Nonlinear Multivariate Analysis of NELS:88

George Michailidis *Jan de Leeuw*

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90024

1 Introduction

The NELS:88 data set is a good example of a very large data set with a plethora of variables used as indicators for similar constructs (such as school climate, or parental support). We would like to use multivariate analysis techniques to study the possibility of scale construction in several subsets of NELS:88 variables. However, some of the variables are measured on interval level (*numerical* variables), some on *ordinal* level and some on *nominal* level. For ordinal data only the order of the categories (per variable) is taken into account, and for nominal data only the classes of objects formed by each variable. This mixing of measurement levels may present problems at analysis time. Standard multivariate analysis techniques tend to treat all variables as numerical (based on the multivariate normal distribution model) or as nominal.

In this paper we review some nonlinear multivariate techniques that allow for mixed measurement levels. The data reduction obtained from these techniques can reveal interesting relationships between some of the components of NELS:88. The constructed composite variables may then be used in regression or multilevel analysis. We illustrate these techniques on three sets of variables that describe time spent on homework (BYS79), student behavior (BYSC49), and parental involvement (BYP58), taken from the student, administrator and parent base-year surveys respectively. A description of the variables in each set along with their coding is included in the Appendix. Moreover, the introduction of background variables such as gender, race etc partitions the students into various subgroups (e.g. males and females for the gender variable). We can then examine whether the results of the original nonlinear multivariate analysis exhibit differences among these subgroups. The frequencies of these background variables are also given in the Appendix.

The paper is organized as follows. In Section 2 we present the theory underlying the nonlinear multivariate analysis techniques. The empirical findings are given in Section 3, while some concluding remarks are drawn in Section 4.

2 Some Technical Background

Consider a $n \times m$ data matrix, with rows corresponding to objects (e.g. students) and columns to variables. Suppose that variable $j = 1, ..., m$ takes $k_j \in \mathcal{Z}_+$ different values (its categories). Let G_j denote the $n \times k_j$ *indicator* matrix corresponding to this variable. It is a binary matrix with entries $g_{il} = 1$ $(i = 1, ..., n,\ l = 1, ..., k_j)$, if object i belongs to category l, and $g_{il} = 0$ if it belongs to another category. The concept of *homogeneity* plays a fundamental role in what follows. In the present context homogeneity is used in a data theoretic sense and is closely related to the concept of data reduction. Homogeneity refers to the extent that different variables measure the same characteristic or characteristics [4]. Hence, homogeneity specifies a type of similarity. Let y_j be a $k_j \times 1$ vector, containing the category *quantifications* of the j^{th} variable. Then, $G_j y_j$ gives a single quantification (transformation) of the n objects, induced by variable j. Without imposing any further conditions on the vector y_j the quantification is solely determined by the ties in the data, i.e. the objects that belong to the same category receive the same quantification. If we decide to work with p simultaneous quantifications for each variable, we can collect them in a $k_j \times p$ matrix Y_j, which is called the *multiple nominal* quantification of variable j. Hence, the matrices $G_j Y_j$ induce m multiple quantifications of the objects. We have perfect homogeneity in case all multiple quantifications of the objects are the same (see [3]). What we would like to achieve is to minimize the loss of homogeneity, with loss defined in terms of squared deviations, over normalized object quantifications

$$\begin{aligned} \sigma(X; Y_1, ..., Y_m) &= m^{-1} \sum_{j=1}^{m} \mathrm{SSQ}(X - G_j Y_j), \qquad (1) \\ \text{subject to} \quad & X'X = nI_p \ \text{ and } \ X'u = 0 \end{aligned}$$

where SSQ denotes the sum of squares, u is the unit vector and I_p the identity matrix. The $n \times p$ matrix X contains the new representation of the n objects in the lower dimensional space (*object scores*), while the $k_j \times p$ matrix Y_j the new category quantifications, where p is the dimensionality of the new space. The condition $X'u = 0$ implies that X is in deviations from the column means, while $X'X = nI_p$ makes the columns of X uncorrelated, with variance equal to 1 (in the absence of missing data). The following *alternating least squares* algorithm (Homals) minimizes the loss function in (1). It starts with an arbitrarily normalized X and then computes the optimal Y_j by

$$Y_j = (G_j'G_j)^{-1}G_j'X \tag{2}$$

and then uses the optimal Y_j to compute the new optimal X by

$$X = m^{-1}\sum_{j=1}^{m} G_jY_j. \tag{3}$$

The optimal X is then orthonormalized by the Gram-Schmidt procedure and the algorithm goes back to (2) until convergence. In words, the optimal coordinates for a variable category is the *centroid* of the (optimal) coordinates of the objects that fall in that category. Similarly, the optimal coordinate of an object is the centroid of the (optimal) coordinates of the categories containing that object. The Homals algorithm also allows us to calculate the *discrimination measures*, one for each variable and each dimension, defined by $\eta_{js} = n^{-1}[y'_{(j)s}(G_j'G_j)y_{(j)s}]$, $j = 1, ..., m$, $s = 1, ..., p$ (where $y_{(j)s}$ is the quantification for variable j in the s^{th} dimension of the solution).

