A Simulation Study on Computation and Inference Accuracy of Factor Loadings for Large Data Mines

CHIH-CHIEN YANG, PhD, Professor
Cognitive NeuroMetrics Laboratory
Graduate School of Educational Measurement & Statistics
National Taichung University
140 MinSheng Road, Taichung 403
TAIWAN

LIANG-TING TSAI
Graduate School of Educational Measurement & Statistics
National Taichung University

Abstract: - This study simulates and examines the weighting effects of using relatively diminutive samples to infer a gigantic data mine. The statistical inference in the simulations is carried out by using a basic factor analysis model. Several stratum sizes and stratum-vs.-population ratios are designed in the data generating procedures; therefore, the experiment can reflect the practical research environments. Estimations are conducted by using the altered maximum likelihood estimation (MLE) algorithm. The preliminary results show the proposed method is promising.

Key-Words: - Numerical Simulation, Factor Analysis, Data Mining, Statistical Models

1 Introduction

The current popularity of (inter-)national large-scale surveys (e.g., TIMSS [1], PISA [2], PIRLS [3], etc.) has raised urgencies on establishing proper approaches of inferring the gigantic populations (data mines), particularly, when using sophisticated factor analysis alike techniques yet only relatively limited sample sizes are available. Specifically, these internationally collected datasets, although, had complexly designed sampling frames to be generalizable to the intended super populations, the datasets may actually contain a small proportion of only far less than 10% of the original targets. Most technical reports of these surveys urged the engagement of sampling weights when researchers report analyses of descriptive statistics; however, less was advised when statistical inference was made by using factor analysis models. On the other hand, methodological research (e.g., [4], [5]) had some evidence that strongly advocated the use of sampling weights in factor analysis procedures on these datasets.

To examine the influence of sampling weights on the accuracy of parameter estimation in factor analyses, we conducted a study to simulate practical large-scale surveys in which various proportions of samples versus population, sampling stratum sizes, and stratum proportions can be occurred. Our preliminary simulation results demonstrate that proper treatments of sampling weights are crucial and incorporating weights into the parameter estimation procedures of factor analysis is a non-ignorable top-priority.

2 Simulation Design

The factor analysis model in the simulation study was a very basic single factor model that has five continuous outcomes (y) and only one latent factor (η). To differentiate the two strata as well as the sub-populations, a factor loading was set to be different between the two strata. The mathematical equation to demonstrate the relations was shown as follows.

$$y_j = \lambda_j \eta_1 + \varepsilon_j, \tag{1}$$

where λ is the factor loading and ε symbolizes the normally distributed random errors. A path diagram was drawn in the following Figure 1 to illustrate the model structure.

In the Fig. 1, numbers shown for the λ's are the designated parameter values for the artificial data generating procedures. Particularly, the second set of λ's has a difference of 0.2 between the two strata. All the rest parameters between the two strata were set to be equal to simplify the model structure; therefore, computing time required for estimation can be reduced. This is crucial, particularly, when the replications is set high to reduce stochastic errors.

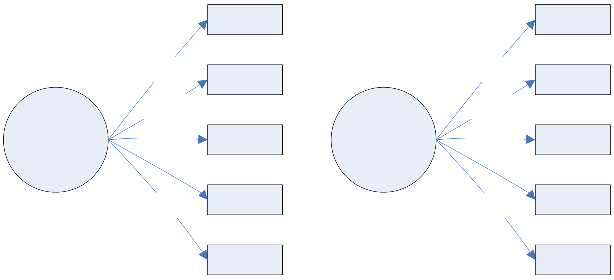

Fig. 1 Path diagram for factor analysis model

The simulation study was designed to reflect the actual large data mines as close as possible, in which two strata combined with five different proportions of strata versus the entire data mine and four sampling stratum sizes were employed. For each condition, a total of 200 replications was performed; therefore, aggregated computation results can be summarized and analyzed to reveal a general and scientific trend. The design was outline in the following Table 1.

