(11) 'Factor Analysis Centennial Symposium' に関する研究報告

Maximum Likelihood Analysis of a Structural Equation Models with Non-normal Data

The Chinese University of Hong Kong Sik-Yum Lee & Xin-Yuan Song

In this paper, we consider maximum likelihood (ML) analysis of some non-normal data in a framework of structural equation modeling. The non-normality under consideration is due to the discreteness of the data, which are ordered categorical, and the non-linearity of the latent variables. To improve the applicability for practical problems, the model framework can accommodate two-level data that can be missing at random (MAR).

Suppose that there is a set of $p \times 1$ random vectors $\mathbf{y}_{gi}$, $i = 1, \cdots, N_g$, that are nested within groups $g = 1, \cdots, G$. A two-level model is proposed as follow: $\mathbf{y}_{gi} = \mathbf{v}_g + \mathbf{v}_{gi}$, $i = 1, \cdots, N_g$, $g = 1, \cdots, G$, where $\mathbf{v}_g$, and $\mathbf{v}_{gi}$ are random vectors such that $\mathbf{v}_g$ and $\mathbf{v}_h$ are independent for $g \neq h$, and $\mathbf{v}_{gi}$ and $\mathbf{v}_{gj}$ are independent for $i \neq j$. We define the measurement equation of the individual-level model as: $\mathbf{v}_{gi} = \mathbf{\Lambda}_1 \boldsymbol{\omega}_{1gi} + \boldsymbol{\varepsilon}_{gi}$, $\boldsymbol{\omega}_{1gi} = (\boldsymbol{\eta}_{gi}^T, \boldsymbol{\xi}_{1gi}^T)^T$ is a vector of latent variables, and $\boldsymbol{\varepsilon}_{gi}$ is a vector of error measurements that is independent with $\boldsymbol{\omega}_{1gi}$ and is distributed as $N[\mathbf{0}, \mathbf{\Psi}_1]$. The structural equation is defined as $\boldsymbol{\eta}_{1gi} = \mathbf{\Pi}_1 \boldsymbol{\eta}_{1gi} + \mathbf{\Gamma}_1 \mathbf{F}(\boldsymbol{\xi}_{1gi}) + \boldsymbol{\delta}_{1gi}$, where $\mathbf{\Pi}_1$ and $\mathbf{\Gamma}_1$ are unknown regression coefficients, $\mathbf{F}(\boldsymbol{\xi}_{1gi})$ is a vector of differentiable functions, $\boldsymbol{\xi}_{1gi}$ is a latent random vector with distribution $N[\mathbf{0}, \mathbf{\Phi}_1]$, and $\boldsymbol{\delta}_{1gi}$ is a vector of error measurements that is independent of $\boldsymbol{\xi}_{1gi}$, and is distributed as $N[\mathbf{0}, \mathbf{\Psi}_{1\delta}]$, where $\mathbf{\Psi}_{1\delta}$ is diagonal. We assume that $|\mathbf{I}_1 - \mathbf{\Pi}_1|$ is a non-zero value that is independent of $\mathbf{\Pi}_1$. We consider a confirmatory factor analysis model for the group-level model. The measurement equation is defined by $\mathbf{v}_g = \mathbf{\Lambda}_2 \boldsymbol{\omega}_{2g} + \boldsymbol{\varepsilon}_g$, $g = 1, \cdots, G$, where $\boldsymbol{\omega}_{2g}$ is a vector of latent variables with distribution $N[\mathbf{0}, \mathbf{\Phi}]$, and $\boldsymbol{\varepsilon}_g$ is a vector of error measurements that is independent with $\boldsymbol{\omega}_{2g}$ and is distributed as $N[\mathbf{0}, \mathbf{\Psi}_2]$, where $\mathbf{\Psi}_2$ is diagonal matrix.

Without lost of generality, let $\mathbf{y}_{gi} = (\mathbf{x}_{gi}^T, \mathbf{w}_{gi}^T)^T$, where $\mathbf{x}_{gi}$ are observable continuous measurements, and $\mathbf{w}_{gi}$ are latent continuous measurements that correspond to the observable ordered categorical measurements. Moreover, we consider $\mathbf{x}_{gi} = \{\mathbf{x}_{gi,obs}, \mathbf{x}_{gi,mis}\}$ and $\mathbf{z}_{gi} = \{\mathbf{z}_{gi,obs}, \mathbf{z}_{gi,mis}\}$, where $\mathbf{x}_{gi,obs}$ and $\mathbf{z}_{gi,obs}$ represent the observed data, while $\mathbf{x}_{gi,mis}$ and $\mathbf{z}_{gi,mis}$ represent the MAR missing data. Let $\mathbf{X}_{obs} = \{\mathbf{x}_{gi,obs}; i = 1, \cdots, N_g, g = 1, \cdots, G\}$ and let $\mathbf{Z}_{obs} = \{\mathbf{z}_{gi,obs}; i = 1, \cdots, N_g, g = 1, \cdots, G\}$, our main objective is to obtain the ML estimates of the unknown parameters, and the BIC for model comparison, on the basis of $\mathbf{X}_{obs}$ and $\mathbf{Z}_{obs}$.

Let $\mathbf{D}_o$ be the observed data set, $\mathbf{D}_u$ be the latent data, $\mathbf{\Omega}$ be the latent variables, $\mathbf{D}_c$ be the complete-data set; and $\boldsymbol{\theta}$ be the vector that contains the unknown parameters, and $L_c(\mathbf{D}_0, \mathbf{D}_u, \mathbf{\Omega}; \boldsymbol{\theta})$ be the complete-data log-likelihood.

The ML estimate of $\boldsymbol{\theta}$ is obtained via the following EM algorithm: E-step: Evaluate $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = E\{L_c(\mathbf{D}_o, \mathbf{D}_u, \mathbf{\Omega}; \boldsymbol{\theta}) | \mathbf{D}_o, \boldsymbol{\theta}^{(r)}\}$, in which the expectation is taken with respect to the conditional distribution of $(\mathbf{D}_u, \mathbf{\Omega})$ given $\mathbf{D}_o$ at $\boldsymbol{\theta}^{(r)}$. M-step: Determine $\boldsymbol{\theta}^{(r+1)}$ by maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$. The E-step is completed via a Gibbs sampler (Geman & Geman, 1984) algorithm, which iteratively draws observations from the components in $p(\mathbf{D}_u | \mathbf{D}_0, \mathbf{\Omega}; \boldsymbol{\theta})$ and $p(\mathbf{\Omega} | \mathbf{D}_0, \mathbf{D}_u; \boldsymbol{\theta})$. The M-step is completed by conditional maximization (Meng & Rubin, 1993).

The following Bayesian information criterion (Kass & Raftery, 1995) is proposed to serve

as a statistic for comparing models M_1 and M_2:

$$BIC_{12} = -2\left\{\log \frac{p(\mathbf{D}_o; M_1, \hat{\boldsymbol{\theta}}_1)}{p(\mathbf{D}_o; M_2, \hat{\boldsymbol{\theta}}_2)}\right\} + (d_1 - d_2)\log N,$$

where $p(\mathbf{D}_o; M_k, \hat{\boldsymbol{\theta}}_k)$ is the observed data likelihood that is evaluated at the ML estimate of $\boldsymbol{\theta}_k$ under M_k, d_k is the dimension of $\boldsymbol{\theta}_k$, and N is the sample size.

For comparing simple SEMs M_1 and M_2, if $p(\mathbf{D}_o; M_1, \hat{\boldsymbol{\theta}}_1)/p(\mathbf{D}_o; M_2, \hat{\boldsymbol{\theta}}_2)$ can be easily calculated, the application of BIC_{12} is straightforward. However, for some complicated models, the computation of this ratio, and hence BIC_{12}, would be difficult. Gelman and Meng (1998) developed a rather efficient and accurate algorithm, called path sampling, for computing normalizing constants of probability models, and showed that it is a generalization of importance sampling and bridge sampling. This algorithm is implemented to compute the observed data likelihood in the BIC of our problem.

Suppose we wish to compute the observed data likelihood $p(\mathbf{D}_o; M_1, \hat{\boldsymbol{\theta}}_1)$. We consider a model M_0 with a parameter vector $\boldsymbol{\theta}_0$ that is nested in M_1 such that $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1^*)$ and $\boldsymbol{\theta}_1$ reduces to $\boldsymbol{\theta}_0$ if $\boldsymbol{\theta}_1^* = \mathbf{0}$. Then, we can link up M_1 and M_0 by a linked model M_t with parameter vector $t\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_0, t\boldsymbol{\theta}_1^*)$ via a continuous path t in $[0, 1]$. Let $p(\mathbf{D}_o, \mathbf{D}_u, \boldsymbol{\Omega}; t\hat{\boldsymbol{\theta}}_1)$ be a complete data density function evaluated at the ML estimate $t\hat{\boldsymbol{\theta}}_1$ for each t, such that at $t = 1$ and 0, $p(\mathbf{D}_o, \mathbf{D}_u, \boldsymbol{\Omega}; \hat{\boldsymbol{\theta}}_1)$ and $p(\mathbf{D}_o, \mathbf{D}_u, \boldsymbol{\Omega}; \hat{\boldsymbol{\theta}}_0)$ are the density functions that correspond to M_1 and M_0, respectively. Let $U(\mathbf{D}_o, \mathbf{D}_u, \boldsymbol{\Omega}; t\hat{\boldsymbol{\theta}}_1) = d\log p(\mathbf{D}_o, \mathbf{D}_u, \boldsymbol{\Omega}; t\hat{\boldsymbol{\theta}}_1)/dt$, and $\lambda_{10} = \log[p(\mathbf{D}_o; M_1, \hat{\boldsymbol{\theta}}_1)/p(\mathbf{D}_o; M_0, \hat{\boldsymbol{\theta}}_0)]$. It can be shown by similar deviation as in Gelman and Meng (1998) that

$$\lambda_{10} \doteq \frac{1}{2}\sum_{s=0}^{S}(t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}) \ ,$$

where $\{t_{(s)}; s = 0, \ldots, S\}$ are ordered grids in $[0, 1]$, $\bar{U}_{(s)} = J^{-1}\sum_{j=1}^{J} U(\mathbf{D}_o, \mathbf{D}_u^{(j)}, \boldsymbol{\Omega}^{(j)}; t_{(s)}\hat{\boldsymbol{\theta}}_1)$, with $\{(\mathbf{D}_u^{(j)}, \boldsymbol{\Omega}^{(j)}); j = 1, \ldots, J\}$ are observations that are simulated at the E-step from the conditional distribution of $(\mathbf{D}_u, \boldsymbol{\Omega})$ given $\mathbf{D}_o$ at $t_{(s)}\hat{\boldsymbol{\theta}}_1$. Hence, we have $\log[p(\mathbf{D}_o, M_1, \hat{\boldsymbol{\theta}}_1)] = \lambda_{10} + \log p(\mathbf{D}_o; M_0; \hat{\boldsymbol{\theta}}_0)$. The similar procedure can be applied to find $\log p(\mathbf{D}_o, M_2, \hat{\boldsymbol{\theta}}_2)$. Finally, BIC_{12} can be obtained.

