MULTILEVEL HOMOGENEITY ANALYSIS WITH CONSTRAINTS ON THE CATEGORY QUANTIFICATIONS

GEORGE MICHAILIDIS AND JAN DE LEEUW

ABSTRACT. We consider a multilevel sampling design framework, where we collect observations on N individual cases grouped (clustered) within K units (clusters). We extend the techniques of homogeneity analysis and principal components analysis with rank-one restrictions to this multilevel 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. The National Educational Longitudinal Study (NELS:88) data set is used to illustrate some of the techniques presented in the paper.

Date: October 25, 1996.

1. Introduction to Homogeneity Analysis

Given a data set comprised of N observations (individuals, objects) on J variables, the concept of *homogeneity* addresses the question to what extent different variables measure the same property or properties in the data. In order to answer this question, a *measure* for the difference or resemblance of the variables is needed. Moreover, the measurement level of the data may allow us to *transform* the variables before comparing them, since different transformations are suitable for different types of data. The problem then becomes to find admissible transformations that maximize the homogeneity of the variables. In case where the variables measure more than one property we may want to identify another orthogonal solution. This is in accordance with the principle of data reduction which advocates that a small number of dimensions should be used to explain a maximum amount of information present in the data.

The technique studied in this paper is known under many different names. For example, we have *principal components of scale analysis* [1, 2], *factorial analysis of qualitative data* [3], *second method of quantification* [7], *multiple correspondence analysis* [15, 6], and *homogeneity analysis* [8, 11]. 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 [5] 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. [9]) 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 [10]). More recently Breiman and Friedman [4] developed similar methods using conditional expectations and exploring the structure of the L_2 space.

In the Gifi system [8] the relationship between multiple correspondence analysis and optimal scaling is explored systematically, with the former being the starting point. The basic idea is to scale the objects (map them into a low dimensional Euclidean space) in such a way that objects (individuals etc) with similar profiles are relatively close together, while objects with different profiles are relatively far apart. The emphasis, as in the French literature, is on the geometry of the problem. Moreover, other multivariate analytic techniques such as principal component analysis, canonical correlation analysis, discriminant analysis, path analysis can be derived from multiple correspondence analysis by imposing appropriate restrictions. These restrictions also have a clear geometric interpretation in the space where the objects are scaled. This is in contrast with the approach taken in Young [10] or Breiman and Friedman [4], who take optimal scaling as their starting point, and emphasize algebraic and analytic properties of the solution.

To formally introduce our framework, suppose that initially we allow only linear transformations of the variables. The difference between variables is expressed in terms of a *loss function*. As a convenient measure for differences between variables we will consider the mean squared Euclidean distance between the (possibly) transformed variables and a common hypothetical variable. The homogeneity of the transformed variables and the hypothetical variable will be maximized if the loss is minimized. This hypothetical variable serves as a new scale for the individuals or objects. The loss function is given by

$$\sigma\big(x;\alpha_1,\ldots,\alpha_J\big) = \frac{1}{J}\sum_{j=1}^{J} \text{SSQ}\,\big(X - \alpha_j s_j\big), \tag{1.1}$$

where x denotes the common variable, α_j the weight of variable s_j and $\text{SSQ}\,(H) = \text{tr}(H'H)$ denotes the Frobenius norm of matrix H (i.e. the sum of squares of the elements of the matrix H). The values of variable x are called *object scores*. In order to avoid the trivial solution corresponding to $x = 0$, and $\alpha_j = 0$ for every $j \in \mathbf{J} = \{1,\ldots,J\}$, we require in addition $x'x = 1$. The solution to the minimization problem given in (1.1) corresponds to performing linear principal component analysis on the correlation matrix of the data set S (see [8, 11, 12]). The common variable x is the first principal component.

A natural question that arises at this point is what happens if we allow nonlinear transformations of the form $\phi_j(s_j),\ j \in \mathbf{J}$, where ϕ_j may be any nonlinear function of the variable s_j. By admitting all nonlinear transformations the problem is rendered trivial, since the space of admissible transformations is enlarge to a N-dimensional space, and thus any quantification of the variables will suffice in order to obtain a perfect fit. Since our main interest is on *categorical* variables, we will consider the indicator matrix G_j as a *basis* of the transformation space. It is a binary matrix with entries $G_j(i,t) = 1,\ i = 1,\ldots,N,\ t = 1,\ldots,\ell_j$ (where ℓ_j represents the number of categories for variable j), if object i belongs to category t, and $G_j(i,t) = 0$ if it belongs to some other category. Incorporating our choice of basis G_j implies that we may write $\phi_j(s_j) = G_j y_j$, for some vector of coefficients y_j. In case of multiple solutions we get $\phi_j(s_j) = G_j Y_j$, where Y_j is the $\ell_j \times p$ matrix of *multiple category quantifications*, where p denotes the dimensionality of the solution.

The loss function then becomes

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

Once again 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 the unit vector of appropriate dimensions.

The setup of the problem we posit in (1.2), (1.3) and (1.4) implies the following:

- A scale consisting of numerical variables is *homogeneous* if all variables in the scale are linearly related.
- A scale consisting of nominal, ordinal, and numerical variables is *homogenizable* if all variables in the scale can be *transformed* or *quantified* in such a way that the resulting scale is homogeneous.
- Homogeneity of a set of (centered) variables is measured by computing the *sum of squares within objects* and the *sum of squares between objects*. Perfect homogeneity corresponds to zero within objects variation. A measure of homogeneity is the ratio of the between objects sum of squares to the total sum of squares.
- Homogeneity analysis *transforms* numerical variables (i.e. assigns numerical values to each of the categories of the variable) or *quantifies* ordinal or nominal variables (i.e. assigns numerical values to each of the categories of the variable) in such a way that homogeneity is maximized.

In this paper, we extend the framework to handle data of a multilevel nature and propose some models that take into consideration the hierarchical structure of such data. However, we do not deal with topics such as sets of variables, additivity restrictions on the variables etc (see chapter 5 in [8]). For making this presentation complete, a brief review of the solution to the problem posed above is also included.

2. The ALS Algorithm

The solution to the minimization problem given in (1.2), (1.3) and (1.4) is found by employing the *Alternating Least Squares* (ALS) algorithm. In the first step, (1.2) is minimized with respect to Y_j for fixed X. The set of normal equations is given by

$$D_j Y_j = G_j' X, \; j \in \mathbf{J}, \tag{2.1}$$

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

$$\hat{Y}_j = D_j^{-1} G_j' X, \; j \in \mathbf{J}. \tag{2.2}$$

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

$$JX = \sum_{j=1}^{J} G_j Y_j, \tag{2.3}$$

so that

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

In the third step of the algorithm the X matrix is column centered and then orthonormalized, so that the normalization restrictions (1.3) and (1.4) are satisfied. These three steps are repeated until the algorithm converges to the global minimum (see chapter 3 in [8]). Hence, the ALS algorithm finds the desired solution to the problem given in (1.2), in the presence of nominal data. This solution is also known in the literature ([8], [11], [12]) as the Homals solution (homogeneity analysis by means of alternating least squares).

Remark 2.1. *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 (2.2) that $Y_j^\sharp = D_j^{-1}G_j'X^\sharp = \hat{Y}_j R$. Thus, any rotation of the object scores and of the category quantifications corresponds to a solution to the problem given in (1.2).

Remark 2.2. *Normalization Issues.* The matrix of the object scores X is column centered by subtracting from each element $x(is)$ the mean of the corresponding column (i.e. set $W = X - u(u'X/N)$). Then, W is orthonormalized by the modified Gram-Schmidt [17]. Finally, set $X = \sqrt{N}W$, and it is easy to see that $X'X = NI_p$.

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

$$\begin{aligned} J^{-1}\sum_{j=1}^{J}\operatorname{tr}(X - G_jY_j)'(X - G_jY_j) &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}(X'X + Y_j'G_j'G_jY_j - 2Y_j'G_j'X) = \\ J^{-1}\sum_{j=1}^{J}\operatorname{tr}(X'X - Y_j'D_jY_j) &= J^{-1}\sum_{j=1}^{J}\operatorname{tr}(NI_p - Y_j'D_jY_j) = Np - J^{-1}\sum_{j=1}^{J}\operatorname{tr}(Y_j'D_jY_j). \end{aligned} \tag{2.5}$$

The sum of the diagonal elements of the matrices $Y_j'D_jY_j$ is called the *fit* of the solution. Furthermore, the *discrimination measures* are given by

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

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 [8]). Hence, the loss function can also be expressed as

$$N\big(p - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2\big). \tag{2.7}$$

The quantities $\gamma_s = J^{-1}\sum_{j=1}^{J}\eta_{js}^2,\ s = 1,\dots,p$ are called the *eigenvalues* and correspond to the average of the discrimination measures.

Remark 2.3. *Homogeneity Analysis as a Singular Value Decomposition Problem.* It can also be shown [11] that the object scores X are the left singular vectors of the matrix $J^{-1/2}(I - uu'/N)GD^{-1/2}$, which is the 'super-indicator' matrix ($G = [G_1|\dots|G_J]$) in deviation from column means and corrected for marginal frequencies. Moreover, the eigenvalues $\gamma_s,\ s = 1,\dots,p$ correspond to the first p singular values of the above matrix. The complete singular value decomposition (SVD) solution has $q = \sum_{j=1}^{J}\ell_j - J$ dimensions. Once the object scores are calculated, the category quantifications can be computed using (2.1) [8]. The advantage of employing the ALS algorithm is that it only computes the first $p << q$ dimensions of the SVD solution, thus increasing the computational efficiency and decreasing computer memory requirements.

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 (2.2).
- ◇ Objects with the same response pattern (identical profiles) receive identical object scores (follows from (2.4)). 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 (2.6)).
- ◇ 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 (2.2)).
- ◇ 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 $u'D_jY_j = u'D_jD_j^{-1}G_j'X = u'G_jX = u'X = 0$.

Remark 2.4. *Missing Data.* The present loss function makes the treatment of missing data a fairly easy exercise. Missing data can occur for a variety of reasons: blank responses, coding errors etc. Let $M_j,\ j \in \mathbf{J}$ denote the $N \times N$ binary diagonal matrix with entries $M_j(ii) = 1$ if observation

i is present for variable j and 0 otherwise. Define $M_* = \sum_{j=1}^{J} M_j$. Notice that since G_j is an incomplete indicator matrix (has rows with just zeros), we have that $M_j G_j = G_j,\ j \in \mathbf{J}$. The loss function then becomes

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}\big(X - G_j Y_j\big)' M_j \big(X - G_j Y_j\big), \tag{2.8}$$

subject to the normalization restrictions $X' M_* X = JNI_p$ and $u' M_* X = 0$. The $\hat{Y}_j$'s are given by (2.2), while the object scores by $\hat{X} = M_*^{-1} \sum_{j=1}^{J} G_j Y_j$. In the presence of missing data, it is no longer the case that $u' D_j Y_j = 0$ (the category quantifications are not centered), because in the weighted summation with respect to the row scores of X, some of the scores are skipped. This option is known in the literature [8] 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.