In homogeneity analysis when $p \geq 2$ we work with multiple quantifications. Each dimension adds another quantification of the categories of each variable, and the different quantifications of the same variable have usually no simple relation to each other. Adding rank-one restrictions allows us to have multidimensional solutions for object scores with only a *single* quantification for the various categories of each variable. The rank-one restriction is given by

$$Y_j = z_ja_j', \tag{4}$$

with the additional requirements that $u'(G_j'G_j')y_j = 0$ and $y_j'(G_j'G_j)y_j = 1$, where z_j is the $k_j \times 1$ vector of *single category quantifications*, and a_j the $p \times 1$ vector of *weights*. Hence, the Y_j quantification matrix is restricted to have rank-one, i.e. its columns are proportional to each other. In the absence of missing data, the elements of a_j can also be interpreted as ordinary *component loadings*. If no further conditions are imposed on the single quantifications z_j, we then deal with *single nominal* quantifications. If we deal with ordinal data we can require that the elements of z_j be in the appropriate order. This defines the *single ordinal* treatment of a variable. Finally, in case we deal with numerical variables, we can impose linear restrictions on the elements of z_j, which in turn defines the *single numerical* treatment of a variable. The Princals algorithm adds one more step to the Homals algorithm, corresponding to relation (4). In the presence of ordinal variables the algorithm uses a monotone regression (see [5]), while in the presence of numerical variables it uses the ordinary linear regression.The combination of homogeneity analysis with the rank-one restrictions defines a form of nonlinear principal components. In this presentation the techniques are interpreted from a geometric point of view. A different starting point for the development of these techniques is given in [1] and [2].

In most empirical applications there are missing data. This leads to incomplete indicator matrices G_j, since for some objects the corresponding row of G_j will have only zero values. In both the Homals and Princals algorithms such missing data are treated as *passive*. This means that when category quantifications are computed as averages of object scores, these averages are taken only over objects with non-missing data. Moreover, in the alternating step where object scores are calculated as means of category quantifications, the averages are taken only over the non-missing categories. Some alternative ways for treating missing data are given in [4] (pages 73-76).

3 Empirical Findings

In this section we report the results of our analysis using the Homals and Princals algorithms on the three sets of variables considered in the present study.

3.1 BYS79 - Student Homework

We analyzed this set of variables using the Homals algorithm (thus treating the variables as multiple nominal). A two dimensional solution is considered, that produces a quite good fit (eigenvalues .44 and .31 respectively). These eigenvalues in the absence of missing data are interpreted as squared canonical correlation coefficients between the optimally quantified variables and the object scores.

A variable *discriminates* better to the extent that its category quantifications are further apart. Figure 1 displays the discrimination measures (whose average equals the eigenvalue in the p^{th} dimension) of the five variables. Geometrically, the discrimination measures give the averaged squared distance of category points weighted by their marginal frequencies to the origin. It can be seen that all variables discriminate equally well in both dimensions. However, variables C (english homework), D (social studies homework) and to a certain extent B (science homework) discriminate slightly better than the other ones. The variable category quantification plot is given in Figure 2. The points in the graph represent the centers of gravity of the object points associated with each category.

Clear regions of student homework patterns are revealed in this plot. In the lower

