Behaviormetrika, Vol. 24, No. 1, 71-83
1997

ON AVERAGING VARIABLES IN A CONFIRMATORY FACTOR ANALYSIS MODEL*

Ke-Hai Yuan, Peter M. Bentler** and Yutaka Kano***

The normal theory maximum likelihood and asymptotically distribution free methods are commonly used in covariance structure practice. When the number of observed variables is too large, neither method may give reliable inference due to bad condition numbers or unstable solutions. The main existing solution to the problem of high dimension is to build a model based on marginal variables. This practice is inefficient because the omitted variables may still contain valuable information regarding the structural model. In this paper, we propose a simple method of averaging proper variables which have similar factor structures in a confirmatory factor model. The effects of averaging variables on estimators and tests are investigated. Conditions on the relative errors of the measured variables are given that verify when a model based on averaged variables can give better estimators and tests than one based on omitted variables. Our method is compared to the method of variable selection based on mean square error of predicted factor scores. Some aspects related to averaging, such as improving the normality of observed variables, are also discussed.

1. Introduction

In the social and behavioral sciences, high dimensional data that are believed to be related to the latent traits of interest are often obtained. Confirmatory factor analysis is regularly brought into use for evaluating the latent structure of high dimensional data. This methodology permits the researcher to specify not only the dimension of the latent factor space, but also which variables are hypothesized to be good indicators of such factors and which variables are unaffected by given factors. This is an appropriate methodology for many studies in which the design of the instrument implies a particular latent structure. For example, in the design of a personality inventory, even before the data are collected, items or variables may have been generated to be indicators of particular factors. At the same time, such variables may be presumed not to be related to other latent traits. A confirmatory factor analysis, with its associated statistical theory, can be used to verify theoretical specifications, as well as to determine areas of weakness in the theory or inadequate specification of the measurement relations. In structural equation modeling practice, typically each variable depends on only one latent common factor, leaving the correlation among the observed variables to be explained by the patterns of correlation among the latent factors. Anderson and

Key Words and Phrases ; variable selection, covariance structures, mean square error

* This work was supported by National Institute on Drug Abuse Grants DA01070 and DA00017.

** Department of Psychology and Center for Statistics, University of California, Los Angeles, Box 951563, Los Angeles, CA 90095-1563, USA.

*** Institute of Mathematics, University of Tsukuba, Tsukuba, Ibaraki, 305 Japan.

Gerbing (1982, 1988) discussed the rationale for such unidimensional measurement (see also Jöreskog, 1971 ; Hattie, 1985 ; Hunter & Gerbing, 1982), indicating that it allows the most unambiguous assignment of meaning to the estimated constructs. Thus motivated, we will mainly consider the unidimensional measurement model in this paper.

In employing statistical theory, maximum likelihood (ML) under assumed normality and asymptotically distribution free (ADF) generalized least squares are the two most frequently used methods for model evaluation. Unfortunately, however, neither of these methods can give reliable inference when the number of variables in a sample becomes large. Even with modern computer programs such as EQS and LISREL (Bentler & Wu, 1995 ; Jöreskog & Sörbom, 1993), the computations involving a very large number of parameters and/or variables become excessive. As a result, it is not unreasonable for practitioners to choose a subset of the observed variables for model fitting and testing. An obvious problem with eliminating variables from a confirmatory factor analysis is that the neglected variables may contain valuable information on estimators and tests regarding the structural model. This suggests that it would be important to quantify the loss obtained from discarding variables, and to discard variables so as to minimize such loss. In a series of papers, Yanai (1980), Tanaka and Kodake (1981), and Tanaka (1983) proposed the use of variable selection based on the configuration of the true factor score $\boldsymbol{f}$ and the predicted factor score $\hat{\boldsymbol{f}}$. Let $\hat{\boldsymbol{f}}_{(i)}$ denote the factor score predictor without the ith variable, they suggested removing the variable which makes

$$\mathrm{MSE}_i=\mathrm{E}(\hat{\boldsymbol{f}}_{(i)}-\boldsymbol{f})'(\hat{\boldsymbol{f}}_{(i)}-\boldsymbol{f}) \tag{1.1}$$

the smallest each time, as in stepwise regression. Tanaka (1983) showed numerically that the above proposed method is superior than the variable selection method proposed by Jolliffee (1972, 1973), who discarded variables in principal component analysis.

An alternative approach to removing variables is averaging or summing variabes, which is a typical procedure used in the social and behavioral sciences. For example, responses to items on psychological tests are typically summed to yield more reliable total scores, and the total number correct is usually used as a basis for assigning grades to students in a classroom. The best rationale for averaging or summing occurs when variables have similar meanings, though in practice averages are computed more generally. For example, the grade-point average is used to represent performance in school or college, even though a student's performance may vary by topics. Even though averaging variables has been regularly used in descriptive statistics, when facing a high dimensional data set, this obvious approach is typically not used and a subset of variables becomes the basis for structural modeling practice. In this paper, we propose to build a factor analysis model based on average variables with similar factor structures.

More specifically, we will explore the information contained in the removed variables by studying the information in the variables before and after averaging. Section 2 will investigate the effect of averaging variables on model estimation and testing, emphasizing especially models with unidimensional measurement structures in which any variable is influenced primarily by a single latent factor. In Section 3, we compare our method with variable selection based on minimizing (1.1). A discussion and our conclusions will be presented at the end of the paper.

2. The effect of averaging variables on estimators and tests

This section will consider models based on averaged variables and those based on selected marginal variables. The effect of averaging variables on estimators and tests associated with the normal theory ML and ADF methods will be studied.