3. **Rank-One Restrictions**

The Homals solution leaves the category quantifications Y_j free. It only places restrictions on the object scores for identification purposes. However, in case we deal with ordinal or numerical data, we also have to take into consideration the restrictions imposed by the measurement level of the variables. This is impossible in the multiple quantification framework outlined in the previous sections. First, consider a multiple treatment of numerical variables. In this case, the quantification of the categories must be the same as the standardized a priori quantification. This implies that multiple numerical quantification contains incompatible requirements. Second, consider a multiple treatment of ordinal variables. This requirement is not contradictory in itself; however, the different quantifications must have the same order as the prior quantifications, thus resulting in being highly intercorrelated. It follows that such an option does not have much to offer. In order to overcome these difficulties, it is also required that the quantifications satisfy a *rank-one* restriction (see chapter 4 in [8]); that is,

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

where, q_j is a ℓ_j-column vector containing the *single category quantifications* and β_j a p-column vector of *component loadings*. In this case the quantifications in p dimensions become proportional to each other. To minimize (1.2) under the restriction (3.1), we start by computing the $\hat{Y}_j$'s as in

(2.2). We then partition the Gifi loss function as follows:

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

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

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

with respect to q_j and β_j. We do this by going to another ALS loop (alternate over q_j and β_j). Minimizing (3.2) for β_j given fixed q_j, we get the following set of normal equations.

$$\xi_j\beta_j = \hat{Y}_j'D_jq_j, \; j \in \mathbf{J}, \tag{3.4}$$

where $\xi_j = q_j'D_jq_j$ is a scalar. Hence, we get

$$\hat{\beta}_j = \big(\hat{Y}_j'D_jq_j\big)/\xi_j, \; j \in \mathbf{J}. \tag{3.5}$$

Write now (3.2) as

$$\sum_{j=1}^{J} \text{tr}\big(q_j(\hat{\beta}_j' + (\beta_j' - \hat{\beta}_j')) - \hat{Y}_j\big)'D_j\big(q_j(\hat{\beta}_j' + (\beta_j' - \hat{\beta}_j')) - \hat{Y}_j\big) = \sum_{j=1}^{J} \text{tr}\big(q_j\hat{\beta}_j' - \hat{Y}_j\big)'D_j\big(q_j\hat{\beta}_j' - \hat{Y}_j\big) + \sum_{j=1}^{J} \text{tr}\big(\beta_j' - \hat{\beta}_j'\big)'D_j\big(\beta_j' - \hat{\beta}_j'\big). \tag{3.6}$$

Since there are no further restrictions regarding the β_j's, we simply set $\beta_j = \hat{\beta}_j, \; j \in \mathbf{J}$ and the first step of the inner iteration loop is complete. So, it remains to minimize (3.2) with respect to q_j for fixed β_j (the other half of the inner iteration loop). We then get the following set of normal equations

$$\psi_jD_jq_j = D_j\hat{Y}_j\beta_j, \; j \in \mathbf{J}, \tag{3.7}$$

where $\psi_j = \beta_j'\beta_j$ is a scalar. Thus, we get that

$$\hat{q}_j = \hat{Y}_j\beta_j/\psi_j, \; j \in \mathbf{J}. \tag{3.8}$$

Write now (3.2) as follows

$$\sum_{j=1}^{J} \text{tr}\big((\hat{q}_j + (q_j - \hat{q}_j))\beta_j' - \hat{Y}_j\big)'D_j\big((q_j + (q_j - \hat{q}_j))\beta_j' - \hat{Y}_j\big) = \sum_{j=1}^{J} \text{tr}\big(\hat{q}_j\beta_j' - \hat{Y}_j\big)'D_j\big(\hat{q}_j\beta_j' - \hat{Y}_j\big)'D_j + \sum_{j=1}^{J} \text{tr}\big(q_j - \hat{q}_j\big)'D_j\big(q_j - \hat{q}_j\big)\psi_j. \tag{3.9}$$

Finding the q_j's means projecting the vector $\hat{q}_j$ on the cone $\mathcal{C}_j$ that represents the feasible region for the y_j vector in the parameter space defined by Y_j of dimensionality ℓ_j. This becomes a monotone regression problem in the *ordinal* case, a linear regression problem in the *numerical* case, and simply setting $q_j = \hat{q}_j$ in the *nominal* case. We then set $\hat{Y}_j = \hat{q}_j\hat{\beta}_j'$ and proceed to compute the object scores. This solution that takes into consideration the measurement level of the variables is referred in the literature ([8], [12]) as the Princals solution (principal component analysis by means of alternating least squares). The Princals model allows the data analyst to treat each variable differently; some may be treated as multiple nominal and some others as single nominal, ordinal or numerical. In that sense, Princals generalizes the Homals model.

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

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

$$\sum_{j=1}^{J}\mathrm{tr}(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) + \sum_{j=1}^{J}\mathrm{tr}(\hat{q}_j\hat{\beta}_j' - \hat{Y}_j)'D_j(\hat{q}_j\hat{\beta}_j' - \hat{Y}_j). \tag{3.10}$$

Using (2.7), the first term in (3.10) can be also written as $N(p - \sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2)$, which is called the *multiple loss*. The discrimination measures are called the *multiple fit* of variable j in dimension s. Imposing the normalization restriction $q_j'D_jq_j = N$, and using the fact that $\hat{Y}_j'D_jq_j\beta_j' = N\beta_j\beta_j'$ (from (3.6)), the second part of (3.10) can be written as

$$\sum_{j=1}^{J}\mathrm{tr}(\hat{Y}_j'D_j\hat{Y}_j - N\beta_j\beta_j') = N(\sum_{j=1}^{J}\sum_{s=1}^{p}(\eta_{js}^2 - \beta_{js}^2)), \tag{3.11}$$

which is called the *single loss*. The quantities β_{js}^2, $s = 1, \ldots, p$ are called *single fit*, and correspond to squared component loadings (see chapter 4 in [8]).

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

$$\begin{aligned}
&\sum_{j=1}^{J}\mathrm{tr}(X - G_jY_j)'M_j(X - G_jY_j) = \\
&\sum_{j=1}^{J}\mathrm{tr}(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j)))'M_j(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))) = \\
&\sum_{j=1}^{J}\mathrm{tr}(X - G_j\hat{Y}_j)'M_j(X - G_j\hat{Y}_j) + \sum_{j=1}^{J}\mathrm{tr}(Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j).
\end{aligned} \tag{3.12}$$

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

4. Multilevel Modeling

In many situations individual objects (level-1 units) can be naturally grouped (*clustered*) into groups (clusters, level-2 units). For example, in educational research students are grouped by class or school, in sociological research individuals are grouped by socioeconomic status, in marketing research consumers are clustered in geographical regions, while in longitudinal studies we have repeated measurements on individuals. In the first example clusters correspond to classes or schools, 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 to time periods. Formally, we collect data on N objects grouped naturally in K clusters, with n_k objects per cluster, $k \in \mathbf{K} = \{1, \dots, K\}(\sum_{k=1}^{K} n_k = N)$. 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 [13] and Van der Heijden and De Leeuw [14] 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 $X = [X_1', \dots, X_K']'$. 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 $Y_j = [Y_{j1}', \dots, Y_{jK}']'$. We collect the K indicator matrices of variable j in the superindicator matrix $G_j = \bigoplus_{k=1}^{K} G_{jk}$, 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 [13] and [14] can be derived from our framework. In our case the Gifi loss function becomes

$$\sigma(X;Y_1,\dots,Y_J)=J^{-1}\sum_{j=1}^{J}\text{SSQ}\left(X-G_jY_j\right)=\sum_{j=1}^{J}\sum_{k=1}^{K}\text{SSQ}\left(X_k-G_{jk}Y_{jk}\right). \tag{4.1}$$

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

$$X_k'X_k=n_kI_p,\ u'X_k=0,\ \text{for every } k\in\mathbf{K}. \tag{4.2}$$

The other possibility $u'X=0$ and $X'X=NI_p$ is briefly discussed later on in Remark 4.2.

The problem in (4.1) is identical to the one presented in (1.2); thus, its solution is given by

$$\hat{Y}_j=D_j^{-1}G_j'X,\ j\in\mathbf{J}, \tag{4.3}$$

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 $\hat{Y}_{jk}=D_{jk}^{-1}G_{jk}'X_k,\ j\in\mathbf{J},\ k\in\mathbf{K}$. For fixed Y_j's, we get

$$\hat{X}=\frac{1}{J}\sum_{j=1}^{J}G_jY_j, \tag{4.4}$$

which gives that $\hat{X}_k=J^{-1}\sum_{j=1}^{J}G_{jk}Y_{jk}$, for every $k\in\mathbf{K}$. We then center and orthonormalize the X_k matrices and repeat these two steps until the algorithm converges.

We define next the *cluster discrimination measures*

$$\eta_{jks}^2\equiv Y_{jk}'(.,s)D_{jk}Y_{jk}(.,s)/n_k,\ j\in\mathbf{J},\ k\in\mathbf{K},\ s=1,\dots,p, \tag{4.5}$$

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 η_{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

$$\eta_{js}^2\equiv Y_j'(.,s)D_jY_j(.,s)/N,\ j\in\mathbf{J},\ s=1,\dots,p, \tag{4.6}$$

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. We have that

$$\eta_{js}^2\equiv\frac{1}{N}Y_j'(.,s)D_jY_j(.,s)=\frac{1}{N}\sum_{k=1}^{K}Y_{jk}'(.,s)D_{jk}Y_{jk}(.,s). \tag{4.7}$$

so, it easy to see that

$$\eta_{js}^2=\frac{1}{N}\sum_{k=1}^{K}n_k\eta_{jks}^2,\ j\in\mathbf{J},\ s=1,\dots,p. \tag{4.8}$$

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.

We can then 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 4.1. *Model Equivalences.* It is worth noting that under normalization (4.2) this model is equivalent to applying the ordinary Homals algorithm (see Section 2) to each of the K clusters separately.

Remark 4.2. *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 $X'X = \sum_{k=1}^{K} X_k'X_k = NI_p$. 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.

