ALL ABOUT VARIABLE SELECTION IN FACTOR ANALYSIS AND STRUCTURAL EQUATION MODELING

Y. KANO

OSAKA UNIVERSITY, JAPAN
kano@hus.osaka-u.ac.jp

1. Introduction. Since variable selection is an important process of multivariate analysis, the topic has been discussed extensively in the literature and many important fruitful consequences have been implemented in statistical programs. Almost all the discussion on variable selection, however, focuses on the selection of independent variables in models with a clear dependent (criteria) variable such as regression analysis, discriminant analysis and time series analysis. This does not mean that variable selection in other models such as factor analysis and principal component analysis is not important. In analyses with those models, variable selection is a very important step. In fact, when one makes analysis of data from a questionnaire, there are usually many items and some of them are selected and analyzed. More importantly, scale construction in social sciences is nothing but variable selection in factor analysis. Thus, the variable selection is often made in those models. The problem is that no well-established procedure exists and no option for variable selection is supplied in statistical programs.

2. Variable selection in a covariance structure model. Yanai (1980) and Tanaka (1983) have used factor score configuration to discuss variable selection in factor analysis. Kano and Ihara (1994) suggested use of goodness of fit measures to select variables to well fit the model considered to a data set in factor analysis. Kano and Harada (2000) took the Lagrange Multiplier test approach to reduce computation of goodness-of-fit measures, and they developed a computer program named *SEFA* to print a list of several goodness-of-fit measures for models that are obtained by deletion or addition of one variable. The *SEFA* runs on a WWW server and can be used by anyone who can access internet. Kano and Harada (2001) employed the same idea to develop a program *SCoFA* for variable selection in confirmatory factor analysis[1].

Let us explain briefly the basic idea of the variable selection procedure in a covariance structure model $\{V(\boldsymbol{X}) = \Sigma(\theta) | \theta \in \Theta\}$. Consider construction of a (approximate) test statistic or goodness-of-fit indices for a model in which the first variable X_1 is deleted from $\boldsymbol{X}$, using only statistics associated with the original model $\Sigma(\theta)$ for $\boldsymbol{X}$. Let T_0, T_2 and $T_{2'}$ be the chi-square likelihood ratio statistics for testing goodness-of-fit of the models $V(\boldsymbol{X}) = \Sigma(\theta)$, $V(\boldsymbol{X}_2) = \Sigma_{22}(\theta)$ and $V(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{21}} & \Sigma_{22}(\theta) \end{bmatrix}$, respectively. Let $T_{02'}$ be the LM statistic for testing

$$H_0 : V(\boldsymbol{X}) = \Sigma(\theta) \quad \text{versus} \quad H_{2'} : V(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{21}} & \Sigma_{22}(\theta) \end{bmatrix}.$$

We then have

$$T_2 \stackrel{a}{=} T_{2'} = T_0 - (T_0 - T_{2'}) = T_0 - T_{02'}.$$

(See Kano & Harada (2000) for details.) This shows that the test statistic for the model for $\boldsymbol{X}_2$ can be formed as $T_0 - T_{02'}$ which is just a function of the statistics associated with the model for $\boldsymbol{X}$. A similar formula holds for any model deleting one variable other than X_1. The LM approach successfully reduce computation, compared to the amount of total computation for p marginal models with $p-1$ variables individually.

3. Certain theoretical justification. Michael Browne (1998) mentioned that the LM statistic could not work if $\boldsymbol{X}_2$ would contain variables inconsistent with the model. For the question, Kano (1999) proved that small misspecification for the $\boldsymbol{X}_2$ does not fatally influence on the performance of the LM test. In fact, he showed that $T_0 - T_{02'}$ converges in law to the *central* chi-square distribution as the sample size n tends to infinity if $d_{11} = 0$ and $\boldsymbol{d}_{12} = \boldsymbol{0}$ in (1), and to the *noncentral* chi-square distribution if $d_{11} \neq 0$ or $\boldsymbol{d}_{12} \neq \boldsymbol{0}$ in (1), even if

$$V(\boldsymbol{X}) = \Sigma(\theta) + \frac{1}{\sqrt{n}} \begin{bmatrix} d_{11} & \boldsymbol{d}_{12} \\ \boldsymbol{d}_{21} & D_{22} \end{bmatrix} \tag{1}$$

with D_{22} possibly nonzero.

