Multilevel Homogeneity Analysis

George Michailidis Jan de Leeuw

Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095

Abstract

We consider a multilevel sampling design framework, where we collect observations on N individual cases grouped (clustered) within K units (clusters). Homogeneity analysis and principal component analysis techniques with rank restrictions have been used successfully in studying the interdependence of sets of mixed measurement (nominal, ordinal and numerical) categorical variables. In this paper, we extend these techniques to the multilevel sampling framework. We also propose some new 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 form the National Educational Longitudinal Study (NELS:88) is used to illustrate some of the techniques presented in the paper.

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.

A system that analyzes categorical data is the Gifi system [5]. The technique of homogeneity analysis represents the cornerstone of the system. 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 is on the geometry of the solution. We have a data set comprised of N observations (objects, individuals) and J categorical variables, with $\ell_j,\ j \in \mathbf{J} = \{1, 2, \ldots, J\}$ categories per variable. In the Gifi system categorical variables are coded by using *indicator* matrices G_j, with entries $G_j(i,t) = 1,\ i = 1, \ldots, N,\ t = 1, \ldots, \ell_j$ if object i belongs to category t, and $G_j(i,t) = 0$ if it belongs to some other category. Since, in this approach we take into account only the fact that some objects are in a particular category while others are in different ones, the treatment of the variables is called *nominal.*

Remark 1.1 *Notation.* In the remainder of the paper we employ the following notational conventions. Upper case letters are used for matrices (e.g. A), and lower case letters for vectors (e.g. a). The $(s,t)^{th}$ element of a matrix is denoted by $A(s,t)$, the s^{th} row by $A(s,.)$ and the t^{th} column by $A(.,t)$. Analogously, the s^{th} element of a vector is denoted by $a(s)$. Finally, u denotes the *unit* vector (vector comprised of only ones), and I_p the identity matrix of order p.

The Gifi loss function is given by

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

$$J^{-1} \sum_{j=1}^{J} \text{tr}(X - G_j Y_j)'(X - G_j Y_j), \quad (1.1)$$

where SSQ (H) denotes the sum of squares of the elements of the matrix H. 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

$$X'X = NI_p, \quad u'X = 0. \quad (1.2)$$

The elements of the X matrix are called the *object scores*, and those of the Y_j matrices the *category quantifications*. Under the loss function (1.1) the difference among the transformed variables G_jY_j is measured by the mean squared distance to one hypothetical (latent) variable X. By definition, perfect consistency exists and consequently zero loss is obtained if

$$X = G_1Y_1 = \ldots = G_jY_j = \ldots = G_JY_J, \qquad (1.3)$$

that is, each linear combination G_jY_j is identical to the common space X. In this case the object scores are perfectly *discriminating* and the category quantifications are perfectly *homogeneous*. In the non-perfect case, the Gifi loss function (1.1) can be minimized by means of an *Alternating Least Squares* (ALS) algorithm, that has the following steps.

Step 1: Minimize (1.1) with respect to Y_j for fixed X. The optimal category quantifications are given by $\hat{Y}_j = D_j^{-1}G_j'X, j \in \mathbf{J}$, where D_j is a diagonal matrix conatining the univariate marginals of variable j.

Step 2: Minimize (1.1) with respect to X for Y_j's. The optimal object scores are given by $\hat{X} = J^{-1}\sum_{j=1}^{J} G_j'Y_j$.

Step 3: Column center and orthonormalize the matrix X, so that the normalization constraints (1.2) are satisfied.

These steps are repeated until the algorithm coverges to the global minimum (see chapter 3 in [5]). Hence, the ALS algorithm finds the desired solution to the problem given in (1.1), in the presence of nominal data.

Once the ALS algorithm has converged, the Gifi loss function can be written as (after some algebra)

$$J^{-1}\sum_{j=1}^{J}\text{tr}(X - G_jY_j)'(X - G_jY_j) = \qquad (1.4)$$
$$Np - J^{-1}\sum_{j=1}^{J}\text{tr}(Y_j'D_jY_j).$$

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* of variable j in dimension s are given by

$$\eta_{js}^2 \equiv Y_j'(.,s)D_jY_j(.,s)/N, \;\; j \in \mathbf{J}, \;\; s = 1, \ldots, p. \qquad (1.5)$$

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 $G_jY_j(.,s)$ in dimension s, and the corresponding column of object scores $X(.,s)$ (see chapter 3 in [5]). Hence, the loss function can also be expressed as

$$N(p - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2) = N(p - \sum_{s=1}^{p}\gamma_s), \qquad (1.6)$$

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

We summarize next some basic properties of the Homals solution.

