Exploratory Multilevel Redundancy Analysis

Takane, Yoshio
McGill University, Department of Psychology
1205 Dr. Pen¯eld Avenue
Montreal Quebec H3A 1B1, Canada
E-mail: takane@psych.mcgill.ca

Zhou, Lixing
McGill University, Department of Psychology
1205 Dr. Pen¯eld Avenue
Montreal Quebec H3A 1B1, Canada
E-mail: lixing-zhou@hotmail.com

1. Introduction

Hierarchically structured data are commonly encountered in many ¯elds of scienti¯c investigation. In educational assessment studies, for example, students' performance in mathematics is measured in various schools. Such data are called hierarchically structured because students are nested within schools. Another example of hierarchically structured data arise in repeated measurement designs where some attributes of subjects are repeatedly measured over time.

Hierarchical (multilevel) linear models (HLM: Bock, 1989; Bryk and Raudenbush, 1992; Goldstein, 1987; Hox, 1995) are often used to analyze such data, explicitly taking into account the hierarchical nature of the data. Interpretations of parameters in such models, however, become increasingly more di±cult as they accommodate more levels, more predictor variables, and more criterion variables. This paper presents a method of multilevel analysis with a dimension reduction feature to facilitate interpretations of model parameters. The proposed method is a multivariate extension of the procedure developed by Takane and Hunter (2002). The method ¯rst decomposes variability in the criterion variables into several orthogonal components using predictor variables at di®erent levels, and then applies singular value decomposition (SVD) to the decomposed parts to ¯nd more parsimonious representations. An example is given to illustrate the method. Some possible extensions of the proposed method are also suggested.

2. The Model

For illustration, let us consider the following situation. (This situation closely resembles the example given later.) Suppose we are interested in assessing what attributes (factors) of students and their environments a®ect their performance in mathematics. For this, we measure students' performance in mathematics. We may use multiple tests to obtain reliable test scores and to capture all important aspects of math performance. Students (the ¯rst-level units) are usually nested within schools (the second-level units). We also collect relevant information about the students and schools they belong to.

We may ask several questions in this context: 1. How much of the students' math performance can or cannot be explained by school di®erences. The former is referred to as the between-school e®ects, and the latter as the within-school e®ects. 2. How much of the between-school e®ects can be explained by known school characteristics (the school-level predictor variables) and in what way do the school-level predictor variables a®ect student performance? 3. How much of the within-school e®ects can be explained by prescribed subject characteristics (the subject-level predictor variables), and in what way do the subject-level predictor variables a®ect student performance? 4. Are there any interactions between the school-level and student-level predictor variables that a®ect student performance? HLM allows us to investigate and answer all of these questions.

student-level predictor variables.) The ¯fth term represents the portion of the within-school e®ects that can be explained by the interactions between the school-level and student-level predictor variables. (The matrix $D_{X^*} W_1^*$ represents the interactions between the two.) The sixth term pertains to the portion of the interactions between schools and the student-level predictor variables that cannot be explained by the fourth and ¯fth terms. Finally, the last term in the model represents residuals left unaccounted for by any systematic e®ects in the model.

There are several important special cases of the full model presented above. When no school-level predictor variables exist, neither terms 2 and 3 nor terms 5 and 6 can be isolated. In this case, the model reduces to a simple analysis of covariance model. When no student-level predictor variables exist, terms 4, 5, 6, and 7 cannot be isolated. When neither the school-level nor student-level predictor variables exist, neither terms 2 and 3 nor terms 4, 5, 6, and 7 can be isolated. In this case, we simply have a one-way ANOVA model.