References

Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163-185.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267-278.

CONSTRUCTING CONFIDENCE INTERVALS FOR COMMUNALITIES IN FACTOR ANALYSIS

Masanori Ichikawa Tokyo University of Foreign Studies
Sadanori Konishi Graduate School of Mathematics, Kyushu University

We consider the problem of constructing accurate confidence intervals for communalities in the normal theory exploratory factor analysis. Under the model, the covariance matrix $\mathbf{\Omega}$ (> 0) of the observed variables $x_1, \ldots, x_p$ is decomposed as $\mathbf{\Omega} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$, where $\mathbf{\Lambda}$ is a $p \times m$ matrix of *factor loadings* and $\mathbf{\Psi}$ is a diagonal matrix. The variance of each x_i is decomposed as the sum of *communality* η_i and *uniqueness* ψ_i. Let $\mathbf{S}$ be the usual unbiased estimator of $\mathbf{\Omega}$ based on a random sample of size $N = n + 1$ from $\mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Omega})$, where $\mathbf{\Omega} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. The maximum Wishart likelihood estimator of the communality $\hat{\eta}_i$ is a function of $\mathbf{S}$ and is Fisher consistent. Hence, the problem of interest here is to construct confidence intervals for a real valued function $h(\cdot)$ of $\mathbf{\Omega}$.

Let $K(x)$ be the distribution function of the Studentized statistic $T = n^{1/2}\{h(\mathbf{S}) - h(\mathbf{\Omega})\}/\hat{\sigma}$, where $\hat{\sigma}^2$ is a consistent estimator of the asymptotic variance σ^2 of $Y = n^{1/2}\{h(\mathbf{S}) - h(\mathbf{\Omega})\}$. Then the exact α endpoint of $h(\mathbf{\Omega})$ based on the distribution of T is given by

$$h_{\mathrm{exact}}[\alpha] = h(\mathbf{S}) - n^{-1/2}\hat{\sigma}K^{-1}(1-\alpha).$$

The usual interval is based on the limiting normal approximation to the distribution of T and its α-level endpoint is, by replacing $K^{-1}(1-\alpha)$ with the percentile point $z_{1-\alpha}$ of the standard normal distribution, given by

$$h_{\mathrm{NORM}}[\alpha] = h(\mathbf{S}) - n^{-1/2}\hat{\sigma}z_{1-\alpha}. \tag{1}$$

By using the Cornish-Fisher expansion for quantiles of T, we have

$$h_{\mathrm{exact}}[\alpha] = h(\mathbf{S}) - n^{-1/2}\hat{\sigma}z_{1-\alpha} - n^{-1}\hat{\sigma}\left\{(m_1 - \tfrac{1}{6}m_3) - (\tfrac{1}{2}m_{11} - \tfrac{1}{6}m_3)z_{1-\alpha}^2\right\} + O(n^{-3/2}), \tag{2}$$

where m_1, m_3, and m_{11} are defined by

$$\begin{aligned} \mathrm{E}(Y) &= n^{-1/2}\sigma m_1 + O(n^{-3/2}), \\ \mathrm{E}\{Y - \mathrm{E}(Y)\}^3 &= n^{-1/2}\sigma^3 m_3 + O(n^{-3/2}), \\ n\mathrm{E}\left[\frac{\{h(\mathbf{S}) - h(\mathbf{\Omega})\}(\hat{\sigma}^2 - \sigma^2)}{\sigma^3}\right] &= m_{11} + O(n^{-1}), \end{aligned}$$

respectively. By replacing m_1, m_3, and m_{11} in (2) with their estimates, we have an approximate α-level endpoint

$$h_{\mathrm{CF}}[\alpha] = h(\mathbf{S}) - n^{-1/2}\hat{\sigma}z_{1-\alpha} - n^{-1}\hat{\sigma}\left\{(\hat{m}_1 - \tfrac{1}{6}\hat{m}_3) - (\tfrac{1}{2}\hat{m}_{11} - \tfrac{1}{6}\hat{m}_3)z_{1-\alpha}^2\right\}. \tag{3}$$

The bootstrap-t method estimates $K(x)$ by $\hat{K}(x)$, the distribution function of the bootstrap version of the statistic T^* and its α-level endpoint is given by

$$h_{\mathrm{BST}}[\alpha] = h(\mathbf{S}) - n^{-1/2}\hat{\sigma}\hat{K}^{-1}(1-\alpha). \tag{4}$$

In practice, the percentile point $\hat{K}^{-1}(\alpha)$ is approximated by a simulation.

Another bootstrap approach is the BC_a interval and its endpoint is defined by

$$h_{\mathrm{BC}_a}[\alpha] = \hat{G}^{-1}\left\{\Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right)\right\}, \tag{5}$$

where $\hat{G}$ is the cumulative distribution function of the bootstrap value $h(\mathbf{S}^*)$ and z_0 and a are, respectively, the bias-correction and acceleration parameters. The BC_a interval is transformation-respecting and range-preserving.

The ABC endpoint is an analytical approximation to the BC_a endpoint and is usually calculated by an evaluation of a statistic. By equating $h(\mathbf{S} + n^{-1/2}\hat{\sigma}\mathbf{E})$ with the expansion of the BC_a endpoint for an appropriate choice of $\mathbf{E}$, we have

$$h_{\mathrm{ABC}}[\alpha] = h(\mathbf{S} + 2n^{-1/2}\ell_\alpha\hat{\sigma}^{-1}\mathbf{S}\hat{\mathbf{H}}^{(1)}\mathbf{S}), \tag{6}$$

where $\ell_\alpha = (\hat{z}_0 + z_\alpha)/\{1 - \hat{a}(\hat{z}_0 + z_\alpha)\}^2$, $\hat{\mathbf{H}}^{(1)} = \left(\frac{\partial h}{\partial s_{ab}}\Big|_{\mathbf{S}=\hat{\mathbf{\Omega}}}\right)$, and $\hat{z}_0$ and $\hat{a}$ are the estimators of z_0 and a, respectively. The ABC interval is transformation-respecting.

An approximate endpoint $h[\alpha]$ is called *first-order accurate* if $\mathrm{Prob}(h(\mathbf{\Omega}) \le h[\alpha]) = \alpha + O(n^{-1/2})$, and *second-order accurate* if $\mathrm{Prob}(h(\mathbf{\Omega}) \le h[\alpha]) = \alpha + O(n^{-1})$. The usual endpoint (1) is first-order accurate whereas the endpoints (3) to (6) are second-order accurate.

Let $\mathbf{\Xi} = \mathbf{\Phi} \odot \mathbf{\Phi}$, where $\mathbf{\Phi} = \mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{\Psi}^{-1}$. We assume that $\mathbf{\Xi}$ is positive definite and define a diagonal matrix $\mathbf{\Theta}_i$ whose diagonal elements are the i-th row (column) of $\mathbf{\Xi}^{-1} = (\xi^{ij})$. By noting the relation $s_{ii} = \hat{\eta}_i + \hat{\psi}_i$, it can be shown that, for the case of $\hat{\eta}_i$, $\sigma^2 = 2\xi^{ii} + 2\omega_{ii}^2 - 4\psi_i^2$. It can be also shown that the quantities m_1, m_3, and m_{11} are given by

$$\begin{aligned}
\sigma m_1 &= \psi_i \mathrm{tr}\,(\mathbf{I}_m + \mathbf{\Gamma}^{-1}) + 2\mathrm{tr}\,\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}(\mathbf{B} \odot \mathbf{\Xi}^{-1}),\\
\sigma^3 m_3 &= -8\mathrm{tr}\,(\mathbf{\Phi}\mathbf{\Theta}_i)^3 + 24\mathrm{tr}\,\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{B}\mathbf{\Theta}_i - 24\psi_i^2[\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}]_{ii} + 8\omega_{ii}^3 - 24\psi_i^3 + 48\psi_i\xi^{ii},\\
\sigma^3 m_{11} &= -8\mathrm{tr}\,(\mathbf{\Phi}\mathbf{\Theta}_i)^3 + 16\mathrm{tr}\,\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{B}\mathbf{\Theta}_i - 8\psi_i^2[\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}\mathbf{\Theta}_i\mathbf{\Phi}]_{ii} + 8\omega_{ii}^3 - 8\omega_{ii}\psi_i^2 - 16\psi_i^3 + 32\psi_i\xi^{ii},
\end{aligned}$$

where $\mathbf{B} = (\beta_{ij}) = \mathbf{\Psi}^{-1}\mathbf{\Lambda}\mathbf{\Gamma}^{-2}\mathbf{\Lambda}'\mathbf{\Psi}^{-1}$.

Various types of intervals described above were compared by application to real data and a Monte Carlo experiment. It was found that all the second-order accurate intervals achieved better balance in the left and right than did the usual intervals. It was also found that the Cornish-Fisher expansion based intervals were too short, whereas the bootstrap-t intervals tended to be long. The BC_a and ABC methods are recommended for constructing confidence intervals for communalities in factor analysis.

References

Anderson, T. W. and Rubin, H. (1956). *Proc. 3rd Berkeley Symp.*, **5**, (pp. 111–150).

DiCiccio, T. and Efron, B. (1992). *Biometrika*, **79**, 231–245.

Efron, B. (1987). *J. Amer. Statist. Assoc.*, **82**, 171–200.

Hall, P. (1988). *Ann. Statist.*, **16**, 927–953.

Ichikawa, M. and Konishi, S. (2002). *J. Mult. Anal.*, **81**, 47–66.

Factor Analysis by Using Higher Order Moments

Leiden University Ab Mooijaart

In Factor Analysis (FA) mostly covariances or correlations are fitted only. These coefficients are called second order mixed moments because they deal with cross-products of two variables. If the variables are normally distributed, these second order moments contain, besides the first order moments, all the necessary information. However, if the observed variables are not normally distributed higher order moments, like cross-products of three variables, contain additional information. For instance, the skewnesses of the variables contain also information about the underlying structure. In this paper we extend the common formulation of FA in terms of second order moments to higher order moments.

