A REGRESSION MODEL FOR MULTILEVEL HOMOGENEITY ANALYSIS

GEORGE MICHAILIDIS AND JAN DE LEEUW

1. **Introduction**

One of the basic techniques that analyzes categorical data is *homogeneity analysis*. The technique originated in the work of Guttman [17] as a method of scale construction using reciprocal averaging. Burt [5] described homogeneity analysis as a principal components analysis of qualitative data. Hayashi [19] stressed homogeneity analysis as one possible way of quantifying categories. It should be noted that earlier work by Hirschfeld [21] and Fisher [13] concentrated on the bivariate case (analysis of contingency tables). Extensive reviews on the history of homogeneity analysis can be found in de Leeuw [8], Nishisato [26], Benz$\acute{e}$cri [3], Tenenhaus and Young [28] and Gifi [14].

The focus of the various derivations and presentations of homogeneity analysis is on a *single* group of observations (individuals, objects etc). However, in many applications the same variables have been administered to multiple groups of objects. Typical examples include multiple choice tests given to students in different schools, personality inventories administered to depressed and 'normal' individuals, marketing survey questionnaires distributed to different socioeconomic groups and so on. In this paper, we extend homogeneity analysis to multilevel data structures, and introduce a model that allows to examine how variables are related across groups and how groups vary. The new techniques are illustrated on a data set from the National Educational Longitudinal Study of 1988 (NELS:88).

Remark 1.1. *Notation.* We denote by upper case letters matrices (e.g. A) and by lower case letters 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, let u denote the *unit* vector (vector comprised of only ones).

2. **Homogeneity Analysis in the Gifi System**

Suppose we have collected data for N *objects* (individuals, products, countries, etc) on J *categorical variables,* with $\ell_j,\ j \in \mathbf{J} = \{1,\dots,J\}$ *categories* per variable. Variables map the objects into a finite set of categories (*profiles*). The categories of each variable have a certain *measurement level*. The measurement level of the variables can be *numerical* (variables measured at non-overlapping intervals), *ordinal* (the order of the categories matters), or *nominal* (only the classes formed by the objects play a role). We are interested in mapping both objects and variables into a joint p-dimensional space ($p < J$) in such a way that (i) objects with similar profiles are close together and (ii) categories with similar contents are close together.

We proceed to give a precise mathematical formulation to the above verbal description of homogeneity analysis. Indicator matrices are used to code the J variables (see [9]). Let $G_j,\ j \in \mathbf{J}$ denote the $N \times \ell_j$ indicator matrix corresponding to variable j. It is a binary matrix with entries $G(i,t) = 1,\ i = 1, \ldots, N,\ t = 1, \ldots, \ell_j$, if object i belongs to category t, and $G(i,t) = 0$ if it belongs to some other category. According to the homogeneity principles, we would like to quantify (transform) the variables to achieve *maximum* homogeneity. Let Y_j denote the $\ell_j \times p$ matrix containing the optimal *multiple category quantifications* of variable $j \in \mathbf{J}$, and let X be a $N \times p$ matrix containing the resulting p *optimal scales*. The elements of the X matrix are also known as the *object scores* [14]. The dimensionality p is determined by the data analyst according to whether she wants the objects to be on a scale ($p = 1$) or in a plane ($p = 2$) etc. In general, it is hard to find a perfect solution; that is, determine the Y_j's and X exhibiting *perfect consistency*, i.e. $X = G_1 Y_1 = \ldots = G_J Y_J$. Hence, we would like to minimize departures from perfect consistency by employing the following loss function

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

where SSQ $(H) = \text{tr}(H'H)$ 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

$$X'X = NI_p, \tag{2.2}$$

$$u'X = 0, \tag{2.3}$$

The goal of homogeneity analysis in the Gifi system is to choose X and the Y_j's so that the loss in (2.1) is minimized. The solution to this minimization problem can be found by using the following *Alternating Least Squares* (ALS) algorithm:

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

Step 1: Estimate the multiple category quantifications by $\hat{Y}_j = D_j^{-1} G_j' X,\ j \in \mathbf{J}$, where $D_j = G_j' G_j$ and contains the univariate marginals. Thus, the optimal quantification of a category is the centroid of the scores of the objects belonging to that category.

Step 2: Estimate the object scores by $\hat{X} = J^{-1} \sum_{j=1}^{J} G_j Y_j$. Thus, the optimal score of an object is the centroid of the quantifications of the categories the object is in.

Step 3: Column center and orthonormalize the matrix of the object scores, so that the normalization restrictions are satisfied.

Step 4: Check the convergence criterion.

Steps 1-4 are repeated until the algorithm converges to the global minimum (see chapter 3 in [14] and Remark 2.2). Hence, the ALS algorithm finds the desired solution to the problem given in (2.1), in the presence of nominal data. This solution is known in the literature ([14], [9], [10]) as the Homals solution (homogeneity analysis by means of alternating least squares). The rules in Steps 1 and 2 are known as the *centroid principles* (*principes barycentriques* [3]), and the ALS algorithm based on them is called *reciprocal averaging*.

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 Step 2 of the algorithm 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 (2.1).

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

$$J^{-1}\sum_{j=1}^{J}\text{tr}\big(X - G_jY_j\big)'\big(X - G_jY_j\big) = J^{-1}\sum_{j=1}^{J}\text{tr}\big(X'X + Y_j'G_j'G_jY_j - 2Y_j'G_j'X\big) = \tag{2.4}$$
$$J^{-1}\sum_{j=1}^{J}\text{tr}\big(X'X - Y_j'D_jY_j\big) = J^{-1}\sum_{j=1}^{J}\text{tr}\big(NI_p - Y_j'D_jY_j\big) = Np - J^{-1}\sum_{j=1}^{J}\text{tr}\big(Y_j'D_jY_j\big).$$

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

$$\eta_{js}^2 \equiv Y_j'(.,s)D_jY_j(.,s)/N,\ j \in \mathbf{J},\ s = 1,\dots,p. \tag{2.5}$$

Geometrically, the discrimination measures give the average squared distance (weighted by the marginal frequencies) of the category quantifications to the origin of the p dimensional space. It can be shown that (assuming there are no missing data) the discrimination measures are equal to the squared correlation between an optimally quantified variable $G_jY_j(.,s)$ and the corresponding column of object scores $X(.,s)$ (see chapter 3 in [14]). Hence, the loss function can also be expressed as

$$N\big(p - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2\big). \tag{2.6}$$

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.

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.

- $\diamond$ Objects with the same response pattern (identical profiles) receive identical object scores. In general, the distance between two object points is related to the 'similarity' between their profiles.
- $\diamond$ A variable discriminates better to the extent that its category points are further apart.
- $\diamond$ If a category applies uniquely to only one object, then the object point and that category point will coincide.
- $\diamond$ 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.
- $\diamond$ 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).
- $\diamond$ 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 $u'D_jY_j = u'D_jD_jG_j'X = u'G_jX = u'X = 0$.

Remark 2.2. *Homogeneity Analysis as an Eigenvalue Problem.* Substituting the optimal $\hat{Y}_j = D_j^{-1}G_j'X$ for given X in the Gifi loss function (2.1), we get

$$\begin{aligned}\sigma\big(X;\star\big) &= J^{-1}\sum_{j=1}^{J}\text{tr}\big(X - G_jD_j^{-1}G_j'X\big)'\big(X - G_jD_j^{-1}G_j'X\big) \\ &= J^{-1}\sum_{j=1}^{J}\text{tr}\big(X'X - X'G_jD_j^{-1}G_j'X\big),\end{aligned} \tag{2.7}$$

where the asterisk has replaced the argument over which the loss function is minimized. Let $P_j = G_jD_j^{-1}G_j'$ denote the orthogonal projector on the subspace spanned by the columns of the indicator matrix G_j. Let $P_\star = J^{-1}\sum_{j=1}^{J}P_j$ be the average of the J projectors. Equation (2.7) can be rewritten as

$$\sigma\big(X;\star\big) = J^{-1}\sum_{j=1}^{J}\text{tr}\big(X - P_jX\big)'\big(X - P_jX\big) = J^{-1}\sum_{j=1}^{J}\text{tr}\big(X'X - X'P_jX\big). \tag{2.8}$$

This together with the normalization constraints (2.2) and (2.3) gives that maximizing (2.8) is equivalent to maximizing tr$(X'\mathcal{L}P_\star\mathcal{L}X)$, where $\mathcal{L} = I - uu'/u'u$ is a centering operator that leaves $\mathcal{L}X$ in deviations from its column means. The optimal X corresponds to the first p eigenvectors of the matrix $\mathcal{L}P_\star\mathcal{L}$. We can then write the minimum loss as follows:

$$\sigma\big(\star;\star\big) = N\big(p - \sum_{s=1}^{p}\lambda_s\big), \tag{2.9}$$

where $\lambda_s,\ s = 1,\ldots,p$ are the first p eigenvalues of $P_\star$. Therefore, the minimum loss of homogeneity analysis is a function of the p largest eigenvalues of the average projector $P_\star$. Notice that the complete eigenvalue solution has $q = \sum_{j=1}^{J}\ell_j - J$ dimensions. The advantage of employing the ALS algorithm is that it only computes the first $p << q$ dimensions of the solution, thus increasing the computational efficiency and decreasing the computer memory requirements.

Remark 2.3. *Missing Data.* The Gifi loss function makes the treatment of missing data a fairly easy exercise. Missing data can occur for a variety of reasons: blank responses, coding errors etc. Let $M_j,\ j \in \mathbf{J}$ denote the $N \times N$ binary diagonal matrix with entries $M_j(i,i) = 1$ if observation i is present for variable j and 0 otherwise. Define $M_* = \sum_{j=1}^{J} M_j$. Notice that since G_j is an incomplete indicator matrix (has rows with just zeros), we have that $M_j G_j = G_j,\ j \in \mathbf{J}$. The loss function then becomes

$$\sigma(X; Y_1, \ldots, Y_J) = J^{-1} \sum_{j=1}^{J} \operatorname{tr}\bigl(X - G_j Y_j\bigr)' M_j \bigl(X - G_j Y_j\bigr), \tag{2.10}$$

subject to the normalization restrictions $X' M_* X = JNI_p$ and $u' M_* X = 0$. The category quantifications are given by $\hat{Y}_j = D_j^{-1} G_j' X$, 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 [14] 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. **Multilevel Homogeneity Analysis**

In many practical 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 section we extend homogeneity analysis to the multilevel sampling framework.