3. Estimation

The seven terms in model (10) are all columnwise orthogonal and so coe±cients in each term can be separately estimated by OLS (Ordinary Least Squares). We thus have

(11) $\hat{c}_{00} = 1'_N Y = N;$

$\hat{C}_{01} = (W_0^{*\prime}$

Putting the estimates of parameters given above into (10), we obtain the following (orthogonal) decomposition of Y:

$$
\begin{aligned}
(22) \qquad Y = {} & P_{1_N} Y + P_{GW_0^*} Y + P_{GA^*} Y \\
& + P_{X^*} Y + P_{D_{X^*} W_1^*} Y + P_{D_{X^*} B^*} Y + Q_{D_{X^*}} Q_G Y;
\end{aligned}
$$

where in general $P_Z = Z(Z'Z)^{-1}Z'$ is the orthogonal projector onto Sp(Z), $A^* = Q_{[1_J;W_0^*]=G'G}$, and $B^* = Q_{[J_r;W_1^*]=D_X X}$. This decomposition of Y entails a more generic decomposition of E^N, the N-dimensional Euclidean space, which is split into the orthogonal direct-sum of the seven subspaces spanned by the orthogonal projectors preceding Y's in (22). This generic decomposition is depicted in the following table.

Table 1. The decomposition of $E^N = Sp(I_N)$.

(1) P_{1_N}	(2) $P_{GW_0^*}$	(3) P

SVD($GW_0^*\hat{C}_{01}$) for the second term in (22), for example, while the latter by GSVD$(\hat{C}_{01})_{W_0^{*\prime}G'GW_0^*}$

Criterion referenced NELS-equated proficiency scores were calculated in the form of probabilities based on a cluster of items that mark certain proficiency levels. There are five levels of proficiency in math which are hierarchically ordered in the sense that mastery of a higher level typically implies proficiency at the lower levels. The NELS-equated proficiency probabilities were computed using IRT-estimated item parameters calibrated in NELS: 88. We use the five proficiency probabilities as our criterion variables. Each proficiency probability represents the probability that a student would pass a given proficiency level. Proficiency at level 1 corresponds to simple arithmetical operations on whole numbers. Level 2 pertains to simple operations with decimals, fractions, powers, and roots. Level 3 represents simple problem solving, requiring the understanding of low level mathematical concepts. Level 4 pertains to understanding of intermediate level mathematical concepts and/or multi-step solutions to word problems. Level 5 concerns complex multi-step word problems and/or advanced mathematics material.

We eliminated students with missing data from our analysis. We also eliminated schools with fewer than 20 students in the data set. This left us with $N = 10{,}939$ students nested within $J = 562$ schools. The school-level predictor variables used are given in Table 3. Each of the statements was rated on a 5-point scale with respect to how accurate the statement was as a description of the school (1. not accurate at all to 5. very accurate). The student-level predictor variables used are shown in Table 4. There were three categorical variables. They were coded into 8 dummy variables altogether prior to the analysis.

Table 2 gives a breakdown of the total SS (SS_T) explained by the different terms in model (10).

Table 2. A breakdown of the total SS.

Between-School SS (SS_B) 17.9%		Within-School SS (SS_W) 82.1%	
SS_2	SS_3	SS_4	SS.l-1.6421t0.91Tr]TJ0NSS

Table 3. The e®ects of school-level predictor variables.

Variables	Component weights (T)
Teachers press students to achieve	.69
Teachers' morale is high	-.00
Students expected to do homework	.42

We next look at the e®ects of student-level predictor variables, which are summarized in Table 4. We estimated C_{10} in (10) and then applied GSVD. Singular values were found to be .33, .08, .01, .01, and .00, so the ¯rst component again accounted for a majority (over 97%) of the SS_4. It may be observed that male students did slightly better than female students. There are larger race di®erences among the three racial groups. White students performed better than black and Hispanic students. (Again, recall that only 2.3% of the SS_T can be accounted for by SS_4.) Contrary to people's common sense, hours spent on homework had relatively small e®ects on students' mathematical pro¯ciency. A moderate amount of time spent on homework has a small positive e®ect, while no hours or too many hours have small negative e®ects. Covariances between this component and the criterion variables (component loadings) were .12, .18, .19, .15, .04, so again the component seems to represent students' overall performance in mathematics.

Table 4. The e®ects of student-level predictors.