Three different kinds of factor analysis are: linear FA (LFA), polynomial FA (PFA) and Interaction FA (INFA). All these types of FA can be used for modeling nonnormally distributed variables. A general formulation of Structural Equation Models (of which factor analysis is a special case) with linear, quadratic and interaction effects is:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}_0\boldsymbol{\eta} + \boldsymbol{\Gamma}_1\boldsymbol{\xi} + \boldsymbol{\Gamma}_2(\boldsymbol{\xi}\otimes\boldsymbol{\xi}) + \boldsymbol{\zeta},$$

where $\boldsymbol{\Gamma}_2$ is here the main matrix of interest. See also Klein and Moosbrugger (2000). By our formulation main effects, quadratic effects and interaction effects are possible in one model. The measurement model is defined as $\mathbf{x} = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$, $\mathbf{y} = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}$.

By defining $\mathbf{z}_\mathbf{x} \equiv \mathbf{x} - E[\mathbf{x}]$, $\mathbf{z}_\mathbf{y} \equiv \mathbf{y} - E[\mathbf{y}]$ and $\mathbf{z}_{\xi\otimes\xi} \equiv (\boldsymbol{\xi}\otimes\boldsymbol{\xi}) - E[\boldsymbol{\xi}\otimes\boldsymbol{\xi}]$ we can write

$$\begin{pmatrix}\mathbf{z}_\mathbf{x}\\ \mathbf{z}_\mathbf{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Lambda}_\mathbf{x} & \mathbf{0}\\ \boldsymbol{\Lambda}_\mathbf{y}\mathbf{B}^{-1}\boldsymbol{\Gamma}_1 & \boldsymbol{\Lambda}_\mathbf{y}\mathbf{B}^{-1}\boldsymbol{\Gamma}_2\end{pmatrix}\begin{pmatrix}\mathbf{z}_\xi\\ \mathbf{z}_{\xi\otimes\xi}\end{pmatrix} + \begin{pmatrix}\mathbf{I}_p & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_q & \boldsymbol{\Lambda}_\mathbf{y}\mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{\delta}\\ \boldsymbol{\varepsilon}\\ \boldsymbol{\zeta}\end{pmatrix},$$

where $\mathbf{B} = \mathbf{I} - \mathbf{B}_0$. This formulation can be written in terms of super vectors and super matrices as

$$\mathbf{z}_O = \boldsymbol{\Lambda}\mathbf{z}_L + \mathbf{K}\mathbf{e},$$

where the subscripts "O" and "L" correspond to "observed" and "latent" vectors.
For the first, second and third order moments we have

$$E[\mathbf{z}_O] = \boldsymbol{\Lambda}E[\mathbf{z}_L]$$
$$E[\mathbf{z}_O\otimes\mathbf{z}_O] = (\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda})E[\mathbf{z}_L\otimes\mathbf{z}_L] + (\mathbf{K}\otimes\mathbf{K})E[\mathbf{e}\otimes\mathbf{e}]$$
$$E[\mathbf{z}_O\otimes\mathbf{z}_O\otimes\mathbf{z}_O] = (\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda})E[\mathbf{z}_L\otimes\mathbf{z}_L\otimes\mathbf{z}_L] + (\mathbf{K}\otimes\mathbf{K}\otimes\mathbf{K})E[\mathbf{e}\otimes\mathbf{e}\otimes\mathbf{e}].$$

A generalized least squares function can be defined as:

$$f(\boldsymbol{\theta}) = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))'\mathbf{W}(\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})),$$

in which $\mathbf{W}$ is some weight matrix. Minimizing $f(\boldsymbol{\theta})$ with respect to the latent parameters $\boldsymbol{\theta}$ gives generalized least squares estimates of $\boldsymbol{\theta}$. For more discussion about this estimation procedure see Browne (1984), Mooijaart (1985).

An example is the interacton model. See for some references Busemeyer and Jones (1983), Kenny and Judd (1984), Ping (1996) Bollen and Paxton (1998), Jackart and Wang (1995), Jöreskog and Yang (1996).

A Monte Carlo study was carried out for the three types of factor models given above with 500 replications and different sample sizes. The results are:
1: Fitting all the third order moments, in addition to all second order moments, gives large biases of the estimates.
2: Fitting all the third order moments gives the smallest standard errors of the estimates (which can be proven in general). But see remark 1.
3: The distribution of the goodness-of-fit statistic X^2 is chi square distributed according to the large sample theory in cases where the number of fitted moments is restricted.

General conclusion: A careful selection of moments to be fitted is necessary.

References

Bollen, K.A. & Paxton, P. (1998). Two-stage least squares estimation of interaction effects. In: R.E. Schumacker & G.A. Marcoulides (1998). *Interaction and nonlinear effects in structural equation modeling,* (p. 125-151). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

Busemeyer, J.R. & Jones, L.E. (1983). Analysis of multiplicative combination rules when causal variables are measured with error. *Psychological Bulletin, 93,* 549-562.

Klein, A. & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with LMS method. *Psychometrika, 65*, 457-474.

Mooijaart, A. (1985). Factor analysis for non-normal variables. *Psychometrika, 50*, 323-342.

Jaccard, J. & Wan, C.K. (1995). Measurement error on the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin, 117,* 348-357.

Jöreskog, K.G. & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G.A. Marcoulides & R.E. Schumacker (Eds.), Advanced structural equation modeling: Issues and techniques (pp. 57-88). Mahwak, NJ: Erlbaum.

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

Ping, R.A. (1996). Latent variable regression: A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research, 31*, 95-120.

Higher-order estimation error in factor analysis and structural equation modeling under nonnormality

Haruhiko Ogasawara
Otaru University of Commerce

1. Abstract

A general formula of the higher-order asymptotic standard error is derived for the estimators of the parameters in structural equation modeling. The formula covers nonnormally distributed data as well as normally distributed ones. For this derivation, the third- and fourth-order asymptotic central moments of sample variances and covariances are provided for nonnormally and normally distributed cases. The formula requires the partial derivatives of an estimator up to the third order with respect to sample variances and covariances, which are shown for the case of the Wishart maximum likelihood estimator. To see the accuracy of the formula, simulations are performed using the exploratory/confirmatory factor analysis models. It is numerically shown that some of the added contributions of the higher-order asymptotic standard errors are substantial with small to modest sample sizes.

2. Main results

Under usual notation, we have the following results.

Lemma 1. The asymptotic third-order central moment of s_{ab}, s_{cd} and s_{ef} with the assumption of its existence is

$$\begin{aligned}
&\mathrm{E}\{(s_{ab}-\sigma_{ab})(s_{cd}-\sigma_{cd})(s_{ef}-\sigma_{ef})\} \\
&= n^{-2}(\sigma_{abcdef}-\sigma_{abcd}\sigma_{ef}-\sigma_{abef}\sigma_{cd}-\sigma_{cdef}\sigma_{ab} \\
&\quad -\sigma_{acd}\sigma_{bef}-\sigma_{bcd}\sigma_{aef}-\sigma_{abc}\sigma_{def}-\sigma_{abd}\sigma_{cef} \\
&\quad -\sigma_{abe}\sigma_{cdf}-\sigma_{abf}\sigma_{cde}+2\sigma_{ab}\sigma_{cd}\sigma_{ef})+O(n^{-3}), \\
&(p \geq a \geq b \geq 1;\ p \geq c \geq d \geq 1;\ p \geq e \geq f \geq 1),
\end{aligned}$$

where σ_{abcdef}, σ_{abcd} and σ_{abc} are the multivariate central moments of the variables corresponding to the subscripts, respectively.

Let $\mathrm{K}=\{\kappa_{abcd}\}$ be the fourth-order cumulant matrix of observable variables. Then, we have

Theorem 1. The bias-corrected third-order (added) asymptotic variance of $\hat{\theta}_i$ with the assumption of the existence of the associated moments of observed variables and three times differentiability of $\hat{\theta}_i$ with respect to sample variances

and covariances is given as

$$\Delta\mathrm{avar}(\hat{\theta}_i;n^{-2}) = n^{-2}\left\{-\frac{\partial\theta_i}{\partial\boldsymbol{\sigma}'}\mathbf{K}\frac{\partial\theta_i}{\partial\boldsymbol{\sigma}}+\sum_{a\geq b}\sum_{c\geq d}\sum_{e\geq f}\frac{\partial\theta_i}{\partial\sigma_{ab}}\frac{\partial^2\theta_i}{\partial\sigma_{cd}\partial\sigma_{ef}}\right.$$

$$\times(\sigma_{abcdef}-\sigma_{ab}\sigma_{cdef}-\sigma_{cd}\sigma_{abef}-\sigma_{ef}\sigma_{abcd}$$

$$-\sigma_{acd}\sigma_{bef}-\sigma_{bcd}\sigma_{aef}-\sigma_{abc}\sigma_{def}-\sigma_{abd}\sigma_{cef}$$

$$-\sigma_{abe}\sigma_{cdf}-\sigma_{abf}\sigma_{cde}+2\sigma_{ab}\sigma_{cd}\sigma_{ef})$$

$$+\sum_{a\geq b}\sum_{c\geq d}\sum_{e\geq f}\sum_{g\geq h}\left(\frac{1}{2}\frac{\partial^2\theta_i}{\partial\sigma_{ab}\partial\sigma_{ef}}\frac{\partial^2\theta_i}{\partial\sigma_{cd}\partial\sigma_{gh}}+\frac{\partial\theta_i}{\partial\sigma_{gh}}\frac{\partial^3\theta_i}{\partial\sigma_{ab}\partial\sigma_{cd}\partial\sigma_{ef}}\right)$$

$$\left.\times(\sigma_{abcd}-\sigma_{ab}\sigma_{cd})(\sigma_{efgh}-\sigma_{ef}\sigma_{gh})\right\}$$

$$=n^{-2}\left\{-\frac{\partial\theta_i}{\partial\boldsymbol{\sigma}'}\mathbf{K}\frac{\partial\theta_i}{\partial\boldsymbol{\sigma}}+\sum_{a\geq b}\sum_{c\geq d}\sum_{e\geq f}\frac{\partial\theta_i}{\partial\sigma_{ab}}\frac{\partial^2\theta_i}{\partial\sigma_{cd}\partial\sigma_{ef}}\right.$$

$$\times(\sigma_{abcdef}-\sigma_{ab}\sigma_{cdef}-2\sigma_{cd}\sigma_{abef}-2\sigma_{acd}\sigma_{bef}$$

$$\left.-2\sigma_{abc}\sigma_{def}-2\sigma_{abd}\sigma_{cef}+2\sigma_{ab}\sigma_{cd}\sigma_{ef})\right\}$$

$$+\mathrm{tr}\left\{\frac{1}{2}\mathrm{acov}\left(\frac{\partial\hat{\theta}_i}{\partial\mathbf{s}}\right)\mathrm{acov}(\mathbf{s})+\mathrm{acov}\left(\hat{\theta}_i,\frac{\partial^2\hat{\theta}_i}{\partial\mathbf{s}\partial\mathbf{s}'}\right)\mathrm{acov}(\mathbf{s})\right\}$$

$(i=1,\ldots,q)$.