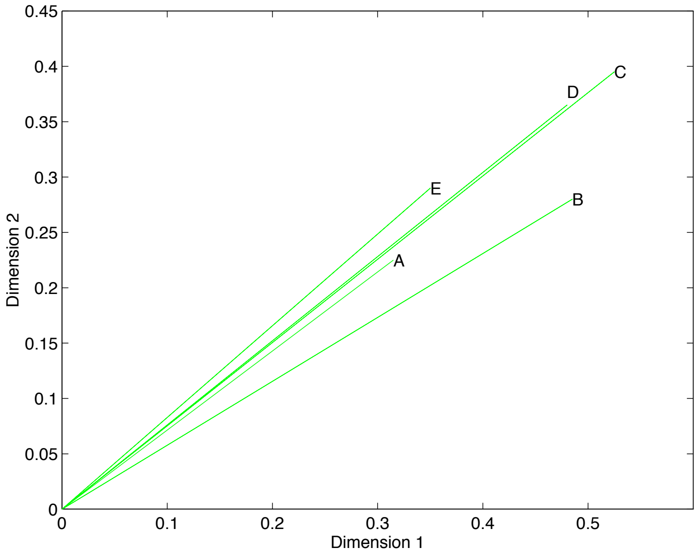

Figure 1: Discrimination measures of homework variables

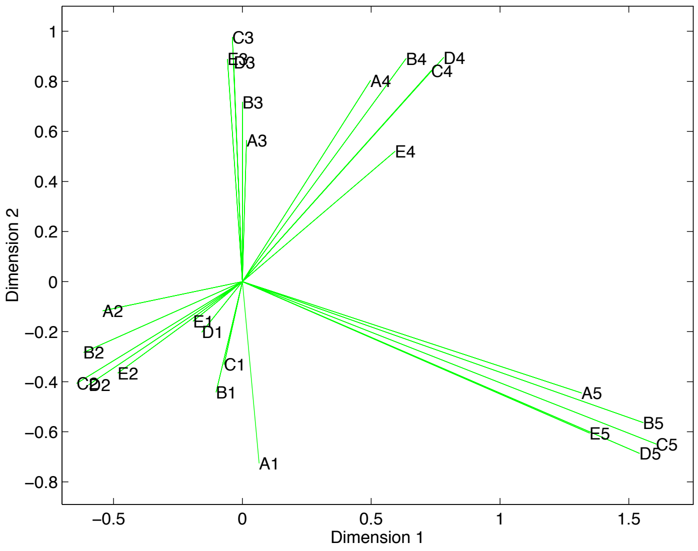

Figure 2: Optimal category quantifications of homework variables

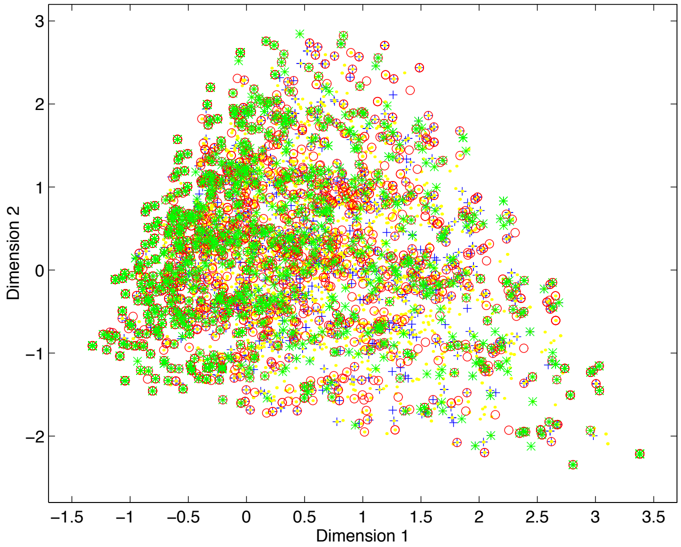

Figure 3: Object scores of homework variables (*=Asian, o=Hispanic, +=Black, .=White)

right quadrant are the variable categories associated with the highest levels of time spent on homework for all subjects. Thus, students in this area of the plot are associated with these categories. To a large extent, the analysis separates the extremely studious from the rest of the students. In the lower left quadrant we identify the students that are associated with the two lower levels of time spent on homework. Finally, in the upper right quadrant we find the variable categories associated with levels 3 and 4 of time spent on homework. It is interesting to observe that in the optimal Homals solution the "clustering" of the students is done according to the same category levels. Thus, students spend approximately equal amounts of time studying the various subjects, or putting it differently, studious students spend a lot of time on each subject's homework, while students that study little do that consistently for each subject. The analysis clearly identifies 3 distinct regions relating to the amount of time students spend on doing homework. Moreover, it also reveals distinctly nonlinear patterns; that is, variable categories are not linear with the dimensions of the space.

Figure 3 contains the plot of student scores obtained from the two dimensional solution. The distance between two student scores is related to the homogeneity of their response patterns (profiles). Hence, students with identical patterns are plotted as identical points. In Figure 3 we used race as a passive variable (did not participate in the initial analysis) to separate the student scores. The plot does not reveal any differences between the various races. This implies that students from different ethnic groups exhibit similar patterns regarding their studying habits. Analogous results are obtained by examining the object score plots with gender and type of school (i.e. public-private or rural-suburban-urban) used as passive variables.