◇ Category quantifications and object scores are represented in a joint space.

◇ A category point is the centroid of objects belonging to that category (see Step 1).

◇ Objects with the same response pattern (identical profiles) receive identical object scores (see Step 3). 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 (1.5)).

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

◇ Category points with low marginal frequencies will be located further away from the origin of the joint space, whereas categories with high marginal frequencies will be located closer to the origin (see Step 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 employed normalization of the object scores , since $u'D_jY_j = u'D_jD_j^{-1}G_j'X = u'G_j'X = u'X = 0$.

◇ The solution is *invariant* under *rotations* of the object scores and of the category quantifications. To see this, suppose we select a different basis for the column space of the object scores X; that is, let $X^\sharp = X \times R$, where R is a rotation matrix satisfying $R'R = RR' = I_p$. We then get from Step 2 that $Y_j^\sharp = D_j^{-1}G_j'X^\sharp = \hat{Y}_jR$. Thus, the axes of the joint space can not be uniquely identified.

1.1 Nonlinear Principal Components Analysis

Principal Components Analysis (PCA) attempts to replace the J coordinates of each row of the data matrix by $p << J$ new coordinates, while retaining as much as possible information contained in the original data. Moreover, in the presence of categorical variables, an optimal transformation of the variables must also be incorporated in the procedure. In the Gifi system nonlinear PCA is derived as homogeneity analysis with restrictions [3].

The starting point for this derivation is the ordinary loss function given in (1.1). However, *rank-one restrictions* of the form

$$Y_j = q_j\beta_j', \ \ j \in \mathbf{J}, \qquad (1.7)$$

are imposed on the multiple category quantifications, with q_j being a ℓ_j-column vector of *single* category quantifications for variable j, and β_j a p-column vector of weights (component loadings). Thus, each quantification matrix Y_j is restricted to be of rank-one, which implies that the quantifications in p dimensional space become proportional to each other. The introduction of the rank-one restrictions allows the existence of multidimensional solutions for object scores with a single quantification (optimal scaling) for the categories of the variables, and also makes it possible to incorporate the measurement level of the variables (ordinal, numerical) into the analysis.

To minimize (1.1) under the restriction (1.7), we start by computing the $\hat{Y}_j$'s as in Step 2 of the ALS algorithm. We then partition the Gifi loss function as follows:

$$\sum_{j=1}^{J} \text{tr}(X-G_j[\hat{Y}_j+(Y_j-\hat{Y}_j)])'(X-G_j[\hat{Y}_j+(Y_j-\hat{Y}_j)]) = \sum_{j=1}^{J} \text{tr}(X-G_j\hat{Y}_j)'(X-G_j\hat{Y}_j)+\sum_{j=1}^{J} \text{tr}(Y_j-\hat{Y}_j)'D_j(Y_j-\hat{Y}_j). \qquad (1.8)$$

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

$$\sum_{j=1}^{J} \text{tr}(q_j\beta_j' - \hat{Y}_j)'D_j(q_j\beta_j' - \hat{Y}_j), \qquad (1.9)$$

with respect to q_j and β_j. This is achieved by an inner ALS loop comprised of the following steps:

Step 1: Minimize (1.9) with respect to β_j for fixed q_j. The optimal component loadings are given by $\hat{\beta}_j = \hat{Y}_j'D_jq_j/q_j'D_jq_j$, $j \in \mathbf{J}$.

Step 2: Minimize (1.9) with respect to q_j for fixed β_j. The optimal single category quantifications are given by $\hat{q}_j = \hat{Y}_j\beta_j/\beta_j'\beta_j$, $j \in \mathbf{J}$.

Step 3: Incorporate the measurement level of the variables. This becomes a monotone regression problem in the *ordinal* case, a linear regression problem in the *numerical* case, and simply keeping $\hat{q}_j$ in the nominal case.