4.1. **Princals.** We briefly present the extension of the basic Princals model to the multilevel framework. We require $y_{jks}(t) = \beta_{jks}q_{jk}(t)$, which implies that if we plot $y_{jks}(t)$ against $q_{jk}(t)$ for different values of $s \in \{1,\dots,p\}$ we see parallel straight lines. In matrix form we have, $Y_{jk} = q_{jk}\beta_{jk}'$, $j \in \mathbf{J}$, $k \in \mathbf{K}$. The rank-one restrictions can be written in compact form as follows:

$$Y_j = \mathcal{Q}_j\mathcal{B}_j',\ j \in \mathbf{J}, \tag{4.9}$$

where $\mathcal{Q}_j = \bigoplus_{k=1}^{K} q_{jk}$ is the $K\ell_j \times K$ matrix of single quantifications and $\mathcal{B}_j = [\beta_1,\dots,\beta_J]$ the $p \times K$ matrix of weights.

Our starting point is to compute the $\hat{Y}_j$'s as in (4.3). We then partition the Gifi loss function similarly to (3.2). Hence, after imposing the rank-one restrictions we have to minimize

$$\sum_{j=1}^{J}\operatorname{tr}\big(\mathcal{Q}_j\mathcal{B}_j' - \hat{Y}_j\big)'D_j\big(\mathcal{Q}_j\mathcal{B}_j' - \hat{Y}_j\big), \tag{4.10}$$

with respect to $\mathcal{Q}_j$ and $\mathcal{B}_j$. We do this by using a second ALS loop (alternate over $\mathcal{Q}_j$ and $\mathcal{B}_j$). Solving (4.10) for $\mathcal{B}_j$ given fixed $\mathcal{Q}_j$, we get

$$\hat{\mathcal{B}}_j = \big(\hat{Y}_j'D_j\mathcal{Q}_j\big)\big(\mathcal{Q}_j'D_j\mathcal{Q}_j\big)^{-1},\ j \in \mathbf{J}, \tag{4.11}$$

which implies $\hat{\beta}_{jk} = (\hat{Y}'_{jk} D_{jk} q_{jk})/\xi_{jk},\ k \in \mathbf{K},\ j \in \mathbf{J}$, where $\xi_{jk} = q'_{jk} D_{jk} q_{jk}$ is a scalar. This is a direct extension of the single cluster case. Minimizing (4.10) with respect to $\mathcal{Q}_j$ for fixed $\mathcal{B}_j$ (the other half of the inner iteration) we get

$$\hat{Q}_j = \left(\text{diag } \hat{Y}_j \mathcal{B}_j\right)\left(\text{diag } \mathcal{B}'_j \mathcal{B}_j\right)^{-1},\ j \in \mathbf{J}, \tag{4.12}$$

which gives that $\hat{q}_{jk} = \hat{Y}_{jk} \beta_{jk}/\psi_{js}$, where $\psi_{jk} = \beta'_{jk}\beta_{jk}$ is a scalar. After taking into account the measurement level restrictions of the variables we set $\hat{Y}_{jk} = \hat{q}_{jk}\hat{\beta}'_{jk}$, and proceed to calculate the object scores.

It is easy to see that the multilevel Princals model corresponds to applying the ordinary Princals model (see Section 3) to each of the K clusters separately (as was the case with the ordinary multilevel Homals).

Imposing the normalizations constraints $q'_{jk} D_{jk} q_{jk} = n_k,\ j \in \mathbf{J},\ k \in \mathbf{K}$, the single loss component is given by

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \text{tr}\left(\hat{Y}'_{jk} D_{jk} \hat{Y}_{jk} - n_k \beta_{jk} \beta'_{jk}\right), \tag{4.13}$$

The quantities given by $\beta^2_{jks},\ s = 1, \ldots, p$ are called *cluster single fit*, while the ones given by $N^{-1} \sum_{k=1}^{K} n_k \beta^2_{jks}$ are called *variable single fit.*

4.2. **NELS:88 Example.** A data set from the National Education Longitudinal Study of 1988 (NELS:88) will be 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.

This set of variables addresses some issues directly related to the school culture and climate, as seen from the students point of view. These variables touch upon day-to-day school experiences that influence the way students, teachers and administrators act, relate to one another and form their expectations and to a certain extent beliefs and values [21, 22].

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. However, we were forced to drop variable J from any subsequent analysis because categories 3 and 4 were empty in the majority of these schools. 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 4.1. 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 (teacher) is interested in attending (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 (teacher) for those 12 schools. Information about other schools is marginally interesting to them. The techniques previously developed are used for descriptive and not for inferential purposes [18, 23] in this example.

The following Table summarizes the student response patterns for the 10 variables included in the 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 4.2. Student Response Patterns (in %, N=498)

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 4.3. School Eigenvalues

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 4.3. The overall fit can be 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 4.1. 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 had 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) had smaller discrimination measures than the total for the majority of the variables; however, for a couple of variables the cluster measures were much larger than the total ones, thus indicating the possible presence of outliers. Figure 4.2 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 4.3. The optimal transformations of all variables in the first dimension are monotone decreasing 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 4.4 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

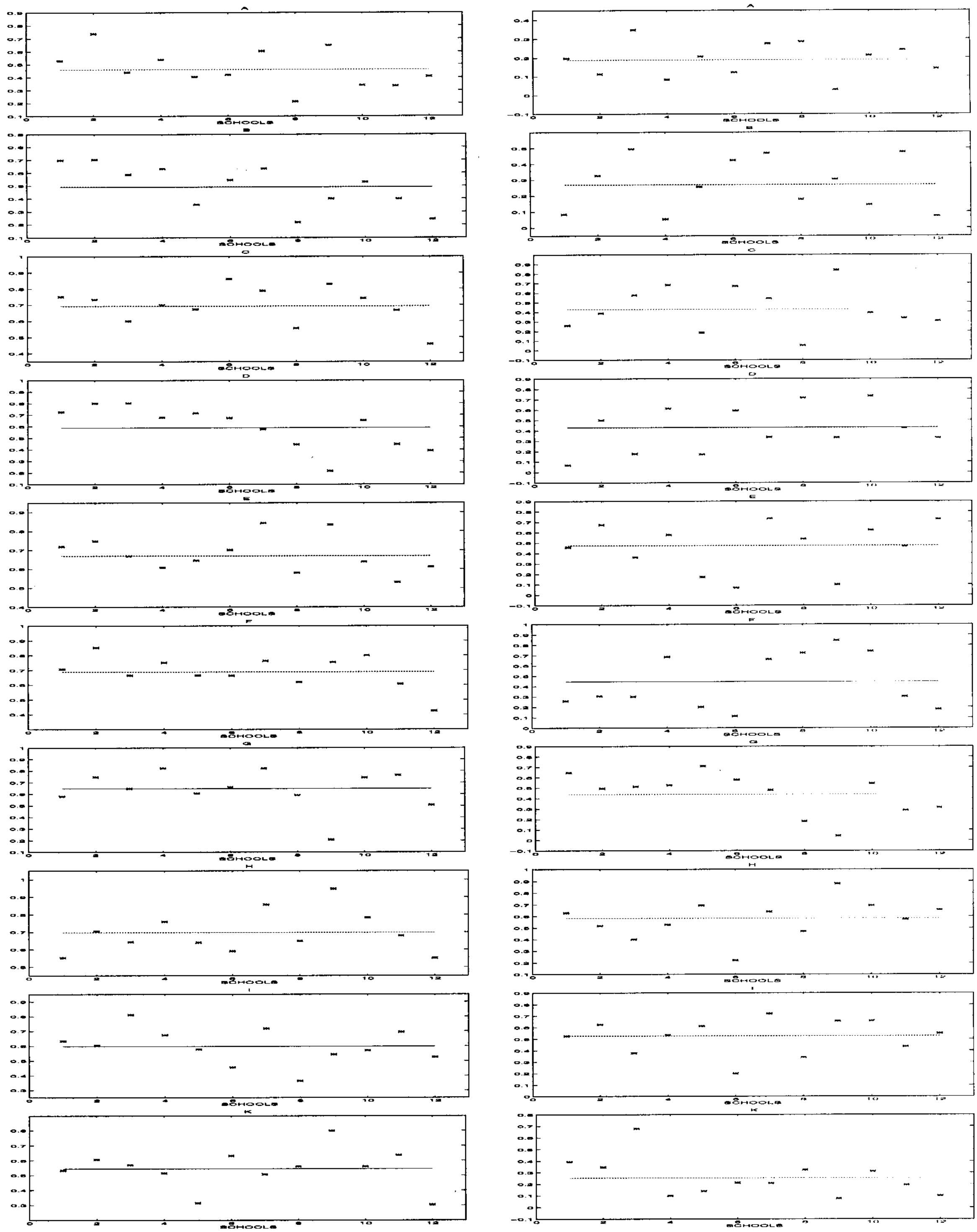

FIGURE 4.1. Discrimination Measures of the Variables for the Schools; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12; the solid line represents the variable's overall discrimination measure (Left: dimension 1, Right: dimension 2).

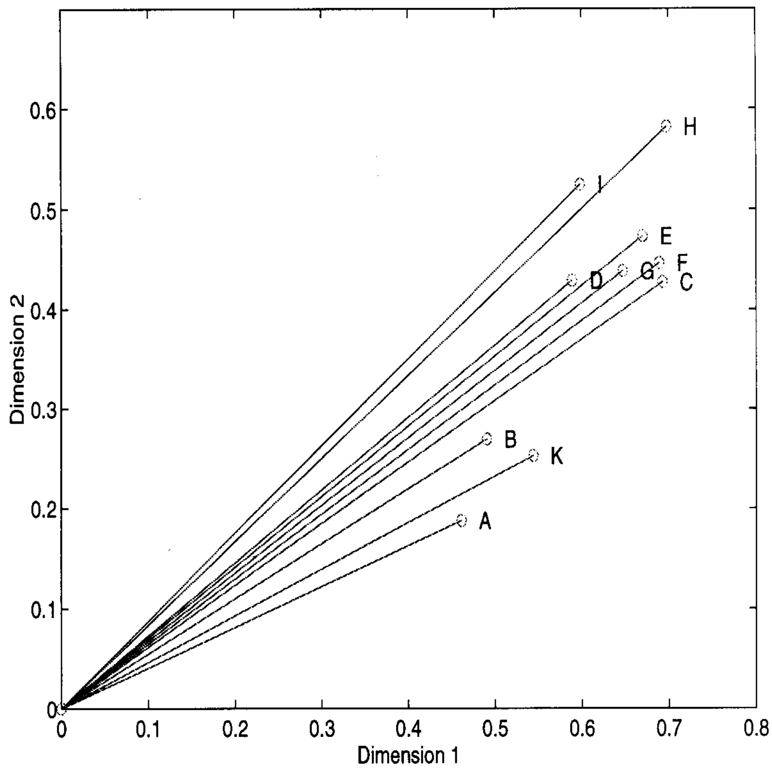

FIGURE 4.2. Total Discrimination Measures

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 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. In

