British Journal of Mathematical and Statistical Psychology (1989) 42, 81–90 Printed in Great Britain

Factor structure in groups selected on observed scores

Bengt O. Muthén

Graduate School of Education, University of California, Los Angeles, Los Angeles, CA 90024, USA

A new method is proposed for estimating factor means and factor covariances in a group of individuals selected on their observed scores. The selection variable is, for example, the total score on an admissions test. Given a factor model for the test items based on the group of test takers, we may be interested in the factor structure for those in the top quartile. The differences in factor means and covariances between this selected group and the full group gives useful information both on successful test performance and on test validity. The new method draws on the classic Pearson–Lawley selection formulas. It avoids the fallacy of factor analysis on the selected group, which would lead to incorrect estimates. The new method is applied to a simple factor structure model for the GMAT test. Although the majority of the GMAT items test verbal skills, it is found that a quantitative factor shows the greatest change in moving from average to top quartile test takers.

1. Introduction

Consider a certain factor analysis model that fits well for a population of individuals. The analyses leading to this estimated model are assumed to be based on random observations from this population. The question to be studied is: what is the factor structure for a subset of this population, when the subset is defined by a variable that is a function of the observed variables?

This problem arises naturally in the analysis of student responses to university admissions tests of various kinds, e.g. SAT, GRE, LSAT, GMAT. Here, the set of variables consists of test items administered to a sample from a population of test takers. The sum of the test items constitutes the test score. This test score plays a major role in decisions to admit students. The test items may be added up ('parcelled') according to different content areas and factor analysed. If a certain simple factor structure is found in this population an interesting investigation is then to try to determine what the factor structure is for students in the upper tail of the test score distribution. How is the factor structure different for this group of students, who are likely to be admitted, as compared to the general population? Do the factor means increase more for certain factors, implying that these are particularly important in a successful test outcome? Do the decreased factor variances indicate a

#23

particularly strong increase in homogeneity for certain factors? What is the change in pattern of factor correlations? Answers to these questions may give useful insights into the validity of the test and the dimensions underlying the test performance. For instance, if a certain factor mean increases relatively little when moving from the general population to the top group, this indicates that the items measuring this factor have little power in discriminating between average and successful test takers. If a factor mean increases relatively strongly, this indicates that the factor is a dominant one in determining successful outcomes. An investigation of this kind is also a valuable complement to a 'validity study', where test responses are used as predictors of success among the select group of admitted students.

To actually obtain an estimated model for the top group of test takers is, however, not a straightforward task. Direct attempts at applying factor analysis on a sample from such a subpopulation will fail. This is because when a factor model holds for the total population it will not in general hold for observations randomly sampled from a subpopulation defined by selection on the total observed scores. The aim of this paper is to provide a new method which, while avoiding the problem of subpopulation factor analysis, still provides an estimated factor model for the selected group.

Section 2 points out the statistical problems of ordinary factor analysis on a selected group and describes the theory for the new method. As an illustration, Section 3 considers test taker data on the Graduate Management Admissions Test (GMAT) administered to students who want to apply to Graduate Schools of Business and Management. After briefly describing previous factor analyses carried out on a large sample of such test takers (Muthén, Shavelson, Hollis, Kao, Muthén, Tam, Wu, Yang, 1988), factor analysis is applied to the top quartile on the GMAT score. The results are constrasted with those obtained by the new method.

2. Factor analysis in selected populations

Consider the factor analysis model for a p vector of observed variables y,

$$y = \boldsymbol{\nu} + \boldsymbol{\Lambda}\eta + \varepsilon, \tag{1}$$

where $\boldsymbol{\nu}$ is a p vector of intercept, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings, η is an m vector of factor scores standardized to zero means, and ε is a p vector of residuals that are independent of η and have zero means. This gives

$$E(y) = \mu_y = \nu, \tag{2}$$

$$\mathbf{V}(y) = \boldsymbol{\Sigma}_{yy} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{3}$$

where $\boldsymbol{\Psi}$ is a factor covariance matrix and $\boldsymbol{\Theta}$ is a residual covariance matrix usually assumed to be diagonal.