The first part of (1.8) is called *multiple loss.* The second part is called *single loss* and can be written, after imposing the normalization constraint $q_j'D_jq_j = N, \ j \in \mathbf{J}$, as

$$\sum_{j=1}^{J} \text{tr}(\hat{Y}_j'D_j\hat{Y}_j - N\beta_j\beta_j'), \qquad (1.10)$$

The quantities given by $\beta_{js}^2, \ s = 1, \ldots, p$ are called *single fit.*

2 Multilevel Modeling

In many situations individual objects can be naturally grouped (*clustered*) into groups (clusters). 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, $k \in \mathbf{K} = \{1, \ldots, K\}(\sum_{k=1}^{K} n_k = N)$. Once again, we want to examine J categorical variables, with $\ell_j, \ j \in \mathbf{J}$ categories each. In this chapter we pursue two goals. The 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 related 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 [4] and van der Heijden and de Leeuw [8] have used these techniques to examine panel and event history data. In this 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, use interactive coding to reduce them to two-way (ordinary) matrices, and apply homogeneity analysis to the collection of such matrices. More recently, Carlier and Kroonenberg [1] apply the PARAFAC model [6, 2] to the three-way matrices. Both approaches are 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 $G_j = \bigotimes_{k=1}^{K} G_{jk}$ 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 and the object scores, we are able to generate interesting and useful models and incorporate prior knowledge. It is also shown that the approach taken in [4] and [8] can be derived from our framework. In this case the Gifi loss function becomes

$$\begin{aligned}\sigma(X; Y_1, \dots, Y_J) &= J^{-1} \sum_{j=1}^{J} \mathrm{SSQ}\left(X - G_j Y_j\right) \qquad (2.1)\\ &= J^{-1} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathrm{SSQ}\left(X_k - G_{jk} Y_{jk}\right).\end{aligned}$$

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

$$X_k' X_k = n_k I_p, \quad u' X_k = 0, \text{ for every } k \in \mathbf{K}. \qquad (2.2)$$

The problem in (2.1) is identical to the one presented in (1.1); thus, the category quantifications are given by $\hat{Y}_j = D_j^{-1} G_j' X,\ j \in \mathbf{J}$, 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 $\hat{Y}_{jk} = D_{jk}^{-1} G_{jk}' X_k,\ j \in \mathbf{J},\ k \in \mathbf{K}$. Similarly, the object scores are given by $\hat{X} = \frac{1}{J} \sum_{j=1}^{J} G_j Y_j$, which gives that $\hat{X}_k = J^{-1} \sum_{j=1}^{J} G_{jk} Y_{jk}$, for every $k \in \mathbf{K}$.

We define next the *cluster discrimination measures*

$$\eta_{jks}^2 \equiv Y_{jk}'(.,s) D_{jk} Y_{jk}(.,s)/n_k, \qquad j \in \mathbf{J},\ k \in \mathbf{K},\ s = 1, \dots, p, \qquad (2.3)$$

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

$$\eta_{js}^2 \equiv Y_j'(.,s) D_j Y_j(.,s)/N, \quad j \in \mathbf{J}, \quad s = 1, \dots, p. \qquad (2.4)$$

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 $\eta_{js}^2 = \frac{1}{N} \sum_{k=1}^{K} Y_{jk}'(.,s) D_{jk} Y_{jk}(.,s)$, so, it easy to see that

$$\eta_{js}^2 = \frac{1}{N} \sum_{k=1}^{K} n_k \eta_{jks}^2, \quad j \in \mathbf{J}, \quad s = 1, \dots, p. \qquad (2.5)$$

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. It can be seen that 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.

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 1) to each of the K clusters separately.

2.1 Multilevel Nonlinear Principal Components Analysis

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 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 component loadings and the single category quantifications are given by $\hat{\beta}_{jk} = \hat{Y}_{jk}' D_{jk} q_{jk})/q_{jk}' D_{jk} q_{jk}$ and $\hat{q}_{jk} = \hat{Y}_{jk} \beta_{jk}/\beta_{jk}' \beta_{jk},\ k \in \mathbf{K},\ j \in \mathbf{J}$ respectively. This is a direct extension of the single cluster case.

In this case the single loss is given by

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \mathrm{tr}\left(\hat{Y}_{jk}' D_{jk} \hat{Y}_{jk} - n_k \beta_{jk} \beta_{jk}'\right), \qquad (2.6)$$

The quantities given by β^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.*

2.2 Model Restrictions

We are interested in generating alternative models that allow us to fully utilize the multilevel structure of the data. We think of the category quantifications as transformations. Thus, they can be plotted against the category number for each triple (j, k, s), $j \in \mathbf{J}$, $k \in \mathbf{K}$, $s = 1, \ldots, p$. We write $y_{jks}(t)$, $t = 1, \ldots, \ell_j$. The models we study are defined by various restrictions on these transformations. Two main families of models are proposed in what follows.