Corollary 1. The bias-corrected third-order (added) asymptotic variance of $\hat{\theta}_i$ with the assumption of multivariate normality in addition to that in Theorem 1 is

$$\Delta\mathrm{avar}(\hat{\theta}_i;n^{-2}) = n^{-2}\left[\sum_{a\geq b}\sum_{c\geq d}\sum_{e\geq f}\frac{\partial\theta_i}{\partial\sigma_{ab}}\frac{\partial^2\theta_i}{\partial\sigma_{cd}\partial\sigma_{ef}}\right.$$

$$\times\{(\sigma_{fa}\sigma_{bc}+\sigma_{fb}\sigma_{ac})\sigma_{de}+(\sigma_{fa}\sigma_{bd}+\sigma_{fb}\sigma_{ad})\sigma_{ce}$$

$$+(\sigma_{ea}\sigma_{bc}+\sigma_{eb}\sigma_{ac})\sigma_{df}+(\sigma_{ea}\sigma_{bd}+\sigma_{eb}\sigma_{ad})\sigma_{cf}\}$$

$$+\sum_{a\geq b}\sum_{c\geq d}\sum_{e\geq f}\sum_{g\geq h}\left(\frac{1}{2}\frac{\partial^2\theta_i}{\partial\sigma_{ab}\partial\sigma_{ef}}\frac{\partial^2\theta_i}{\partial\sigma_{cd}\partial\sigma_{gh}}+\frac{\partial\theta_i}{\partial\sigma_{gh}}\frac{\partial^3\theta_i}{\partial\sigma_{ab}\partial\sigma_{cd}\partial\sigma_{ef}}\right)$$

$$\left.\times(\sigma_{ac}\sigma_{bd}+\sigma_{ad}\sigma_{bc})(\sigma_{eg}\sigma_{fh}+\sigma_{eh}\sigma_{fg})\right],$$

$(i=1,\ldots,q)$.

Key words: Mean square errors, asymptotic standard errors, asymptotic biases, higher-order analysis, structural equation modeling, nonnormal distributions, asymptotic robustness.

Factor analysis, reliability and unidimensionality

University of Groningen, Jos M.F. ten Berge

Factor analysis and test theory go back a long way. Not only do researchers want reliable and unidimensional tests and scales. Factor analysis and test theory also share the same mathematical framework, revolving around a *set of feasible tautologies*. That is, for a fixed observed covariance matrix $\mathbf{\Sigma}$, there exists a set of diagonal matrices $\mathbf{D}$ such that both $\mathbf{\Sigma}-\mathbf{D}$ and $\mathbf{D}$ psd. In the context of factor analysis, the diagonal matrices are solutions for $\mathbf{\Psi}$, the nonnegative diagonal matrix of unique variances, entailing a so-called reduced covariance matrix $\mathbf{\Sigma}-\mathbf{\Psi}$. Each point in the feasible set is one possible factor analysis solution. In the context of test theory, the diagonal matrices are solutions for $\mathbf{\Sigma}_E$, the nonnegative diagonal matrix of error variances, entailing a true score covariance matrix $\mathbf{\Sigma}-\mathbf{\Sigma}_E$: Each point in the set now refers to one possible solution for $\mathbf{\Sigma}_E$. This paper gives an overview of those points in the feasible set having direct psychometric interpretations.

The ideal of factor analysis is to find a solution for $\mathbf{\Psi}$ such that we can decompose

$$\mathbf{\Sigma} = (\mathbf{\Sigma}-\mathbf{\Psi}) + \mathbf{\Psi} = \mathbf{FF}' + \mathbf{\Psi}$$

with $\mathbf{\Sigma}-\mathbf{\Psi}$ of low rank, which can be factored as $\mathbf{\Sigma}-\mathbf{\Psi}=\mathbf{FF}'$, with $\mathbf{F}$ $(m\times r)$ a matrix of loadings, for some small number of common factors r. A key issue has been to what extent can communalities (diagonal elements of $\mathbf{\Sigma}-\mathbf{\Psi}$) reduce the rank. The answer has a long history, revolving around Ledermann's bound (LB). Ledermann (1937) proposed the function $\varphi(m) = [2m+1-\sqrt{(8m+1)}]/2$ as an *upper bound* to the number of common factors. Counterexamples were given by Wilson and Worcester (1939) and by Guttman (1958). Shapiro (1982) proved that LB is almost surely a *lower bound* to number of factors, and that the minimum reduced rank is unstable below LB. Shapiro (1985) and Bekker and Ten Berge (1997) also settled the identification issue generically: The unique variances are almost surely identified when the rank is below the LB. It may be noted that the LB also plays a key role in the so-called inverse PCA problem. From an $m\times r$ PCA loading matrix, the entire correlation matrix can be retrieved when r is at or above LB. (Ten Berge & Kiers, 1999).

In test theory, each point of the feasible set is characterized by its implied reliability. The most interesting points are the origin (implying that perfect reliability is a possibility), and the maximum sum of coordinates point, entailing the greatest lower bound to reliability (Bentler, 1972; Jackson & Agunwamba, 1977; Bentler & Woodward, 1980; Ten Berge, Snijders & Zegers, 1981).

In factor analysis, each point of the feasible set is characterized by its implied reduced eigenvalues and reduced rank. Although low reduced rank (below LB) occurs with probability zero, some points in the set are closer to the ideal of low rank than others, in the sense that their eigenvalues beyond the first r are small. Minimum Rank Factor Analysis (MRFA, Ten Berge & Kiers, 1991) identifies the most r-

dimensional point, r=1,2,..., in the set, by minimizing the sum of the last $m-r$ reduced eigenvalues. A quadratic variant (MRFA-Q) minimizes the sum of squares of the last $m-r$ reduced eigenvalues. That is the same objective function as in Least Squares Factor Analysis, except that MRFA-Q is constrained to stay within the feasible set. The most important practical difference between methods that are and those that are not inside the feasible set is that the former allow evaluating a percentage of explained common variance, analogous to the percentage of explained observed variance in PCA. The MRFA solution for r=1 is the most unidimensional (congeneric) point in feasible set. A natural measure of unidimensionality of a test is the percentage of common variance explained by that solution, see Ten Berge & Sočan (in press). The least unidimensional point in the feasible set is the origin, which corresponds to PCA.

References

Bekker, P.A. & Ten Berge, J.M.F. (1997). Generic global identification in factor analysis. **Linear Algebra and its Applications**, 264, 255-264.

Bentler P.M. (1972). A lower-bound method for the dimension-free measurement of reliability. **Social Science Research**, 1, 343-357.

Bentler, P.M. & Woodward, J.A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. **Psychometrika**, 45, 249-267.

Guttman, L. (1958). To what extent can communalities reduce rank? **Psychometrika**, 23, 297-308.

Jackson, P.H. & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds. **Psychometrika**, 42, 567-578.

Ledermann, W. (1937). On the rank of reduced correlation matrices in multiple factor analysis. **Psychometrika**, 2, 85-93.

Shapiro, A. (1982). Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. **Psychometrika**, 47, 187-199.

Shapiro, A. (1985). Identifiability of factor analysis: Some results and open problems. **Linear Algebra and its Applications**, 70, 1-7.

Ten Berge, J.M.F. & Kiers, H.A.L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. **Psychometrika**, 56, 309-315.

Ten Berge, J.M.F. & Kiers, H.A.L. (1999). Retrieving the correlation matrix from a truncated PCA solution: The inverse principal component problem. **Psychometrika**, 64, 317-324.

Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. **Psychometrika**, 46, 357-366.

Ten Berge, J.M.F. & Sočan, G. The greatest lower bound to reliability of a test and the hypothesis of unidimensionality. **Psychometrika**, *in press*.

Wilson, E.B. & Worcester, J. (1939). The resolution of six tests into three general factors. **Proceedings National Academy of Sciences**, 25, 73-77.

An Uncompensated Factor Analysis Model

The National Center for University Entrance Examinations Kojiro Shojima[1]

In factor analysis, the jth variable can be described as $X_j = \mu_j + \boldsymbol{\lambda}_j' \boldsymbol{f} + E_j$, where $\boldsymbol{f} = [F_1 \cdots F_M]'$ is a M-dimensional random vector, $\boldsymbol{\lambda}_j = [\lambda_{j1} \cdots \lambda_{jM}]'$ is a vector of factor pattern for item j, μ_j is an intercept, and E_j is a random error assumed to follow a normal distribution with mean 0 and variance $1/\psi_j$; i.e., $E_j \sim N(0, 1/\psi_j)$.

In the factor analysis model, the conditional distribution of X_j when $\boldsymbol{f} = \boldsymbol{\theta}$, $X_j|\boldsymbol{\theta}$, follows a normal distribution with mean $\mu_j + \boldsymbol{\lambda}_j'\boldsymbol{\theta}$ and variance $1/\psi_j$. That is,

$$h(X_j = x_j|\boldsymbol{\theta}) = \left(\frac{\psi_j}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\psi_j}{2}(x_j - \mu_j - \boldsymbol{\lambda}_j'\boldsymbol{\theta})^2\right) \tag{1}$$

$$= \frac{\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\alpha}_j'(\boldsymbol{\theta} - \boldsymbol{\beta}_j - \boldsymbol{\gamma}_j x_j)\right)^2\right) \tag{2}$$

$$= \frac{\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\beta}_j - \boldsymbol{\gamma}_j x_j)'\boldsymbol{A}_j(\boldsymbol{\theta} - \boldsymbol{\beta}_j - \boldsymbol{\gamma}_j x_j)\right). \tag{3}$$

(3) becomes very similar to the multidimensional continuous response model (Samejima, 1974, Psychometrika), where $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are M-dimensional vectors. Furthermore, using the following transformations,

$$\boldsymbol{\lambda}_j = (\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j)^{-1}\boldsymbol{\alpha}_j, \qquad \mu_j = -(\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j)^{-1}\boldsymbol{\alpha}_j'\boldsymbol{\beta}_j \qquad \text{and} \quad \psi_j = (\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j)^2,$$

(2) reduces into (1). Note that $\boldsymbol{A}_j = \boldsymbol{\alpha}_j\boldsymbol{\alpha}_j'$ $(M \times M)$ in (3), and the mn element of $\boldsymbol{A}_j$ is $\alpha_{jm}\alpha_{jn}$. Then, the expected score of X_j conditioned on $\boldsymbol{\theta}$, $E[X_j|\boldsymbol{\theta}]$ is represented as Fig. 1.