Very little work has be done on applying homogeneity analysis techniques to multilevel data. De Leeuw, van der Heijden and Kreft [11] and van der Heijden and de Leeuw [29] have used these techniques to examine panel and event history data. In their 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) indicator matrices, and apply homogeneity analysis to the collection of such matrices. More recently, Carlier and Kroonenberg [7] apply the PARAFAC model (see Remark 4.2) 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

$$G_j = \begin{pmatrix} G_{j1} & 0 & 0 & \\ 0 & G_{j2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & G_{jK} \end{pmatrix}$$

which is called the *design* matrix. The Gifi loss function becomes

$$\sigma(X; Y_1, \dots, Y_J) = J^{-1} \sum_{j=1}^{J} \text{SSQ}\left(X - G_j Y_j\right) = \sum_{j=1}^{J} \sum_{k=1}^{K} \text{SSQ}\left(X_k - G_{jk} Y_{jk}\right). \tag{3.1}$$

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

$$X_k' X_k = n_k I_p,\ u' X_k = 0,\ \text{for every } k \in \mathbf{K}. \tag{3.2}$$

The other possibility $u'X = 0$ and $X'X = NI_p$ is briefly discussed later on in Remark 3.3.

The problem in (3.1) is identical to the one presented in (2.1); thus, its solution is given by

$$\hat{Y}_j = D_j^{-1} G_j' X,\ j \in \mathbf{J}, \tag{3.3}$$

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

$$\hat{X} = \frac{1}{J} \sum_{j=1}^{J} G_j Y_j, \tag{3.4}$$

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*

$$\eta^2_{jks} \equiv Y_{jk}'(.,s) D_{jk} Y_{jk}(.,s)/n_k,\ j \in \mathbf{J},\ k \in \mathbf{K},\ s = 1, \dots, p. \tag{3.5}$$

Since the category quantifications have a weighted sum equal to zero, they are interpreted the usual way; the larger the η^2_{jks}, 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

$$\eta^2_{js} \equiv Y_j'(.,s)D_jY_j(.,s)/N,\ j \in \mathbf{J},\ s = 1,\dots,p, \tag{3.6}$$

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

$$\eta^2_{js} \equiv \frac{1}{N}Y_j'(.,s)D_jY_j(.,s) = \frac{1}{N}\sum_{k=1}^{K} Y_{jk}'(.,s)D_{jk}Y_{jk}(.,s). \tag{3.7}$$

so, it easy to see that

$$\eta^2_{js} = \frac{1}{N}\sum_{k=1}^{K} n_k\eta^2_{jks},\ j \in \mathbf{J},\ s = 1,\dots,p. \tag{3.8}$$

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. It can be seen that larger clusters are weighted more in the total.

We can then define *cluster eigenvalues* given by $\gamma_{ks} = J^{-1}\sum_{j=1}^{J}\eta^2_{jks}$, and *total eigenvalues* given by $\gamma_s = J^{-1}\sum_{j=1}^{J}\eta^2_{js}$. The cluster and the total eigenvalues are related by $\gamma^2_s = N^{-1}\sum_{k=1}^{K} n_k\gamma^2_{ks}$, similarly to the discrimination measures.

Remark 3.1. *Model Equivalences.* It is worth noting that under normalization (3.2) this model is equivalent to applying the ordinary Homals algorithm (see Section 2) to each of the K clusters separately.

Remark 3.2. *Comparing Clusters.* As we have seen in section 2 the Homals solution is rotationally invariant. The latter combined with the fact that the multilevel Homals solution amounts to calculating K separate solutions (see Remark 3.1), introduces the problem of making meaningful comparisons between clusters, since different clusters may have different orientations of the axes. We would like to make the clusters as similar as possible by rotating their axes to a target solution. Any of the K solutions can be used as the target one. This amounts to solving a *Procrustes orthogonal rotation problem* [15]. Let $X_k(t)$ be the matrix of object scores of the cluster designated as the target solution, and X_k the object scores of some other cluster. We then have to minimize tr$(X_k(t)-X_kR)'(X_k(t)-X_kR)$ with respect to the $p\times p$ rotation matrix R (i.e. $R'R = RR' = I_p$). The solution is given by first calculating the singular value decomposition of $X_k'X_k(t) = U\Lambda V'$, and then setting $R = UV'$, the orthogonal polar factor of $X_k'X_k(t)$.

Remark 3.3. *On another possible normalization.* Instead of normalizing the object scores locally (within every cluster $k \in \mathbf{K}$), we might require a global scaling given by $u'X = \sum_{k=1}^{K} u'X_k = 0$ and $X'X = \sum_{k=1}^{K} X_k'X_k = NI_p$. 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.

Remark 3.4. *On the Design Matrices.* In the single cluster case the indicator matrix $G_j,\ j \in \mathbf{J}$ is considered to be a *basis* of the transformation space. It corresponds to the Kronecker basis, since the basis isomorphism is given by the identity. Thus, the objects are classified according to the elements of this basis (in other words the categories). In the multilevel framework, we want the objects to be classified by both the categories of variable j and the fact that they belong to cluster k. This requirement automatically translates to a product transformation space. However, the form of the transformation space we adopt, namely $G_j = \bigoplus_{k=1}^{K} G_{jk}$, implies that the subspaces $G_{jk},\ k \in \mathbf{K}$ are independent and span G_j, and that the dimensionality of the transformation space is equal to $\sum_{k=1}^{K} \dim G_{jk} = K\ell_j$. This fact allows us to look at different transformations for each cluster separately.

3.1. **NELS:88 Example.** A data set from the base year survey of the student population of the National Educational Longitudinal Study has been selected to illustrate our techniques. The variables in this data set deal 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 [27, 1].

The students in the data set are clustered in 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 Table 3.1.

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 3.1. 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 [12, 25] in this example.

Table 3.2 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 3.2. 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 3.3. 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 sample as a whole is given in Table 3.3. 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 3.1. 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 have discrimination measures larger than the total ones for some of the variables, and smaller than the total measures for the rest of the variables. Finally, some schools (e.g. 8, 9, and to a certain extent 11) have smaller discrimination measures than the total for the majority of the variables; however, for a couple of variables the cluster measures were much larger than the total ones, thus indicating the possible presence of outliers. Figure 3.2 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.

Figure 3.3 displays the category quantifications of the variables for each school. The points in the graph represent the centers of gravity of the object points associated with each category. Several different patterns can be observed between the variable categories. For example for some schools (1, 4, 6, 10, 11 and 12) the following pattern emerges. In the lower left quadrant of the graph we find the ’serious problem’ categories for cutting class, physical conflicts, robbery and vandalism, use of alcohol and drugs, possession of weapons and physical and verbal abuse of teachers. However, the ’serious problem’ category for student tardiness and absenteeism (variables A and B) was located at different places in different schools. Thus, students in this area of the map are associated with these categories, which implies that they consider their school to be seriously affected by these problems. In the upper half of the graph, we find the ’minor/moderate problem’ categories for almost all the variables. Students associated with these categories believe that these problem areas are present only to a certain degree in their schools. Finally, in the lower right quadrant of the graph we find the ’not a problem’ categories for all the variables; hence, students in that area of the graph think that there are no problem areas in their schools. It is interesting to observe that the ’clustering’ of the students is done according to the same category levels. Thus, students consider all the areas representing either a serious, or a minor/moderate or not a problem in their school. In principle, in this set of schools we do not have students that indicate some areas as being a serious problem and some other areas as not a problem. Hence, to a large extent the analysis cleanly separates the students that think there exist serious problems in their schools, from the ones that think their schools are problem free (as far as the areas identified in the data set are concerned). Moreover, the analysis reveals distinctly nonlinear student response patterns; that is, variable categories are not linear with 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. Some other interesting points arising from examining the category quantifications plots are (i) the fact that use of alcohol is a ’serious’ problem in the rural schools 8 and 9 (but not in

7) (see the position of G1 in the respective graphs), and (ii) the fact that student tardiness and to a certain degree absenteeism are ’serious’ problems in the private schools (position of A1 and B1, especially in school 12). In general, these 12 schools exhibit a wide range of student response patterns. By closely examining the optimal category quantification plots we have identified three ’main’ groups of schools: those where the majority of the students believe there are problems, those where most of the students believe there are no 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 3.4 (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. In general, it can be seen that the response profiles of these 12 schools differ from the profile given by examining all 1,050 schools (see Section 2.6.1). Some of the schools (notably, the public suburban and private schools) exhibit almost a quadratic profile (’horseshoe’), although in most of them the ’horseshoe’ is not filled. The latter fact indicates that students are cleanly clustered in three groups in those schools. Examining each school separately has provided a better understanding of the variety of response patterns existing in this NELS:88 data set.

4. **Regression Model**

The NELS example shows that in the presence of many grouping units unconstrained multilevel homogeneity analysis leads to estimating a large number of model parameters, which in turn introduces instabilities in the solution. In this section we introduce a model that restricts the category quantifications to be the same across groups, but weights them by different factors for each group. 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 $Q_j,\ j \in \mathbf{J}$ are the restricted (overall) category quantifications and $B_k,\ k \in \mathbf{K}$ the slope matrices, and $\alpha_{jk},\ j \in \mathbf{J},\ k \in \mathbf{K}$ are parameters that ensure that the category quantifications have a weighted sum over categories equal to zero. In this model the slope matrices$B_k,\ k \in \mathbf{K}$ are required to be *diagonal*.