A. *Restrictions on the Category Quantifications.*

1. In this model, we impose *equality restrictions* on the category quantifications between clusters. Let $\Gamma^j_{\mathbf{K}}$ denote a partition of the clusters $k \in \mathbf{K}$ for variable j. We then require $\tilde{y}_{jk}(t, s) = \alpha_{jk}(s) + z_{j\mathcal{K}}(t, s)$, $t = 1, \ldots, l_j$ or in matrix form $\tilde{Y}_{jk} = u\alpha'_{jk} + Z_{j\mathcal{K}}$, $j \in \mathcal{J}$, $k \in \mathcal{K}$, $\mathcal{K} \in \Gamma_{\mathbf{K}}$, where $Z^{\mathcal{K}}_j$ is the $\ell_j \times p$ matrix of *restricted* category quantifications for clusters $k \in \mathcal{K}$. The parameters α_{jks} are employed to ensure that the $\tilde{y}_{jk}(t, s)$'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 [3]. For the princpal components model we require $\tilde{Y}_{jk} = u'\alpha'_{jk} + q_{j\mathcal{K}}\beta'_{j\mathcal{K}}$, $k \in \mathcal{K}$, $\mathcal{K} \in \Gamma^j_{\mathbf{K}}$.

2. In this model, we impose *linear restrictions* on the category quantifications between clusters. We require $\tilde{y}_{jk}(t, s) = \alpha_{jk}(s) + \sum_{r=1}^{R_j} w_{kr} z_{jr}(t, s)$, $t = 1, \ldots, l_j$, where R_j denotes the number of restrictions, and w_{kr} *given* cluster weights.

B. *Regression Model.* We require that $\tilde{y}_{jk}(t, s) = \alpha_{jk}(s) + \beta_k(s) q_j(t)$, with additional optional restriction $\beta_k(s) = \beta^{\mathcal{K}}(s)$ for $k \in \mathcal{K}$, $\mathcal{K} \in \Gamma_{\mathbf{K}}$. This restriction can also be expressed using matrix notation as $\tilde{Y}_{jk} = u\alpha'_{jk} + Q_j B_k$, with Q_j, $j \in \mathbf{J}$ being a $\ell_j \times p$ matrix of category quantifications and B_k, $k \in \mathbf{K}$ a $p \times p$ diagonal matrix of *slopes*. In this case we fit the same model to all K clusters, but we allow different loadings for each cluster (different B_k's), so as to be able to examine group (cluster) differences. For the principal components models we require that $\tilde{y}_{jk}(t, s) = \alpha_{jk}(s) + \beta_k(s) q_j(t)$, or using matrix notation $\tilde{Y}_{jk} = u\alpha'_{jk} + q_j\beta'_k$. We can also have $\beta_{ks} = \beta^{\mathcal{K}}_s$, $k \in \mathcal{K}$, $\mathcal{K} \in \Gamma_{\mathbf{K}}$. All models deal with the category quantifications. The objective is twofold: examine differences between the clusters and reduce the number of model parameters to enhance the stability of the solution.

However, we can also consider *equality restrictions* on the *object scores.* In case we deal with K sets of observations on the same n individuals (panel data), we can require $X_k = \mathcal{X}$ for every $k \in \mathbf{K}$, and the normalization restriction $u'\mathcal{X} = 0$, $\mathcal{X}'\mathcal{X} = nI_p$. This analysis will produce a single set of object scores but K different sets of category quantifications. The solution to this problem would correspond to the analysis of the BROAD matrix introduced by van der Heijden and de Leeuw in [8].

3 Equality Restrictions on the Category Quantifications

We begin by introducing the *constraint* matrix C_j, $j \in \mathbf{J}$ that maps $\mathbf{K} \rightarrow \Gamma^j_{\mathbf{K}}$. 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:

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

In the first example the 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. In the second 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

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

where $Z_j = [Z'_{j1}, \ldots, Z'_{jR_j}]'$. We employ the ALS algorithm to minimize (3.1) with respect to Z_j and X. Fixing first X we get $\hat{Z}_j = (H'_j D_j H_j)^{-1} H'_j G'_j X$, $j \in \mathbf{J}$. 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 (3.1) with respect to X we get $\hat{X} = J^{-1} \sum_{j=1}^{J} G_j H_j Z_j$. Note than in case D_j^{-1} exists, we can also write the restricted category quantifications as

$$\begin{aligned} \hat{Z}_j &= \left(H'_j D_j H_j\right)^{-1} H'_j D_j D_j^{-1} G'_j X \qquad (3.2) \\ &= \left(H'_j D_j H_j\right)^{-1} H'_j D_j Y_j, \ \ j \in \mathbf{J}. \end{aligned}$$