1. An Uncompensated Factor Analysis Model

Assuming that every nondiagonal element in $\boldsymbol{A}_j$ is set to be 0 in (3), we can obtain

$$g(x_j|\boldsymbol{\theta}) = \frac{\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}\{\boldsymbol{\alpha}_j \odot (\boldsymbol{\theta} - \boldsymbol{\beta}_j - \boldsymbol{\gamma}_j x_j)\}'\{\boldsymbol{\alpha}_j \odot (\boldsymbol{\theta} - \boldsymbol{\beta}_j - \boldsymbol{\gamma}_j x_j)\}\right)$$

$$= \frac{\boldsymbol{\alpha}_j'\boldsymbol{\gamma}_j}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}\sum_{m=1}^{M} \alpha_{jm}^2(\theta_m - \beta_{jm} - \gamma_{jm}x_j)^2\right). \tag{4}$$

The expected score of (4), $E[X_j|\theta]$, becomes Fig. 6. By the way, it is obvious from (1) and (4) that the compensated model (ordinary FA model) yields no difference as compared with the uncompensated model if the simple structure is completely realized in the factor pattern.

[1]The author is grateful to Dr. Kano (Osaka University) and Dr. Yanai (The National Center for University Entrance Examination) for their useful comments and suggestions on earlier versions of this paper.

2. A Mixture Model of Compensated and Uncompensated Relationships

If we introduce a dummy variable d_{jmn} $(= d_{\mathsf{jnm}})$ which is 1 if the relationship between the mth and the nth is compensated, and is 0 otherwise. That is,

$$d_{\mathsf{jmn}} = d_{\mathsf{jnm}} = \begin{cases} 1 & \text{when the relationship is compensated} \\ 0 & \text{when the relationship is uncompensated} \end{cases}. \tag{5}$$

Then preparing a symmetric matrix $\boldsymbol{D}_{\mathsf{j}} = \{d_{\mathsf{jmn}}\} = \{d_{\mathsf{jnm}}\}$, we can express various situations of compensated or uncompensated relationships among factors, using $\boldsymbol{D}_{\mathsf{j}} \odot \boldsymbol{A}_{\mathsf{j}} = \{d_{\mathsf{jmn}}\alpha_{\mathsf{jm}}\alpha_{\mathsf{jn}}\}$ instead of $\boldsymbol{A}_{\mathsf{j}}$.

3. A Model That Continuously Considers The Degree of Compensation

To take the degree of compensation (uncompensation) into account, each element of $\boldsymbol{D}_{\mathsf{j}}$ can be regarded as continuum with its region [0,1]. Continuously operating the element of $\boldsymbol{D}_{\mathsf{j}}$, we can consider medium models between the compensated and the uncompensated (see Figures 2–5). It is certain that the model with d=1.0 and d=0.0 becomes the compensated model and the uncompensated model, respectively.

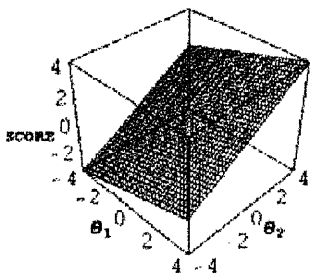

Fig. 1: FA (d=1.0)

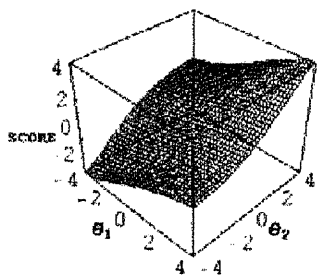

Fig. 2: d=0.8

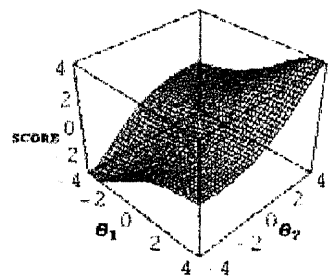

Fig. 3: d=0.6

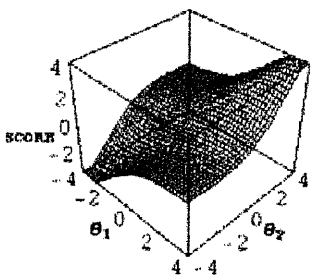

Fig. 4: d=0.4

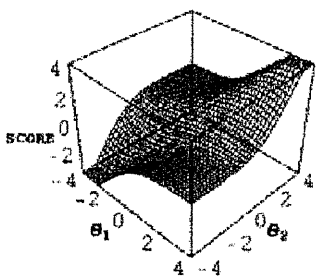

Fig. 5: d=0.2

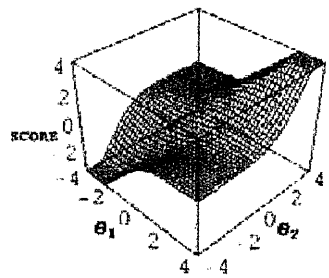

Fig. 6: Uncomp. FA (d=0.0)

Mean Comparison: Manifest Variable versus Latent Variable

Ke-Hai Yuan (University of Notre Dame) and Peter M. Bentler (UCLA)

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ be a random sample from a p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume that $\mathbf{y}_i$ is generated by the factor model

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{1a}$$

where $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$ are independent with $E(\boldsymbol{\xi}_i) = \boldsymbol{\tau}$, $\mathrm{Cov}(\boldsymbol{\xi}_i) = \boldsymbol{\Phi}$, and $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$ is a diagonal matrix. So the mean vector and covariance matrix of $\mathbf{y}_i$ are

$$\boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\tau} \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{1b}$$

In classical multivariate analysis there is a great interest in testing the mean of the manifest variables (MV), the typical hypothesis is $H_{01} : \boldsymbol{\mu} = \mathbf{0}$. The well-known Hotelling T^2 is designed for such a purpose, which is given by $T^2 = n\bar{\mathbf{y}}'\mathbf{S}^{-1}\bar{\mathbf{y}}$, where $\bar{\mathbf{y}}$ is the sample mean and $\mathbf{S}$ is the sample covariance matrix.

In the setup of model (1), testing $\boldsymbol{\mu} = \mathbf{0}$ can also be accomplished by testing the means of the latent vector $\boldsymbol{\xi}$. For example, when $\boldsymbol{\nu} = \mathbf{0}$ in (1b), $\boldsymbol{\mu} = \mathbf{0}$ is equivalent to $\boldsymbol{\tau} = \mathbf{0}$. More generally, suppose the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are further parameterized as $\mathbf{m}(\boldsymbol{\theta})$ and $\mathbf{C}(\boldsymbol{\theta})$. When correctly specified, $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta}_0)$, $\boldsymbol{\Sigma} = \mathbf{C}(\boldsymbol{\theta}_0)$ and we call $\boldsymbol{\theta}_0$ the population value of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1$ is a $q_1 \times 1$ vector containing all the parameters that appear only in $\mathbf{m}(\boldsymbol{\theta})$, and $\boldsymbol{\theta}_2$ contains the q_2 remaining parameters. When $\boldsymbol{\mu} = \dot{\mathbf{m}}_1\boldsymbol{\theta}_{01}$, which generally holds when $\mathbf{m}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}_1$, then testing for H_{01} is equivalent to testing $H_{02} : \boldsymbol{\theta}_{01} = \mathbf{0}$. Two commonly used test statistics for testing H_{02} are the Wald statistic T_W and the likelihood ratio statistic T_{LR}. Their distributions are described by asymptotics and they are asymptotically equivalent under idealized conditions (Engle, 1984). An interesting question is which statistic provides the most powerful test for $\boldsymbol{\mu} = \mathbf{0}$. We study the different test statistics asymptotically and empirically.

Under a sequence of local alternatives, we have

$$T^2 \xrightarrow{\mathcal{L}} \chi_p^2(\delta_1) \ \text{ with } \ \delta_1 = n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}; \tag{2}$$

and

$$T_W \xrightarrow{\mathcal{L}} \chi_{q_1}^2(\delta_2), \tag{3}$$

where δ_2 depends on the conditions discussed below. Because T_{LR} and T_W are asymptotically equivalent (Eagle, 1984), T_{LR} can also be described by (3). When the factor loadings $\boldsymbol{\Lambda}$ in (1) are known, Kano (2001) showed that $\delta_1 = \delta_2 = \delta$. Because $p > q_1$, using $T_{LR} \sim \chi_{q_1}^2(\delta)$ corresponds to a greater statistical power than using $T^2 \sim \chi_p^2(\delta)$. Without assuming $\boldsymbol{\Lambda}$ known and without a proof, Hancock (2001) also concluded that $\delta_2 = \delta_1$. Using rather rigorous derivations, we found that $\delta_1 \geq \delta_2$ and the equality sign holds only when the orthogonal

condition $\dot{\mathbf{m}}_1'\boldsymbol{\Sigma}^{-1}\dot{\mathbf{m}}_2 = \mathbf{0}$ holds, where $\dot{\mathbf{m}}_1 = (\partial\mathbf{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_1'|\boldsymbol{\theta}_0)$ and $\dot{\mathbf{m}}_2 = (\partial\mathbf{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_2'|\boldsymbol{\theta}_0)$. It is obvious that Kano's (2001) assumption satisfies the orthogonal condition. When the factor loading matrix $\boldsymbol{\Lambda}$ is unknown, $\delta_2 - \delta_1 > 0$ in general and their difference increases as $\boldsymbol{\tau}$ departs from the null hypothesis H_{02}. However, $\delta_2 > \delta_1$ does not imply that T_W or T_{LR} are more powerful than T^2 because p can be much bigger than q_1. Their specific powers depend on p, q_1 and the magnitude of $\boldsymbol{\tau}$. Similar conclusion holds in multiple-group mean comparison, that is, the noncentrality parameter (NCP) corresponding to T_W or T_{LR} for testing the latent variable means can be much smaller than that corresponding to Hotelling T^2 for testing the manifest means.