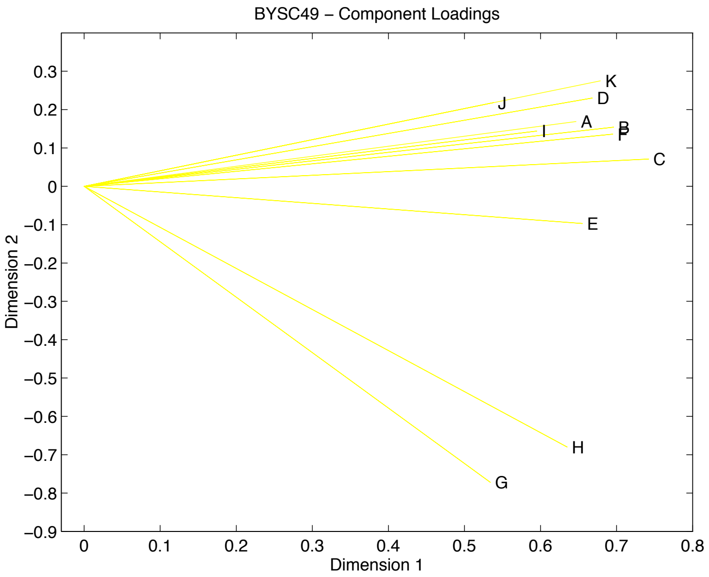

Figure 4: Component loadings of student behavior variables

3.2 BYSC49 - Student behavior

We analyzed this set of variables using the Princals algorithm. One reason is that we wanted to treat the variables as single ordinal and the other to avoid the familiar horseshoe pattern in the object scores plot (see [4] pages 147-148) that the Homals solution produced. The fit of a two dimensional solution was satisfactory given the large number of variables involved (eigenvalues .44 and .16 respectively). In Figure 4 the component loadings are displayed. The component loadings are the correlations of the optimally quantified variables with the object scores in the absence of missing data (hence, discrimination measures can also be interpreted as squared component loadings). In case the arrows (vectors) in the plot are (almost) of unit length (in the usual Euclidean norm) the angle between any two of them reflects the value of the correlation coefficient between the two corresponding quantified variables. The component loadings plot shows that variables G and H (use of alcohol and illegal drugs) discriminate very well along the second dimension and satisfactorily along the first dimension, while the remaining variables discriminate well only along the first dimension. Hence, we can say that the second dimension reflects students' perceptions on whether the use of alcohol and drugs presents a problem at their school, while the first dimension mainly summarizes their perception regarding issues such as absenteeism, tardiness, vandalism, physical and verbal abuse of teachers etc. Moreover, variables G and H are highly correlated (since their component loading vectors have approximately unit length) and lowly correlated with the remaining variables. The variable category quantification plot is given in Figure 5. The first thing to notice is that quantifications of each variable are ordered and lie on a line that goes through the origin; a property of the Princals solution for single ordinal variables. On the right side of the graph are "clustered" students that think that none of

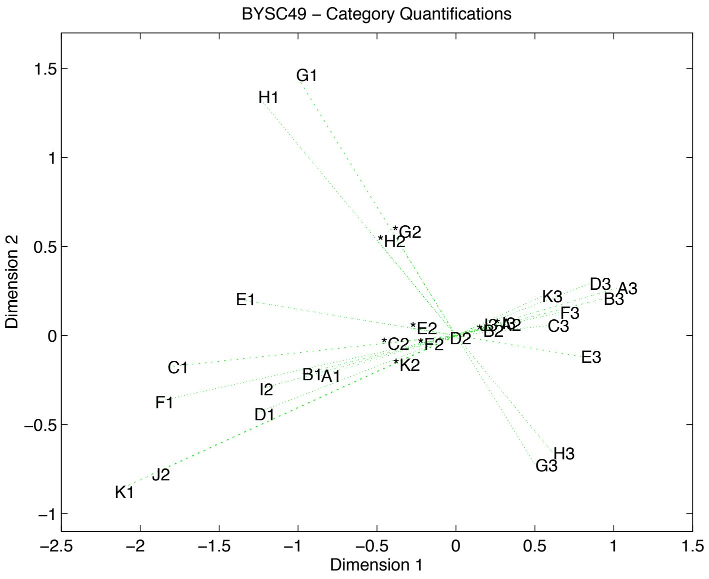

Figure 5: Optimal category quantifications of student behavior variables

the issues addressed by variables A through K is a problem at their school, while on the left students who consider them as presenting a moderate to a serious problem. Finally, in the center of the graph (around the origin) are the students that think these issues represent a minor problem at their school (middle categories). The Princals solution manages to identify several patterns regarding the type of issues students perceive to present a problem at their school.