To minimize the usual Gifi loss function, we start by computing the $\hat{Y}_j$'s as in (3.3). We then partition the Gifi loss function as follows:

(4.2)
$$J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\operatorname{tr}\bigl(X_k - G_{jk}\hat{Y}_{jk}\bigr)'\bigl(X_k - G_{jk}\hat{Y}_{jk}\bigr) + J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\operatorname{tr}\bigl(Y_{jk} - \hat{Y}_{jk}\bigr)'D_{jk}\bigl(Y_{jk} - \hat{Y}_{jk}\bigr),$$

and after imposing the restriction on the Y_{jk}'s we have to minimize

$$J^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\operatorname{tr}\bigl(Q_jB_k - \hat{Y}_{jk}\bigr)'D_{jk}\bigl(Q_jB_k - \hat{Y}_{jk}\bigr), \tag{4.3}$$

with respect to Q_j and B_k. Since the B_k's are diagonal, (4.3) can also be written as

$$J^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{s=1}^{p}\bigl(q_j^s\beta_k^s - \hat{y}_{jk}^s\bigr)'D_{jk}\bigl(q_j^s\beta_k^s - \hat{y}_{jk}^s\bigr), \tag{4.4}$$

where $q_j^s = Q_j(.,s),\ s = 1,\dots,p$ is an ℓ_j row vector and $\beta_k^s = B_k(s,s)$ a scalar. In what follows we will use both forms of the loss function. We minimize (4.4) using an ALS algorithm, by alternating over q_j^s and β_k^s in an *inner* iteration loop. For fixed q_j^s the optimal β_k^s's are given by

$$\hat{\beta}_k^s = \bigl(\sum_{j=1}^{J}(\hat{y}_{jk}^s)'D_{jk}q_j^s\bigr)/\bigl(\sum_{j=1}^{J}(q_j^s)'D_{jk}q_j^s\bigr),\ k \in \mathbf{K},\ s = 1,\dots,p. \tag{4.5}$$

This completes the first step of the inner iteration loop of the ALS algorithm. So, it remains to minimize (4.4) with respect to q_j^s for fixed β_k^s. We then get the following set of normal equations,

$$\sum_{k=1}^{K}(\beta_k^s)^2D_{jk}q_j^s = \sum_{k=1}^{K}\beta_k^sD_{jk}\hat{y}_{jk}^s,\ j \in \mathbf{J},\ s = 1,\dots,p. \tag{4.6}$$

From (4.6) we get that

$$\hat{q}_j^s = \bigl(\sum_{k=1}^{K}(\beta_k^s)^2D_{jk}\bigr)^{-1}\sum_{k=1}^{K}\beta_k^sD_{jk}\hat{y}_{jk}^s,\ j \in \mathbf{J},\ s = 1,\dots,p. \tag{4.7}$$

In the absence of any further restrictions on the category quantifications we set $Q_j = \hat{Q}_j,\ j \in \mathbf{J}$. So, we get $\hat{Y}_{jk} = \hat{Q}_j\hat{B}_k$ and the inner iteration loop is complete. Then, we proceed to minimize the Gifi loss function with respect to X, which is done as shown in section 3.

Once the algorithm has converged, we want to center the cluster category quantifications in order to facilitate the interpretation of the joint plot of object scores and category quantifications. For this purpose we use the intercept parameters. We set $\tilde{Y}_{jk} = u\hat{\alpha}_{jk}' + \hat{Q}_j\hat{B}_k$, where

$$\hat{\alpha}_{jk}' = -(u'D_{jk}\hat{Q}_j\hat{B}_k)/n_k,\ j \in \mathbf{J},\ k \in \mathbf{K}. \tag{4.8}$$

Thus we get $u'D_{jk}\tilde{Y}_{jk} = 0$ as required.

The complete ALS algorithm for the regression model has the following steps:

Step 0: Initialize X, so that $u'X = 0$ and $X'X = NI_p$.
Step 1: Estimate the unrestricted multiple category quantifications by $\hat{Y}_j = D_j^{-1}G_j'X,\ j \in \mathbf{J}$.
Step 2: Estimate the slope coefficients $\hat{\beta}_k^s = (\sum_{j=1}^J \hat{y}_{jk}^{s'} D_{jk} q_j^s)/(\sum_{j=1}^J q_j^{s'} D_{jk} q_j^s),\ k \in \mathbf{K}$, $s = 1, \ldots, p$.
Step 3: Estimate the overall category quantifications by $\hat{q}_j^s = (\sum_{k=1}^K (\beta_k^s)^2 D_{jk})^{-1} \sum_{j=1}^J \hat{\beta}_k^s D_{jk} \hat{y}_{jk})$, $j \in \mathbf{J},\ s = 1, \ldots, p$.
Step 4: Update the unrestricted multiple category quantifications by $\hat{Y}_{jk} = \hat{Q}_j \hat{B}_k,\ k \in \mathbf{K}$, $j \in \mathbf{J}$.
Step 5: Estimate the object scores by $\hat{X} = J^{-1} \sum_{j=1}^J G_j Y_j$.
Step 6: Column center and orthonormalize the matrices $X_k,\ k \in \mathbf{K}$ of object scores.
Step 7: Check the convergence criterion.
Step 8: Once the algorithm has converged, center the group category quantifications $Y_{jk},\ j \in \mathbf{J},\ k \in \mathbf{K}$.

In principle, to obtain the minimum over Q_j and B_k steps 2-4 that constitute the inner iteration loop should be repeated until (4.3) is minimized. However, since the value of the loss function will be smaller after a single iteration of the inner ALS loop, inner iteration upon convergence is not necessary in practice.

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 $Y_{jk}^\sharp = D_{jk}^{-1} G_{jk} X^\sharp = \hat{Y}_{jk} R$. Write (4.7) in compact form as $\sum_{k=1}^K D_{jk} Q_j B_k B_k' = \sum_{k=1}^K D_{jk} \hat{Y}_{jk} R B_k$. However, the matrices B_k and R do not in general commute (i.e. $RB_k \neq B_k R$), so that the matrix Q_j of category quantifications 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 [18, 6, 2]

$$Z(.,.,k) = \Phi \Delta_k \Psi',\ k \in \mathbf{K}, \tag{4.9}$$

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. In case the k slices are symmetric matrices, then the model is written as

$$Z(.,.,k) = A\Delta_k A',\ k \in \mathbf{K}. \tag{4.10}$$

Models (4.9) and (4.10) are known as Parallel Factors model (PARAFAC) and Individual Differences Scaling model (INDSCAL) respectively, and a more extensive discussion about them and other models for three-way data can be found in chapter 7. It can be seen that the regression model can be casted in this framework. In particular, we have

$$Y_{jk} \equiv Z(.,.,k) = Q_j B_k I_p,\ k \in \mathbf{K}. \tag{4.11}$$

Therefore, the regression model can be considered as a constrained form of the PARAFAC model.

Remark 4.3. *On the ALS Algorithm.* Unlike the unconstrained problem of section 3, the constrained problem given by (4.3) does not admit a close form solution (see also Remark 4.5 later on). What we do in the inner iteration loop of the ALS algorithm is successive improvements of the value of the loss function by holding in each steps a subset of the parameters fixed (the quantifications and the weights). The value of the function will never be higher than before, because the optimal values for the parameters that we compute in each step can not worse than their previous values. Since (4.3) is bounded below by zero, more often than not, the algorithm will end up close to the global minimum. However, this is not guaranteed. There may be cases where the algorithm might get caught in a local minimum. Experience with the ALS algorithm for the PARAFAC model suggests [2] that good initial values for the inner iteration loop are essential. So our strategy in practice is to skip steps 2-4 for the first few iterations, so that we start with more stable $\hat{Y}_{jk}$'s the inner iteration loop.

Remark 4.4. *Other Slope Matrices.* One can consider other types of slope matrices (i.e. non-diagonal). An interesting case arises if the slope matrices are upper triangular. Consider the situation where the columns of Q_j are the monomials (linear, quadratic, etc). Then $Q_j B_k$ with B_k upper triangular will make the columns of Y_{jk} polynomials of increasing degree, but with different coefficients. For general slope matrices the model becomes strange and of no particular interest.

4.1. **Loss and Fit.** In the regression model the loss function is partitioned into two parts,

$$\begin{aligned} &J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathrm{tr}\big(X_k - G_{jk}\hat{Y}_{jk}\big)'\big(X_k - G_{jk}\hat{Y}_{jk}\big)' + \\ &J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathrm{tr}\big(\hat{Q}_j\hat{B}_k - \hat{Y}_{jk}\big)'D_{jk}\big(\hat{Q}_j\hat{B}_k - \hat{Y}_{jk}\big). \end{aligned} \tag{4.12}$$

Using the fact $\hat{Y}_{jk}'D_{jk}\hat{Y}_{jk} = \hat{Y}_{jk}'D_{jk}(D_{jk}^{-1}G_{jk}'X_k) = \hat{Y}_{jk}'G_{jk}'X_k$ we can rewrite the first part of (4.12) as

$$J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathrm{tr}\big(X_k'X_k - \hat{Y}_{jk}'D_{jk}\hat{Y}_{jk}\big) = J^{-1}\sum_{j=1}^{J}\mathrm{tr}\big(X'X - \hat{Y}_j'D_j\hat{Y}_j\big), \tag{4.13}$$

which is called *multiple loss*. The diagonal elements of the matrices $\hat{Y}_j' D_j \hat{Y}_j / N$ are called *multiple fit*.

By examining (4.1) it can be seen that there is a built-in indeterminacy in the model. Thus, in order for the Q_j's to be identified, we require $\sum_{j=1}^{J} Q_j' \tilde{D}_j Q_j = JNI_p,\ k \in \mathbf{K}$, 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. Notice that this normalization constraint can also be written as $\sum_{j=1}^{J} Q_j' \tilde{D}_j Q_j = \sum_{j=1}^{J} \sum_{k=1}^{K} Q_j' D_{jk} Q_j = \sum_{k=1}^{K} W_k = JNI_p$, where $W_k = \sum_{j=1}^{J} Q_j' D_{jk} Q_j,\ k \in \mathbf{K}$. Using the fact that $\sum_{j=1}^{J} \big(Q_j' D_{jk} Q_j\big) B_k = \sum_{j=1}^{J} Q_j' D_{jk} \hat{Y}_{jk}$ the second part of (4.12) can be rewritten as

$$J^{-1} \sum_{j=1}^{J} \sum_{k=1}^{K} \operatorname{tr}\big(\hat{Y}_{jk}' D_{jk} \hat{Y}_{jk} - \hat{B}_k' \hat{Q}_j' D_{jk} \hat{Q}_j \hat{B}_k\big) = J^{-1} \big[\sum_{j=1}^{J} \operatorname{tr}\big(\hat{Y}_j' D_j \hat{Y}_j\big) - \sum_{k=1}^{K} \operatorname{tr}\big(W_k \hat{B}_k^2\big)\big], \tag{4.14}$$

where $B_k^2 = B_k B_k'$ since B_k is diagonal. Let $\overline{W}_k = W_k / JN,\ k \in \mathbf{K}$ *after imposing* the normalization constraint, so that $\sum_{k=1}^{K} W_k = JN \sum_{k=1}^{K} \overline{W}_k$. We can then write (4.14) as

$$J^{-1} \sum_{j=1}^{J} \operatorname{tr}\big(\hat{Y}_j' D_j \hat{Y}_j\big) - J^{-1} \sum_{k=1}^{K} \sum_{s=1}^{p} \overline{W}_k(s,s) B_k^2(s,s)\big). \tag{4.15}$$

The quantity $\sum_{s=1}^{p} \sum_{k=1}^{K} \overline{W}_k(s,s) B_k^2(s,s)$ is called *regression fit*, while the expression given in (4.15) is called *regression loss*.