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 $D_{jk}(t, t) / \left(\sum_{k \in \mathcal{K}} D_{jk}(t, t)\right)$ 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 $k \in \mathcal{K}$. In the presence of equality constraints on the category quantifications the ALS algorithm becomes: (i) estimate the restricted category quantifications, (ii) calculate $Y_j^\star = H_j \hat{Z}_j$, (iii) estimate the object scores, 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

$$\hat{\alpha}'_{jk} = -(u' D_{jk} Y_{jk}^\star)/n_k,\ j \in \mathbf{J},\ k \in \mathbf{K}. \tag{3.3}$$

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.

The cluster and total discrimination measures are defined as before. Some algebra shows that the total discrimination measures of a resticted solution decrease compared to those of an unrestricted solution, which in turn impies that the fit of the restricted solution also decreases.

Remark 3.1 *Analysis of Panel Data.* Van der Heijden and de Leeuw [8] 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 [8] 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 *Linear Restrictions.* Fuzzy coding has been extensively used in multiple correspondence analysis [9] 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 objects, 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 the *combination* matrix C_j, $j \in \mathbf{J}$ that maps the set of clusters $\mathbf{K}$ into R_j, $j \in \mathbf{J}$ groups. Its entries satisfy the restriction

$$\sum_{r=1}^{R_j} C(k,r) = 1,\ \ k \in \mathbf{K}. \tag{3.4}$$

The restriction (3.4) 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. We then can proceed as before.

4 Regression Model

Equality restrictions on the category quantifications allow to reduce the number of parameters to be fitted. However, in many cases the possibilities are practically infinite. A more restrictive model is the regression model presented next, where the category quantifications are restricted across clusters, but they are weighted by different factors (the slope matrices B_k's) for each cluster. The model formally is given by

$$\tilde{Y}_{jk} = u'\alpha_{jk} + Q_j B_k,\ \ j \in \mathbf{J},\ \ k \in \mathbf{K}, \tag{4.1}$$

where the slope matrices B_k are required to be *diagonal*, and where α_{jk}, $j \in \mathbf{J}$, $k \in \mathbf{K}$ are parameters that ensure that the category quantifications have a weighted sum over categories equal to zero.

To minimize the usual Gifi loss function, we start by computing the $\hat{Y}_j$'s as before. We then partition the Gifi loss function as in (1.8), and after imposing the restriction on the Y_{jk}'s we have to minimize

$$J^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\text{tr}(Q_jB_k-\hat{Y}_{jk})'D_{jk}(Q_jB_k-\hat{Y}_{jk}), \qquad (4.2)$$

with respect to Q_j and B_k. This is done by ALS again (alternate over Q_j and B_k in the *inner* iteration loop). The steps of the inner loop are given next.

Step 1: Estimate the slope matrices by $\hat{B}_k = (\text{diag } W_k)^{-1}(\text{diag } (\sum_{j=1}^{J} Q_j' D_{jk}\hat{Y}_{jk})),\ k \in \mathbf{K}$.

Step 2: Estimate the overall category quantifications by $\hat{Q}_j(y,.) = S_j(t,.)\ (\sum_{k=1}^{K} D_{jk}(t,t)V_k)^{-1},\quad t = 1,\ldots,\ell_j,\ j \in \mathbf{J}$.

Step 3: Update the unrestricted multiple category quantifications by $\hat{Y}_{jk} = \hat{Q}_j\hat{B}_k,\ k \in \mathbf{K},\ j \in \mathbf{J}$.

Remark 4.1 *Absence of Rotational Invariance.* The solutions of the regression model are no longer rotationally invariant (contrary to the multilevel Homals solution). To see this, let R be a rotation matrix, and let $X^\sharp = X \times R$. We then get that $Y_{jk}^\sharp = D_{jk}^{-1}G_{jk}X^\sharp = \hat{Y}_{jk}R$. Thus, the normal equation we solve to calculate the category quantifications Q_j in Step 2 becomes $\sum_{k=1}^{K} D_{jk}Q_jB_kB_k' = \sum_{k=1}^{K} D_{jk}\hat{Y}_{jk}RB_k$. However, the matrices B_k and R do not in general commute (i.e. $RB_k \neq B_kR$), so that the matrix Q_j of category quantification is not rotational invariant. Under the regression model the axes become identified, and as a consequence of this we are able to look at more dimensions.