Let $\boldsymbol{X}=(x_{1,1}, \cdots, x_{1,m_1}; x_{2,1}, \cdots, x_{2,m_2}; \cdots; x_{p,1}, \cdots, x_{p,m_p})'$ be a mean zero random vector of length $\sum_{j=1}^{p} m_j$. Assume that the factor structure of X can be expressed as

$$x_{j,i}=\lambda_{j,i}f_{k(j)}+\varepsilon_{j,i},\ i\in T_j,\ 1\leq j\leq p, \tag{2.1}$$

where $T_j=\{1, \cdots, m_j\}$, $\lambda_{j,i}$ are factor loadings, $f_{k(j)}$ are common factors with $\mathrm{E}f_{k(j)}=0$, and $\varepsilon_{j,i}$ are errors or unique factors with $\mathrm{E}\varepsilon_{j,i}=0$. In model (2.1) we have assumed that for each j, the observed variables $x_{j,i}$, $i\in T_j$ only depend on one common factor $f_{k(j)}$. It is possible that $k(j_1)=k(j_2)$ even if $j_1\neq j_2$. This is from the consideration of identifiability of model (2.2) in the following. Let

$$y_j=\frac{1}{m_j}\sum_{i\in T_j} x_{j,i}=\bar{\lambda}_j f_{k(j)}+\bar{\varepsilon}_j, \tag{2.2}$$

where

$$\bar{\lambda}_j=\frac{1}{m_j}\sum_{i\in T_j}\lambda_{j,i},$$
$$\bar{\varepsilon}_j=\frac{1}{m_j}\sum_{i\in T_j}\varepsilon_{j,i}.$$

So the averaged variables still keep the same factor structure. Assuming that the structural model (2.2) can still be identified, we will investigate the effect of averaging on estimators and tests based on the observed variables $\boldsymbol{Y}=(y_1, \cdots, y_p)'$. In assuming a factor analysis model (2.1) on $\boldsymbol{X}$, the common interest is to get good estimators of factor loadings $\lambda_{j,i}$, while the variances of $\varepsilon_{j,i}$ can be regarded as nuisance parameters. The magnitudes of $\lambda_{j,i}$ and $\mathrm{var}(\varepsilon_{j,i})$ only have relative meaning, since we can multiply (2.1) by a constant without changing the factor structure. For easy comparison of the effect of averaging, we now also assume $\lambda_{j,i}=\lambda_j$, $i\in T_j$. Considering the possiblilty of $k(j_1)=k(j_2)$ and assuming r common factors, we can write (2.2) as

$$\boldsymbol{Y}=\boldsymbol{\Lambda}\boldsymbol{f}+\bar{\boldsymbol{\varepsilon}}, \tag{2.3}$$

where $\boldsymbol{\Lambda}$ is a $p\times r$ matrix with the lth column consisting of the elements λ_j such that $k(j)=l$ for $j\in S_l$, an index set, $\boldsymbol{f}=(f_1, \cdots, f_r)'$, and $\bar{\boldsymbol{\varepsilon}}=(\bar{\varepsilon}_1, \cdots, \bar{\varepsilon}_p)'$. Assuming independence between f and $\varepsilon_{j,i}$ and among $\varepsilon_{j,i}$, then $\bar{\varepsilon}_j$ are also independent and are independent with $\boldsymbol{f}$. Let $\boldsymbol{\Phi}=\text{var}(\boldsymbol{f})$, $\psi_{j,i}=\text{var}(\varepsilon_{j,i})$, we have the following moment structure on $\boldsymbol{Y}$:

$$\text{var}(\boldsymbol{Y})=\boldsymbol{\Lambda\Phi\Lambda}'+\bar{\boldsymbol{\Psi}}, \tag{2.4}$$

where $\bar{\boldsymbol{\Psi}}=\text{diag}(\bar{\psi}_1, \cdots, \bar{\psi}_p)$ with $\bar{\psi}_j=\frac{1}{m_j^2}\sum_{i\in T_j}\psi_{j,i}$. Suppose another method is to use marginal variables $\boldsymbol{X}_m=(x_1, \cdots, x_p)$ with x_j being one of the $x_{j,i}$, $i\in T_j$. Then the moment structure of $\boldsymbol{X}_m$ is

$$\text{var}(\boldsymbol{X}_m)=\boldsymbol{\Lambda\Phi\Lambda}'+\boldsymbol{\Psi}, \tag{2.5}$$

where $\boldsymbol{\Psi}=\text{diag}(\psi_1, \cdots, \psi_p)$ and ψ_j is one of the $\psi_{j,i}$, $i\in T_j$.

We first compare the efficiency of estimators of the structural parameters based on model (2.4) and model (2.5). Suppose the observed data are normal and we use normal theory ML to estimate the unknown parameters. Let vec($\cdot$) be an operator which transform a matrix into a vactor by stacking the columns of the matrix. Then for the ML estimator (MLE) $\hat{\boldsymbol{\theta}}_n$ of the unknown parameter vactor $\boldsymbol{\theta}_0$ based on a sample of size n, the asymptotic covariance of $\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ is given by $2\{\dot{\boldsymbol{\sigma}}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\dot{\boldsymbol{\sigma}}\}^{-1}$, where $\dot{\boldsymbol{\sigma}}=(\partial\text{vec}\,(\boldsymbol{\Sigma}(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}' \mid \boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the structured covariance of the observed variables in either model (2.4) or (2.5). Since $\dot{\boldsymbol{\sigma}}$ is the same based on (2.4) or (2.5), we only need to compare the $\boldsymbol{\Sigma}$ matrix in order to compare the efficiency. It is obvious that the estimator $\hat{\boldsymbol{\theta}}_n$ based on a smaller covariance matrix $\boldsymbol{\Sigma}$ will be more efficient. Let $\boldsymbol{\Sigma}_m=\text{var}(\boldsymbol{X}_m)$ and $\bar{\boldsymbol{\Sigma}}=\text{var}(\boldsymbol{Y})$, then the extra efficiency that may be gained by averaging variables is really decided by $\boldsymbol{\Psi}-\bar{\boldsymbol{\Psi}}=\text{diag}(\psi_1-\bar{\psi}_1, \cdots, \psi_p-\bar{\psi}_p)$. Since ψ_j is one of the $\psi_{j,i}$, we will get a better estimator if