4.2. **A Special Case.** We examine next the special case where the matrices containing the group univariate marginals are of the form

$$D_{jk} = \gamma_k \Psi_j,\ ,j \in \mathbf{J},\ k \in \mathbf{K}, \tag{4.16}$$

where $\gamma_k,\ k \in \mathbf{K}$ are given scalars, and $\Psi_j,\ j \in \mathbf{J}$ diagonal. This implies that the distribution of the categories of all variables are proportional across clusters. The problem then becomes to minimize

$$\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{s=1}^{p} \gamma_k \big(q_j^s \beta_k^s - \hat{y}_{jk}^s\big)' \Psi_j \big(q_j^s \beta_k^s - \hat{y}_{jk}^s\big), \tag{4.17}$$

with respect to q_j^s and $\beta_k^s,\ j \in \mathbf{J},\ k \in \mathbf{K},\ s = 1, \ldots, p$, subject to the normalization constraint

$$\sum_{j=1}^{J} (q_j^s)' \big(\sum_{k=1}^{K} D_{jk}\big) q_j^s = N \Longleftrightarrow \sum_{j=1}^{J} (q_j^s)' \Psi_j q_j = N / \big(\sum_{k=1}^{K} \gamma_k\big),\ s = 1, \ldots, p. \tag{4.18}$$

The optimal β_k^s's are given by (using the normalization constraint)

$$\hat{\beta}_k^s = c^{-1} \sum_{j=1}^{J} (\hat{y}_{jk}^s)' \Psi_j q_j^s,\ k \in \mathbf{K},\ s = 1, \ldots, p, \tag{4.19}$$

where $c = N/\sum_{k=1}^{K}\gamma_k$. Plugging the optimal β_k^s's in (4.17) and after some algebra we get that it remains to minimize

$$\sum_{k=1}^{K}\sum_{s=1}^{p}\gamma_k\big(\sum_{j=1}^{J}(\hat{y}_{jk}^s)'\Psi_j q_j^s\big)^2, \tag{4.20}$$

with respect to $q_j^s,\ j \in \mathbf{J}$, subject to the normalization restriction (4.18). We form the Langrangean

$$\mathcal{L} = \sum_{s=1}^{p}\sum_{k=1}^{K}\gamma_k\big(\sum_{j=1}^{J}(\hat{y}_{jk}^s)\Psi_j q_j^s\big)^2 - \sum_{s=1}^{p}\lambda_s\big(\sum_{j=1}^{J}(q_j^s)'\Psi_j q_j^s - c\big). \tag{4.21}$$

Write (4.21) as

$$\begin{aligned}\sum_{s=1}^{p}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{J}(q_j^s)\Psi_j\hat{y}_{jk}^s(\hat{y}_{ik}^s)'\Psi_i q_i^s - \sum_{s=1}^{p}\lambda_s\big(\sum_{j=1}^{J}(q_j^s)'\Psi_j q_j^s - c\big) = \\ \sum_{s=1}^{p}\sum_{j=1}^{J}\sum_{i=1}^{J}(q_j^s)'\Psi_j V_{ji}^s\Psi_i q_i^s - \sum_{s=1}^{p}\lambda_s\big(\sum_{j=1}^{J}(q_j^s)'\Psi_j q_j^s - c\big).\end{aligned} \tag{4.22}$$

where $V_{ji}^s = \sum_{k=1}^{K}\hat{y}_{jk}^s(\hat{y}_{ik}^s)',\ i,j \in \mathbf{J}$. Notice that $(V_{ji}^s)' = V_{ij}^s$. Let $q^s = [(q_1^s)'\ (q_2^s)'\ldots(q_K^s)']'$ and $\Psi = \bigoplus_{j=1}^{J}\Psi_j$ and define

$$V^s = \begin{pmatrix} V_{11}^s & \ldots & V_{1J}^s \\ \ldots & \ldots & \ldots \\ V_{J1}^s & \ldots & V_{JJ}^s \end{pmatrix}$$

Then, (4.22) can be written as

$$\sum_{s=1}^{p}(q^s)'\Psi V^s\Psi q^s - \lambda_s\big((q^s)'\Psi q^s - c\big). \tag{4.23}$$

Setting $\partial\mathcal{L}/\partial q^s = 0$ we get

$$\Psi V^s\Psi q^s = \lambda_s\Psi q^s \iff V^s\Psi q^s = \lambda_s q^s,\ s = 1,\ldots,p, \tag{4.24}$$

which shows that λ_1 is the largest eigenvalue of the $V^s\Psi$ matrix and q^1 the corresponding eigenvector, λ_2 the second largest eigenvalue and so on. Therefore, in this special case the problem admits a closed form solution, and the inner loop of the ALS algorithm will converge to the global minimum.

Remark 4.5. *The General Case.* In light of the above result let us examine the case for general D_{jk} matrices. To ease the presentation we examine the problem in *one* dimension (so, the superscripts s is dropped). Then, after substituting the optimal β_k's in the loss function, we get that we want to maximize

$$\sum_{k=1}^{K}\frac{\sum_{j=1}^{J}\sum_{i=1}^{J}q_j'D_{jk}\hat{y}_{jk}\hat{y}_{ik}'D_{ik}q_i}{\sum_{j=1}^{J}q_j'D_{jk}q_j}, \tag{4.25}$$

with respect to $q_j,\ j \in \mathbf{J}$, subject to the normalization constraint

$$\sum_{j=1}^{J} q_j' (\sum_{k=1}^{K} D_{jk}) q_j = N. \tag{4.26}$$

Letting $D_k = \bigoplus_{j=1}^{J} D_{jk},\ k \in \mathbf{K}$ and $A_k = D_k V_k D_k$, where

$$V_k = \begin{pmatrix} V_{11}^k & \dots & V_{1J}^k \\ \dots & \dots & \dots \\ V_{J1}^k & \dots & V_{JJ}^k \end{pmatrix}$$

and $V_{ji}^k = \hat{y}_{jk}\hat{y}_{ik}'$, (4.25) can be written as

$$\sum_{k=1}^{K} \frac{q' A_k q}{q' D_k q}. \tag{4.27}$$

The additive nature of (4.27) shows that maximization corresponds to solving the following system

$$A_k q - \lambda_k D_k q,\ k \in \mathbf{K}, \tag{4.28}$$

subject to the normalization constraint (4.26). Therefore, we are looking at the simultaneous diagonalization of the $A_k - \lambda_k D_k$i's. One obvious but fairly uninteresting condition is that $\{A_k - \lambda_k D_k\}_{k=1}^K$ is a family of commuting matrices (see [22]). However, notice that $\{A_k - \lambda_k D_k\}_{k=1}^K$ is also a family of symmetric positive definite pencils, and the question is whether they admit a simultaneous diagonalization.

4.3. **NELS:88 Example (continued).** We continue with the example introduced in the previous section. A two dimensional solution incurred a total loss of 589.4, with the multiple loss component contributing 507.4 (86.1contributing 82.0 (13.9%). The following graph summarizes the contribution of every school to the total fit (multiple and regression fit). The public urban schools (1, 2, 3) exhibit the poorest fit as a group, while the public rural (7, 8, 9) the best one. The public suburban schools (4, 5, 6) had very similar fits, while the private schools (10, 11, 12) exhibit a large variation in terms of fit. The plot of the overall category quantifications (Q_j's) is given next. The two more striking features of figure 4.2 are the quadratic pattern of the category quantifications, and the clustering according to the prior classification. Hence, the category quantifications of the category 'not a problem' form a separate cloud of points, and the same holds true for the remaining categories. Therefore, four separate groups of response profiles are formed. The regression solution recovers the quadratic profile that is present when examining all students in the data set (23,248) as a single group.

In figure 4.3 the values of the slope (weight) matrices $B_k,\ k \in \mathbf{K}$ are given. As expected, all the variables are weighted more heavily in the first dimension, than in the second one. This implies that the solution exhibits a better fit in the first dimension, compared to the second dimension. Regarding the schools, it can be seen that private school # 12 has very low weights in both dimension, followed by public suburban school # 5. The public urban schools (1, 2, 3) receive very similar weights, as well as the the public rural ones (7, 8, 9). On the other hand, the public suburban and especially the private schools exhibit larger differences in the patterns of their weight matrices. Figure 4.3 suggests the following grouping of the 12 schools in this example: schools 1-4, 8, 10 and 11 form one group, schools 5, 6 and 12 a second one and schools 7 and 9 a third one. In general, there is a great deal of consistency in terms of weighting in the two dimensions; that is, schools with a high weight in the first dimension usually have a high one in the second one and vice versa.