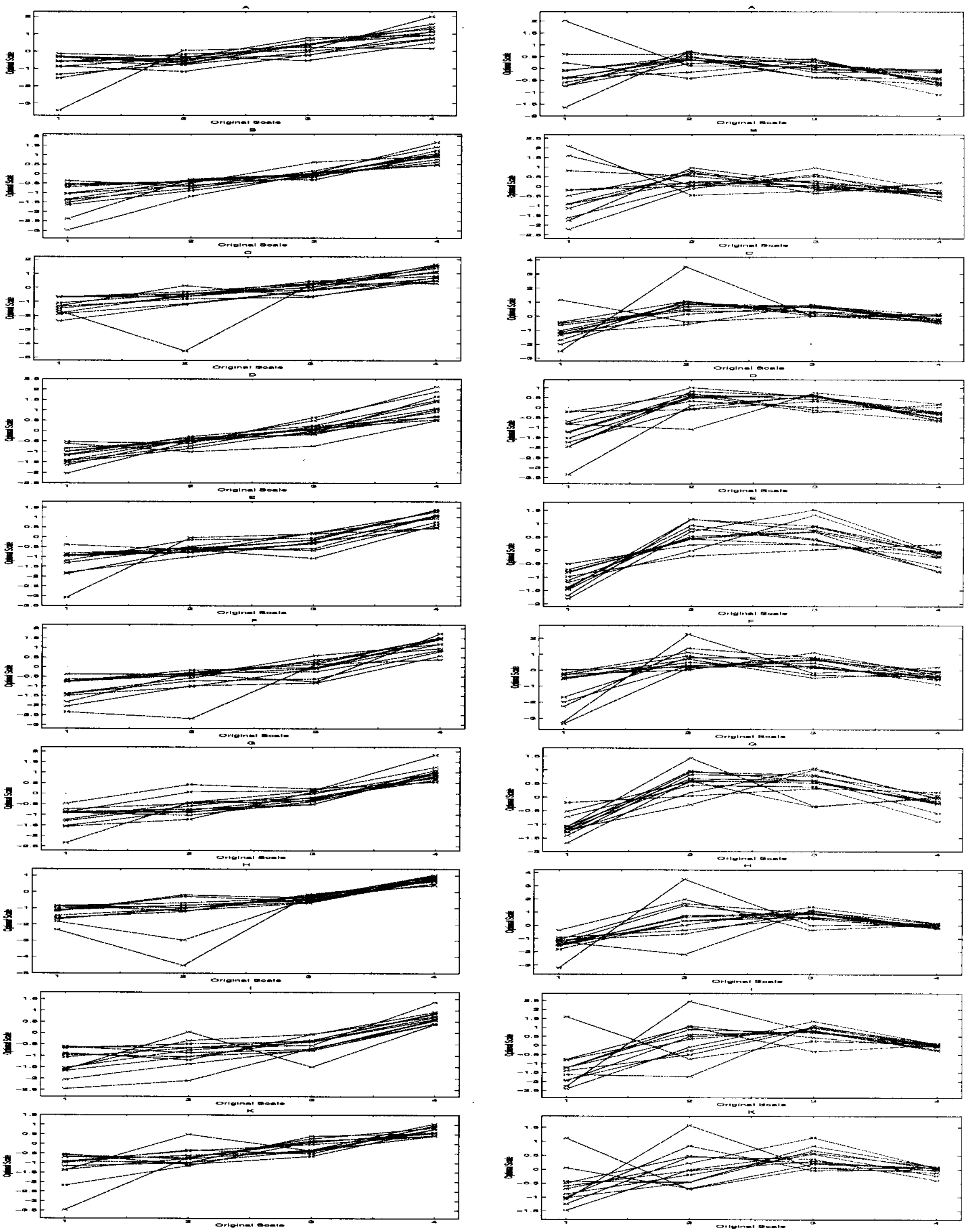

FIGURE 4.3. Optimal Transformations of the Variables (Left: dimension 1, Right: dimension 2)

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 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 4.5 (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.

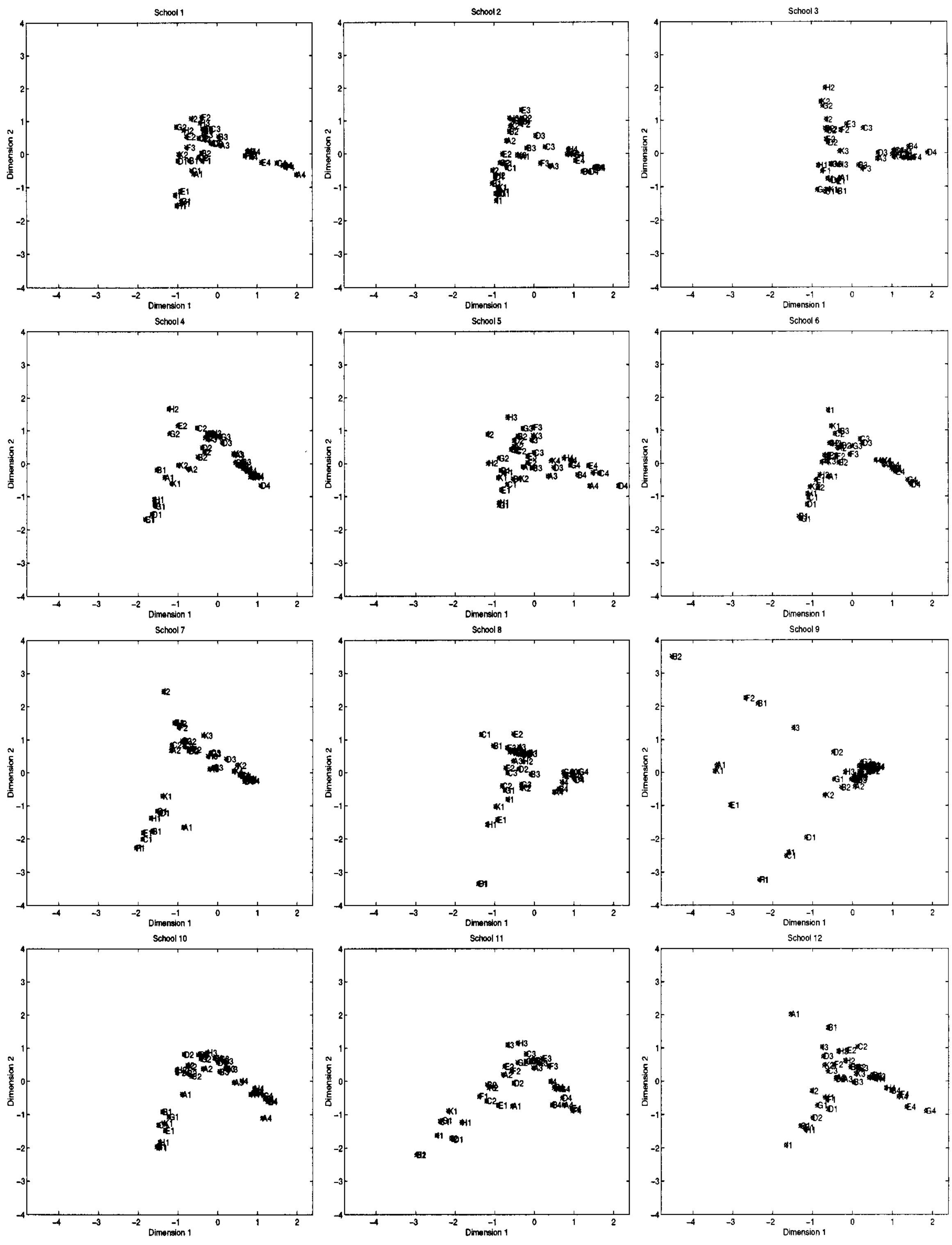

FIGURE 4.4. Optimal Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

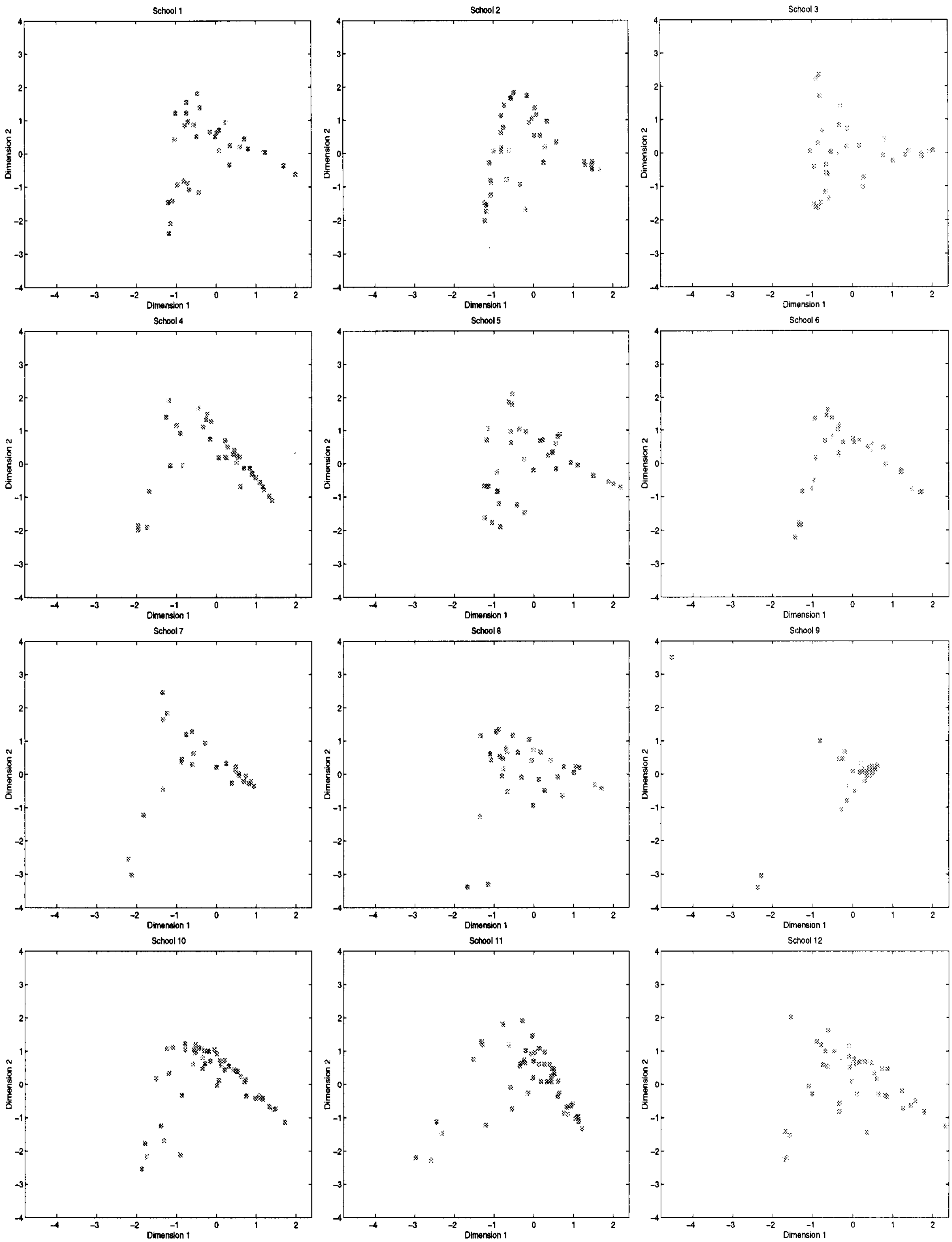

FIGURE 4.5. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

5. 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 4.