$$\bar{\psi}_j\leq\min_{i\in T_j}\psi_{j,i},\ 1\leq j\leq p. \tag{2.6}$$

If all the $\psi_{j,i}$, $i\in T_j$ are equal, it is obvious that (2.6) holds. More generally, (2.6) holds if

$$\max_{i\in T_j}\psi_{j,i}\leq(m_j+1)\min_{i\in T_j}\psi_{j,i},\ 1\leq j\leq p. \tag{2.7}$$

Since we can always rearrange the order of variables in each group, we will assume $\psi_{j,1}\leq\cdots\leq\psi_{j,m_j}$ in the rest of this paper. Suppose T_1 has two elements while all the other T_j have only one element. Then we will get a better estimator based on model (2.4) than based on model (2.5) if $3\psi_{1,1}>\psi_{1,2}$. Condition (2.7) is very conservative, since we may not use the variable with the smallest $\psi_{j,i}$ in model (2.5). When there is not a great difference among $\psi_{j,i}$, $i\in T_j$, condition (2.7) will be easily satisfied, and the estimator $\hat{\theta}_n$ based on model (2.4) will be more efficient. Assuming that the $\psi_{j,i}$ are approximately of the same magnitude, the efficiency of $\hat{\theta}_n$ based on model (2.4) will increase as the number of averaged variables increases.

The effect of averaging on the efficiency of the ADF estimator is more complicated to see. Let vech$(\cdot)$ be an operator which transforms a symmetric matrix into a vector by picking the nonduplicated elements of the matrix, $\dot{\boldsymbol{\sigma}}=(\partial \text{vech}\,(\boldsymbol{\Sigma}(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}' \mid \boldsymbol{\theta}_0)$. Then for the ADF estimator $\tilde{\boldsymbol{\theta}}_n$ based on model (2.4) or (2.5), the asymptotic covariance of $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ is given by $(\dot{\boldsymbol{\sigma}}'\boldsymbol{V}^{-1}\dot{\boldsymbol{\sigma}})^{-1}$, where $\boldsymbol{V}$ is the covariance matrix of vech$(\boldsymbol{ZZ}')$ and $\boldsymbol{Z}$ is a vector of the centralized observed variables used in the structural model. As for the case of the normal theory MLE, we need to compare the matrix V obtained from each method. Let $\bar{\boldsymbol{V}}=\text{var}\,[\text{vech}\,(\boldsymbol{YY}')]$ and $\boldsymbol{V}_m=\text{var}[\text{vech}(\boldsymbol{X}_m\boldsymbol{X}_m')]$. We will get better estimator based on model (2.4) if $\bar{\boldsymbol{V}}\le\boldsymbol{V}_m$. Direct calculation shows

$$\begin{aligned}\boldsymbol{V}_m=&\boldsymbol{D}_p^+(\Lambda\otimes\Lambda)\,\text{var}[\text{vec}(\boldsymbol{ff}')](\Lambda'\otimes\Lambda')\boldsymbol{D}_p^{+\prime}+2\boldsymbol{D}_p^+(\boldsymbol{\Psi}\otimes\Lambda\boldsymbol{\Phi}\Lambda')\boldsymbol{D}_p^{+\prime}\\&+2\boldsymbol{D}_p^+(\Lambda\boldsymbol{\Phi}\Lambda'\otimes\boldsymbol{\Psi})\boldsymbol{D}_p^{+\prime}+\text{var}[\text{vech}(\boldsymbol{\varepsilon\varepsilon}')],\end{aligned}\tag{2.8}$$

and

$$\begin{aligned}\bar{\boldsymbol{V}}=&\boldsymbol{D}_p^+(\Lambda\otimes\Lambda)\,\text{var}[\text{vec}(\boldsymbol{ff}')](\Lambda'\otimes\Lambda')\boldsymbol{D}_p^{+\prime}+2\boldsymbol{D}_p^+(\bar{\boldsymbol{\Psi}}\otimes\Lambda\boldsymbol{\Phi}\Lambda')\boldsymbol{D}_p^{+\prime}\\&+2\boldsymbol{D}_p^+(\Lambda\boldsymbol{\Phi}\Lambda'\otimes\bar{\boldsymbol{\Psi}})\boldsymbol{D}_p^{+\prime}+\text{var}[\text{vech}(\bar{\boldsymbol{\varepsilon}}\,\bar{\boldsymbol{\varepsilon}}')],\end{aligned}\tag{2.9}$$

where $\boldsymbol{D}_p$ is the duplication matrix defined in Magnus and Neudecker (1988, p. 49) and $\boldsymbol{D}_p^+$ is the Moore-Penrose generalized inverse of $\boldsymbol{D}_p$. The first terms on the right hand side (RHS) of (2.8) and (2.9) are the same. If condition (2.7) is satisfied, the second and third terms on the RHS of (2.9) will be smaller than those of (2.8). We need to calculate var[vech$(\boldsymbol{\varepsilon\varepsilon}')$] in order to compare $\boldsymbol{V}_m$ and $\bar{\boldsymbol{V}}$. Denote $\boldsymbol{A}=\text{diag}(\varepsilon_1^2,\cdots,\varepsilon_p^2)$, $\boldsymbol{B}=\boldsymbol{\varepsilon\varepsilon}'-\boldsymbol{A}$, then

$$\text{vech}(\boldsymbol{\varepsilon\varepsilon}')=\text{vech}(\boldsymbol{A})+\text{vech}(\boldsymbol{B}).$$