Finally, the cluster category quantifications (after centering) and the object scores are given in figures 4.4 and 4.5, respectively. Figure 4.4 summarizes the information presented in figures 4.2 and 4.3. Hence, the public urban and the public rural schools are fairly homogeneous, while the public suburban and to a greater degree the private schools exhibit larger differences between them. Moreover, it is interesting to observe the very similar patterns that all schools exhibit for the category quantifications. The clustering according to the prior classification that we saw in figure 4.2 is present here as well, although the ’minor’ and ’moderate’ categories appear to be mixed. On the other hand, the object scores have greater variability in their patterns. For example, we have an almost perfect quadratic pattern by school # 4, to a almost filled ’horseshoe’ by school # 11, with the remaining schools somewhere in between. These two graphs suggest that in school # 4 none of the students gave responses mixing categories 1 and 4, or even 2 and 4. However, this seems to be the case with students in schools 5, 11 and possibly the public urban schools. The object score plot also reveals the presence of outliers in schools 7 and 9, something known from the unconstrained multilevel Homals solution presented in section 3. The solution shows that the public rural schools are the most problem free, followed by the private schools, while the public urban schools seem to be rather ’rough.’ A similar conclusion was reached by imposing a different type of restriction on the category quantifications (see [24]). The solution from the regression model filters most of the noise present in the data. It borrows strength from all the clusters, and therefore tends to reduce the variability between clusters and produce a more uniform pattern. When examining a relative small number of clusters it might filter the data a bit too much, thus eliminating some possibly interesting features. In such cases, different weights for different subsets of variables (see section 5.1) might offer the most interesting alternative.

5. **Extensions of the Regression Model**

In this section we examine various possible extensions and generalizations of the model presented in section 4.

5.1. **Sets of Variables.** The basic restriction on the group category quantification matrices $Y_{jk} = Q_j B_k$ implies that all the variables in the same cluster receive the same weighting provided by the elements of the slope matrix B_k. This weighting is, naturally, a function of all J variables, with the weights given by the corresponding marginals. However, in many data analytic situations when examining the interdependence of a set of variables one might want to have a different weighting scheme for different subsets of variables. This may be due to prior knowledge regarding the nature of the variables under consideration. For example, in case one examines the relationship of grades received by a student in various subjects, with the amount of time the student spent studying these subjects, it is reasonable to give a different weighting to the set of grade variables and a different one to the set of study variables. Analogous situations occur in many other fields in the physical, social and life sciences. In order to accommodate the above described situation in our regression framework, we partition the set of variables $\mathbf{J}$ into H subsets, $J(h),\ h = 1, \dots, H$, so that $\sum_{h=1}^{H} |J(h)| = J$, where $|A|$ denotes the cardinality of set A. We then require

$$Y_{jk} = Q_j B_k^h,\ j \in J(h),\ J(h) \subseteq \mathbf{J},\ k \in \mathbf{K}. \tag{5.1}$$

The estimation of the category quantifications $Q_j,\ j \in \mathbf{J}$ and the slope matrices $B_k^h,\ k \in \mathbf{K},\ h = 1, \dots, H$, is done by a small modification in the inner ALS loop. More precisely the slope matrices are estimated by

$$\hat{B}_k^h = \big(\text{diag} \sum_{j \in J(h)} Q_j' D_{jk} Q_j\big)^{-1} \big(\text{diag} \,(\sum_{j \in J(h)} Q_j' D_{jk} \hat{Y}_{jk})\big),\ k \in \mathbf{K},\ h = 1, \dots, H, \tag{5.2}$$

and the category quantifications by

$$\hat{Q}_j(t, .) = S_j^h(t, .)\big(\sum_{k=1}^{K} D_{jk}(t, t) V_k^h\big)^{-1},\ t = 1, \dots, \ell_j,\ j \in J(h),\ h = 1, \dots, H, \tag{5.3}$$

where $V_k^h = B_k^h B_k^h$ and $S_j^h = \sum_{k=1}^{K} D_{jk} \hat{Y}_{jk} B_k^h$. Therefore, the weighting of the cluster category quantifications becomes a function only of the variables that belong to the set $J(h)$, as expected.

It is worth noting that in the extreme case where each set $J(h)$ contains exactly one variable, then there are no further restrictions imposed.

5.2. **Restrictions on the Slope Matrices.** The slope matrices $B_k,\ k \in \mathbf{K}$ determine the weighting the quantifications of all the variables receive for cluster k. Since, these matrices are largely influenced by the marginal frequencies (see (4.5)), we can say that they express in a certain sense the importance of each cluster. In other words, large clusters are weighted more heavily than small ones. However, in many cases there are other important variables that describe the importance or the peculiarities of the clusters and which we would like to incorporate in the analysis.

Notice that the situation we just described presents many similarities to what goes on in the hierarchical linear models (HLM) literature (see for example [4, 23]). The basic idea in the HLM literature is that individuals in the same group (e.g. classroom, school, socioeconomic status) are closer or more similar than individuals in different groups. Thus, for example students, in the same class share values on many of the variables used in a particular regression model. One way to formalize this idea is to fit a separate regression model (with its own intercept and slope) for each group, the so-called *first level* model. We can then build another regression model (*second level*) for the slopes, thus making them depend on class variables such as class size, teacher's philosophy etc. There are linear models on both levels, and if there are more levels (e.g. students within classes, within schools, within school districts) there are more nested linear models. Therefore, a new class of regression models can be built that takes into account the hierarchical structure of the data and makes it possible to incorporate variables from all levels. The basic assumptions for regression models are linearity, normality, homoskedasticity and independence. In the HLM framework the first two are maintained, but the last two are adapted to more complex situations, e.g. independence of individuals across groups, but dependence within the group. Such adaptations have yielded an extensive body of theoretical results for estimation procedures, prediction, regression diagnostics (see [4, 20]).

In our case we have the regression model $Y_{jk} = Q_j B_k$, that corresponds to the first level model in the HLM literature, and at this point we are interested in modeling the slope matrices B_k (which would correspond to the second level model). However, the focus of this study is on the representation of the data and no specific stochastic assumptions are made. Thus, error terms will be absent from the subsequent discussion (a big contrast to the HLM literature). In order to introduce the second level part of the model, we define the K-row vectors $\zeta_s,\ s = 1, \ldots, p$, where each element $\zeta_s(k) = B_k(s, s)$; thus, we gather all the K elements of the slope matrices in the s^{th} dimension in a single vector. Then, we require

$$\zeta_s = \Phi\delta_s,\ s = 1, \ldots, p, \tag{5.4}$$

where Φ is a $K \times \nu$ design matrix and δ_s a $\nu \times 1$ vector of regression coefficients, where ν is the number of second level variables. The component of the Gifi loss function given by (4.3) can be decomposed after estimating the slope matrices by (4.6) as follows:

$$J^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\text{tr}\big(Q_j\hat{B}_k - \hat{Y}_{jk}\big)' D_{jk}\big(Q_j\hat{B}_k - \hat{Y}_{jk}\big) + J^{-1}\sum_{k=1}^{K}\text{tr}\big(B_k - \hat{B}_k\big)' W_k\big(B_k - \hat{B}_k\big), \tag{5.5}$$

where $W_K = \sum_{j=1}^{J} Q_j' D_{jk} Q_j$. Thus, we must minimize the second component of (5.5) after imposing the restriction (5.4). Let Σ_s be a $K \times K$ diagonal matrix containing the (s, s) elements of the $W_k,\ k \in \mathbf{K}$ matrices; that is, $\Sigma_s =$ diag $[W_1(s, s),\ W_2(s, s), \ldots, W_K(s, s)]$. It is easy to see that we can write

$$\sum_{k=1}^{K}\big(B_k - \hat{B}_k\big)' W_k\big(B_k - \hat{B}_k\big) = \sum_{s=1}^{p}\big(\Phi\delta_s - \hat{\zeta}_s\big)\Sigma_s\big(\Phi\delta_s - \hat{\zeta}_s\big). \tag{5.6}$$

Hence, we need to minimize (5.6) with respect to $\delta_s,\ s = 1, \ldots, p$. The minimum is given by the generalized least squares estimate

$$\hat{\delta}_s = \big(\Phi'\Sigma_s\Phi\big)^{-1}\Psi\Sigma_s\hat{\zeta}_s,\ s = 1, \ldots, p. \tag{5.7}$$

The complete ALS algorithm for the regression model with linear restrictions on the slope matrices contains the same steps as before. However, between steps 2 and 3, an additional step must be inserted, that estimates the parameters δ_s and then updates the slope matrices by computing $\hat{\zeta}_s = \Psi\delta_s,\ s = 1, \ldots, p$.

The restriction (5.4) allows to bridge ideas from the Gifi system (e.g. optimal scaling of categorical variables) and from the HLM literature (e.g. linear restrictions on first level model parameters).

6. **Concluding Remarks**

REFERENCES

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

[2] Arabie, P., Carroll, J.D. and DeSarbo, W.S. (1987), *Three Way Scaling and Clustering*, Newbury Park: Sage University Press

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

[4] Bryk, A.S. and Raudenbush, S.W. (1992), *Hierarchical Linear Models. Applications and Data Analysis Methods*, Newbury Park: Sage Publications

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

[6] 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

[7] Carlier, A. and Kroonenberg, P.M. (1996), ”Decompositions and Biplots in Three-Way Correspondence Analysis,” *Psychometrika*, **61**, 355-373

[8] de Leeuw, J. (1984), *Canonical Analysis of Categorical Data*, Leiden: DSWO Press

[9] 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

[10] 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

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

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

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

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

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

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

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

[18] 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

[19] 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

[20] Hilden-Minton, J.A. (1995), *Multilevel Diagnostics in Mixed and Hierarchical Linear Models*, Unpublished Doctoral Dissertation, UCLA

[21] Hirschfeld, H.O. (1935), ”A Connection between Correlation and Contingency,” *Proceedings of the Cambridge Philosophical Society*, **31**, 520-524

[22] Horn, R.A. and Johnson, C.R. (1990), *Matrix Analysis*, Cambridge: Cambridge University Press

[23] Longford, N.T. (1993), *Random Coefficients Model*, New York: Oxford University Press

[24] Michailidis, G. and de Leeuw, J. (1996), ”Constrained Homogeneity Analysis, with Applications to Hierarchical Data,” UCLA Statistics Series, # 207

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

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

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

[28] Tenenhaus, M. and Young, F. (1985), ”An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data,” *Psychometrika*, **50**, 90-119

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

DEPARTMENT OF ENGINEERING-ECONOMIC SYSTEMS & OPERATIONS RESEARCH, STANFORD UNIVERSITY, STANFORD, CA 94305

E-mail address: gmichail@stanford.edu

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA AT LOS ANGELES, LOS ANGELES, CA 90095