Variables	Categories	Component weights (T)
1.Gender	male	.28
	female	-.28
2. Race	Black	-1.90
	Hispanic	.18
	White	1.71
3. Homework	0 hours	-.13
	1-4 hours	.28
	5 or more hours	-.15

6. Discussion and Future Work

In this paper, we proposed a method for multilevel redundancy analysis. This method is particularly attractive since OLS estimates of regression parameters can be obtained in closed form. The estimated regression parameters are then subjected to rank reduction by GSVD. Reduced-rank approximations of regression parameters are useful, particularly when the dimensionality of the parameter space is high. An application of the proposed method was empirically demonstrated through a real example.

There are a number of possible extensions that can make the proposed method even more useful:

1. Although only the two-level model has been discussed in this paper, similar methods can be developed for higher-level multivariate data. The number of terms in the model, however, grows very quickly. For example, a full three-level HLM with predictor variables at all levels, there are 15 terms altogether.

2. Bootstrap (e.g., Efron, 1982) or other resampling techniques could be used to assess the stability of individual parameters, which may in turn be used to test their signi¯cance. Since the normality assumption is almost always in suspect in survey data, the bootstrap methods may also be useful to benchmark the distribution of the conventional statistics used in HLM.

3. The number of components to be retained in dimension reduction may be determined by permutation tests in a manner similar to Takane and Hwang (2002), who developed a permutation procedure

for testing the number of significant canonical correlations.

4. Additional (linear) constraints can be readily incorporated in the OLS estimation procedure. This allows the statistical tests of the hypotheses represented by the constraints.

5. When the U parameters are assumed to be random rather than fixed, observations obtained from subjects in the same schools are no longer statistically independent. The dependence structure among the observations may be estimated from the initial estimates of parameters (obtained under the independence assumption), which may then be used to re-estimate the parameters, and so on. This leads to an iterative estimation procedure for full maximum likelihood estimation (MLE) of parameters (Goldstein, 1987). A simpler method called REML (REstricted Maximum Likelihood: e.g., LaMotte, 2007) may also be of interest in this context.

6. The ridge type of regularized LS (RLS) estimation may be used instead of OLS. The RLS is easy to apply and is known to provide estimates of regression parameters which are on average closer to population parameters (Takane and Hwang, 2007; Takane and Jung, 2008).

7. Interesting special cases arise when we set $Y = Q_{1_N} X = \tilde{X}$ and or $Y = D_{X^*}$. The former leads to

$$\tilde{X} = P_G \tilde{X} + Q_G \tilde{X} = P_{GW_0^*} \tilde{X} + P_{GA^*} \tilde{X} + Q_G \tilde{X}; \tag{31}$$

and the latter to

$$D_{X^*} = P_{X^*} D_{X^*} + P_{D_{X^*} Q_{J_r = D_{XX}}} D_{X^*} = P_{X^*} D_{X^*} + P_{D_{X^*} W_1^*} D_{X^*} + P_{D_{X^*} B^*} D_{X^*}; \tag{32}$$

where A^* and B^* were introduced shortly after (22) above Table 1. The SVD of terms in these decompositions may be called multilevel PCAs (Principal Component Analyses).

7. References

Bock, R. D. (1989). *Multilevel Analysis of Educational Data*. San Diego, CA: Academic Press.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Goldstein, H. I. (1987) *Multilevel Models in Educational and Social Research*. London: Oxford University Press.

Hox, J. J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Ingels, S. J., Planty, M., and Bozick, R. (2005). *A Profile of the American High School Senior in 2004: A First Look-Initial Results from the First Follow-up of the Education Longitudinal Study of 2002 (ELS:2002) (NCES 2006348)*. U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

LaMotte, L. R. (2007). A direct derivation of the REML likelihood function. *Statistical Papers*, **48**, 321–327.

Takane, Y., and Hunter, M. A. (2002). Dimension reduction in hierarchical linear models. In S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), *Measurement and Multivariate Analysis* (pp. 145–154). Tokyo: Springer Verlag.

Takane, Y., and Hwang, H. (2002). Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research*, **37**, 163–195.

Takane, Y., and Hwang, H. (2007). Regularized linear and kernel redundancy analysis. *Computational Statistics and Data Analysis*, **52**, 394–405.

Takane, Y., and Jung, S. (2008). Regularized partial and/or constrained redundancy analysis. *Psychometrika*, **73**, 671–690.

Van den Wollenberg, A. L. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.