At this point it is interesting to examine the object scores stratified by some background variables (treated as passive ones). Similarly, to the BYS79 set of variables we do not observe any differences for gender and race. However, as Figures 6 and 7 indicate there are significant differences regarding perceptions of students in public and private schools, and also among schools in rural, suburban and urban areas. We observe that the majority of the students attending private schools thinks that the issues addressed by the BYSC49 set of variables represent at most a minor problem at their schools, contrary to the beliefs of the students in public schools. Similar findings hold for the majority of students attending schools in suburban and rural areas. On the other hand, students attending schools in urban areas think that especially issues such as physical and verbal abuse of their teachers, as well as student possession of weapons, physical conflicts among students, robbery and theft are serious problems at their schools. These findings suggest that separate analyses for private and public schools (or for urban, suburban and rural schools) might provide a better insight, or that techniques that accommodate several sets of variables such as Overals (see [4], chapter 5) might prove useful.

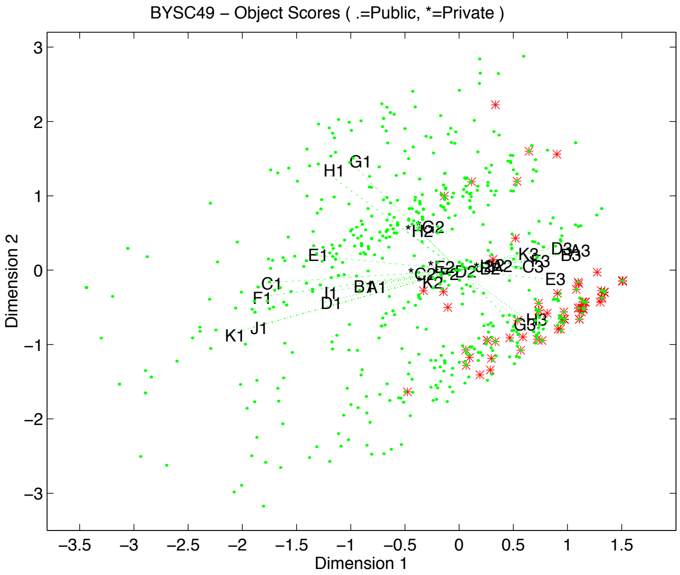

Figure 6: Object scores of student behavior variables (.=Public School,*=Private School)

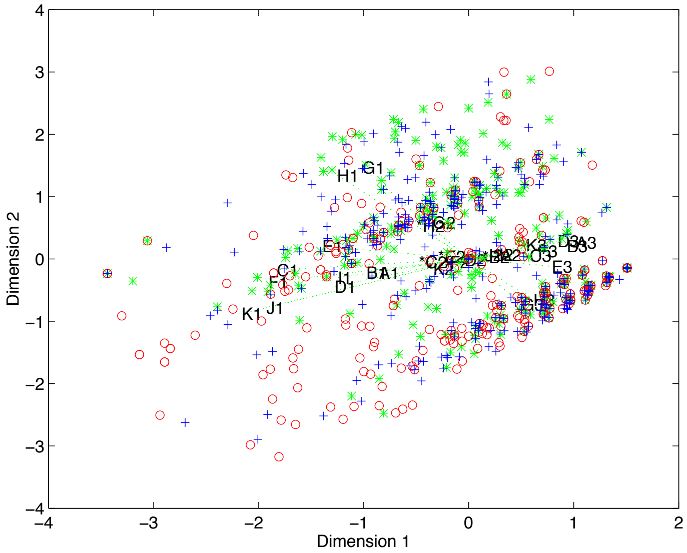

Figure 7: Object scores of student behavior variables (o=Urban School,+=Suburban School, *=Rural School)

