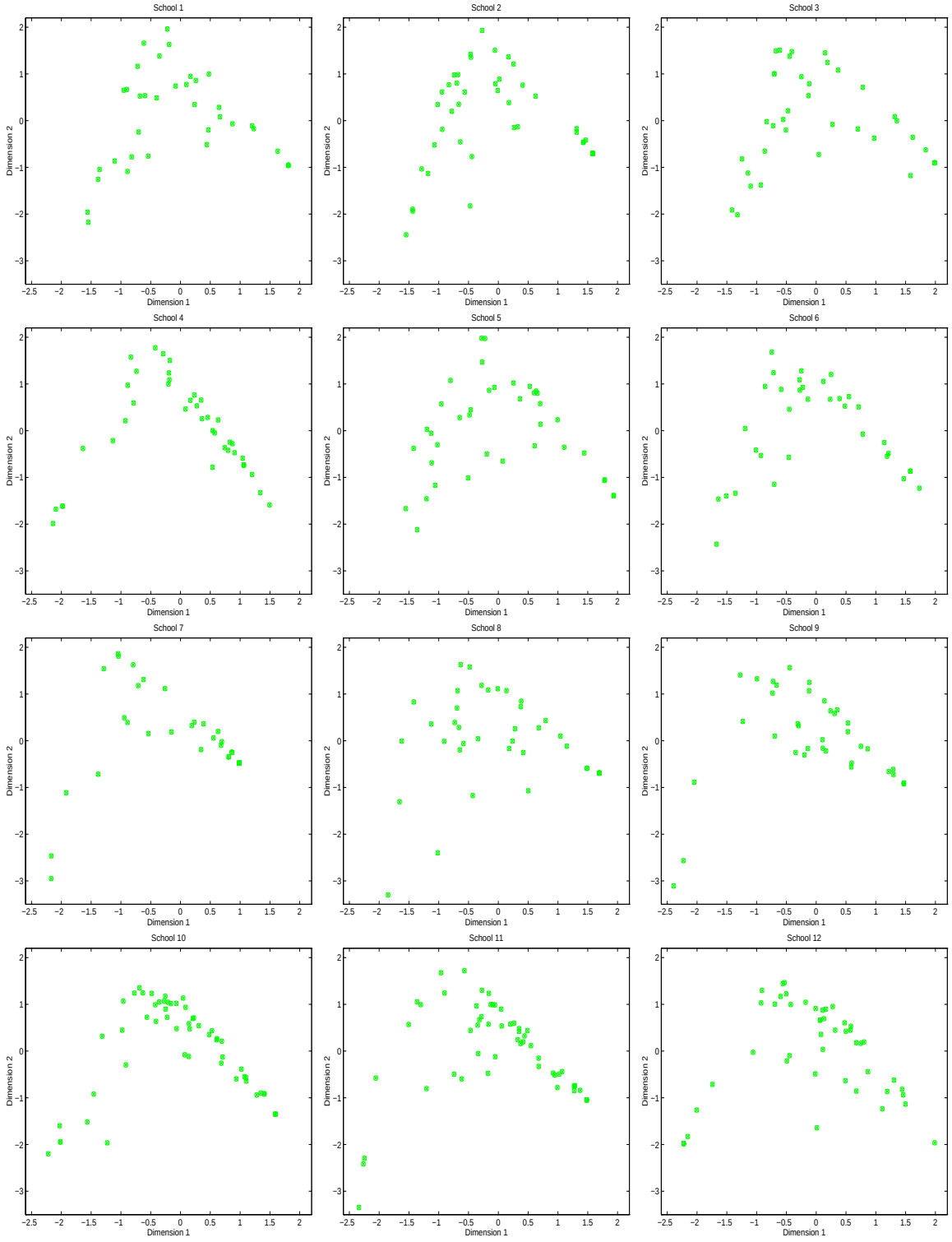


FIGURE 4.2. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

[3]. The former is a model based technique and therefore more appropriate for classical statistical inference. However, it does not allow for the simultaneous analysis of nominal, ordinal and numerical data as the multilevel Princals model does [30], and is not as flexible as our approach in imposing constraints across different clusters for different variables. The latter is best suited for categorical variables where a mathematical function relating the probability of a particular response on a certain question (variable) to an underlying concept can be safely assumed. Moreover, item response analysis is by design a one dimensional technique and hence does not allow the data practitioner to examine interactions of the variables in multiple dimensions. Nevertheless, in case we were dealing with a data set containing the results of a multiple choice test for students from different schools, it would be of great interest to compare the analyses obtained by applying multilevel homogeneity analysis and multigroup item response analysis.