Table 1 Design for the simulation study

Strata-Population Ratio	N = 100		N = 200	
	Strata Size		Strata Size	
w1 : w2	N1	N2	N1	N2
8 : 1	33	67	67	133
6 : 1	40	60	80	120
4 : 1	50	50	100	100
2 : 1	67	33	133	67
1 : 1	80	20	160	40
	N = 400		N = 600	
8 : 1	133	267	200	400
6 : 1	160	240	240	360
4 : 1	200	200	300	300
2 : 1	267	133	400	200
1 : 1	320	80	480	120

The entire data mine has 10,000 individual subjects. As a result, all the artificially generated strata owned less than 10 percents of the original population size. This is designed to reflect the common conditions often seen in the research practice. The pseudo-random number generator in the computer software Matlab was used and programmed to generate the datasets. After the data generating procedure was completed, the datasets were estimated by the altered maximum-likelihood estimation (MLE) algorithm [6].

3 Results & Analyses

Simulation and estimation results are summarized in the following Table 2, in which it reports the averages of 200 replications. Specifically, two separate columns are named “non-weighted” and “weighted” to provide comparisons of the estimations that were completed by non-weighted and weighted factor analysis estimating procedures, respectively.

Specifically, the total sample sizes and ratios of the two strata in the corresponding sample are listed in the first two columns. The “average estimate” columns summarize the 200 estimated factor loadings that are first averaged by the five factor loadings (λ’s) in each replication. The 95% coverage rates recoded the percentages of the 95% confidence intervals (C.I.) covering the true populations values. Ideally, a coverage rate needs to be near the theoretical 95% value to demonstrate an acceptable performance for inferring the population-wise level, i.e., the level of entire data mine. Comparing the average estimates and coverage rates between non-weighted and weighted procedures illustrates the lapse between the proposed method and a potentially flawed practice. A greater lapse shows an increased urgency of the proposed method in the specifically designed condition.

Table 2 Averages of estimating simulated datasets

Sampling Sizes	w1 : w2	Non-Weighted		Weighted	
		Average estimate	95% coverage rates	Average estimate	95% coverage rates
100	8 : 1	0.6804	0.860	0.7725	0.915
	6 : 1	0.6993	0.885	0.7729	0.905
	4 : 1	0.7110	0.875	0.7640	0.925
	2 : 1	0.7442	0.905	0.7691	0.915
	1 : 1	0.7734	0.960	0.7734	0.960
200	8 : 1	0.6842	0.835	0.7774	0.930
	6 : 1	0.6941	0.850	0.7780	0.915
	4 : 1	0.7175	0.925	0.7819	0.940
	2 : 1	0.7340	0.935	0.7596	0.940
	1 : 1	0.7744	0.955	0.7744	0.955
400	8 : 1	0.6825	0.705	0.7710	0.930
	6 : 1	0.6919	0.730	0.7724	0.960
	4 : 1	0.7100	0.845	0.7713	0.970
	2 : 1	0.7494	0.950	0.7756	0.945
	1 : 1	0.7762	0.955	0.7762	0.955
600	8 : 1	0.6872	0.655	0.7809	0.920
	6 : 1	0.6899	0.590	0.7684	0.935
	4 : 1	0.7161	0.810	0.7755	0.950
	2 : 1	0.7447	0.920	0.7720	0.950
	1 : 1	0.7696	0.955	0.7696	0.955
Population	1 : 1	0.7727	0.955	0.7727	0.955

4 Conclusion

The preliminary results show the proposed method for statistically inferring a large data mine by using a relatively diminutive sample is promising. The gained accuracy by using the proposed method has a considerable effect over the non-weighted method. Further research is being pursued to extend the research design, for example, adding more stratum sizes and stratum-vs.-population ratios, so that more generalizable findings can be established.

References:

[1] Gonzalez, E.J., & Miles, J.A. (eds.), *TIMSS 1999 user guide for the International database: IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Lynch School of Education, Boston College, 2001.

[2] Organization for Economic Co-Operation and Development (OECD), *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. OECD, 2003.

[3] Martin, M.O., Mullis, I.V.S., & Kennedy, A.M. (eds.), *PIRLS 2001 Technical Reports*. Publisher: International Study Center, Lynch School of Education, Boston College, 2003.

[4] Grilli, L., & Pratesi, M., Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs, *Survey Methodology*, Vol.30, 2004, pp. 4-14.

[5] Kaplan, D., & Ferguson, A.J., On the utilization of sample weights in latent variable models. *Structural Equation Modeling*, Vol.6, 1999, pp. 305-321.

[6] Asparouhov, T., Sampling Weights in Latent Variable Modeling. *Structural Equation Modeling*, Vol.12, 2005, pp. 411-434.