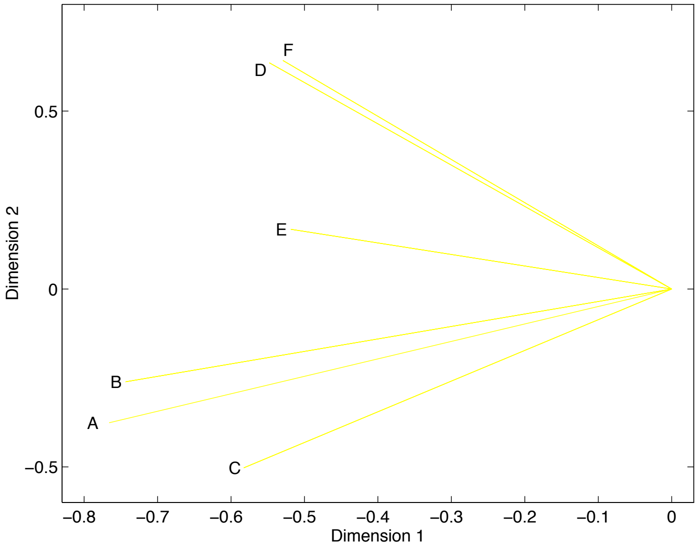

Figure 8: Component loadings for parental involvement variables

3.3 BYP58 - Parental Involvement

A two dimensional Princals analysis was performed, and a satisfactory fit was obtained (eigenvalues .39 and .23 respectively). In Figure 8 the component loadings of the parental involvement variables are presented. It can be seen that the first dimension summarizes variables associated with student's performance and behavior at school, while the second dimension with variables associated with parental involvement in general school activities. Moreover, variables A, B and C are highly correlated among themselves and so are variables D and F, while these two groups of variables are not highly correlated. In Figure 9 the optimal category quantifications are given. On the right side of the graph are "clustered" parents that showed no interest in their child's academic performance and school behavior and that were not involved in general school activities. On the other hand, on the lower left quadrant are parents who contact the school often about their children performance, while on the upper left quadrant parents that are actively involved only in general school activities. Finally, it should be noted that the object scores stratified by gender, race and type of school variables did not reveal significant differences among the various categories.

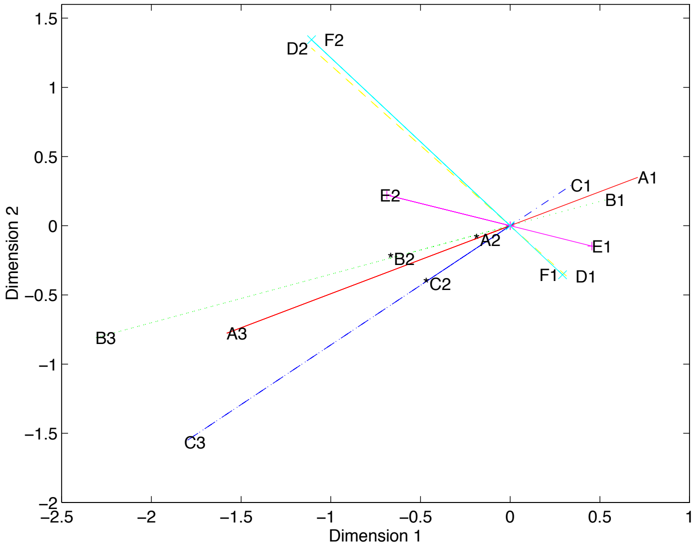

Figure 9: Optimal category quantifications for parental involvement variables

3.4 Stability Analysis

There has been a lot of work on stability analysis of the Homals and Princals algorithms (see chapter 12 in [4]). The first procedure to assess the stability of our solution relies on *permutation* methods. The basic idea is to destroy the existing *structure* in the data by permuting the objects within the columns of the data matrix, and then study the *permutation distribution* of the eigenvalues. We then examine the position of the eigenvalues computed from the original data matrix on the permutation distribution. If they are located far out in the tails, we can conclude that the solution was stable. This is one way of formalizing the notion of "no structure". The results of the permutation method are given in the following Table.

Table I
Permutation Averages and Standard Errors of Eigenvalues
based on 100 replications

	Dimension 1			Dimension 2		
Variable	Princals Value	Average Value	Standard Error	Princals Value	Average Value	Standard Error
BYS79	.440	.037	.0063	.310	.001	.0029
BYSC49	.437	.110	.0019	.155	.099	.0016
BYP58	.390	.030	.0030	.228	.002	.0023

It is easy to see that the solutions we obtained were due to some "structure" in the data and could not be attributed to chance.

Another procedure to assess the stability of the solution is based on the *bootstrap* method. Here the idea is to sample with replacement objects and then compute the bootstraped statistic of interest (category quantifications and component loadings for the Homals and Princals algorithms respectively). We then examine the position of the statistic computed from the original data matrix on the bootstrap distribution and proceed as before. The results of the bootstrap method for the three sets of variables are given in Tables 8, 9, and Figure 10 in the Appendix and indicate that our results were very stable.

4 Concluding Remarks

In the present study we have reviewed some nonlinear multivariate techniques and applied them to subsets of the NELS:88 data set. Our examples attempted to illustrate the most important geometrical features of these techniques. Our findings indicate the presence of strong nonlinear patterns among the variables and justify the use of these techniques.

It is worth noting that the data set was treated as a simple random sample from the student population. However, the sample was stratified by school and we did not attempt to take into account this fact. There are several ways to introduce this additional information in the analysis, but none of them seems easy to implement. One way is to incorporate the school variable in the analysis and treat it as multiple nominal. However, the results were not very satisfactory. Hence, this issue remains open for future research.