The chi-square distributions in (2) and (3) are justified under a sequence of local alternative hypotheses. In practice, both the alternative hypothesis and the sample size are fixed. It is interesting to know whether (2) and (3) are still valid. Our empirical results indicate that (2) poorly describes the distribution of T^2 and (3) poorly describes the distribution of T_W and T_{LR} unless the null hypothesis is trivially violated. A large sample size does not make the approximation better. The magnitude of the alternative hypothesis plays a much stronger role than the sample size n. We also found that T_{LR} can be stochastically much greater than T_W as the alternative hypothesis moves away from the null hypothesis.

Notice that the statistic T_W or T_{LR} are based on the assumption that the factor model (1) is correctly specified. When model (1) is misspecified, then the NCPs in both T_W and T_{LR} can be smaller or greater than that corresponding to a correctly specified base model. A misspecified covariance structure has a relatively small effect on mean comparison. A misspecified mean structure has a quite significant effect on mean comparison.

When data do not follow a multivariate normal distribution, under a sequence of local alternative hypotheses T_W still asymptotically follows the chi-square distribution as in (3) when $\dot{\mathbf{m}}_1'\boldsymbol{\Sigma}^{-1}\dot{\mathbf{m}}_2 = \mathbf{0}$. However, we can only show that T_{LR} follows a central chi-square distribution under the null hypothesis and $\dot{\mathbf{m}}_1'\boldsymbol{\Sigma}^{-1}\dot{\mathbf{m}}_2 = \mathbf{0}$.

References

Engle, R. (1984). Wald, likelihood ratio and Lagrange multiplier tests in econometrics. In Z. Griliches, &M. D. Intrilligator (Eds.), *Handbook of econometrics, vol II* (pp. 775–826). Amsterdam: North Holland.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373–388.

Kano, Y. (2001). Structural equation modeling for experimental data. In R. Cudeck, S.H.C. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 381–402). Lincolnwood, IL: Scientific Software International.

Independent component analysis and its application to causal analysis

Shohei Shimizu
Osaka University

Aapo Hyvärinen
University of Helsinki

Yutaka Kano
Osaka University

Keywords: Independent component analysis, nonnormality, independence, causal inference, nonexperimental data

Abstract

We study the application of independent component analysis to discovery of a causal ordering between observed variables. Path analysis is a widely-used method for causal analysis. It is of confirmatory nature and can provide statistical tests for assumed causal relations based on comparison of the implied covariance matrix with a sample covariance. Estimated path coefficients are used to evaluate the magnitude of causality. However, it is based on the assumption of normality and only uses the covariance structure, which is why it has several problems.

A very simple illustration of the problem of finding the direction of causality is given by two regression models, called Model 1 and Model 2 here:

$$\text{Model 1:} \quad x_1 = b_{12}x_2 + \xi_1 \tag{1}$$

$$\text{Model 2:} \quad x_2 = b_{21}x_1 + \xi_2, \tag{2}$$

where the explanatory variable is assumed to be uncorrelated with the disturbance ξ_1 or ξ_2. We cannot say anything about which model is better from the two conventional regression analyses based on the two models above in the framework of SEM. Using the SEM terminology, the both models are saturated on the covariance matrix of $[x_1, x_2]$.

Kano and Shimizu (2003) and Shimizu and Kano (2003) showed that use of nonnormality (higher-order moments) of observed variables makes it possible to distinguish between Model 1 and Model 2. In this article, we extend the method to more than two variables and develop a new statistical method for discovery of a causal ordering using nonnormality of observed variables.

In the following, we actually assume that the data follows such a model so that the causal ordering is possible to find. Thus, we assume the following data model:

$$x_{i(j)} = \sum_{k<j} b_{i(j),i(k)} x_{i(k)} + \xi_{i(j)}. \tag{3}$$

We also assume that the disturbance variables $\xi_{i(j)}$ are nonnormal, and mutually independent. This implies that $\xi_{i(j)}$ is independent from $\xi_{i(k)}$ for all $k < j$.

To investigate the causal structure of the x_i, we would like to find the correct ordering $i(j)$. Thus, the problem is finding the permutation of the observed variables that reflects the causal structure of the data. In what follows, we will show how such an ordering can be identified.

Let us normalize the equation (3) so that the disturbance variables ξ_i have unit norm. Denoting

$$w_{i(j),i(j)} = 1/\sqrt{\mathrm{var}(\xi_{i(j)})} \tag{4}$$

$$w_{i(j),i(k)} = -b_{i(j),i(k)}/\sqrt{\mathrm{var}(\xi_{i(j)})} \quad \text{for } k \neq j, \tag{5}$$

the equation (3) can be expressed as:

$$w_{i(j),i(j)}x_{i(j)} = \sum_{k<j} -w_{i(j),i(k)}x_{i(k)} + \xi^*_{i(j)}, \tag{6}$$

where $\xi^*_{i(j)}$ are the disturbance variables standardized to have unit variance.

Let us denote by $\tilde{\boldsymbol{x}}$ the vector where the observed variables are ordered according to $i(j)$. In matrix form, equation (3) can be expressed as

$$\tilde{\boldsymbol{x}} = B\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\xi}} \tag{7}$$

where the matrix B is lower triangular. Using W, this becomes

$$\mathrm{diag}(W)\tilde{\boldsymbol{x}} = -\mathrm{offdiag}(W)\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\xi}}^* \quad \text{or equivalently} \quad W\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{\xi}}^* \tag{8}$$

where W is still lower triangular, for the correct permutation of the observed variables. This corresponds to the correct permutation of the columns of W. From the theory of ICA, we know that this W can be estimated up to a permutation of its rows, using standard ICA methods.

Thus, the model (7) can be estimated by 1) estimating an initial W by ICA; 2) finding a combination of permutations of the rows and the columns of $\tilde{W}$ so that $\tilde{W}$ becomes as close to lower triangular as possible, using the algorithm above; (3) estimating B by $I_n - \mathrm{diag}(\tilde{W})^{-1}\tilde{W}$. The $\tilde{W}$ denotes a correctly permuted version of W.

Now we shall show the models 1 and 2 can be expressed in this framework. In Model 1, the causal order of observed variables is $(i(1), i(2)) = (2, 1)$. Model 1 can be rewritten as:

$$\text{Model 1:} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & b_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \tag{9}$$

$$\Leftrightarrow \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b_{12} & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} \xi_2 \\ \xi_1 \end{bmatrix}. \tag{10}$$

Here $\tilde{\boldsymbol{x}}$ and B are

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ b_{12} & 0 \end{bmatrix}. \tag{11}$$

One can see that the B is lower triangular when the observed variables are ordered according to $i(j)$. Also in Model 2, one can see the lower triangularity of B in the same manner.

References

Kano, Y. and Shimizu, S. (2003). Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of modeling -The 30th Anniversary of the Information Criterion (AIC)-* (Higuchi, T., Iba, Y. & Ishiguro, M., Eds.), 261-270, ISM Report on Research and Education, No.17, The Institute of Statistical Mathematics, Tokyo.

Shimizu, S. and Kano, Y. (2003). Nonnormal structural equation modeling. (submitted).

Why would factors or components be non-normal and square-correlated?

Aapo Hyvärinen
HIIT Basic Research Unit and Dept of Computer Science
University of Helsinki, Finland
aapo.hyvarinen@helsinki.fi
www.cs.helsinki.fi/aapo.hyvarinen/

Independent component analysis (ICA) [7, 4, 6] is a multivariate linear latent variable model closely related to the classic factor analysis model. Its actual estimation methods, on the other hand, are very similar to projection pursuit. The key difference between ICA and ordinary factor analysis is that the latent factors are assumed to be non-Gaussian, i.e. to have non-normal distributions. This seemingly small difference in the model definition leads to huge differences in the estimation procedure and the applications of the model. In fact, non-normality allows us to separate several linearly mixed independent latent signals, and also to *uniquely determine the factor rotation* without traditional factor rotation methods such as varimax.

The basic definition of ICA is quite simple. Let $x_1, x_2, ..., x_n$ denote n observed random variables. These are modelled as a linear transformation of n latent variables $s_1, s_2, ..., s_n$:

$$x_i = \sum_{j=1}^{n} a_{ij} s_j, \text{ for } i = 1, 2, ..., n. \tag{1}$$

The a_{ij} are constant unknown parameters to be estimated, not unlike factor loadings. We make the following assumptions on the latent variables or independent components s_i:

1. The s_i are mutually (statistically) independent.

2. The s_i are non-Gaussian, i.e. have non-normal distributions.

Then the ICA model can be estimated [4].

In addition to the basic ICA model, several researchers have recently considered the case where the components s_j are *not independent*. Many different variants can be considered: the components might be divided into groups so that components inside a group are dependent but components in different groups are independent [3, 6], the dependencies might follow a fixed topographic organization [6], the structure of trees [2, 8], or some general parametric forms [5, 9]. There is an infinite number of different dependencies possible, so any parametric model has to be restricted to modelling some limited family of dependencies. Most authors have considered one particular type of

dependencies: *correlation of squares* because this kind of dependency seems to be quite dominant in many kinds of real data found in signal processing.

Both non-normality and correlation of squares may seem to be quite odd assumptions in the context of classical factor analysis. Non-normality may even seem to be in contradiction with the classical Central Limit Theorem. In this paper, I discuss a simple framework that might explain why many kinds of real data are non-normal. Interestingly, the same framework leads to correlation of squares as the natural form of dependency between the components. I shall also briefly discuss the case of time series, where this framework leads to an interesting kind of temporal correlations similar to what is seen in models based on by autoregressive conditional heteroscedasticity (ARCH). The framework is based on scale mixtures of normal distributions [1, 10].

References

[1] D. Andrews and C. Mallows. Scale mixtures of normal distributions. *J. Royal Statistical Society B*, 36:99–102, 1974.

[2] F. R. Bach and M. I. Jordan. Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

[3] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998.

[4] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[5] J. Hurri and A. Hyvärinen. Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems*, 14(3):527–551, 2003.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[7] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[8] Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14(3):483–500, 2003.

[9] H. Valpola, M. Harva, and J. Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84:267–282, 2004.

[10] M. West. Outlier models and prior distributions in bayesian linear regression. *J. Royal Statistical Society B*, 46:431–439, 1984.