Remark 4.2 *The Relationship of the Regression Model to the INDSCAL-PARAFAC Model.* Suppose we collect the category quantification matrices Y_{jk} into a three-way array Z, where the categories represent the first dimension of the array, the dimensionality of the solution the second and the clusters the third dimension. In the psychometric literature, where these models originated, the dimensions of the array are called *modes*. For data structures of this form the following model has been suggested in the literature [6, 2]

$$Z(.,.,k) = \Phi\Delta_k\Psi',\ \ k \in \mathbf{K}, \qquad (4.3)$$

where $Z(.,.,k)$ represents one of the k slices of the three-way array Z, Φ is an $\ell_j \times s$ matrix of factor (components) loadings for the first mode, Ψ is an $s \times p$ matrix of factor loadings for the second mode and Δ_k is an $s \times s$ diagonal matrix of weights for each $k \in \mathbf{K}$. The elements of the Δ_k matrix step up or down the sizes of the columns of Φ (or, equivalently, the rows of Ψ'). Therefore, they represent the effect of the changes in the relative importance or influence of the s factors on cluster k. Models (4.3) is known as the Parallel Factors model (PARAFAC)-Individual Differences Scaling model (INDSCAL). It can be seen that the regression model can be thought of as a constrained version of the PARAFAC-INDSCAL model.

In the regression model the loss function is partitioned into two parts,

$$J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\text{tr}\big(X_k - G_{jk}\hat{Y}_{jk}\big)'\big(X_k - G_{jk}\hat{Y}_{jk}\big)' + \qquad (4.4)$$

$$J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\text{tr}\big(\hat{Q}_j\hat{B}_k - \hat{Y}_{jk}\big)'D_{jk}\big(\hat{Q}_j\hat{B}_k - \hat{Y}_{jk}\big).$$

The first part of (4.4) is called *multiple loss*, while the second part *regression loss* and corresponds to the additional loss incurred by imposing restrictions (4.1) to the category quantifications.

Finally, notice that in order for the Q_j's to be identified, we require $\sum_{j=1}^{J} Q_j'\tilde{D}_jQ_j = JNI_p,\ k \in \mathbf{K}$, where $\tilde{D}_j = \sum_{k=1}^{K} D_{jk}$ is the $\ell_j \times \ell_j$ diagonal matrix containing the univariate marginals of variable j for all K clusters combined.

The derivation of the results for the principal components regression model follows analogous steps.

5 NELS:88 Example

We employ a data set from the NELS:88 base year to demonstrate the multilevel homogeneity analysis technique with equality constraints on the category quantifications. The variables deal with student responses on the following problem areas in their schools: (A) Student tardiness, (B) Student absenteeism, (C) Students cutting class, (D) Physical conflicts among students, (E) Robbery or theft, (F) Vandalism of school property, (G) Student use of alcohol, (H) Student use of illegal drugs, (I) Student possession of weapons, and (K) Verbal abuse of teachers. 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 [7].

For this example we selected a sample of 12 schools out of a total 1,052 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. Schools 1, 2 and 3 are public urban, 4, 5 and 6 public suburban, 7, 8 and 9 public rural and 10, 11, 12 private schools.

We first ran an unrestricted solution (not presented here) which 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 a distorting representation of the basic patterns in the data. In order to overcome these shortcomings, we impose constraints on the category quantifications of some of the variables. In particular, variables A, B and C are constrained across the following four school groups (public urban, public suburban, public rural and private), while variables H, I and K are constrained across public and private schools. The remaining variables re left unrestricted.

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 solution and at the same time incorporate prior information.

A two-dimensional solution produced a satisfactory fit, with total eigenvalues .573 and .360 respectively. The category quantification plots and the object scores plot for each of the schools are given in Figures 1 and 2, respectively. In most of the schools, the category quantifications exhibit a quadratic pattern. In the lower left quadrant of the graph we find the 'serious problem' categories for the variables. 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). However, there are variations in this general pattern. For example, we have most category quantification points in the public rural schools being concentrated to the right of the graph (minor and not a problem categories), while the public urban schools are primarily distributed around the serious and moderate categories. 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 (compare schools 8 and 9 for example). The object scores plot reveals the presence of some outlying observations in the public rural schools. It also confirms our previous finding that the majority of the students in the public suburban and private schools responded 'not a problem' or a 'minor' problem, while the majority of the students in the public urban schools answered a 'moderate/serious' problem.

The constraints managed to 'filter' most of the 'noise' present in the unconstrained 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 the 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.