Since $\boldsymbol{\varepsilon}_i$ are independent and $\text{E}\boldsymbol{\varepsilon}_i=0$, we have $\text{E}\boldsymbol{B}=0$, $\text{Evech}(\boldsymbol{\varepsilon\varepsilon}')=\text{Evech}(\boldsymbol{A})$ and

$$\text{Evech}(\boldsymbol{\varepsilon\varepsilon}')\,\text{vech}'(\boldsymbol{\varepsilon\varepsilon}')=\text{Evech}(\boldsymbol{A})\,\text{vech}'(\boldsymbol{A})+\text{Evech}(\boldsymbol{B})\,\text{vech}'(\boldsymbol{B}).\tag{2.10}$$

Notice that no two elements in the vector vech $(\boldsymbol{B})$ are the same, hence

$$\text{var}[\text{vech}(\boldsymbol{B})]=\text{Evech}(\boldsymbol{B})\,\text{vech}'(\boldsymbol{B})$$

will be a diagonal matrix. Similarly, var[vech$(\boldsymbol{A})$] is also a diagonal matrix. So

$$\text{var}[\text{vech}(\boldsymbol{\varepsilon\varepsilon}')]=\text{diag}(a_1,\cdots,a_{p*})$$

is a diagonal matrix with

$$a_i=\begin{cases}\text{var}(\varepsilon_k^2),\ \text{if}\ i=1+\sum_{j=1}^{k-1}[p-(j-1)],\ k=1,\cdots,p\\ \text{E}\varepsilon_j^2\text{E}\varepsilon_k^2,\ j<k,\ \text{elsewhere.}\end{cases}$$

Similarly,

$$\text{var}[\text{vech}(\bar{\boldsymbol{\varepsilon}}\,\bar{\boldsymbol{\varepsilon}}')]=\text{diag}(b_1,\cdots,b_{p*})$$

with

$$b_i=\begin{cases}\mathrm{var}(\bar{\varepsilon}_k^2), \text{ if } i=1+\sum_{j=1}^{k-1}[p-(j-1)],\ k=1,\cdots,p\\ \mathrm{E}\,\bar{\varepsilon}_j^2\mathrm{E}\,\bar{\varepsilon}_k^2,\ j<k, \text{ elsewhere.}\end{cases}$$

It is obvious that

$$\mathrm{E}\,\bar{\varepsilon}_j^2\mathrm{E}\,\bar{\varepsilon}_k^2\le\mathrm{E}\varepsilon_j^2\mathrm{E}\varepsilon_k^2$$

if condition (2.7) holds. In order to compare $\mathrm{var}(\varepsilon_j^2)$ and $\mathrm{var}(\bar{\varepsilon}_j^2)$, we need to express $\mathrm{var}(\bar{\varepsilon}_j^2)$ in the form of $\mathrm{var}\,(\varepsilon_{j,i}^2)$. Direct calculation gives

$$\mathrm{var}(\bar{\varepsilon}_j^2)=\frac{1}{m_j^4}\sum_{i=1}^{m_j}\mathrm{var}(\varepsilon_{j,i}^2)+\frac{2}{m_j^4}\sum_{i,k\in T_j,i\neq k}\mathrm{E}\varepsilon_{j,i}^2\mathrm{E}\varepsilon_{j,k}^2. \tag{2.11}$$

For further simplification, we assume the coefficient of variations (CV) of $\varepsilon_{j,i}^2$ satisfy

$$CV^{-1}(\varepsilon_{j,i}^2)=\frac{\mathrm{E}\varepsilon_{j,i}^2}{\sqrt{\mathrm{var}\,(\varepsilon_{j,i}^2)}}\le C_j,\ i\in T_j, \tag{2.12}$$

and we reorder $\mathrm{var}(\varepsilon_{j,i}^2)$ such that

$$\mathrm{var}(\varepsilon_{j,1}^2)\le\mathrm{var}(\varepsilon_{j,2}^2)\le\cdots\le\mathrm{var}(\varepsilon_{j,m_j}^2).$$

Then from (2.11) and (2.12) we have

$$\mathrm{var}(\bar{\varepsilon}_j^2)\le\frac{1}{m_j^4}\,[(m_j-1)+2C_j^2m_j(m_j-1)]\,\mathrm{var}(\varepsilon_{j,m_j}^2)+\frac{1}{m_j^4}\,\mathrm{var}\,(\varepsilon_{j,1}^2). \tag{2.13}$$

It can be verified from (2.13) that

$$\mathrm{var}(\bar{\varepsilon}_j^2)\le\mathrm{var}\,(\varepsilon_{j,1}^2)$$

when the following condition is satisfied:

$$\mathrm{var}(\varepsilon_{j,m_j}^2)\le K_j\,\mathrm{var}(\varepsilon_{j,1}^2), \tag{2.14}$$

where

$$K_j=(m_j^2+1)(m_j+1)/(2m_jC_j^2+1).$$

For a set of given $\mathrm{var}(\varepsilon_{j,i}^2)$, condition (2.14) will be more easily satisfied if C_j is small. There is a close relation between $CV(\varepsilon_{j,i}^2)$ and the kurtosis $\gamma(\varepsilon_{j,i})$ of $\varepsilon_{j,i}$, that is

$$CV^{-1}(\varepsilon_{j,i}^2)=\frac{1}{\sqrt{2+\gamma(\varepsilon_{j,i})}}. \tag{2.15}$$

From (2.15), it can be seen that the larger the kurtosis $\gamma(\varepsilon_{j,i})$, the easier to satisfy condition (2.14). We list the kurtosis and the corresponding $CV^{-1}(\xi^2)$ of some commonly used distributions in the following table.