5 Appendix

5.1 BYS79

”Time spent on homework each week”

A Mathematics homework

B Science homework

C English homework

D Social studies homework

E Homework for all other subjects

Table 1 (%) (N=24,599)

Variable	Categories					
	1	2	3	4	5	Missing
A	7.9	39.4	21.7	10.0	16.2	4.8
B	16.0	42.6	19.4	8.8	8.1	5.2
C	10.1	43.6	21.1	9.7	10.1	5.4
D	12.8	39.1	21.6	10.4	10.7	5.5
E	13.4	38.0	19.4	11.3	12.6	5.2

Coding: 1 = Less than an hour, 2 = 1 hour, 3 = 2 hours, 3 = 3 hours, 5 = 4 or more hours, 9 = Missing

5.2 BYSC49

"Indicate the degree to which each of the following matters is a problem in your school"

A Student tardiness

B Student absenteeism

C Student class cutting

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

J Physical abuse of teachers

K Verbal abuse of teachers

Table 2 (%) (N=24,599)

Variable	Categories			
	1	2	3	Missing
A	32.2	51.4	14.8	1.6
B	28.8	46.2	23.3	1.7
C	8.6	34.7	55.2	1.4
D	15.7	56.4	26.4	1.5
E	8.4	55.1	35.1	1.4
F	7.6	50.4	40.6	1.4
G	9.2	39.3	50.1	1.4
H	7.1	42.7	48.8	1.4
I		18.5	80.0	1.5
J		7.3	91.1	1.6
K	5.2	42.1	51.1	1.6

Coding: 1 = Moderate - serious, 2 = Minor, 3 = Not a Problem, 9 = Missing

5.3 BYP58

”Since your eighth grader’s school opened last fall, how many times HAVE YOU OR YOUR SPOUSE/PARTNER CONTACTED the school about each of the following?”

A Your eighth grader’s academic performance

B Your eighth grader’s academic program for this year

C Your eighth grader’s behavior in school

D Participating in school fund raising activities

E Providing information for school records (address, work telephone number)

F Doing volunteer work (supervising lunch, chaperoning a field trip)

Table 3 (%) (N=24,599)

Variable	Categories			
	1	2	3	Missing
A	41.0	30.9	14.6	13.4
B	55.6	25.3	5.2	13.8
C	61.5	17.9	13.6	
D	67.0	19.2		13.8
E	52.0	34.3		13.6
F	68.2	18.1		13.7

Coding: 1 = None, 2 = Once or twice, 3 = Three or more, 9 = Missing

5.4 Background Variables

Table 4
School Location
(%) (N=24,599)

Urban	31.0
Suburban	41.7
Rural	27.4

Table 5
School Type
(%) (N=24,599)

Public	78.8
Private	21.2

Table 6
Gender
(%) (N=24,599)

Male	49.8
Female	50.2

Table 7
Race
(%) (N=24,599)

Asian or Pacific Islander	6.2
Hispanic, regardless of race	12.9
Black, not of Hispanic origin	12.2
White, not of Hispanic origin	66.3
Other, missing	2.3

5.5 Stability Analysis

Table 8
Bootstrap Averages and Standard Errors of Component Loadings
based on 100 replications for the BYSC49 set of variables

	Dimension 1			Dimension 2		
Variables	Princals	Bias Corrected	Standard	Princals	Bias Corrected	Standard
	Value	Value	Error	Value	Value	Error
A	.647	.650	.0142	.169	.173	.0119
B	.696	.701	.0150	.154	.151	.0126
C	.742	.734	.0236	.071	.075	.0298
D	.668	.676	.0185	.230	.226	.0221
E	.655	.651	.0153	-.097	-.093	.0164
F	.695	.702	.0209	.136	.141	.0442
G	.534	.538	.0159	-.772	-.775	.0145
H	.635	.641	.0154	-.680	-.685	.0273
I	.595	.598	.0136	.144	.148	.0157
J	.538	.542	.0112	.218	.213	.0176
K	.679	.684	.0169	.275	.269	.0195

Table 9
Bootstrap Averages and Standard Errors of Component Loadings
based on 100 replications for the BYP58 set of variables

	Dimension 1			Dimension 2		
Variables	Princals	Bias Corrected	Standard	Princals	Bias Corrected	Standard
	Value	Value	Error	Value	Value	Error
A	-.764	-.766	.0121	-.377	-.376	.0211
B	-.744	-.744	.0143	-.262	-.261	.0141
C	-.582	-.583	.0112	-.504	-.503	.0151
D	-.550	-.548	.0135	.635	.636	.0168
E	-.521	-.519	.0228	.167	.168	.0147
F	-.532	-.529	.0194	.637	.642	.0267