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 α_{jk} a p column vector of intercepts. The parameters $\alpha_{jk}(s),\ s = 1, \dots, p$ 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 [12].

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 $C_j(k, r) = 1,\ k = 1, \dots, K,\ r = 1, \dots, R_j$ (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 'supercluster', 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 supercluster 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 $H_j = C_j \bigotimes I_{l_j},\ j \in \mathbf{J}$. We can then write the Gifi loss function as

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\left(X - G_j H_j Z_j\right), \tag{5.1}$$

where $Z_j = [Z'_{j1}, \ldots, Z'_{jR_j}]'$. We employ the ALS algorithm to minimize (5.1) with respect to Z_j and X. Fixing first X we get

$$\hat{Z}_j = \left(H'_j D_j H_j\right)^{-1} H'_j G'_j X,\ j \in \mathbf{J}. \tag{5.2}$$

Some algebra shows that $Z_{j\mathcal{K}} = (\sum_{k \in \mathcal{K}} D_{jk})^{-1} \sum_{k \in \mathcal{K}} G'_{jk} X_k,\ \mathcal{K} \in \Gamma^j_{\mathbf{K}}$. Minimizing (5.1) with respect to X we get

$$\hat{X} = J^{-1} \sum_{j=1}^{J} G_j H_j Z_j. \tag{5.3}$$

Note than in case D_j^{-1} exists, we can also write (5.2) as

$$\hat{Z}_j = \left(H'_j D_j H_j\right)^{-1} H'_j D_j D_j^{-1} G'_j X = \left(H'_j D_j H_j\right)^{-1} H'_j D_j Y_j,\ j \in \mathbf{J}. \tag{5.4}$$

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 (5.3) equivalent to (4.4). In the presence of equality constraints on the category quantifications the ALS algorithm becomes: (i) estimate the restricted category quantifications using (5.2), (ii) calculate $Y_j^\star = H_j \hat{Z}_j$, (iii) estimate the object scores using (5.3), 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 $\tilde{Y}_{jk} = u\hat{\alpha}'_{jk} + Y^\star_{jk}$, where

$$\hat{\alpha}'_{jk} = -\left(u' D_{jk} Y^\star_{jk}\right)/n_k,\ j \in \mathbf{J},\ k \in \mathbf{K}. \tag{5.5}$$

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 apparatus is present 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 (5.4). Some algebra shows that the total discrimination measures of the unrestricted solution can also be written as

$$\eta_{js}^2 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s),\ j \in \mathbf{J},\ s = 1,\dots,p, \tag{5.6}$$

while the cluster discrimination measures as

$$\eta_{jks}^2 = \frac{1}{n_k}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s),\ j \in \mathbf{J},\ k \in \mathbf{K},\ s = 1,\dots,p. \tag{5.7}$$

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

$$\begin{aligned}\tilde{\eta}_{js}^2 &= \frac{1}{N}\text{tr}(\tilde{Y}_j' D_j \tilde{Y}_j) = \frac{1}{N}\sum_{k=1}^{K}\text{tr}(\tilde{Y}_{jk}' D_{jk}\tilde{Y}_{jk}) = \frac{1}{N}\sum_{k=1}^{K}\text{tr}(\tilde{Y}_{jk}' D_{jk}(u\hat{\alpha}_{jk}' + Y_{jk}^\star)) = \\ &\frac{1}{N}\sum_{k=1}^{K}\text{tr}(\tilde{Y}_{jk}' D_{jk}u\hat{\alpha}_{jk}' + \tilde{Y}_{jk}' D_{jk}Y_{jk}^\star) = \frac{1}{N}\sum_{k=1}^{K}(Y_{jk}^{\star\prime} D_{jk}Y_{jk}^\star + \hat{\alpha}_{jk}u' D_{jk}Y_{jk}^\star) = \\ &\frac{1}{N}\sum_{k=1}^{K}\text{tr}(Y_{jk}^{\star\prime} D_{jk}Y_{jk}^\star - n_k\hat{\alpha}_{jk}\hat{\alpha}_{jk}') = \frac{1}{N}\text{tr}(Y_j^{\star\prime} D_j Y_j^\star) - \frac{1}{N}\sum_{k=1}^{K} n_k \text{tr}(\hat{\alpha}_{jk}\hat{\alpha}_{jk}').\end{aligned} \tag{5.8}$$

But we also have that (using (5.4))

$$\begin{aligned}\frac{1}{N}\text{tr}(Y_j^{\star\prime} D_j Y_j^\star) &= \frac{1}{N}\text{tr}(Z_j' H_j' D_j H_j Z_j) = \frac{1}{N}\text{tr}(Y_j' D_j H_j (H_j' D_j H_j)^{-1} H_j' D_j Y_j) = \\ &\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)\Big(\sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}(i,s)\Big)^2.\end{aligned} \tag{5.9}$$

An application of Jensen's inequality gives that

$$\Big(\sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}(i,s)\Big)^2 \le \sum_{k=1}^{K}\frac{D_{jk}(i,i)}{\sum_{k=1}^{K} D_{jk}(i,i)}Y_{jk}^2(i,s). \tag{5.10}$$

Combining relations (5.8), (5.9) and (5.10) and after some algebra we get that

$$\tilde{\eta}_{js}^2 \le \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{\ell_j} D_{jk}(i,i)Y_{jk}^2(i,s) = \eta_{js}^2. \tag{5.11}$$

Therefore, we also get

$$\tilde{\gamma}_s = \frac{1}{J}\sum_{j=1}^{J}\tilde{\eta}_{js}^2 \le \frac{1}{J}\sum_{j=1}^{J}\eta_{js}^2 = \gamma_s. \tag{5.12}$$

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 (5.10)). However, nothing can be said on the effect of the restrictions on the cluster discrimination measures and fit.

Remark 5.1. *Analysis of Panel Data.* van der Heijden and de Leeuw [14] used correspondence analysis techniques to examine panel data. If in our approach we take $n_k = n$ for every $k \in \mathbf{K}$, $\Gamma^j_{\mathbf{K}} \equiv \{1\}$ 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 [14] 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 5.2. *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 (5.2), while in the second by (2.2). Notice that in general we have $G_j'^{\mathcal{K}} X^{\mathcal{K}} \neq \sum_{k \in \mathcal{K}} G'_{jk} X_k$, which 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 5.3. *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.

5.1. **Equality Constraints in the Princals Model.** The starting point is to impose rank-one restrictions to the constrained category quantification matrices $Z_{j\mathcal{K}}$. We then have

$$Z_{j\mathcal{K}} = v_{j\mathcal{K}} \theta'_{j\mathcal{K}},\ j \in \mathbf{J},\ \mathcal{K} \in \Gamma^j_{\mathbf{K}}, \tag{5.13}$$

where $v_{j\mathcal{K}}$ is a ℓ_j column vector of single constrained category quantifications and $theta_{\mathcal{K}}$ a p column vector of component loadings. Following analogous steps as in the estimation of the multilevel Princals model, it is easy to see that we also have to minimize with respect to $v_{j\mathcal{K}}$

and $\theta_{j\mathcal{K}}$

$$\sum_{j=1}^{J} \sum_{\mathcal{K}\in\Gamma_{\mathbf{K}}^{j}} \operatorname{tr}\big(v_{j\mathcal{K}}\theta'_{j\mathcal{K}} - \hat{Z}_{j\mathcal{K}}\big)' D_{j\mathcal{K}} \big(v_{j\mathcal{K}}\theta'_{j\mathcal{K}} - \hat{Z}_{j\mathcal{K}}\big), \tag{5.14}$$

where $D_{j\mathcal{K}} \equiv \sum_{k\in\mathcal{K}} D_{jk}$. We employ once again another ALS loop (alternate over $v_{j\mathcal{K}}$ and $\theta_{j\mathcal{K}}$). Solving (5.14) for fixed $v_{j\mathcal{K}}$ we get

$$\hat{\theta}_{j\mathcal{K}} = \big(\hat{Z}'_{j\mathcal{K}} D_{j\mathcal{K}} v_{j\mathcal{K}}\big)/\xi_{j\mathcal{K}},\ j\in\mathbf{J},\ \mathcal{K}\in\Gamma_{\mathbf{K}}^{j}, \tag{5.15}$$

where $\xi_{j\mathcal{K}} = v'_{j\mathcal{K}} D_{j\mathcal{K}} v_{j\mathcal{K}}$ is a scalar. Solving (5.14) for fixed $\theta_{j\mathcal{K}}$ we get

$$\hat{v}_{j\mathcal{K}} = \big(\hat{Z}_{j\mathcal{K}}\theta_{j\mathcal{K}}\big)/\psi_{j\mathcal{K}},\ j\in\mathbf{J},\ \mathcal{K}\in\Gamma_{\mathbf{K}}^{j}, \tag{5.16}$$

where $\psi_{j\mathcal{K}} = \theta'_{j\mathcal{K}}\theta_{j\mathcal{K}}$ is a scalar.

The single component of the loss function is given by

$$\sum_{j=1}^{J} \sum_{\mathcal{K}\in\Gamma_{\mathbf{K}}^{j}} \operatorname{tr}\big(\hat{Z}'_{j\mathcal{K}} D_{j\mathcal{K}} \hat{Z}_{j\mathcal{K}} - n_{\mathcal{K}}\theta_{j\mathcal{K}}\theta'_{j\mathcal{K}}\big), \tag{5.17}$$

where $n_{\mathcal{K}} \equiv \sum_{k\in\mathcal{K}} n_k$.