Among the distributions listed in Table 1, the uniform distribution has the smallest kurtosis; the next one is the normal distribution. Let us take the uniform distribution and check how condition (2.14) be satisfied. When $CV^{-1}=\sqrt{5}/2$, $K_j=2.5$. If we average two variables in the jth group with uniform distributions, then

Table 1
$\gamma(\xi)$ and $CV^{-1}(\xi^2)$ of Some Distributions

ξ	$\gamma(\xi)$	$CV^{-1}(\xi^2)$
$N(\mu, \sigma^2)$	0	$1/\sqrt{2}$
t_n	$6/(n-4)$	$1/\sqrt{2+6/(n-4)}$
χ^2_n	$12/n$	$1/\sqrt{2+12/n}$
$\log N(0, 1)$	$e^4+2e^3+3e^2-6$	$1/\sqrt{e^4+2e^3+3e^2-4}$
Laplace	3	$1/\sqrt{5}$
U[a, b]	$-6/5$	$\sqrt{5}/2$

(2.14) will be satisfied unless $\mathrm{var}(\varepsilon^2_{j,2})$ is 2.5 times of $\mathrm{var}(\varepsilon^2_{j,1})$. All the other distributions beside the U[a, b] in Table 1 have nonnegative kurtosis. For these distributions with positive kurtosis, it is easily verified that when we average two variables in the jth group, condition (2.14) will be satisfied unless $\mathrm{var}(\varepsilon^2_{j,2})$ is 5 times $\mathrm{var}(\varepsilon^2_{j,1})$. For all the distributions listed in Table 1, K_j will increase as m_j increases. When the $\mathrm{var}(\varepsilon^2_{j,i})$ are roughly the same, condition (2.14) will be more easily satisfied when we average more variables. Since we may not choose a marginal variable with the smallest $\mathrm{var}(\varepsilon^2_{j,1})$, like condition (2.7), condition (2.14) is also rather conservative. From the above discussion, we can see that the ADF estimator $\tilde{\boldsymbol{\theta}}_n$ based on model (2.4) will be more efficient than those based on model (2.5) when conditions (2.7) and (2.14) are satisfied.

We have discussed the effect of averaging variables on the efficiency of estimators. The effects on test statisics are also characterized by conditions (2.7) and (2.14). Suppose all the variables $x_{j,i}$, $i \in T_j$ have the same factor structure but are misspecified, that is $\min_\theta |\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \mathrm{var}(\boldsymbol{X})| = |\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) - \mathrm{var}(\boldsymbol{X})| > 0$. Assuming the regularity condition $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) - \mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Delta}/\sqrt{n}$, we next consider the effect of averaging variables on tests. When we use the normal theory likelihood ratio test statistic to judge the adequacy of a structural model, the test statistic will asymptotically follow $\chi^2_{p^*-q}(\delta)$, where q is the number of unknown free parameters in $\boldsymbol{\theta}_0$ and

$$\delta = \frac{1}{2}\,\mathrm{vec}'(\boldsymbol{\Delta})\{(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}) - (\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\dot{\boldsymbol{\sigma}}[\dot{\boldsymbol{\sigma}}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\dot{\boldsymbol{\sigma}}]^{-1}\dot{\boldsymbol{\sigma}}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\}\,\mathrm{vec}(\boldsymbol{\Delta}). \tag{2.16}$$

Using Lemma 1 of Khatri (1966), we can rewrite (2.16) as

$$\delta = \frac{1}{2}\,\mathrm{vec}'(\boldsymbol{\Delta})\{\dot{\boldsymbol{\sigma}}_c[\dot{\boldsymbol{\sigma}}_c'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\dot{\boldsymbol{\sigma}}_c]^{-1}\dot{\boldsymbol{\sigma}}_c'\}\,\mathrm{vec}(\boldsymbol{\Delta}), \tag{2.17}$$

where $\dot{\boldsymbol{\sigma}}_c$ is a $p^2 \times (p^2 - q)$ matrix of full column rank whose columns are orthogonal to those of $\dot{\boldsymbol{\sigma}}$. Since the diagonal elements in both $\boldsymbol{\Psi}$ and $\bar{\boldsymbol{\Psi}}$ are free parameters in the estimation process, the $\boldsymbol{\Delta}$ will be the same based on either model (2.4) or model (2.5). From (2.17), the magnitude of the noncentrality parameter will be decided by $\boldsymbol{\Sigma}$. It is obvious that the model with larger $\boldsymbol{\Sigma}$ will correspond to a smaller δ. Thus, when condition (2.7) is met, the test statistic based on model (2.4) is more powerful in detecting the misspecifications.

When model is misspecified as stated above, the ADF test statistic asymptotically follows $\chi^2_{p^*-q}(\delta)$ with

$$\begin{aligned}\delta &= \mathrm{vech}'(\boldsymbol{\Delta})\{\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\dot{\boldsymbol{\sigma}}(\dot{\boldsymbol{\sigma}}'\boldsymbol{V}^{-1}\dot{\boldsymbol{\sigma}})^{-1}\boldsymbol{V}^{-1}\}\mathrm{vech}(\boldsymbol{\Delta})\\ &= \mathrm{vech}'(\boldsymbol{\Delta})\dot{\boldsymbol{\sigma}}_c(\dot{\boldsymbol{\sigma}}_c'\boldsymbol{V}\dot{\boldsymbol{\sigma}}_c)^{-1}\dot{\boldsymbol{\sigma}}_c'\mathrm{vech}(\boldsymbol{\Delta}),\end{aligned} \tag{2.18}$$

where $\dot{\boldsymbol{\sigma}}_c$ is a $p^*\times(p^*-q)$ matrix of full column rank whose columns are orthogonal to those of $\dot{\boldsymbol{\sigma}}$. From (2.18), it is obvious that the noncentrality parameter will increase when $\boldsymbol{V}$ gets smaller. Consequently, the ADF test statistic based on model (2.4) will be more powerful when the conditions (2.7) and (2.14) are met.