We are interested in a q vector of selection variables s where an observational unit is selected depending on elements of s falling below or exceeding certain thresholds. Let us consider the two cases:

(i) $s = \mathbf{W}_\eta \eta$ (selection related to the factors)
(ii) $s = \mathbf{W}_y y$ (selection directly on the observed variables)

Here, the **W**s contain known weights.

The classic Pearson–Lawley selection formulas (see, for example, Johnson & Kotz, 1972; Lawley, 1943) give the mean vector and covariance matrix of a vector z in a subpopulation (denoted by an asterisk) selected on s. Let $\mathbf{B} = \boldsymbol{\Sigma}_{zs}\boldsymbol{\Sigma}_{ss}^{-1}$. Then

$$\mu_z^* = \mu_z + \mathbf{B}(\mu_s^* - \mu_s), \tag{4}$$

$$\boldsymbol{\Sigma}_{zz}^* = \boldsymbol{\Sigma}_{zz} + \mathbf{B}(\boldsymbol{\Sigma}_{ss}^* - \boldsymbol{\Sigma}_{ss})\mathbf{B}'. \tag{5}$$

$$\boldsymbol{\Sigma}_{zs}^* = \mathbf{B}\boldsymbol{\Sigma}_{ss}^{*-1}. \tag{6}$$

These formulas are valid if the regression of z on s is linear and homoscedastic, for instance, when the vector $(s'z')$ is multivariate normal.

2.1. The problem of factor analysis in selected subgroups

Case (i) of selection related to the factors was studied, e.g., by Meredith (1964), see also Muthén & Jöreskog, (1983), where it was shown that such selection retains the full population factor model in the subpopulation with invariant ν, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Theta}$, while $\mathbf{E}(\eta)$ and $\mathbf{V}(\eta)$ are changed. This is utilized in the simultaneous factor analysis of several groups, see e.g. Jöreskog (1971).

This paper considers selection case (ii) using a single selection variable $s = w'y$, where s may be thought of as a total test score used for admissions. We are interested in the subpopulations for which $s \geqq c$, where c is for instance the upper quartile value in the full population. The goal of our analysis is to correctly estimate the factor mean vector and covariance matrix for the subpopulation and compare these quantities to the corresponding ones for the full population, $\mathbf{0}$ and $\boldsymbol{\Psi}$, respectively.

Let us first consider why attempting to estimate the subpopulation factor covariance matrix by the naive approach of regular factor analysis in the subpopulation will fail.

The invariance of the factor model under selection case (i) is due to the indirect selection on y via η. In contrast, for case (ii) of direct selection on y, the full poulation factor model generally does not hold for subjects randomly drawn from the selected subpopulation. Such situations were studied, e.g. in Muthén, Kaplan & Hollis (1987). Using the Pearson–Lawley selection formula (5) with $z = y$, we find a distorted factor structure in the subpopulation,

$$\boldsymbol{\Sigma}_{yy}^* = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} + \omega(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta})ww'(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}), \tag{7}$$

where

$$\omega = \sigma_{ss}^{-1}(\sigma_{ss}^* - \sigma_{ss})\sigma_{ss}^{-1}. \tag{8}$$

Furthermore, since selection takes place on a variable that is directly related to the dependent variable vector y in the linear regressions of (1), the assumed linearity and homoscedasticity of these regressions will not hold in the subpopulation; see, e.g., Muthén & Jöreskog, (1983).

We conclude that factor analysis of a sample covariance matrix that consistently estimates Σ^*_{yy}, i.e., regular factor analysis on a random sample in the subpopulation, will be incorrectly applied. The assumptions of the regular factor model do not hold and if the analysis is nevertheless applied, it will consider a covariance matrix for which no simple and meaningful structure generally exists.

2.2. *A new approach*

We will propose a different approach which avoids direct factor analysis in the selected subpopulation. While we have established that an individual chosen randomly from the selected subpopulation does not obey (3), individuals in the subpopulation are part of the full population and therefore the factors of (1) also operate in the subpopulation.

The new approach uses the Pearson–Lawley formulas to achieve the goal of correctly estimating the subpopulation factor mean and covariance matrix. We assume that estimates are available for the full population factor model parameters and that w contains known weights.