On the other hand, if the nature of the data set allows the data analyst to specify stronger relationships between the variables (e.g. outcome and predictor variables), then there is a very rich literature on available techniques that allow for stronger inferences and also take into consideration the multilevel nature of the data. For example, we have hierarchical linear models (both fixed and random effects models) [5], as well as their Bayesian extensions [28, 25]. Such models have been widely used in applied educational research in studying various issues on students clustered within schools [5, 38]. Moreover, for path type of models Muthén has extended his structural equation methodology that combines ideas from regression and factor analysis to multilevel [35] and longitudinal data structures [34].

The multilevel framework introduced in this paper proves flexible enough to allow mixing equality and linear restrictions on different variables. The NELS:88 school example shows that imposing restrictions reveals the common patterns between the schools, improves the stability of the solution by eliminating outlier objects, and allows the data analyst to obtain a better understanding of the school differences. However, by imposing restrictions across all schools there is the danger of eliminating some interesting school specific patterns. A problem with our technique is that due to its descriptive nature and heavy reliance on plots, it becomes difficult to examine too many groups simultaneously (e.g. $K \geq 50$). However, something similar can be said for other multilevel/hierarchical techniques, where the large number of plots is replaced by large amounts of output containing estimated coefficients, standard errors and other statistics of interest. We are currently working in devising "summary statistics" and ways of plotting them, that would reveal unusual clusters, or interesting common patterns between clusters more easily, and thus enable us to examine more groups. Finally, it is worth mentioning that in case we are dealing with longitudinal data (a special form of a multilevel data structure) stronger types of constraints that explore the temporal ordering of the groups (time periods) can be considered.

## 6. **Appendix**

A data set from the National Education Longitudinal Study of 1988 (NELS:88) is used throughout this study to demonstrate the techniques. The sampling design of the NELS:88 data set is as follows. The base year (BY) sample of $8^{th}$ grade students in 1988 was constructed by a two-stage process. The first stage involved stratified sampling of approximately 1,050 public and private schools from a population of 40,000 schools containing $8^{th}$ graders.The second stage included random samples of students from each school. Some 24,500 students and their parents, their teachers and their school principals were surveyed. Three followup surveys of the student cohort were conducted in 1990 (first followup (FF)), in 1992 (second followup (SF)) and in 1994 (third followup (TF)). Student, parent, teacher, school administrator as well as dropout questionnaires were administered to the students still attending school and to the dropouts. The BY, FF, SF and TF data sets contain some 6,500 variables in total.

In this example we focus on a set of variables from the BY that deals with student responses on problem areas in their schools (the BYS58 set of variables in the NELS:88 codebook). A description of the variables is given next.

**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 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 response categories are: (1) Serious, (2) Moderate, (3) Minor and (4) Not a problem.

The following Table summarizes the student response patterns for the 10 variables included in the analysis. Due to the fact that some of the categories of variable $J$ were empty for many of the 12 schools, variable $J$ was omitted from any subsequent analysis.

| Variable | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 14.1 | 30.9 | 32.1 | 22.9 |
| B | 10.0 | 28.5 | 33.1 | 28.3 |
| C | 17.7 | 18.5 | 25.5 | 38.4 |
| D | 14.5 | 24.5 | 32.5 | 28.5 |
| E | 15.5 | 17.5 | 31.3 | 35.7 |
| F | 21.9 | 21.7 | 25.9 | 30.5 |
| G | 19.7 | 19.7 | 24.5 | 36.1 |
| H | 15.3 | 11.0 | 23.1 | 50.6 |
| I | 13.5 | 9.6 | 22.7 | 54.2 |
| K | 15.3 | 14.7 | 26.5 | 43.6 |

TABLE 6.1. Student Response Patterns (in %, N=498)

## REFERENCES


[1] Anderson, C.S. (1982), "The Search for School Climate: A Review of the Research," *Review of Educational Research*, **52**, 368-420