In all the above discussion, we assumed $\lambda_{j,i}=\lambda_j$, $i\in T_j$. When this assumption is not appropriate, $\psi_{j,i}$ should be replaced by $\psi_{j,i}/\lambda^2_{j,i}$ in condition (2.7); and $\mathrm{var}(\varepsilon^2_{j,i})$ should be replaced by $\mathrm{var}(\varepsilon^2_{j,i})/\lambda^4_{j,i}$ in condition (2.14). Assuming that $\mathrm{var}(f_{k(j)})=1$, the quantity $\varepsilon_{j,i}/\lambda_{j,i}$ reflects the relative error in $x_{j,i}$ as an indicator of latent variable $f_{k(j)}$. When relative errors among $x_{j,i}$ have great differences, better estimators and tests based on model (2.5) may be obtained, assuming that the marginal variables selected have the smallest relative errors. However, when we have no idea about which variables have the smallest errors, or when all the variables have roughly equal relative errors, estimators based on model (2.4) will be much more efficient.

3. The effect of averaging variables on mean square error of predicted factor scores

In this section, we will compare variable selection based on minimizing the prediction error proposed by Yanai (1980) and Tanaka (1983) and the averaging variables procedure discussed in the last section. Specifically, we will compare the MSE defined in (1.1) based on models (2.4) and (2.5). Two methods of factor score prediction are commonly used, one is the regression method, and the other is Bartlett's method. These factor score estimates are given respectively by $\hat{\boldsymbol{f}}_R=\boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ and $\hat{\boldsymbol{f}}_B=(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}$. The MSE of the predicted factor scores by the regression method is

$$\begin{aligned}\mathrm{MSE}_R &= \mathrm{E}(\hat{\boldsymbol{f}}_R-\boldsymbol{f})'(\hat{\boldsymbol{f}}_R-\boldsymbol{f})\\ &= \mathrm{tr}(\boldsymbol{\Phi}-\boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi})\\ &= \mathrm{tr}(\boldsymbol{\Phi}^{-1}+\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}.\end{aligned} \tag{3.1}$$

It is easily seen from (3.1) that the model with smaller diagonal elements in $\boldsymbol{\Psi}$ gives a smaller MSE. Suppose that condition (2.7) holds. Then the MSE of the predicted factor score $\hat{\boldsymbol{f}}_R$ based on model (2.4) is always smaller than that based on model (2.5). The MSE of Bartlett's predicted factor scores is

$$\begin{aligned}\mathrm{MSE}_B &= \mathrm{E}(\hat{\boldsymbol{f}}_B - \boldsymbol{f})'(\hat{\boldsymbol{f}}_B - \boldsymbol{f}) \\ &= \mathrm{tr}(\Lambda' \boldsymbol{\Psi}^{-1} \Lambda)^{-1}.\end{aligned} \tag{3.2}$$

So the conclusion for Bartlett's factor scores is the same as for the regression factor scores. Suppose $T_1 = \{1, 2\}$ and all the other T_j only contain one element. If $\psi_{1,2}$ is less than three times of $\psi_{1,1}$, we should average $x_{1,1}$ and $x_{1,2}$ instead of removing $x_{1,2}$ from the analysis, even though the MSE corresponding to removing $x_{1,2}$ is the smallest. When condition (2.7) does not hold, we may get a better predicted factor score by using only marginal variables. This comparison also gives us some insight into variable selection procedures: if we really need to remove some variables from a factor analysis model, we should remove variables with the largest relative errors. This observation is also intuitively natural. Suppose we can observe f_1 exactly by $x_{1,1} = f_1$, then the best predicted factor score for f_1 is to use $\hat{f}_1 = x_{1,1}$. If we average $x_{1,1}$ with any other variables with nonzero errors, we will not get the best predicted factor score anymore.

Let MSE($\boldsymbol{X}$) and MSE($\boldsymbol{X}_m$) denote the predicted error of $\boldsymbol{f}$ by $\hat{\boldsymbol{f}}$ based on all the observed variables and some marginal variables, respectively. It can be shown that

$$\mathrm{MSE}(\boldsymbol{X}) \leq \mathrm{MSE}(\boldsymbol{X}_m) \tag{3.3}$$

generally ; the inequality is strict in most cases. Let MSE($\boldsymbol{Y}$) denote the predicted error of $\hat{\boldsymbol{f}}$ based on the averaged variables. We want to compare MSE($\boldsymbol{Y}$) with MSE($\boldsymbol{X}$). We will assume that the model structure is the same as in Section 2. For simplicity, we also assume that $\boldsymbol{\Phi} = \boldsymbol{I}$. Then for the Bartlett's factor score predictor, we have

$$\mathrm{MSE}_B(\boldsymbol{X}) = \sum_{l=1}^{r} \left\{ \sum_{j \in S_l} \left(\lambda_j^2 \sum_{i=1}^{m_j} \psi_{j,i}^{-1} \right) \right\}^{-1}, \tag{3.4}$$

$$\mathrm{MSE}_B(\boldsymbol{X}_m) = \sum_{l=1}^{r} \left\{ \sum_{j \in S_l} (\lambda_j^2 \psi_{j,1}^{-1}) \right\}^{-1}, \tag{3.5}$$

$$\mathrm{MSE}_B(\boldsymbol{Y}) = \sum_{l=1}^{r} \left[\sum_{j \in S_l} \left\{ \lambda_j^2 \left(\frac{1}{m_j^2} \sum_{i=1}^{m_j} \psi_{j,i} \right)^{-1} \right\} \right]^{-1}. \tag{3.6}$$