With $s = w'y$, we obtain

$$\mu_s = w\mu_y, \tag{9}$$

$$\sigma_{ss} = \mathbf{w}'(\mathbf{\Lambda\Psi\Lambda}' + \mathbf{\Phi})\mathbf{w}, \tag{10}$$

$$\sigma_{\eta s} = \mathbf{\Psi\Lambda}'\mathbf{w}. \tag{11}$$

Let ω be defined as in (8), while

$$\kappa = \sigma_{ss}^{-1}(\mu_s^* - \mu_s). \tag{12}$$

Assuming that the regression of η on s is linear and homoscedastic, the Pearson–Lawley formulas applied to $z = \eta$ then yield the factor mean vector

$$\boldsymbol{\mu}_\eta^* = \kappa\mathbf{\Psi\Lambda}'\mathbf{w}, \tag{13}$$

and factor covariance matrix

$$\mathbf{\Sigma}_{\eta\eta}^* = \mathbf{\Psi}^* = \mathbf{\Psi} + \omega\mathbf{\Psi\Lambda}'\mathbf{ww}'\mathbf{\Lambda\Psi}. \tag{14}$$

Assuming further that s is normally distributed, κ and ω may be computed using the mean and variance of a truncated normal variable s, $s \geqq c$ (see, for example, Johnson & Kotz, 1970)

$$\mu_s^* = \mu_s + \phi(d)\pi^{-1}\sigma_s, \tag{15}$$

$$\sigma^*_{ss} = \sigma_{ss}[1 + d\phi(d)\pi^{-1} - (\phi(d)\pi^{-1})^2], \tag{16}$$

where $\phi(d)$ is the standard normal density at d,

$$d = (-c + \mu_s)\sigma_{ss}^{-1/2} \tag{17}$$

and π^{-1} is the area of the normal curve exceeding the truncation point c. For a chosen c and known weights w, factor model parameter estimates obtained from a sample from the full population may then be inserted in (13) and (14) to give the desired factor mean vector and factor covariance matrix estimates for the subpopulation of $s \geqq c$.

3. An application

As an illustration we use data from the GMAT test of October 1984. Muthén *et al.* (1988) analysed a sample of 55 279 test takers from this occasion. The 150 test items were first subjected to an item factor analysis based on a random subsample of test takers, suggesting five interpretable factors. Parcels of items corresponding to the factors were created as proportion correct for a set of items. This gave 24 continuous variables based on 5 to 7 items each. A simple structure, five-factor model was then estimated for these 24 variables using a covariance matrix for the full sample of test takers and found to fit well. The estimated loadings and factor covariance matrix are given in standardized form in Table 1.

Of the 24 variables there are 14 verbal and 10 quantitative ones, corresponding to the 85 and 65 verbal and quantitative items. There is one major verbal factor corresponding to sentence correction and reading comprehension skills, and two minor verbal factors corresponding to item format ('minor' answer key and 'other' answer key). There are two quantitative factors, where the major one reflects accuracy and the other one also reflects speed. The simple structure involves zero loadings in cases where variables and factors have no substantively meaningful relationship. For instance, verbal items are not allowed to load on the quantitative factors. Using confirmatory factor analysis maximum-likelihood estimation, the chi-square fit value for this model was 255 with 218 degrees of freedom when normed to a sample size of 1000, reflecting a very good fit.

In this application we focus on the group of students belonging to the top quartile of the total GMAT score, where the top quartile was estimated as 570. As a first analysis, exploratory factor analysis was carried out on the corresponding sample of 13 504 students. The top portion of Table 2 shows the drop in eigenvalues and maximum-likelihood chi-square for increasing number of factors. We conclude that three factors are probably sufficient in explaining the correlation structure. The three factors are interpreted as a verbal factor corresponding to the general one of Table 1 for the full sample (1), a verbal factor corresponding to 'analysis of situations' items (from the remaining verbal section) (2), and a general quantitative factor (3). The four factor solution splits the verbal 'analysis of situations' factor into the full sample answer key factors (2 and 3). The five factor solution splits the general verbal factor

Table 1. Standardized estimates for GMAT simple structure factor model ($n = 55\,279$)