Improvements of LVM Techiniques by Bayesian Hierarchical Modeling

Kazuo Shigemasu,Takahiro Hoshino
The University of Tokyo, Institute of Statistical Mathematics

1.Introduction

We proposed the Bayesian Analysis of Structural Equation Model using Gibbs Sampler (Shigemasu, Hoshino & Ohmori, 2002). In this analysis, latent variables are included as a step of Gibbs Sampler rather they are integrated out. One advantage of this approach is to make the analysis flexible enough to handle the various kinds of data (e.g., dichotomous, polytomous, multiple selection, ordering). In this study, we extend the analysis to apply the Bayesian method to analyze paired comparison data.

2.Model Distribution

The model distribution for the SEM, is given as follows. Notations and distributional assumtions used in this equation is the same as Shigemasu, Ohmori and Hoshino paper.

$$
\begin{aligned}
&p(\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{\Lambda}, \boldsymbol{B}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) \\
&\propto \ |\Omega|^{-\frac{n}{2}}|\Psi|^{-\frac{n}{2}}|\Phi|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{F}_2\boldsymbol{\Phi}^{-1}\boldsymbol{F}_2^t + (\boldsymbol{X} - \boldsymbol{F}\boldsymbol{\Lambda}^t)\boldsymbol{\Psi}^{-1}(\boldsymbol{X} - \boldsymbol{F}\boldsymbol{\Lambda}^t)^t \right. \\
&\qquad \left. + (\boldsymbol{F}_1\boldsymbol{B}^t - \boldsymbol{F}_2\boldsymbol{\Gamma}^t)\boldsymbol{\Omega}^{-1}(\boldsymbol{F}_1\boldsymbol{B}^t - \boldsymbol{F}_2\boldsymbol{\Gamma}^t)^t]\right\}
\end{aligned}
$$

3.Modification of the Model

The model explained above is changed to include a typical situation where paired comparison method was used. One specification is as follows:

$$\boldsymbol{x}^* = \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{e}_1, \quad \boldsymbol{f}_1 = \boldsymbol{\nu} + \Gamma\boldsymbol{f}_2 + \boldsymbol{e}_2.$$

$$\boldsymbol{\mu}_{\boldsymbol{x}^*} = \boldsymbol{\Lambda}_1\boldsymbol{\nu}, \quad \boldsymbol{\Sigma}_{\boldsymbol{x}^*} = \boldsymbol{\Lambda}_1(\boldsymbol{\Gamma}^t\boldsymbol{\Phi}\boldsymbol{\Gamma} + \Omega)\boldsymbol{\Lambda}_1^t + \boldsymbol{\Psi}.$$

The $\boldsymbol{x}^*$, which was assumed to be observed so far, now is latent, and the actual data of paired comparison is given by the relationships among the latent variables $\boldsymbol{x}^*$.

For each paired comparion, the actual data is obtained only for a particular pair of the variables, and the other combinations are supposed to be missing. If we treat the all paired comparison data simultaneously, the inconsistent data cannot be analyzed so that some authors had to introduce additional error terms to analyze inconsistent data.

4. Algorithm for pair comparison data

Note that $\boldsymbol{x}^*$ is the vector of latent random variables. Now the algorithm is changed as follows.
Prepare starting values of parameters.

Step 0 : generate $\boldsymbol{F}_1^{(0)}$, $\boldsymbol{F}_2^{(0)}$

Step 1 : generate $\boldsymbol{B}^{(k)}$ and $\boldsymbol{\Gamma}^{(k)}$ based on other parameters and latent variables.

Step 2 : generate $\boldsymbol{\Lambda}^{(k)}$

Step 3 : generate $\boldsymbol{\nu}^{(k)}$

Step 4 : generate $\boldsymbol{F}_1^{(k)}$, $\boldsymbol{F}_2^{(k)}$

Step 5 : generate $\boldsymbol{X}^{*(k)}$ until the generated variables meet the constraints given by the observed paired comparison data.

Step 6 : Go to Step 1 with $l = l + 1$.

5.Simulation Data

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & \lambda_1(0.6) & \lambda_3(0.3) & 0 & 0 & 0 & 0 \\ 0 & \lambda_2(0.3) & \lambda_4(0.6) & \lambda_5(0.8) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_6(0.6) & \lambda_7(0.8) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^t,$$

$$\boldsymbol{\Gamma} = (\gamma_1(0.4), \gamma_2(0.3), \gamma_3(0.5))^t, \quad \boldsymbol{\Omega} = diag(\omega_1(0.2), \omega_2(0.3), \omega_3(0.4)), \quad \boldsymbol{\nu} = (0, \nu_2(0.5), \nu_3(1))^t.$$

The values in parentheses are true values of parameters. With seven variables, there are ${}_7C_2 = 21$ pairs. Each pair was observed with probability 0.5, so the expected number of pairs each unit made judgement on was $0.5 \times 21 = 10.5$. We deleted units that make no judgement.

In this simulation, the number of units (subjects) generated was 200 or 400.

6.Results of simulation data

The ten replicated data sets were generated from the true model and were analyzed by the proposed method. The means of estimates given by the proposed method are shown in Table 1.

Table 1: Resulting estimates

N=200	True value	Estimate	MSE	N=400	True value	Estimate	MSE
λ_1	0.6	0.6323	.008127	λ_1	0.6	0.6208	.005314
λ_2	0.3	0.2882	.005364	λ_2	0.3	0.3078	.003222
λ_3	0.3	0.2925	.004801	λ_3	0.3	0.2980	.004038
λ_4	0.6	0.5862	.006230	λ_4	0.6	0.6142	.004105
λ_5	0.8	0.8169	.007218	λ_5	0.8	0.7912	.005670
λ_6	0.6	0.5893	.006294	λ_6	0.6	0.5925	.003601
λ_7	0.8	0.7904	.005787	λ_7	0.8	0.7962	.004244
γ_1	0.4	0.3875	.004648	γ_1	0.4	0.4051	.003098
γ_2	0.3	0.3116	.005204	γ_2	0.3	0.3055	.003142
γ_3	0.5	0.4908	.005433	γ_3	0.5	0.5062	.003019
ω_1	0.2	0.1890	.003770	ω_1	0.2	0.2079	.003232
ω_2	0.3	0.3101	.004019	ω_2	0.3	0.2978	.002651
ω_3	0.4	0.3878	.004938	ω_3	0.4	0.3951	.003575
ν_2	0.5	0.5843	.029473	ν_2	0.5	0.5407	.013085
ν_3	1.0	1.1355	.040560	ν_3	1.0	1.1031	.023741

列数の異なる目標行列に近似する因子パタンを得るための斜交プロクラステス回転

村上　隆

（名古屋大学　大学院教育発達科学研究科）

問題　プロクラステス回転は，何らかの因子分析的方法によって得られた $p \times r\,(p > r)$ の因子負荷行列 $\mathbf{A}$ を，仮説的に与えられる同じ大きさのターゲット行列 $\mathbf{B}$ に最小２乗法の意味で近似させるような回転行列 $\mathbf{T}$ を求める一連の方法群であり，通常，次の関数の最小化が目指される。

$$f(\mathbf{T} \mid \mathbf{A}, \mathbf{B}) = \| \mathbf{B} - \mathbf{AT} \|^2 \tag{1}$$

他の因子回転と同様，斜交解と直交解が区別され，制約条件が $\mathbf{T'T} = \mathbf{I} = \mathbf{TT'}$ の場合が直交解である。斜交解は，さらに回転の結果得られる行列による区別があり，回転の結果として得られる負荷行列がパタン（各変数の「因子」への標準回帰係数を要素とする行列）である場合には $\mathrm{diag}\,(\mathbf{T'T})^{-1} = \mathbf{I}$，構造（各変数と「因子」との相関行列）の場合には $\mathrm{diag}\,\mathbf{T'T} = \mathbf{I}$ である。規準 (1) の最小化のためのアルゴリズムは，パタンの回転に関しては大変複雑である（Browne, 1972）が，より簡潔で，負荷行列の列数 r がターゲット行列の列数 q を上回る場合に適用可能な方法を提案する。

アルゴリズムの概要　式 (1)を列ごとに，次のように書く。

$$f(\mathbf{t}_k \mid \mathbf{A}, \mathbf{b}_k) = \| \mathbf{b}_k - \mathbf{At}_k \|^2 \tag{2}$$

$\mathbf{T}$ からその第 k 列を除いた行列を $\mathbf{T}_{-k}$ とすると，$\mathrm{diag}\,(\mathbf{T'T})^{-1} = \mathbf{I}$ から若干の計算により，

$$\mathbf{t}_k'(\mathbf{I} - \mathbf{T}_{-k}(\mathbf{T}_{-k}'\mathbf{T}_{-k})^{-1}\mathbf{T}_{-k}')\mathbf{t}_k = 1 \tag{3}$$

が導かれ，これが $\mathbf{t}_k$ に関する制約条件を与える。ここで，$\mathbf{M}_k = \mathbf{I} - \mathbf{T}_{-k}(\mathbf{T}_{-k}'\mathbf{T}_{-k})^{-1}\mathbf{T}_{-k}'$ と定義すると，この行列はランク $r-q+1$ の冪等行列で，$r-q+1$ 個の値 1 の固有値と $q-1$ 個の 0 固有値を持つ。したがって，$\mathbf{U}_{1k}, \mathbf{U}_{2k}$ をそれぞれ，$r \times (r-q+1), r \times (q-1)$ の正規直交行列で，$\mathbf{U}_{1k}'\mathbf{U}_{2k} = \mathbf{O}_{(r-q-1)\times(q-1)}$ を満たすものとすると，

$$\mathbf{M}_k = [\mathbf{U}_{1k} \quad \mathbf{U}_{2k}] \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{1k}' \\ \mathbf{U}_{2k}' \end{bmatrix}$$

と分解できる。そこで，$\mathbf{P}_{1k} = \mathbf{AU}_{1k}, \mathbf{P}_{2k} = \mathbf{AU}_{2k}, \mathbf{v}_{1k} = \mathbf{U}_{1k}'\mathbf{t}_k, \mathbf{v}_{2k} = \mathbf{U}_{2k}'\mathbf{t}_k$ と定義すると，最小化基準 (2) は，

$$f(\mathbf{v}_{1k}, \mathbf{v}_{2k} \mid \mathbf{P}_{1k}, \mathbf{P}_{2k}, \mathbf{b}_k) = \| \mathbf{b}_k - \mathbf{P}_{1k}\mathbf{v}_{1k} - \mathbf{P}_{2k}\mathbf{v}_{2k} \|^2 \tag{4}$$

となり，制約条件は，$\mathbf{v}_{1k}'\mathbf{v}_{1k} = 1$ となる。以上の reparametrization により，パラメータは制約つきの $\mathbf{v}_{1k}$ と制約なしの $\mathbf{v}_{2k}$ に分けられることになった。