It is worth noting that in this case there is no analogous expression to (5.4) linking the single constrained category quantifications to the unconstrained ones. This happens because the rank-one restrictions are imposed on the $Z_{j\mathcal{K}}$'s, that are second-level parameters.

5.2. **NELS:88 Example (continued).** We continue with the example discussed in Section 3. The unrestricted Homals solution revealed some common patterns among the schools; 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 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 $a = A, B, C$ and $b = H, I, K$ respectively. The first set of variables was selected after a close examination of the variable transformation plot (Figure 4.3) and the category quantification plot (Figure 4.4). 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 5.1). There is only some small loss in the fit (as expected from (5.12)). 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 5.1). As before, most variables discriminate equally well on both dimensions (Figure 5.2), 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 5.3. 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.

The main advantage of the constrained solution can be seen in Figures 5.4 and 5.5 (both Figures use the same scale). First, observe that all the graphs are nicely centered. Second, by

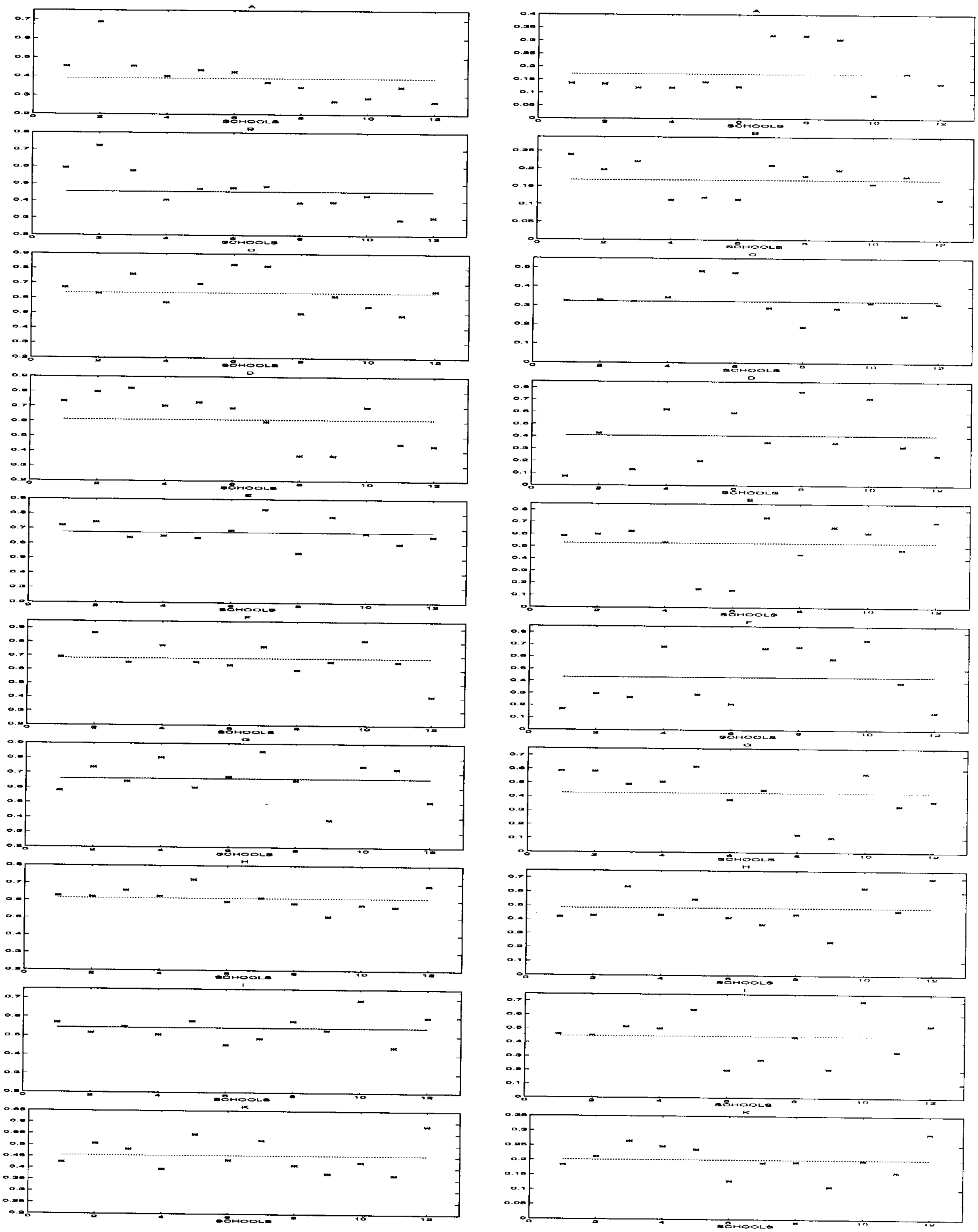

FIGURE 5.1. Discrimination Measures of the Variables for the Schools; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12; the solid line represents the variable's overall discrimination measure (Left: dimension 1, Right: dimension 2).

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 5.1. School Eigenvalues

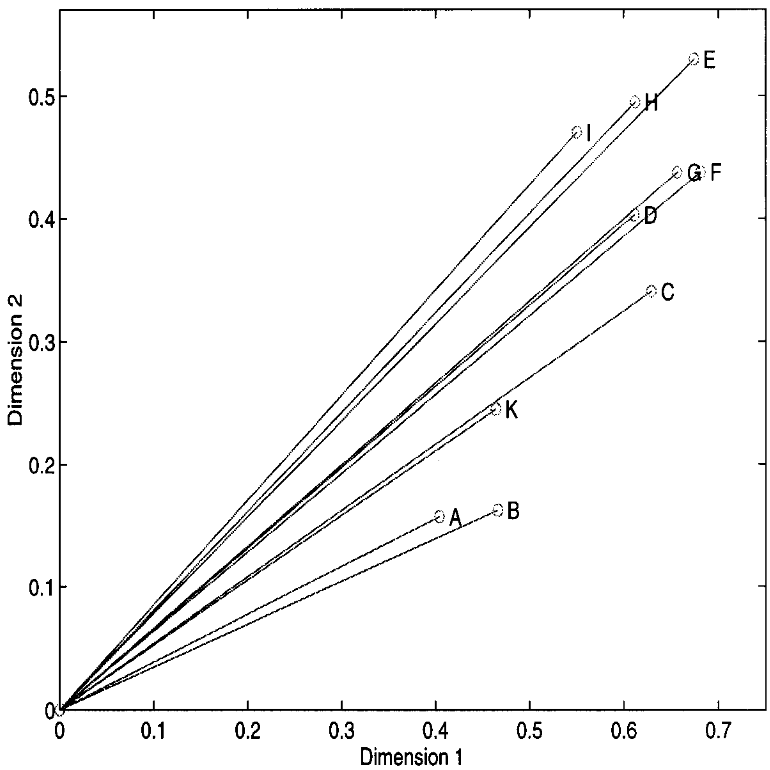

FIGURE 5.2. Total Discrimination Measures

examining the object scores, we immediately notice that the outliers in schools 7, 8 an 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

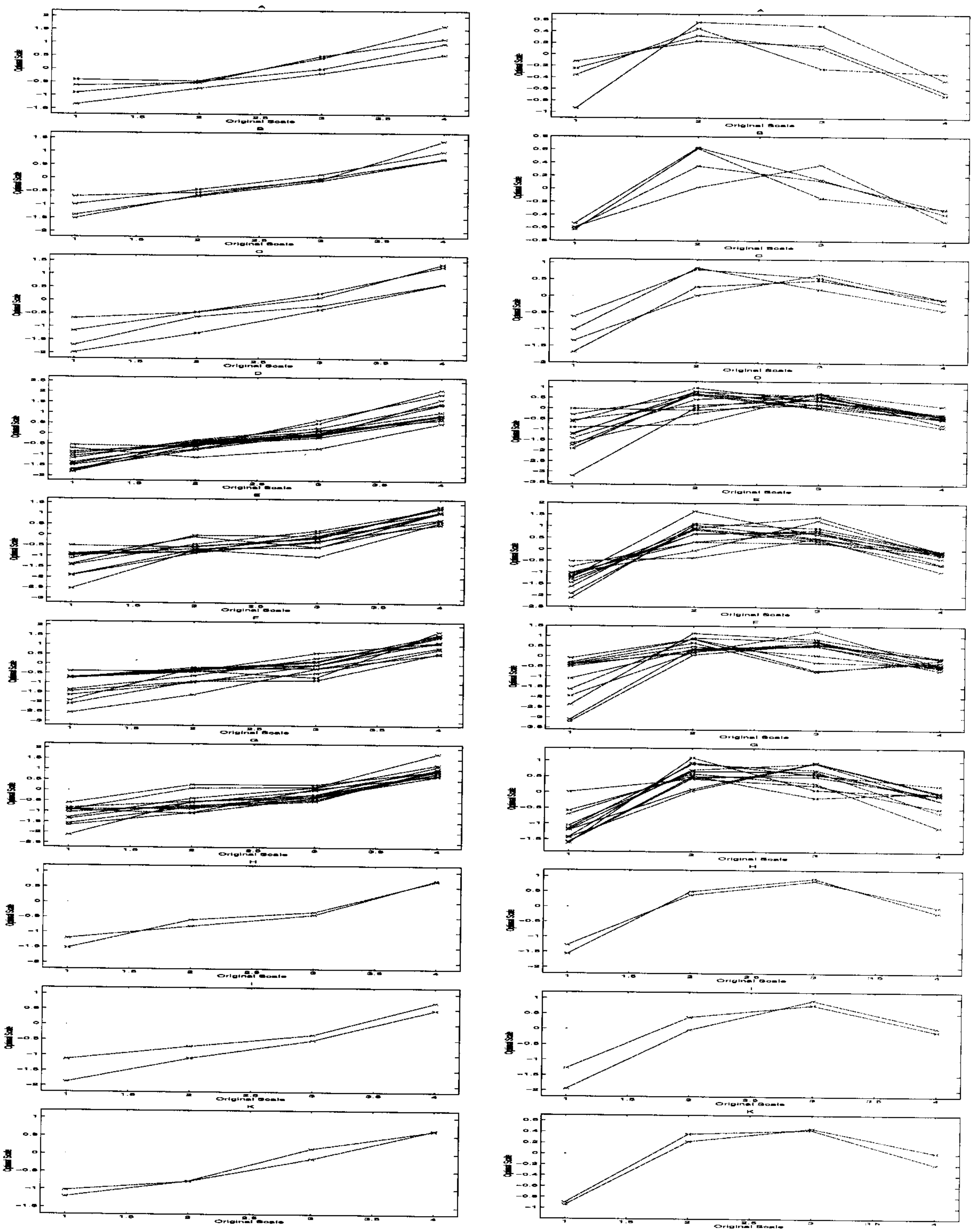

FIGURE 5.3. Optimal Transformations of the Variables (Left: dimension 1, Right: dimension 2)

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 quantification 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 of the public rural schools together through the constraints, there seems to remain enough variation in their response patterns. 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 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. 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 (teacher) has to undertake is which group of schools (public urban, etc.) to attend, since the within group school differences appear to be small.