		Factor				
		Verbal			Quantitative	
		General	Specific		General	Specific
Section		Sentence Corr. & Reading Comp.	Minor Key	Other Key	Accuracy	Speed & Accuracy
Verbal						
Sentence correction	V1	0.588				
	V2	0.599				
	V3	0.625				
	V4	0.656				−0.024
Analysis of situations	V5	0.044	0.598			
	V6		0.662			
	V7	0.356	0.454			−0.002
	V8	−0.044		0.647		
	V9			0.765		
	V10	0.076		0.603		0.109
Reading comprehension	V11	0.573				
	V12	0.630				
	V13	0.622				
	V14	0.597				0.080
Quantitative						
Problem solving 1	Q1	−0.123			0.711	0.084
	Q2	−0.047			0.525	0.335
	Q3	0.046			−0.148	0.838
Data sufficiency	Q4				0.635	
	Q5	0.119			0.492	0.034
	Q6	−0.026			0.244	0.530
	Q7	0.178			−0.088	0.657
Problem solving 2	Q8	−0.054			0.739	−0.071
	Q9	−0.026			0.645	0.218
	Q10					0.846
Factor correlations		1.000				
		0.577	1.000			
		0.668	0.674	1.000		
		0.583	0.620	0.557	1.000	
		0.268	0.411	0.433	0.687	1.000

into variables corresponding to sentence correction items (1) and variables corresponding to reading comprehension (5), the answer key factors are as for the full sample (2 and 3), and there is a general quantitative factor (5).

The fact that fewer factors are observed in the top 25 per cent group than in the total group could be taken as an indication of the full-sample factors being highly correlated in the selected group, collapsing the factor space. However, Table 2 instead shows lower factor correlations than Table 1, probably reflecting the observed variable correlation attenuation. We conclude that the full sample factors could not be recovered in the selected group.

Table 2. Exploratory factor analysis in top 25 per cent group ($n = 13\,504$)

Number of factors	1	2	3	4	5	6
Eigenvalues	3.505	2.354	1.832	1.234	1.058	1.028
Chi-square[a]		760.4	276.1	170.7	115.2	72.6
d.f.		229	207	186	166	147
Chi-square/d.f.		3.32	1.33	0.92	0.69	0.49
			Factor correlation matrix			
3-Factor	1.000					
	0.021	1.000				
	−0.267	0.177	1.000			
4-Factor	1.000					
	0.122	1.000				
	0.193	0.300	1.000			
	−0.267	−0.056	0.026	1.000		
5-Factor	1.000					
	0.096	1.000				
	0.212	0.322	1.000			
	−0.228	−0.055	0.025	1.000		
	0.594	0.152	0.118	−0.243	1.000	

[a]Normed to a sample of 1000.

In using the proposed method, we first observe that the GMAT score is not the same as, but highly correlated (0.99) with, a weighted sum of the 24 observed variables. The weights simply correspond to the number of items in each parcel. Hence, for the selection relation $s = w'y$ we have known, fixed weight w and a negligible residual. The selection variable s appears close to normal with small skew and kurtosis as would be expected since it is a sum of many variables. Likewise, the 24 observed variables have univariate skews and kurtoses that are mostly in the range of −0.5 to 0.5. The assumptions underlying the Pearson–Lawley selection formulas therefore seem plausible.

Applying (13) and (14) of the proposed method, we obtain the estimated factor mean vector and factor covariance matrix given in Table 3.

In Table 3 we first note that the five factors are in different metrics and are not directly comparable. In terms of factor means we obtain comparability by considering the means divided by the corresponding factor standard deviations. We choose to use the standard deviation of the full population and assess the standardized mean increase from the zero value of that population.

Our first observation is that the mean increases are similar, indicating a very desirable test property of equal importance of the different parts of the GMAT in determining a successful outcome. This is a new and important component of test validity. However, the general quantitative factor (accuracy) is the dominant factor in distinguishing average from top quartile test takers. This occurs despite the fact that in comparing verbal vs. quantitative test content, verbal content is reflected in both more items (85 vs. 65) and more item parcels (14 vs. 10). It is also interesting to note