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

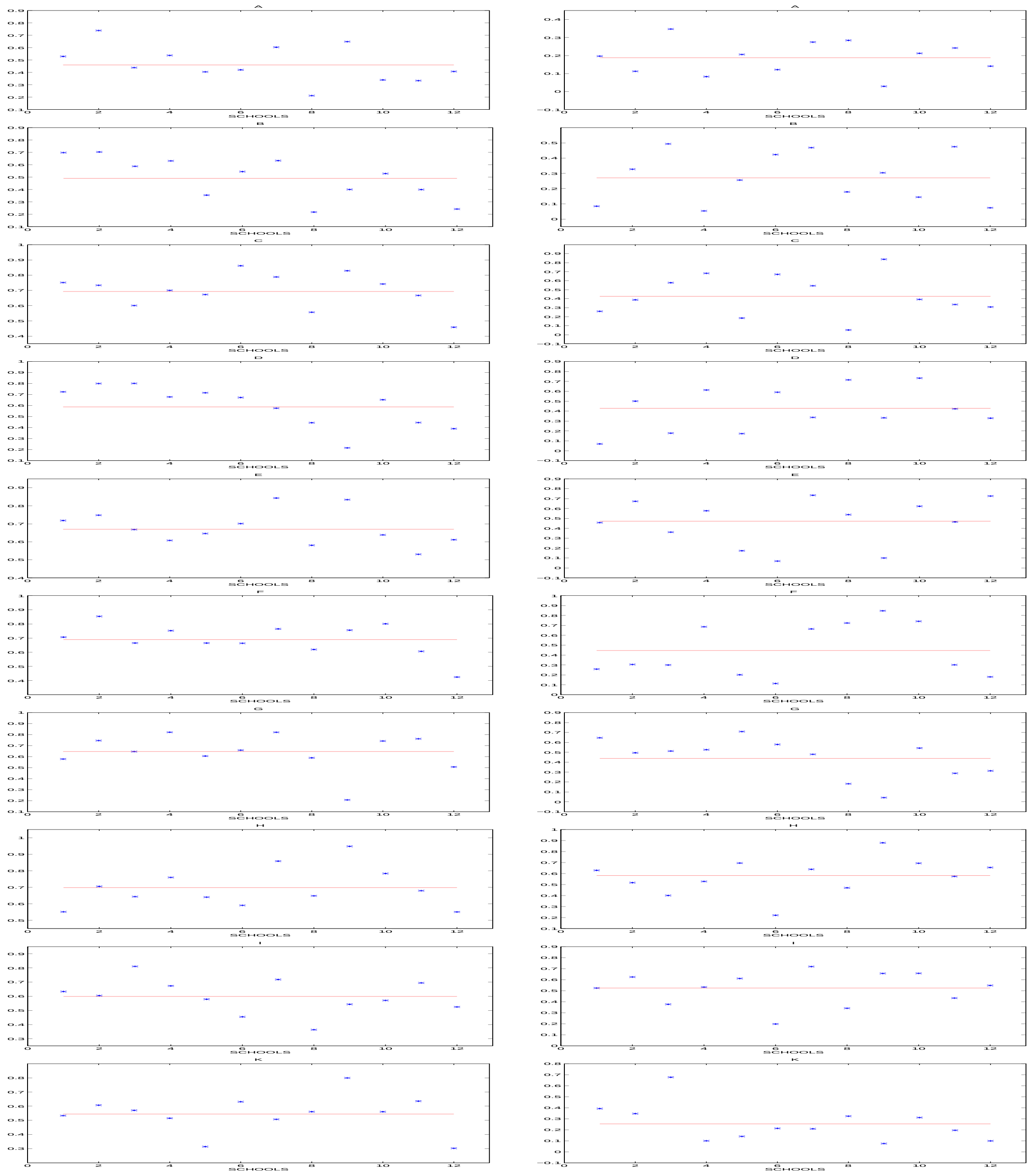

FIGURE 3.1. Discrimination Measures of the Variables for the Schools; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12; the solid line represents the variable's overall discrimination measure (Left: dimension 1, Right: dimension 2).