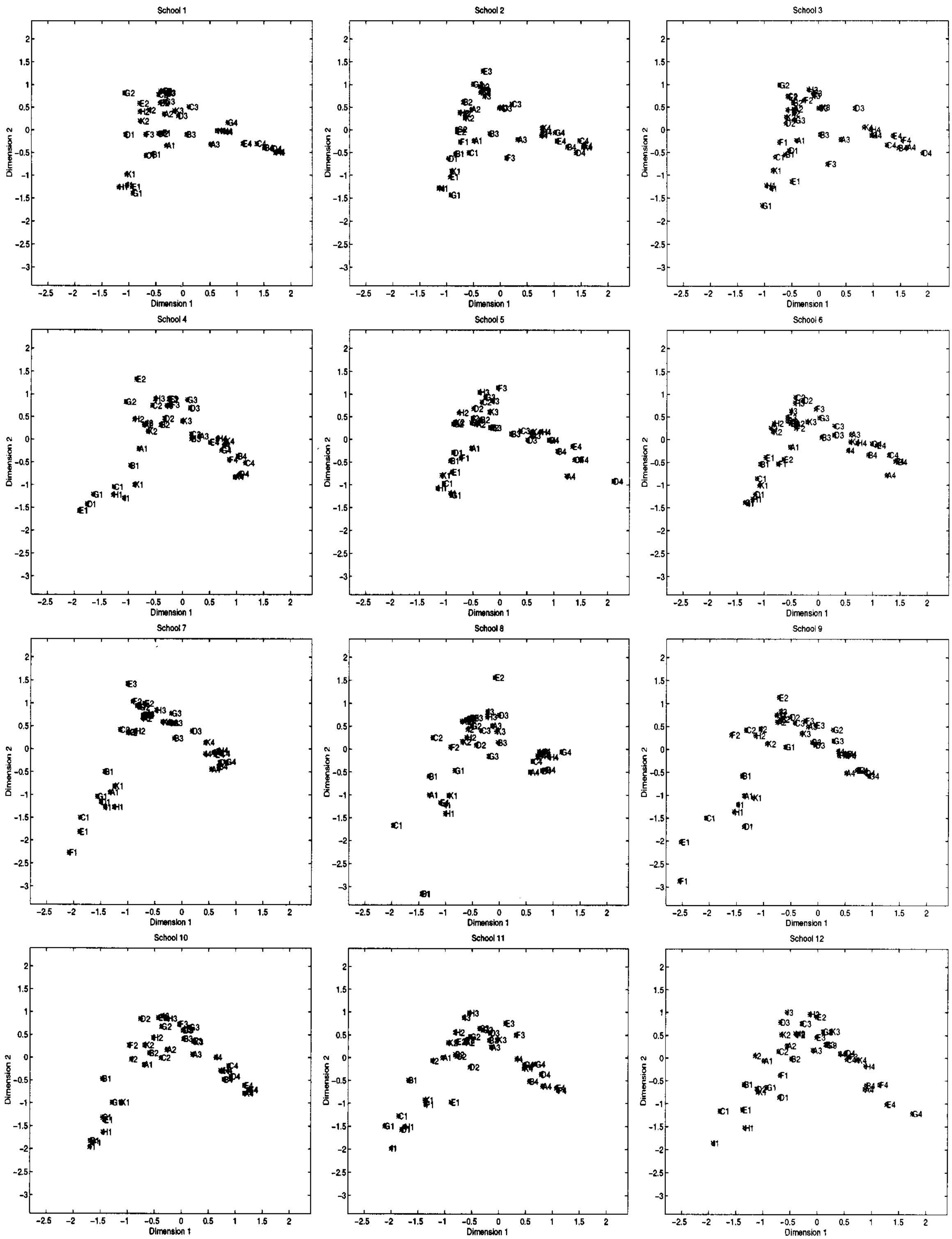

FIGURE 5.4. Optimal Constrained Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

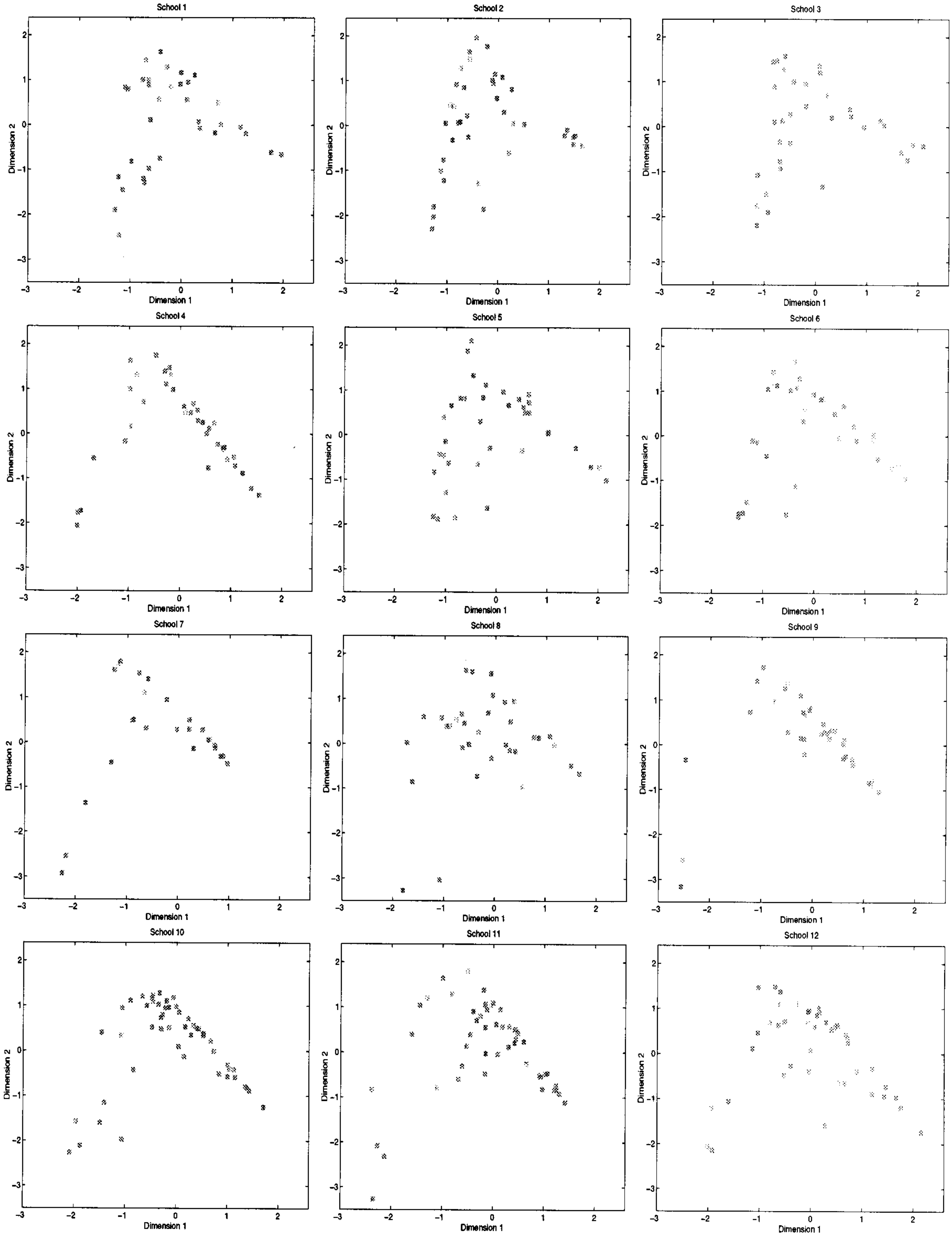

FIGURE 5.5. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

6. More General Coding Schemes for the Constraint Matrices

The use of the constraint matrices C_j (see Section 4) 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" coding $(0, 1)$ 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 [19, 20] 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 $\mathbf{K} \rightarrow \Gamma^j_{\mathbf{K}}$. Its entries satisfy the restriction

$$\sum_{r=1}^{R_j} C(k,r) = 1,\ k \in \mathbf{K}, \tag{6.1}$$

where R_j denotes the cardinality of the set $\Gamma^j_{\mathbf{K}}$. The restriction (6.1) 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 Gifi loss function can be written as

$$\sigma\big(X; Z_1, \ldots, Z_J\big) = J^{-1} \sum_{j=1}^{Z} \text{SSQ}\left(X - G_j H_j Z_j\right), \tag{6.2}$$

where $Z_j = [Z'_{j1}, \ldots, Z'_{jR_j}]'$. We employ the ALS algorithm to minimize (6.2) with respect to Z_j and X. Minimizing (6.2) for fixed X we get

$$\hat{Z}_j = \left(H'_j D_j H_j\right)^{-1} H'_j G'_j X, \; j \in \mathbf{J}, \tag{6.3}$$

while minimizing (6.2) with respect to X for fixed Z_j's we get

$$\hat{X} = J^{-1} \sum_{j=1}^{J} G_j H_j Z_j. \tag{6.4}$$

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. The following Lemma resolves the issue.

Lemma 6.1. *If A is a real positive definite matrix of order n and B a $n \times k$ real matrix of rank k, then $B'AB$ is also a real positive definite matrix of order k.*

Let $y = Bx$ then $y \neq 0$ and $y'Ay = y'B'ABy > 0$

Proof: We argue by contradiction. Suppose that $B'AB$ is not a positive definite matrix. The latter implies that for all real k-column vectors $x \neq 0$ (i.e. all their elements are not identically zero), we must have $x'B'ABx = (Bx)'A(Bx) \leq 0$. Therefore we get $Bx = 0$, and since B is of full column rank, we have that $x = 0$. ∎

The above lemma applied to our case implies 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 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 (5.4))

$$\hat{Z}_j = \left(H'_j D_j H_j\right)^{-1} H'_j D_j D_j^{-1} G'_j X = \left(H'_j D_j H_j\right) H'_j D_j Y_j, \; j \in \mathbf{J}. \tag{6.5}$$

The latter implies that element $Y_{jk}(t,s),\ k \in \mathcal{K}$ participates with a weight $C(k,\mathcal{K})D_{jk}(t,s)/(H_j'D_jH_j)^{-1}(t,t)$ 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 $H_j'D_jH_j$ matrix.