From (3.4) to (3.6), we can see that $\mathrm{MSE}_B(\boldsymbol{X}) < \mathrm{MSE}_B(\boldsymbol{X}_m)$ unless $\psi_{j,i} = \infty$, $i \geq 2$, $j = 1, \cdots, p$. Using the Cauchy-Schwarz inequality $(\boldsymbol{x}'\boldsymbol{y})^2 \leq (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{y})(\boldsymbol{x}'\boldsymbol{A}^{-1}\boldsymbol{y})$, we have

$$\left(\sum_{i=1}^{m_j} \psi_{j,i}^{-1} \right)^{-1} \leq \frac{1}{m_j^2} \sum_{i=1}^{m_j} \psi_{j,i},$$

by letting $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{1}$, a vector of 1 with length m_j ; and $\boldsymbol{A} = \mathrm{diag}(\psi_{j,1}, \cdots, \psi_{j,m_j})$. So it generally holds that

$$\mathrm{MSE}_B(\boldsymbol{X}) \leq \mathrm{MSE}_B(\boldsymbol{Y}). \tag{3.7}$$

But when $\psi_{j,1} = \psi_{j,2} = \cdots = \psi_{j,m_j}$, $j = 1, \cdots, p$, we have $\mathrm{MSE}_B(\boldsymbol{X}) = \mathrm{MSE}_B(\boldsymbol{Y})$. This means that the predicted factor scores based on averaged variable will have the same accuracy as that based on all the variables, when the error variances of the averaged variables are roughly equal. It will never attain this property for the predicted scores based on marginal variables.

For the regression factor score predictor, we have

$$\mathrm{MSE}_R(\boldsymbol{X})=\sum_{l=1}^{r}\left\{1+\sum_{j\in S_l}\left(\lambda_j^2\sum_{i=1}^{m_j}\psi_{j,i}^{-1}\right)\right\}^{-1}, \tag{3.8}$$

$$\mathrm{MSE}_R(\boldsymbol{X}_m)=\sum_{l=1}^{r}\left\{1+\sum_{j\in S_l}(\lambda_j^2\psi_{j,1}^{-1})\right\}^{-1}, \tag{3.9}$$

$$\mathrm{MSE}_R(\boldsymbol{Y})=\sum_{l=1}^{r}\left[1+\sum_{j\in S_l}\left\{\lambda_j^2\left(\frac{1}{m_j^2}\sum_{i=1}^{m_j}\psi_{j,i}\right)^{-1}\right\}\right]^{-1}. \tag{3.10}$$

By comparing the coefficient of λ_j in (3.8) to (3.10), we come to the same conclusion for the regression score predictor as that for Bartlett's score predictor. That is $\mathrm{MSE}_R(\boldsymbol{X})<\mathrm{MSE}_R(\boldsymbol{X}_m)$, $\mathrm{MSE}_R(\boldsymbol{X})\le\mathrm{MSE}_R(\boldsymbol{Y})$ while $\mathrm{MSE}_R(\boldsymbol{X})=MSE_R(\boldsymbol{Y})$ when $\psi_{j,1}=\psi_{j,2}=\cdots=\psi_{j,m_j}$, $j=1,\cdots,p$.

Now, let us consider weighted averages instead of simple unweighted averages. Let $l_{j,i}$ be positive numbers satisfying $\sum_{i=1}^{m_j}l_{j,i}=1$. Corresponding to (2.2), let

$$y_j=\sum_{i\in T_j}l_{j,i}x_{j,i}=\bar{\lambda}_jf_{k(j)}+\bar{\varepsilon}_j \tag{3.11}$$

with

$$\bar{\lambda}_j=\sum_{i\in T_j}l_{j,i}\lambda_{j,i}$$
$$\bar{\varepsilon}_j=\sum_{i\in T_j}l_{j,i}\varepsilon_{j,i}.$$

Corresponding to (3.6), we have

$$\mathrm{MSE}_B(\boldsymbol{Y})=\sum_{l=1}^{r}\left(\sum_{j\in S_l}\bar{\lambda}_j^2\bar{\psi}_j^{-1}\right)^{-1}, \tag{3.12}$$

where

$$\bar{\psi}_j=\sum_{i\in T_j}l_{j,i}^2\psi_{j,i}.$$

We will identify the optimal $l_{j,i}$ which minimize (3.12). Let $l_{j,i}\lambda_{j,i}=w_{j,i}$, then $l_{j,i}=w_{j,i}/\lambda_{j,i}$ and

$$\bar{\lambda}_j^{-2}\bar{\psi}_j=\left\{\sum_{i=1}^{m_j}\frac{\psi_{j,i}}{\lambda_{j,i}^2}w_{j,i}^2\right\}\Big/\left(\sum_{i=1}^{m_j}w_{j,i}\right)^2. \tag{3.13}$$

The constraint $\sum_{i=1}^{m_j}l_{j,i}=1$ is equivalent to $\sum_{i=1}^{m_j}w_{j,i}=1$ in minimizing (3.13). Standard calculation using Lagrange multipliers obtains $w_{j,i}\sim\lambda_{j,i}^2/\psi_{j,i}$ and $l_{j,i}\sim\lambda_{j,i}/\psi_{j,i}$ with a minimum mean square error of

$$\mathrm{MSE}_B(\boldsymbol{Y})=\sum_{l=1}^{r}\left\{\sum_{j\in S_l}\left(\sum_{i=1}^{m_j}\frac{\lambda_{j,i}^2}{\psi_{j,i}}\right)\right\}^{-1}. \tag{3.14}$$

When assuming $\lambda_{j,i}$, $i\in T_j$ are equal, $l_{j,i}\sim\psi_{j,i}^{-1}$. This suggests that larger weights should be given to variables with smaller relative measurement errors, conforming with our earlier observations. Also note that (3.14) equals the MSE based on the whole sample in (3.4)!