6 Concluding Remarks

In this paper we present a framework that bridges the gap between the Gifi system (and consequently multivariate analysis techniques) and hierarchical data structures. Various models that utilize the multilevel nature of the data and allow the data analyst to incorporate prior knowledge are introduced, and their properties and implementation presented. Some of these models present some interesting connections with multimode factor analysis models.

References

[1] Carlier, A. and Kroonenberg, P.M. (1996), "Decompositions and Biplots in Three-Way Correspondence Analy-

sis," *Psychometrika*, **61**, 355-373

[2] Carroll, J.D. and Chang, J.J. (1970), "Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization Eckart-Young Decomposition," *Psychometrika*, **35**, 283-319

[3] 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

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

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

[6] Harshman, R.A. (1970), "Foundations of the PARAFAC Procedure: Models and Conditions for an Explanatory Multi-modal Factor Analysis," *UCLA Working Papers in Phonetics*, **16**, 1-84

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

[8] van der Heijden, P., and de Leeuw, J. (1987?), "Correspondence Analysis, with Special Attention to the Analysis of Panel Data and Event History Data,"

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

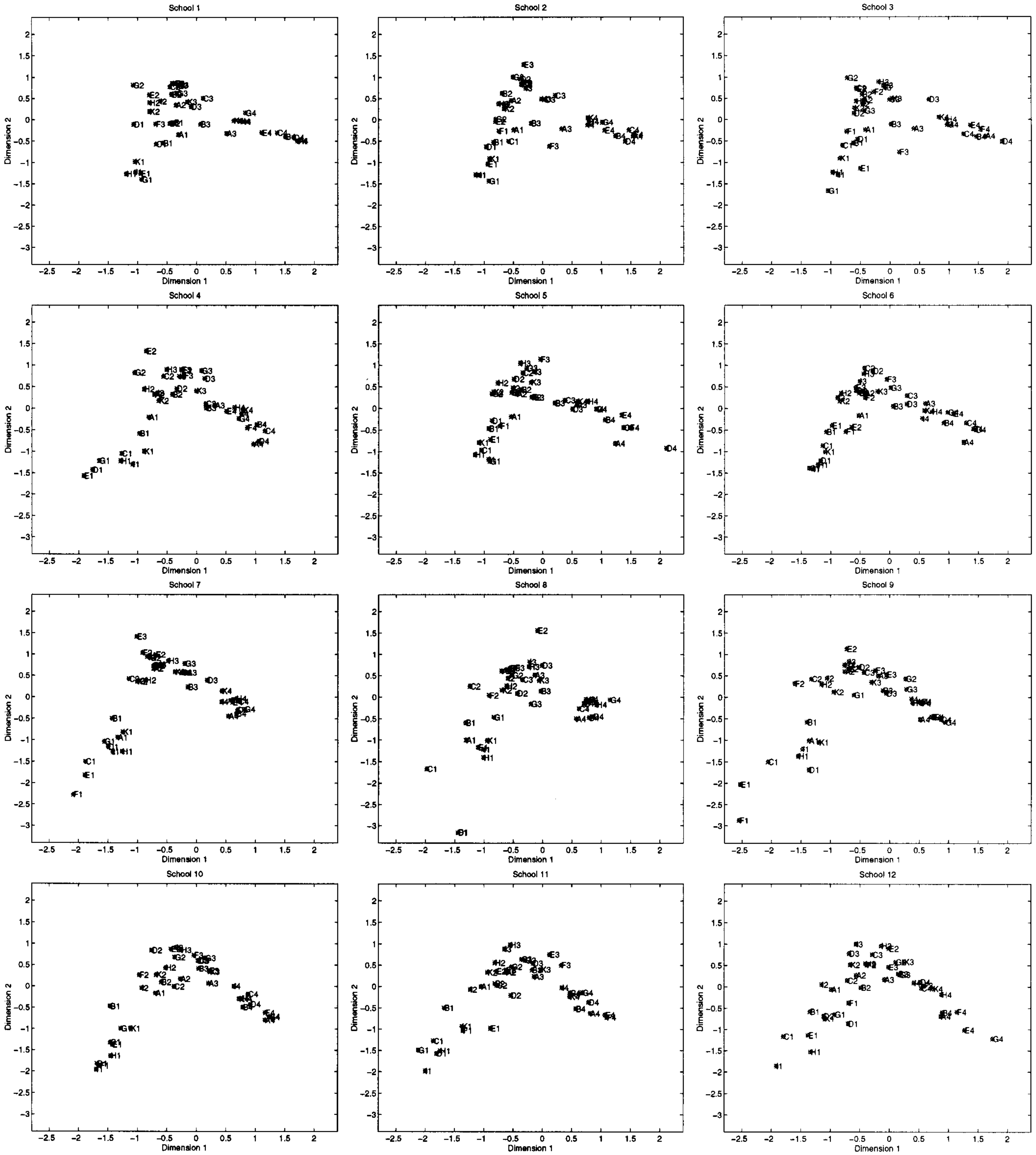

Figure 1: Optimal Constrained Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