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.

6.1. **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. 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 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 $a = A - G$. 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 (1,2,2) are more evenly distributed over the three income categories. 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 restricting all the variables. All schools experience a loss in fit, but schools exhibiting a good fit in the previous solution continue to do so (see Table 6.1).

The category quantifications and object scores plots are given in Figures 6.1 and 6.2 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

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 6.1. School Eigenvalues

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. 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.

7. Concluding Remarks

REFERENCES

[1] Guttman, L. (1941), "The Quantification of a Class of Attributes: A Theory and a Method of Scale Construction," *The Prediction of Personal Adjusment*, Horst et al. (eds.), New York: Social Science Research Council

[2] Guttman, L. (1950), "The Principal Components of Scale Analysis," *Mesurement and Prediction*, Stouffer (ed.), Princeton: Princeton University Press

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

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

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

[7] 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

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

[9] 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

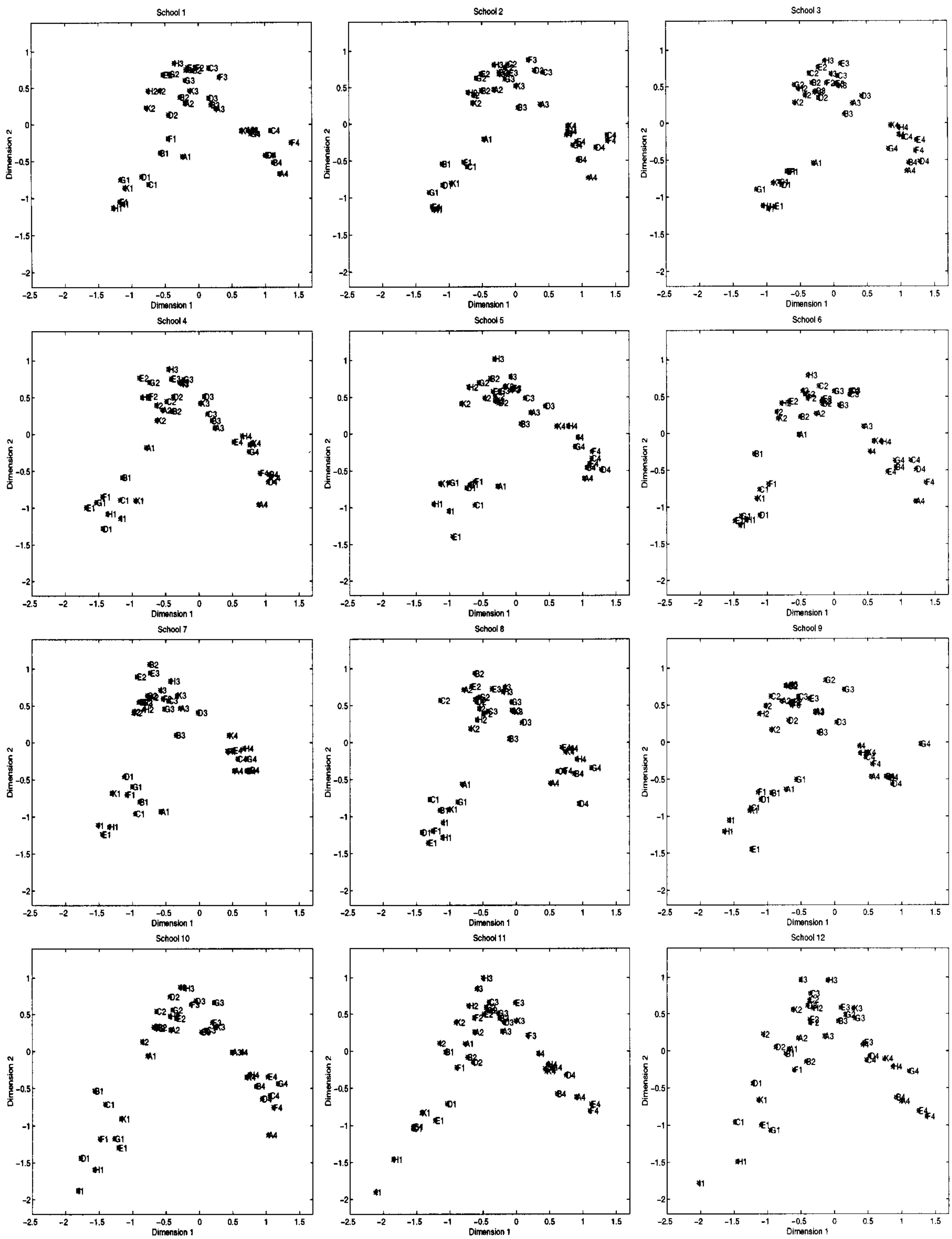

FIGURE 6.1. Optimal Constrained Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

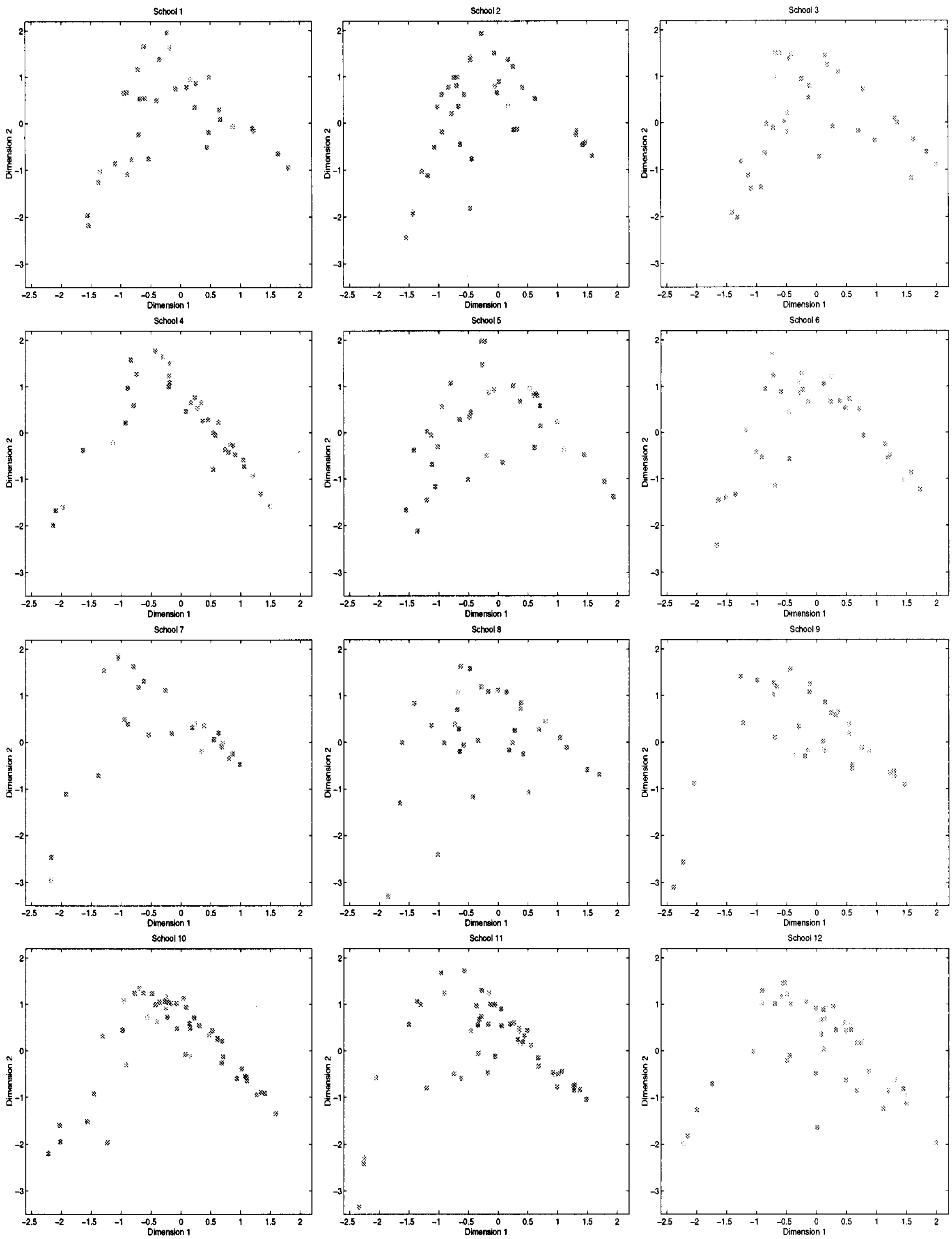

FIGURE 6.2. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

[10] Young, F.W. (1981), "Quantitative Analysis of Qualitative Data," *Pshychometrika*, **46**, 357-388
[11] 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
[12] 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
[13] de Leeuw, J., van der Heijden, P., and Kreft, I. (1985), "Homogeneity Analysis of Event History Data", *Methods of Operations Research*, 50, 299-316
[14] van der Heijden, P., and de Leeuw, J. (1987?), "Correspondence Analysis, with Special Attention to the Analysis of Panel Data and Event History Data,"
[15] Benzécri, J.P. (1973), *Analyse des Données*, Paris: Dunod
[16] Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press
[17] Golub, G.H. and van Loan C.F. (1989), *Matrix Computations*, Baltimore: Johns Hopkins University Press
[18] de Leeuw, J. (1988), "Models and Techniques," *Statistica Neerlandica*, **42**, 91-98
[19] Greenacre, M. and Hastie, T. (1987), "The Geometric Interpretation of Correspondence Analysis," *Journal of the American Statistical Association*, **82**, 437-447
[20] van Rijckevorsel, J.L.A. (1987), *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*, Leiden: DSWO Press
[21] Oakes, J. (1989), "What Educational Indicators? The Case for Assessing the School Context," *Educational Evaluation and Policy Analysis*, **11**, 181-199
[22] Anderson, C.S. (1982), "The Search for School Climate: A Review of the Research," *Review of Educational Research*, **52**, 368-420
[23] Molenaar, I.W. (1988), "Formal Statistics and Informal Data Analysis, or why Laziness Should Be Discouraged," *Statistica Neerlandica*, **42**, 83-90

UCLA Center for Statistics, Los Angeles, CA 90095

E-mail address: gmichail@stat.ucla.edu

UCLA Center for Statistics, Los Angeles, CA 90095

E-mail address: deleeuw@stat.ucla.edu