Theoretically $l_{j,i} \sim \lambda_{j,i}/\psi_{j,i}$ gives the smallest MSE, while in practice we need to estimate $\lambda_{j,i}$ and $\psi_{j,i}$ in order to use the optimal weights. When facing a data set with very discordant relative errors, one may fit a one factor model on each subset with similar structures. After obtaining estimators of $\lambda_{j,i}$ and $\psi_{j,i}$, one can form variables using the optimal linear combinations. Since this will incur rounding errors besides increased computation time, in practice, we would suggest using the simple ordinary average when empirical knowledge indicates that not much difference in relative measurement errors exists. When a large number of variables is to be averaged, differential weighting would not be expected to make much of a difference in any case. See Kano (1986) for evaluation of the MSE of predicted factor scores with estimated parameters.

4. Discussion and conclusions

Motivated by the practical difficulty of analyzing high dimensional data, instead of discarding variables as has been proposed and is commonly done, we propose to average variables with similar latent structures and then to keep those averaged variables for subsequent analyses. In much of the practice of structural equation modeling, observed variables are assumed to depend only on one latent variable. When this is the case, averaged variables will keep the same factor structure as that of the original variables being averaged. Sometimes, a proper scaling may be necessary before averaging, e.g., to make the factor loadings of the averaged variables have the same sign. Then it can be expected that the factor loading based on averaged variables will be the average of the factor loadings of the variables being averaged. If necessary, a preliminary exploratory factor analysis or principal component analysis can be undertaken before averaging those variables expected to have a similar factor structure. Based on the factor loadings obtained, a rescaling to make the averaged variables have roughly equal factor loadings can be performed. Also, through a preliminary analysis, it is possible to identify those variables whose relative errors are too large to average with other variables. Since each latent factor needs more than one indicator to insure that a model is identified, one can classify the variables with similar factor structures into several groups based on their relative errors ; then, variables in each group that have roughly equal relative errors can be averaged. This will make the final estimators the most efficient. If, based on empirical knowledge, we have confidence that several variables are good indicators of a latent factor, a preliminary analysis may not be necessary. Since good indicators mean small relative errors, averaging these variables will result in more accurate estimates than omitting any of them.

Our results also apply to more complicated structural models than the basic confirmatory factor analysis model. In the development of models for interaction effects among latent variables, Kenny and Judd (1984) used four indicators for each quadratic latent variable. Because of a rank deficiency in the weight matrix when using such multiple indicators, Jöreskog and Yang (1996), in contrast, recommended that only one indicator should be used for each quadratic latent variable. If one indicator is prefered, the results of this paper verify that the one formed by averaging the available indicators can increase the accuracy of estimators and tests.

Quite another rationale should be remembered for averaging variables. Even though ADF theory makes it possible to estimate and test covariance structure models without assuming any specific underlying distribution, normal theory maximum likelihood is still the most popular method that practitioners use. It can generally be believed that the latent common factors are normally distributed, and that the nonnormality of the observed data results from the nonnormality of unique factors or errors. Since the error of the averaged variables is the average error of the variables being averaged, by the central limit theorem, the normality of the averaged error will be improved as the number of variables being averaged increases. For example, when the indicator variables are high dimensional categorical variables, normal theory maximum likelihood can not be applied directly. However, it is still possible that normal theory can give a good approximation to the model based on averaged variables.

REFERENCES

Anderson, J.C. and Gerbing, D.W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research*, **19**, 453-460.

Anderson, J.C. and Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, **103**, 411-423.

Bentler, P.M. and Wu, E.J.C. (1995). *EQS for Windows (Macintosh) User's Guide.* Encino, CA: Multivariate Software.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, **9**, 139-164.

Hunter, J.E. and Gerbing, D.W. (1982). Unidimensional measurement, second-order factor analysis, and causal models. In B.M. Staw and L.L. Cummings (eds.), *Research in organizational behavior* (Vol. 4, pp. 267-299). Greenwich, CT: JAI Press.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis I. Artificial data. *Applied Statistics*, **21**, 160-173.

Jolliffe, I.T. (1973). Discarding variables in a principal component analysis II. Real data. *Applied Statistics*, **22**, 21-31.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, **36**, 109-133.

Jöreskog, K.G. and Sörbom, D. (1993). *New Features in LISREL 8*. Chicago: Scientific Software.

Jöreskog, K.G. and Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G.A. Marcoulides and R.E. Schumacker (Eds.), *Advanced Structural Equation Modeling: Issues and Techniques* (pp. 57-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Kano, Y. (1986). A condition for the regression predictor to be consistent in a single common factor model. *British Journal of Mathematical and Statistical Psychology*, **39**, 221-227.

Kenny, D.A. and Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, **96**, 201-210.

Khatri, C.G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, **18**, 75-86.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York.

Tanaka, Y. (1983). Some criteria for variable selection in factor analysis. *Behaviormetrika*, **13**, 31-45.

Tanaka, Y. and Kodake, K. (1981). A method of variable selection in factor analysis and its numerical investigation. *Behaviormetrika*, **10**, 49-61

Yanai, H. (1980). A proposition of generalized method for forward selection of variables. *Behaviormetrika*, **7**, 31-45.

(Received May, 1996, Revised November, 1996)