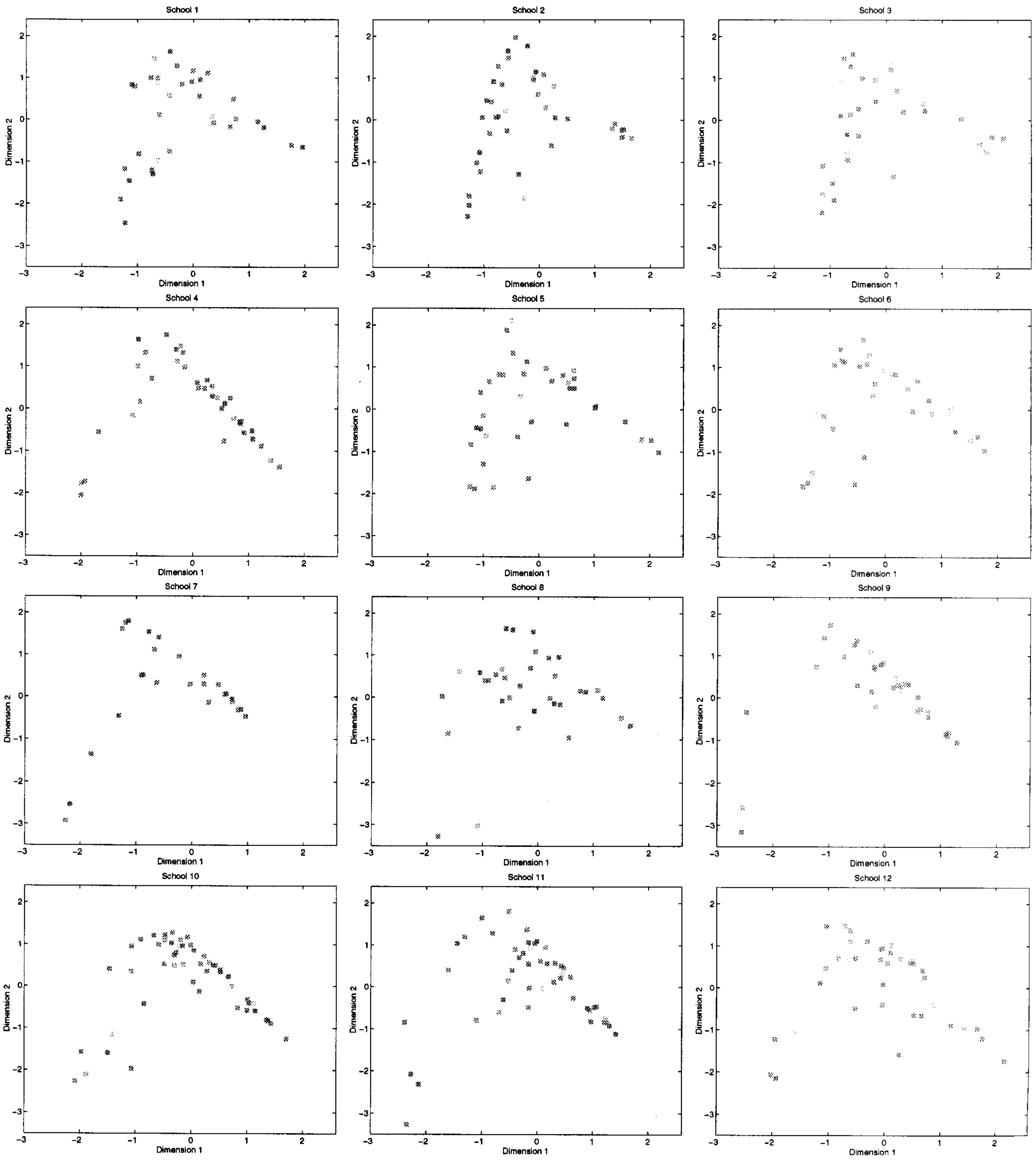

Figure 2: Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12