Table 3. Estimated factor structure in top 25 per cent group vs. total group

	Factor				
	Verbal			Quantitative	
	General	Specific		General	Specific
Section	Sentence Corr. & Reading Comp.	Minor Key	Other Key	Accuracy	Speed & Accuracy
Factor means in subpopulation	0.144	0.145	0.173	0.152	0.207
Relative factor means[a]	1.013	0.936	0.970	1.099	0.895
Factor SD in subpopulation	0.103	0.119	0.134	0.092	0.183
Factor SD in total population	0.143	0.154	0.179	0.138	0.231
Relative factor SD[b]	0.723	0.769	0.751	0.663	0.793
Factor var–cov matrix in subpopulation	0.011				
	0.005	0.014			
	0.005	0.007	0.018		
	0.001	0.003	0.002	0.008	
	−0.005	0.001	0.001	0.007	0.033
Factor var–cov matrix in total population	0.020				
	0.013	0.024			
	0.017	0.019	0.032		
	0.011	0.013	0.014	0.019	
	0.009	0.015	0.018	0.022	0.053
Factor corr. matrix in subpopulation	1.000				
	0.244	1.000			
	0.389	0.437	1.000		
	0.138	0.277	0.125	1.000	
	−0.267	0.037	0.051	0.438	1.000
Factor corr. matrix in total population	1.000				
	0.577	1.000			
	0.668	0.674	1.000		
	0.583	0.620	0.557	1.000	
	0.268	0.411	0.433	0.687	1.000

[a]Factor means in subpopulation divided by factor SD in total population.
[b]Factor SD in subpopulation divided by factor SD in total population.

that the speed aspect of the quantitative factors discriminates the least among average and top quartile test takers.

The general quantitative factor is also the one increasing the most in terms of homogeneity, as measured by the ratio of subpopulation to full population variance. Again, quantitative speed shows the least change. The factor correlation matrices clearly show that the various skills represented by the five factors have much weaker

relationships for the homogeneous top quartile group than for the full population. For instance, the full population correlation of 0.58 between the major verbal and the major quantitative factors reduces to 0.14 in the top quartile group.

4. Discussion

The proposed approach gives a powerful indirect way of estimating factor structure in a strongly selective group that cannot be uncovered by conventional analysis. A strong feature is that the estimates are obtained from full population estimates which may build on a much larger sample than the select group under consideration. It was shown that this approach provided new and useful information on the differences between average and particularly successful test takers, while at the same time providing insights about test validity.

An alternative approach to estimate the subpopulation factor means and covariance matrix would be to simply estimate factor scores for the selected group and calculate the desired quantities from those variables. It is well known, however, that estimated factor scores do not provide unbiased estimates of the means and covariances (see, e.g., Lawley & Maxwell, 1971). Also, if the subgroup sample is small, unstable estimates would be obtained.

A weakness of the proposed approach is that it is based on assumptions of linearity and homoscedasticity of the factors regressed on the selection variable. These assumptions must be evaluated in any given application. An aid in judging their plausibility would be to compare the sample mean vector and covariance matrix for the observed variables in the selected group with the corresponding quantities predicted from the selection formulas, not using the underlying factor structure. However, the results for the factors may be more robust than for the observed variables. Further research in terms of simulation studies can shed some light on such robustness.

Acknowledgements

Presented at the AERA meeting in New Orleans, Louisiana, 5–9 April, 1988. The research described in this paper was funded by the Graduate Management Admissions Council. The GMAC encourages researchers to formulate and freely express their own opinions, and the opinions expressed here are not necessarily those of the GMAC. The author would like to thank Jin-Wen Yang, Mike Hollis, Chih-fen Kao, and Wai-Yan Tam for helpful assistance.

References

Johnson, N. L. & Kotz, S. (1970). *Continuous Univariate Distributions-1: Distributions in Statistics.* Chichester: Wiley.

Johnson, N. L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Chichester: Wiley.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, **36,** 409–426.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society Edinburgh, Section A*, **62,** 28–30.

Lawley, D. N. & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method.* London: Butterworths.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, **29,** 177–185.

Muthén, B. O. & Jöreskog, K. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, **7,** 139–173.

Muthén, B. O., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, **52,** 631–462.

Muthén, B. O., Shavelson, R. J., Hollis, M., Kao, C-F., Muthén, L., Tam, W-Y., Wu, S-T. & Yang, J-W. (1988). Relationship between applicant characteristics, MBA program attributes, and student performance: The psychometric study. Graduate Management Admission Council, Los Angeles, CA.

Received 27 October 1987; revised version received 12 August 1988