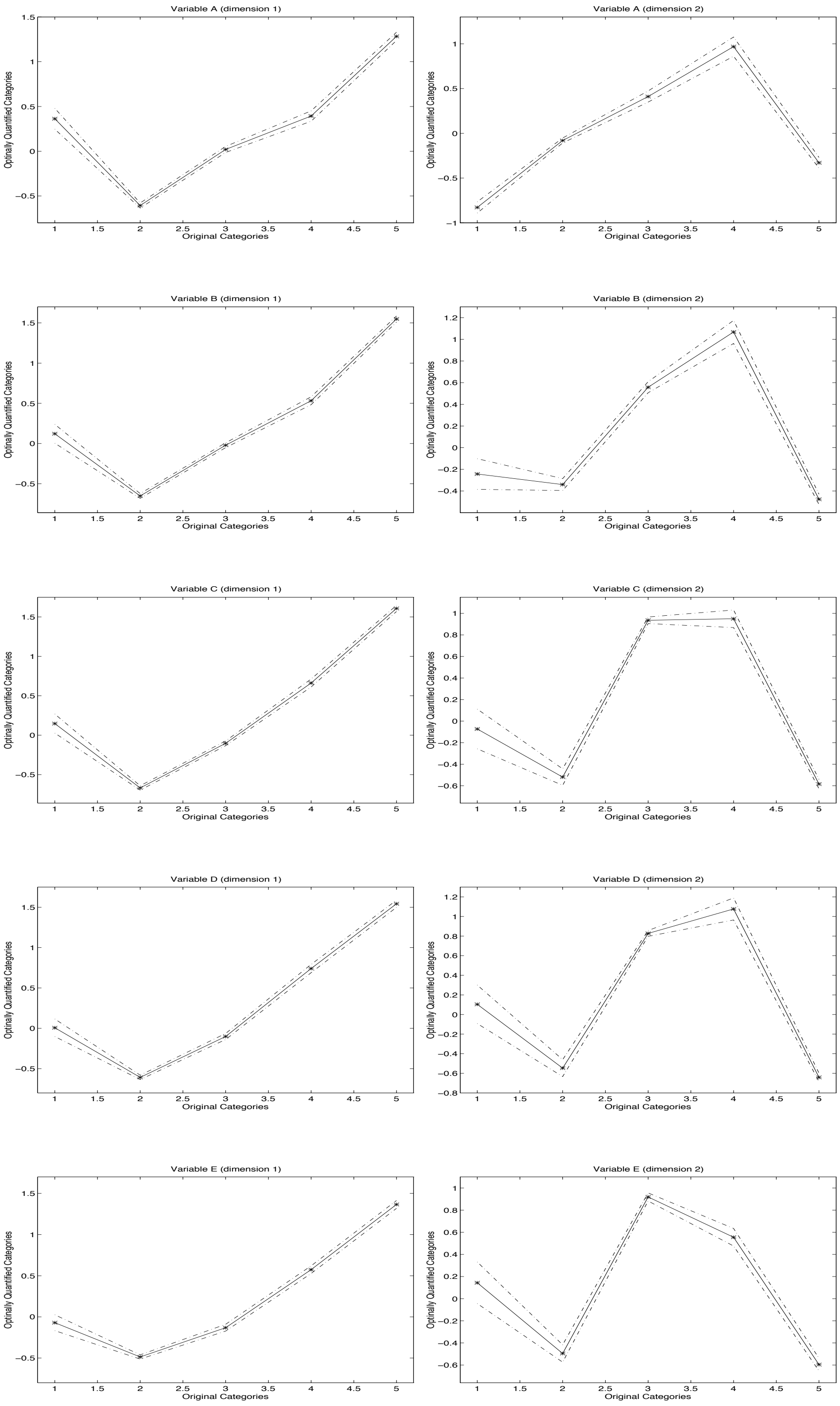

Figure 10: Bias corrected bootstrap averages and 95% confidence intervals of category quantifications based on 100 replications for the BYS79 set of variables

References

[1] De Leeuw, J. (1980), "Homals and Princals: Some Generalizations of Principal Component Analysis", *Data Analysis and Informatics*, E. Diday et al. (eds), North-Holland, Amsterdam

[2] De Leeuw, J. (1982), "Nonlinear Principal Component Analysis", *COMPSTAT*, Physica-Verlag, Vienna

[3] De Leeuw, J. (1984), *Canonical Correlation Analysis of Categorical Data*, DSWO Press, Leiden

[4] Gifi, A. (1990), *Nonlinear Multivariate Analysis*, John Wiley & Sons, Chichester

[5] Kruskal, J.B. (1964), "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis", *Psychometrica*, **29**, 1-27