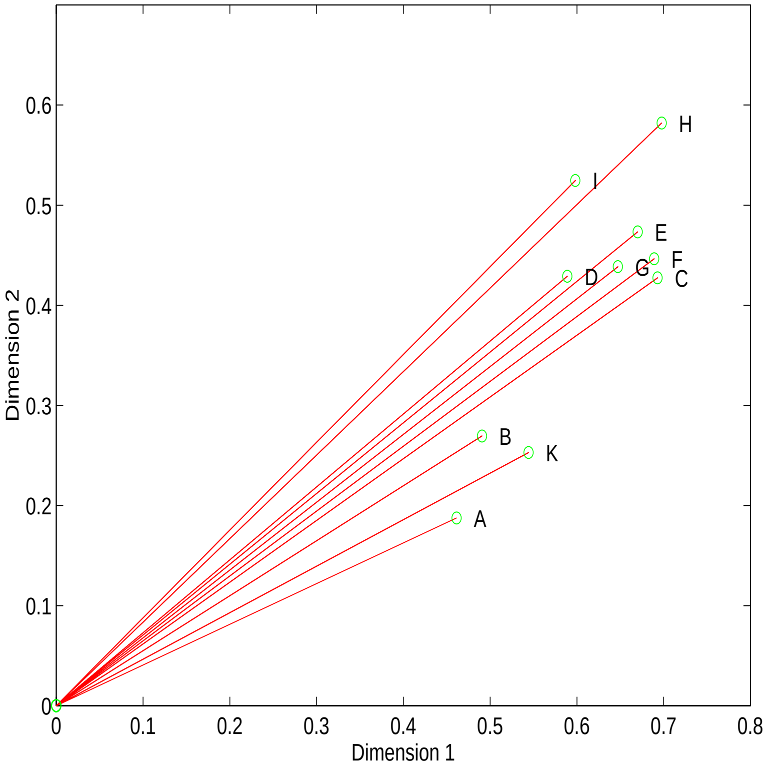

FIGURE 3.2. Total Discrimination Measures

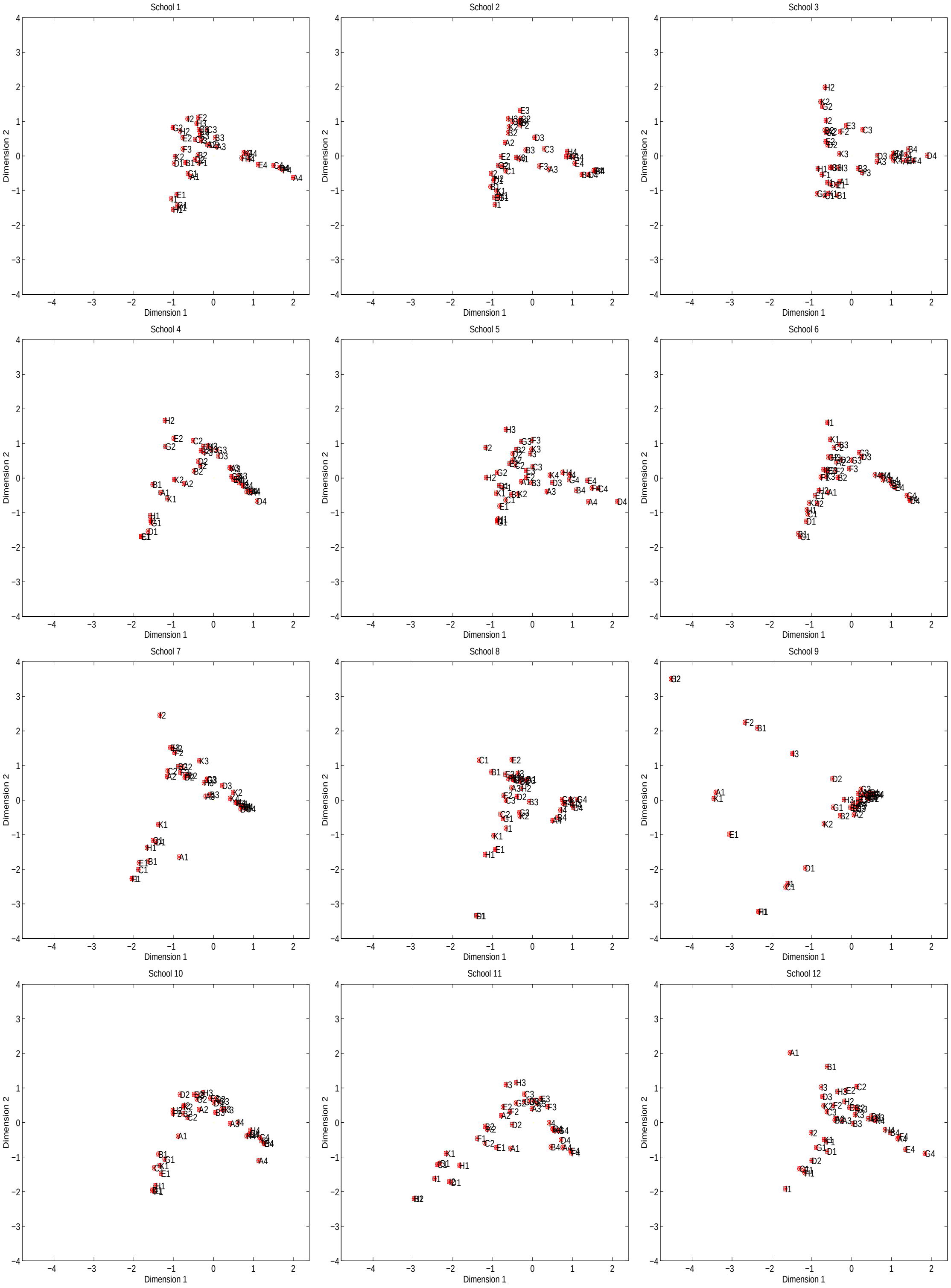

FIGURE 3.3. Optimal Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

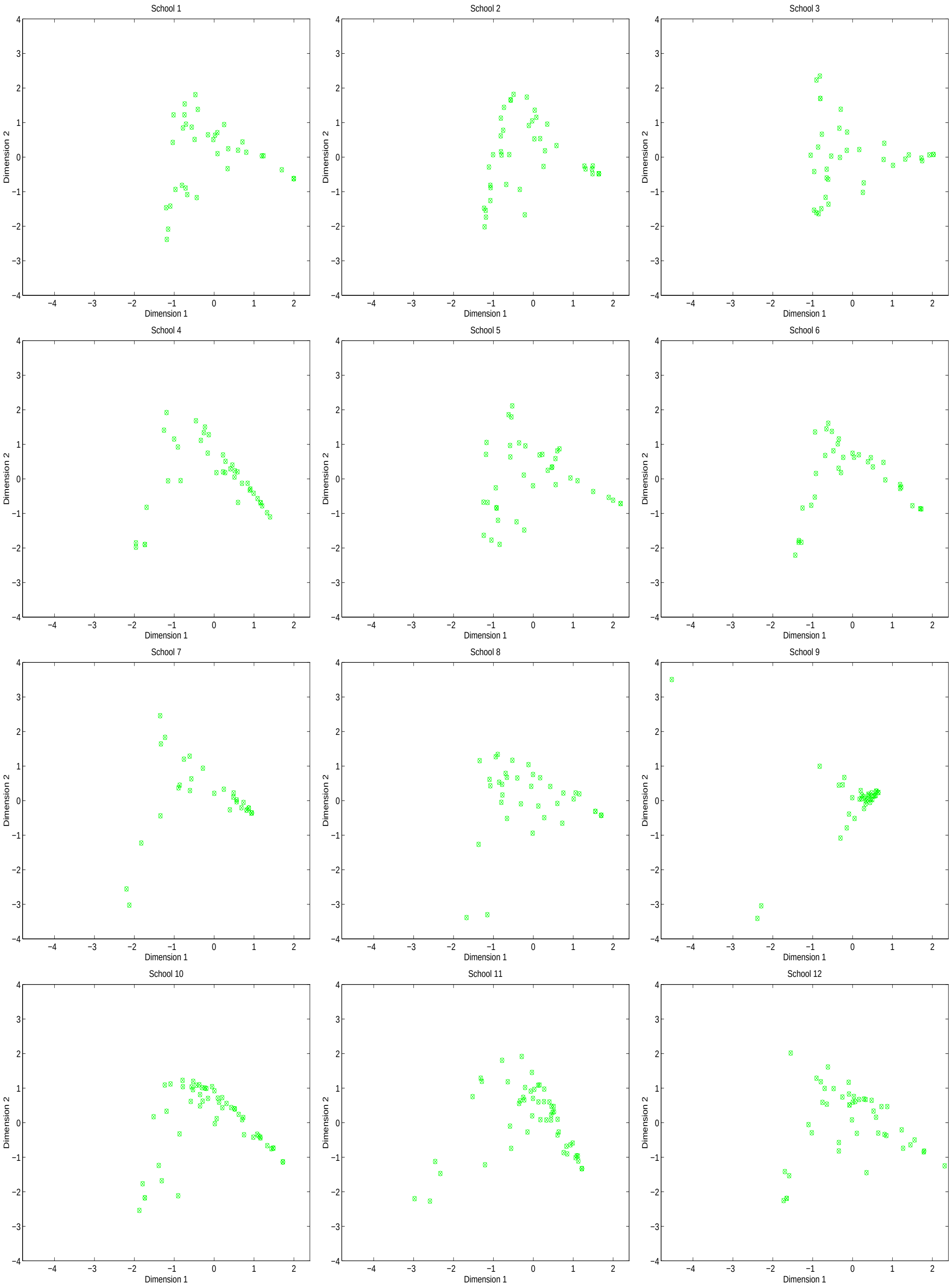

FIGURE 3.4. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

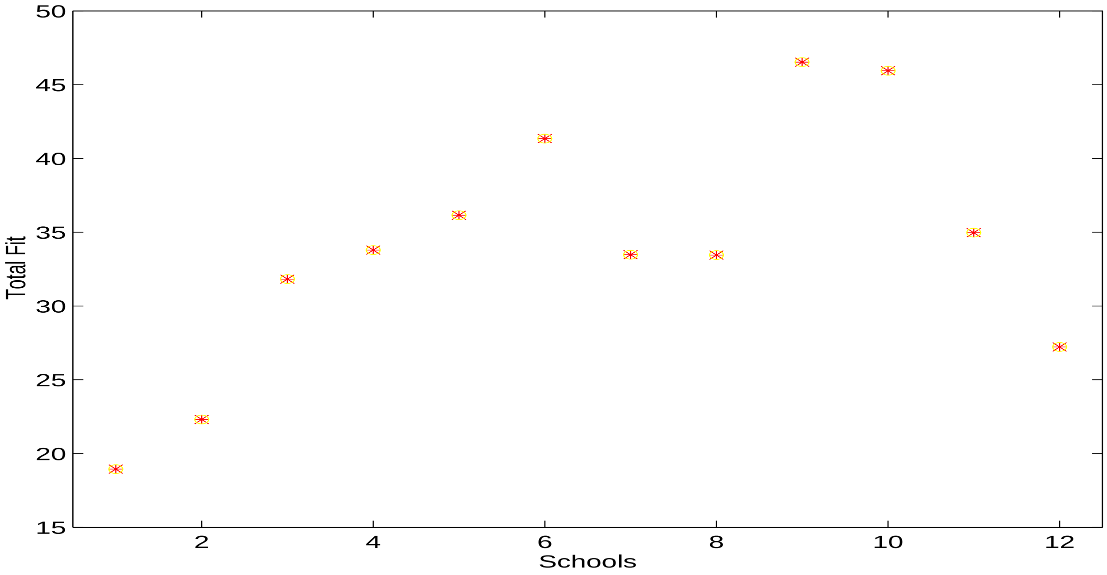

FIGURE 4.1. Total Fit by School

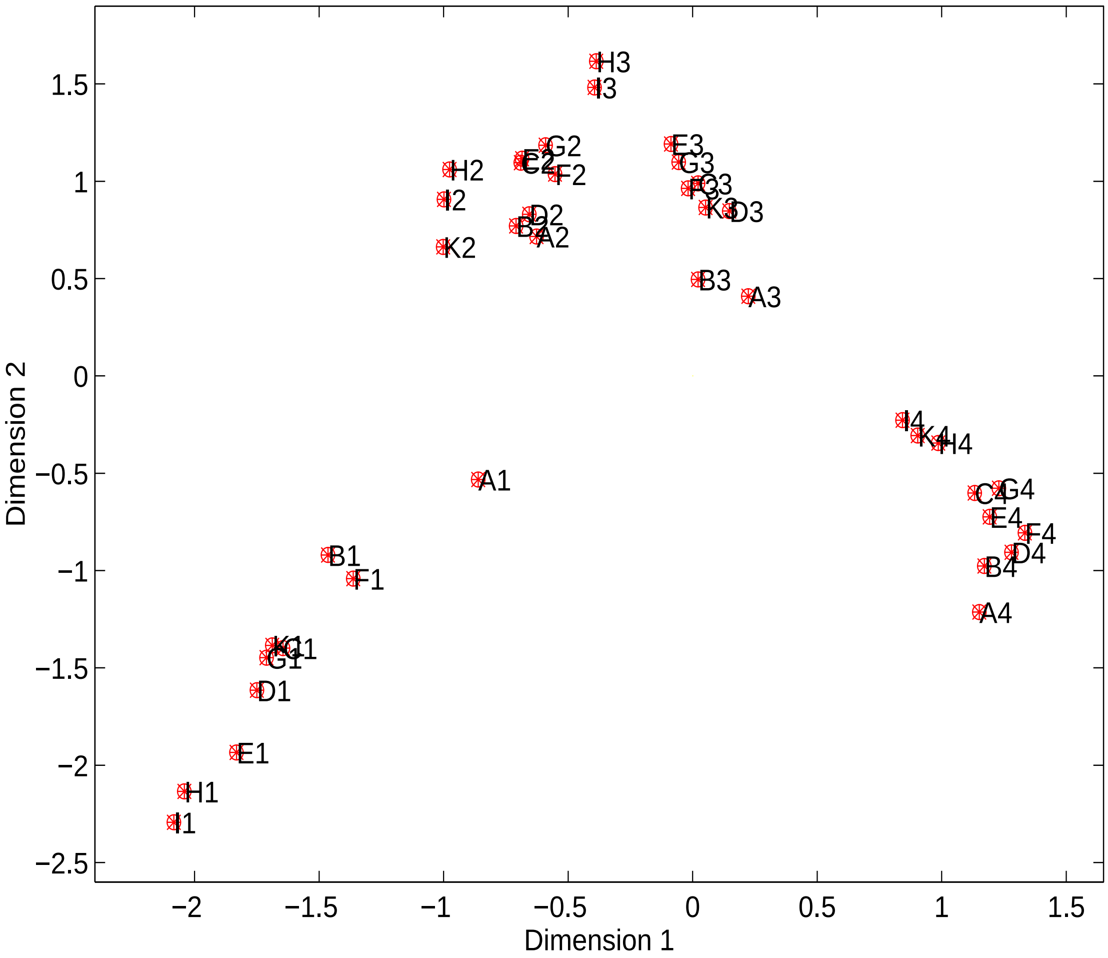

FIGURE 4.2. Category Quantifications ($Q_j,\ j \in \mathbf{J}$)