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

[3] Bock, D.R., Zimowski, M.F. (1997), "Multiple Group IRT," *Handbook of Modern Item Response Theory*, van der Linden et al. (eds), 433-448, New York: Springer

[4] Breiman, L. and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, **80**, 580-598

[5] Bryk, A.S., Raudenbush, S.W. (1992), *Hierarchical Linear Models*, Newbury Park: Sage

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

[7] Burt, C. (1950), "The Factorial Analysis of Qualitative Data," *British Journal of Statistical Psychology*, **3**, 166-185

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

[9] Clogg, C.C., Goodman, L.A. (1985), "Simultaneous Latent Structure Analysis in Several Groups," *Sociological Methodology*, **15**, 81-110, Tuma (ed.), San Francisco: Jossey-Bass

[10] 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

[11] 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

[12] De Leeuw, J., van der Heijden, P., and Kreft, I. (1985), "Homogeneity Analysis of Event History Data", *Methods of Operations Research*, 50, 299-316

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

[14] Fisher, R. A. (1938), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd

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

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

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

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

[19] 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

[20] Guttman, L. (1950), "The Principal Components of Scale Analysis," *Measurement and Prediction*, Stouffer (ed.), Princeton: Princeton University Press
[21] 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
[22] Kaufman, P., Bradby, D. (1992), *Characteristics of at Risk Students in NELS:88*, Report 92-042, Washington: National Center for Education Statistics
[23] Kendall, M.G. (1980), *Multivariate Analysis*, 2nd edition, London: Griffin
[24] Kruskal, J.B. (1965), "Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data," *Journal of the Royal Statistical Society, B*, **27**, 251-263
[25] Laird, N. and Ware, J. (1982), "Random-effects Models for Longitudinal Data," *Biometrics*, **38**, 963-974
[26] Lawler, E. (1976), *Combinatorial Optimization*, New York: Holt, Rinehart and Winston
[27] Lebart, L., Morineau, A., Tabard, N. (1977), *Technique de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*, Paris: Dunod
[28] Lindley, D.V., and Smith, A.F.M. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, Series B*, **34**, 1-41
[29] Lippman, L., Burns, .S., McArthur, E. (1996), *Urban Schools: The Challenge of Location and Poverty*, Report # 96-184, Washington: National Center for Education Statistics
[30] Michailidis, G. (1996), *Multilevel Homogeneity Analysis*, UCLA Dissertation
[31] Michailidis, G., de Leeuw, J. (1996), "The Gifi System of Nonlinear Multivariate Analysis," UCLA Statistics Series, #204
[32] Michailidis, G., de Leeuw, J. (1996), "Constrained Homogeneity Analysis with Applications to Hierarchical Data," UCLA Statistics Series, #207
[33] Molenaar, I.W. (1988), "Formal Statistics and Informal Data Analysis, or why Laziness Should Be Discouraged," *Statistica Neerlandica*, **42**, 83-90
[34] Muthén, B. (1997), "Latent Variable Modeling of Longitudinal and Multilevel Data," *Sociological Methodology*, **27**, 453-480
[35] Muthéen, B., and Sattora, A. (1995), "Complex Sample Data in Structural Equation Modeling," *Sociological Methodology*, **25**, 267-316
[36] The National Education Goals Panel (1992), *The National Education Goals Report: Building a Nation of Learners*, Washington
[37] Oakes, J. (1989), "What Educational Indicators? The Case for Assessing the School Context," *Educational Evaluation and Policy Analysis*, **11**, 181-199
[38] Raudenbush, S.W. (1988), "Educational Applications of Hierarchical Linear Models: A Review," *Journal of Educational Statistics*, **13**, 85-116
[39] Seltzer, M.H. (1993), "Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach," *Journal of Educational Statistics*, **18**, 207-235
[40] Van der Heijden, P., and de Leeuw, J. (1990), "Correspondence Analysis, with Special Attention to the Analysis of Panel Data and Event History Data," *Sociological Methodology*, **20**, 43-87
[41] Van Rijckevorsel, J.L.A. (1987), *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*, Leiden: DSWO Press
[42] Young, F.W. (1981), "Quantitative Analysis of Qualitative Data," *Pshychometrika*, **46**, 357-388

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