---

[1] http://koko15.hus.osaka-u.ac.jp/~harada/sefa2001/stepwise/
http://koko16.hus.osaka-u.ac.jp/~harada/scofa/input.html

4. What is actually done in variable selection with model fit. A FAQ (frequently asked question) for the variable selection procedure from practitioners is what the meaning is of choosing variables through model fit. Also we have often been asked what is actually done in variable selection with model fit or how to do in case where the program indicates that a variable with substantial communality is inconsistent with the model.

One typical answer to the first question is that examination of model adequacy is the first step of statistical analysis and that a model with poor fit is often misleading. To answer all the questions, let us take reliability analysis as an example. One purpose of factor analysis is to construct a scale of a psychological construct and to measure the reliability of the scale constructed.

Consider a one-factor model:

$$X_i = \mu_i + \lambda_i f + u_i \qquad (i = 1, \ldots, p),$$

where $E(f) = E(u_i) = 0$, $V(f) = 1$, $V(u_i) = \psi_i$, $\mathrm{Cov}(f, u_i) = 0$ and $\mathrm{Cov}(u_i, u_j) = 0$ $(i \neq j)$. The scale score is defined as the total sum of X_i, i.e., $X = \sum_{i=1}^{p} X_i$. The scale reliability is then defined as

$$\rho = \frac{V(\sum_{i=1}^{p} \lambda_i f)}{V(X)} = \frac{(\sum_{i=1}^{p} \lambda_i)^2}{(\sum_{i=1}^{p} \lambda_i)^2 + \sum_{i=1}^{p} \psi_i}. \tag{2}$$

What if the factor analysis model fails to fit to the data? As an example, we consider the case where unique factors are correlated, say, $\mathrm{Cov}(u_1, u_2) = \psi_{12} \neq 0$. Then the reliability is not given as in (2) any more, and the presice one is

$$\rho' = \frac{(\sum_{i=1}^{p} \lambda_i)^2}{(\sum_{i=1}^{p} \lambda_i)^2 + \sum_{i=1}^{p} \psi_i + 2\psi_{12}} \quad (\neq \rho). \tag{3}$$

Use of the formula in (2) necessitates good fit of a factor analysis model. If the factor model assumption is violated, practitioners have to use formulas of reliability such as that in (3). When $\psi_{12} > 0$, we have $\rho' < \rho$, and hence ρ in (2) overestimates the reliability. Raykov (2001) among others discussed bias of reliability coefficients or Cronbach's α caused by error correlations. The examination of model fit gives relevant criterion to the problem whether the ρ in (2) can be used. Kano and Azuma (2001) discussed use of the formula (3) to precisely measure reliability and also suggest a procedure how to identify the pairs of unique variables that are correlated. Of course, examining adequacy of one-factor model gives a useful information on unidimensionality of the scale constructed.

Related with the discussion above is whether variables with large communalities but inconsistent with the model should be dropped. In general, removal of variables with large communality causes reduction of reliability. On the other hand, the variable unfitted to a one-factor model can reduce reliability as shown above. Thus, one can not mention anything about whether such a variable should be included, without examination by the formula (3), as far as reliability is concerned.

A similar problem is whether a variable should be included which is consistent with the model but whose communality is small. Harada and Kano (2001) further developed the *SEFA* to print results of testing whether the communality is zero.

Key words: Structural model, LM test, misspecification, goodness of fit index, chi-square statistic.

References

Browne, M. W. (1998). Personal communications.

Harada, A. & Kano, K. (2001) Variable selection and test of communality in exploratory factor analysis. Paper presented at the IMPS2001. Osaka, Japan.

Kano, Y. & Azuma, Y. (2001). Use of SEM programs to precisely measure scale reliability. Paper presented at the IMPS2001. Osaka, Japan.

Kano, Y. (1999/April). Variable selection for structural models. *Technical Report DATA99-01.* Osaka, Japan. (To appear in *Journal of Statistical Planning and Inference*)

Kano, Y. & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika, 65*, 7–22.

Kano, Y. & Harada, A. (2001). *SCoFA*. Stepwise variable selection in confirmatory factor analysis. In preparation.

Kano, Y. & Ihara, M. (1994). Identification of inconsistent variates in factor analysis. *Psychometrika, 59*, 5–20.

Raykov, T. (2001). Bias of Cronbach's coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 26*, 69–76.

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

Yanai, H. (1980). A proposition of generalized method for forward selection of variables. *Behaviormetrika, 7*, 95-107.