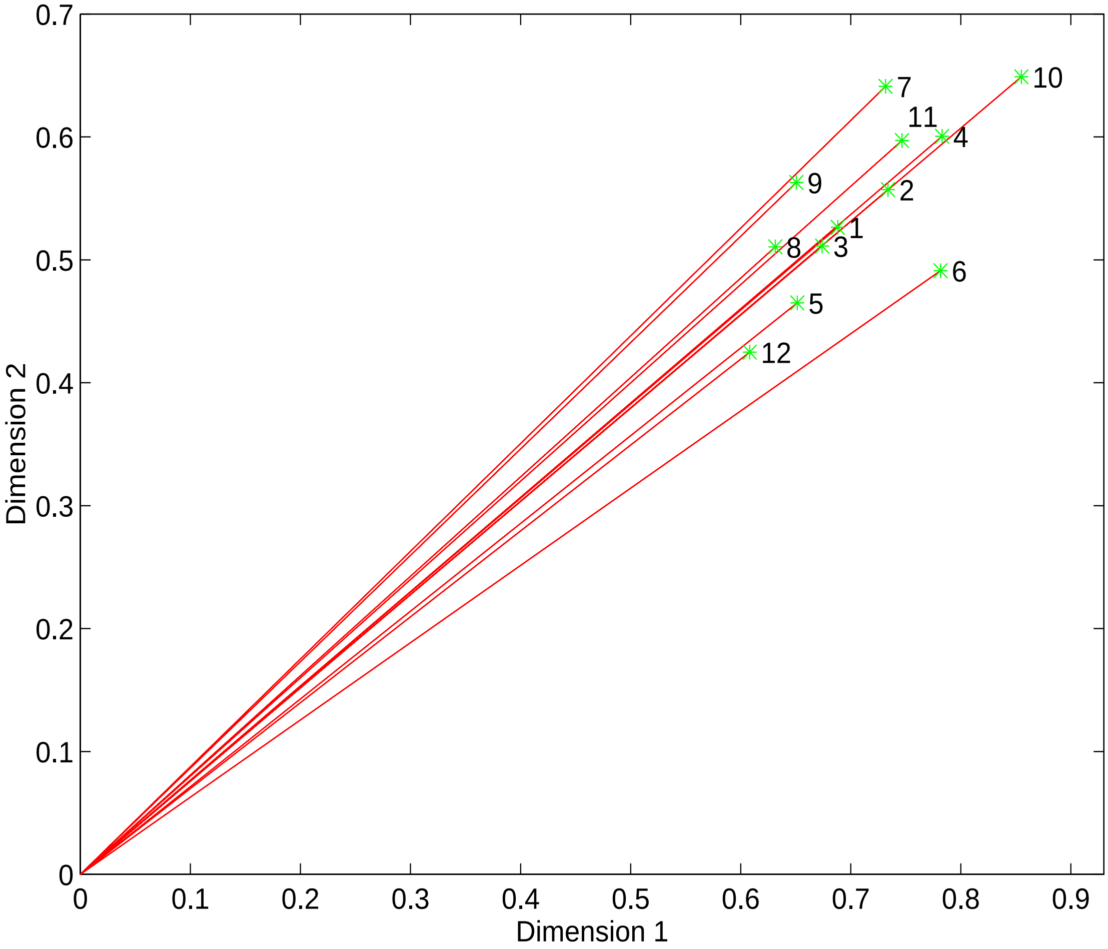

FIGURE 4.3. Slope Matrices $B_k,\ k \in \mathbf{K}$

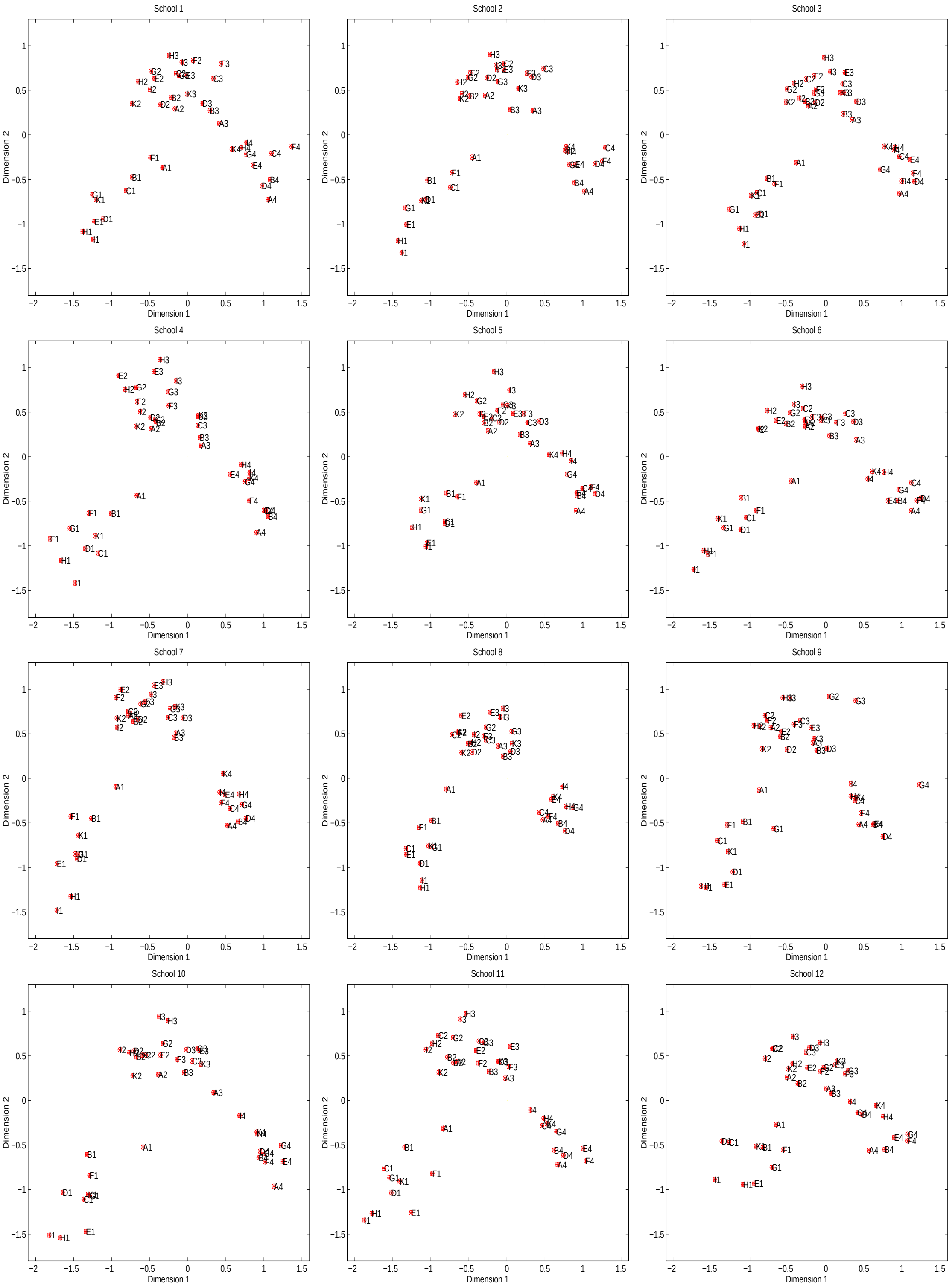

FIGURE 4.4. Optimal Cluster Category Quantifications; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12

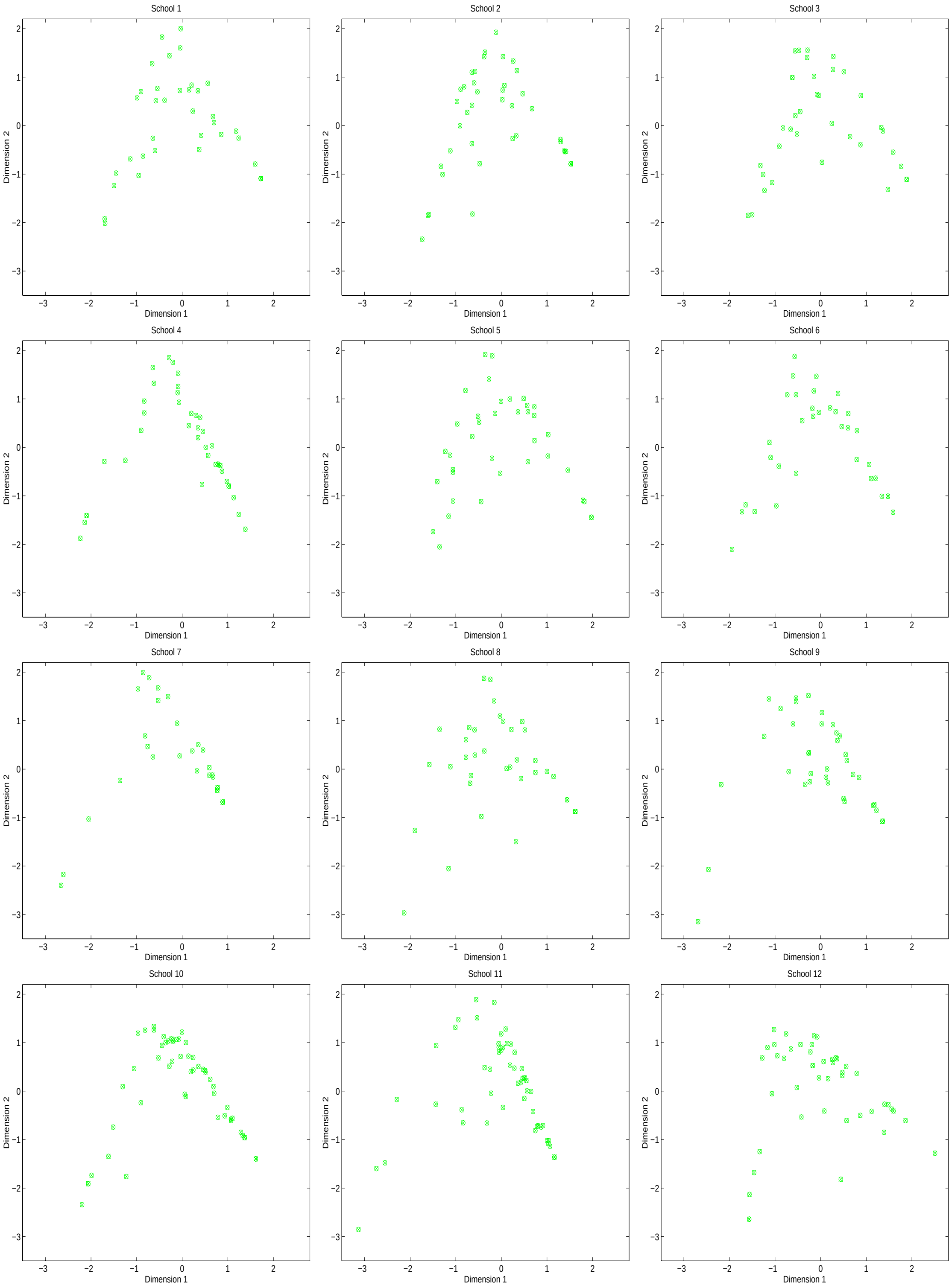

FIGURE 4.5. Object Scores; Public Urban: 1,2,3, Public Suburban: 4,5,6, Public Rural: 7,8,9, Private: 10,11,12