そこで，k ごとに $\mathbf{v}_{2k}$ を所与として，(4) を最小化する $\mathbf{v}_{1k}$ を Kiers & Ten Berge (1992) の majorization method によって求める第１ステップと，$\mathbf{v}_{1k}$ を所与として通常の最小２乗法によって $\mathbf{v}_{2k}$ を求める第２ステップを，すべての k にわたって実行し，それを収束するまで反復する。

適用例　中村他 (2000) による IOS (interpersonal orientation scale, Hill, 1986) のデータに適用する。この尺度は，P, E, A, S の４つの（下位）尺度得点を生み出ことになっているが，日本語版におけるプロマックス回転の結果は，原論文のものとは異なっていた。そのことは，Table 1 左に示した通常の斜交プロクラステス回転の結果においても変わらない。しかし，

$r=6$ の条件では，Table 1 右のように，ほとんど原論文と一致した構造が得られ，主成分間相関についても納得し得るものが得られている。

Table 1 Principal component pattern of IOS in two conditions

		$r=4$					$r=6$				
	Item	I	II	III	IV	Expld.Var	I	II	III	IV	Expld.Var.
	3	**0.51**	**0.42**	0.06	−0.21	0.49	**0.43**	0.33	0.01	−0.06	0.44
	6	**0.51**	**0.41**	−0.23	−0.17	0.45	**0.49**	0.16	0.09	−0.10	0.38
	10	**0.75**	0.01	−0.16	0.05	0.64	**0.71**	0.00	−0.07	0.01	0.49
	11	**0.62**	0.38	−0.27	−0.04	0.61	**0.59**	0.24	−0.07	−0.03	0.52
P	13	**0.51**	−0.08	0.29	0.13	0.38	**0.50**	−0.04	0.11	0.14	0.29
	20	0.40	0.06	**0.45**	−0.06	0.38	**0.52**	−0.17	0.20	0.22	0.34
	24	**0.63**	0.08	0.26	−0.14	0.44	**0.60**	0.06	−0.01	0.04	0.40
	25	**0.52**	0.02	**0.42**	−0.22	0.43	**0.60**	−0.14	0.05	0.13	0.33
	26	**0.66**	−0.26	−0.02	0.13	0.56	**0.69**	−0.27	0.00	0.10	0.37
	1	0.24	**0.69**	−0.09	−0.01	0.55	0.16	**0.61**	−0.05	0.02	0.48
	4	0.27	**0.58**	0.13	0.03	0.51	0.15	**0.58**	0.06	0.03	0.48
E	9	0.16	**0.54**	−0.18	0.27	0.43	0.03	**0.57**	0.07	0.03	0.38
	15	0.08	**0.59**	0.24	0.28	0.53	−0.06	**0.70**	0.09	0.17	0.57
	17	0.08	**0.62**	0.05	0.33	0.51	−0.10	**0.73**	0.17	0.07	0.57
	23	0.21	**0.59**	0.18	0.12	0.53	−0.01	**0.76**	0.11	−0.01	0.60
	5	0.16	0.20	0.33	0.00	0.22	0.17	−0.02	**0.47**	0.05	0.29
	8	−0.22	0.25	0.17	**0.65**	0.36	−0.09	0.14	0.21	**0.47**	0.34
A	16	0.25	0.04	**0.65**	0.26	0.54	0.09	0.17	**0.57**	0.08	0.48
	19	0.36	−0.16	**0.66**	0.14	0.53	0.17	−0.03	**0.69**	−0.06	0.55
	21	0.36	0.08	**0.62**	0.09	0.55	0.14	0.25	**0.51**	−0.04	0.46
	22	−0.11	−0.09	**0.44**	0.37	0.22	−0.19	−0.06	**0.68**	0.07	0.43
	2	−0.06	**0.43**	0.09	0.31	0.25	0.23	0.07	−0.08	**0.53**	0.36
	7	−0.24	0.25	0.28	**0.75**	0.48	−0.02	0.08	0.17	**0.65**	0.52
S	12	−0.09	0.14	0.24	**0.70**	0.43	0.10	0.08	0.01	**0.61**	0.43
	14	0.00	0.19	0.42	**0.56**	0.43	0.08	0.19	0.14	**0.49**	0.40
	18	−0.14	0.01	0.47	**0.67**	0.46	0.05	−0.12	0.32	**0.58**	0.48
Correlations between 'factor scores'											
		I	II	III	IV		I	II	III	IV	
	I	1.00	0.14	−0.03	0.45		1.00	0.51	0.29	0.03	
	II	0.14	1.00	0.27	−0.10		0.51	1.00	0.21	0.21	
	III	−0.03	0.27	1.00	−0.24		0.29	0.21	1.00	0.17	
	IV	0.45	−0.10	−0.24	1.00		0.03	0.21	0.17	1.00	

文　献

Browne, M.W. (1972). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, **25**, 207-212.

Hill, C.A. (1986). Affiliation motivation: People who need people ... but in different ways. *Journal of Personality and Social Psychology*, **22**, 1008-1018.

Kiers, H.A.L. & Ten Berge, J.M.F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika*, **57**, 371-382.

中村陽吉（編著）（2000）対人場面における心理的個人差　測定対象についての分類を中心として　ブレーン出版

Testing Homogeneity of Covariances with Infrequent Missing Data Patterns

Peter M. Bentler (UCLA) Ke-Hai Yuan (Notre Dame) Kevin Kim (Pittsburgh)

When the data for factor analysis or structural equation modeling consists of many subsamples with possibly different characteristics, one may question whether all samples are from a single population with a common mean vector μ and common covariance matrix Σ. It may not make sense to entertain a single structural model in heterogeneous samples, although if only the means are not homogeneous, a single covariance structure would still be appropriate. This problem of homogeneity of means and covariances arises in the context of missing data, where different sets of subjects exhibit different patterns of the presence or absence of scores. When there are few patterns, the multiple group methodology of Allison (1987) and Muthén, Kaplan, and Hollis (1987) can be used to evaluate homogeneity of means and covariances. This methodology becomes impractical to implement when the number of patterns is large. Building on the work of Little (1988) and Tang and Bentler (1998), Kim and Bentler (2002) developed a generalized least squares (GLS) test of homogeneity of means (a minor variant of Little's test), evaluating whether various sample means $\bar{X}_i$ could be considered as samples from a common population with a common mean vector μ; a GLS test of homogeneity of covariance matrices, testing whether various sample covariance matrices S_i could be considered as samples from a common population with a common covariance matrix Σ, and a combined GLS test of homogeneity of means and covariance matrices, that tests both hypotheses simultaneously. In limited sampling studies, they found these tests to perform reasonably well.

With multivariate normal data, these tests are tests of missing completely at random (MCAR). We study the test of homogeneity of covariance matrices under an extreme MCAR condition. The covariance matrices test is based on the null hypothesis that the population covariance matrices $\Sigma_i(\theta)$ for the various patterns of incomplete data with sample size n_i are subsets of a single population covariance matrix $\Sigma(\theta)$ ($\Sigma_i(\theta) \subseteq \Sigma(\theta)$ for all $i = 1, \ldots, m$). It is computed as

$$G_2(\hat{\theta}) = \sum_{i=1}^{m} \frac{k_i}{2} \mathrm{tr}\left(\left[\mathrm{S_i} - \Sigma_\mathrm{i}(\hat{\theta})\right] \Sigma_\mathrm{i}^{-1}(\hat{\theta})\right)^2. \tag{1}$$

This test is based on the p_2^* available sample covariances and the $q_2 = p(p+1)/2$ common covariances estimated, and has $df_2 = (p_2^* - q_2)$ degrees of freedom. Asymptotically, under the null hypothesis, $NG_2(\hat{\theta}) \sim \chi^2_{df_2}$. However, a technical analysis verifies that the homogeneity test statistic requires that n_i go to infinity, and that the relative sample size proportion n_i/N has to converge to a nonzero limit $k_i \to \gamma_i$. Neither of these conditions could be reasonably assumed to hold for any sample missing data pattern that exhibits n_i=1, which we find in practice when a huge number of missing data patterns is allowed in a sample. Even with very small n_i the asymptotic assumptions underlying the tests are not met. As a result, the performance of this test can be disrupted when there are a large number of data patterns

that are exhibited by only one or very few subjects.

Theoretical analysis is supported by simulation under some extreme conditions where up to 90% of cases have some missing data. We find that sample covariance matrices for patterns with n_i=1 are not defined and with small n_i may be very badly estimated. As a result, the GLS test hardly ever rejects the null hypothesis of homogeneity. A minimal correction to the covariance test was proposed and evaluated. This involves removing all patterns that exhibit only one observed case from the computations. As a consequence, the number of observed data covariances is reduced, and the degrees of freedom of the test is substantially reduced as well. This simple revision yielded substantially improved performance. This test variant is now included in the EQS 6 program (Bentler, 2004).

Kim and Bentler (2002) had found their test to perform well, whereas our uncorrected test did badly. A key difference between the current and prior simulations is that in their original study they held the number of missing data patterns to a maximum of 32 patterns. As a result, as sample size increases, more and more information becomes available about all missing data patterns. In the current study, the number of missing data patterns was hugely increased compared to the earlier study, and also, the number of missing data patterns increased with sample size. As compared to 32 patterns, we obtained 257 patterns (on average) at N=1000, an 8-fold increase, and with N=5000, there were almost twice as many, with 493 patterns of missing data (on average). Although performance of the revised test based on deleting cases with n_i=1 is substantially improved, the null hypothesis of homogeneity is now rejected at a slightly higher rate than is desirable. Thus we studied the effect of removing from two up to ten cases per missing data pattern. Overrejection was somewhat reduced, but still remained at a higher level than ideal.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71-103). San Francisco: Jossey Bass.

Bentler, P. M. (2004). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software (www.mvsoft.com).

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika, 67*, 609-624.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Tang, M., & Bentler, P. M. (1998). Theory and method for constrained estimation in structural equation models with incomplete data. *Computational Statistics & Data Analysis, 27*, 257-270.

On the identification problem of the factor analysis model: a review and an application for an estimation of air pollution source profiles and amounts*

SATO, Manabu†

Abstract

The identification problem of the factor analysis model is reviewed. We try to clarify the relation between Anderson-Rubin's and Williams' theorem. Moreover, an application of the identification problem for estimating air pollution source profiles and amounts is introduced.

1 Introduction

The identification problems (P1) and (P2) may be stated as follows:
(P1) Existence of a decomposition
For any p-order positive definite symmetric matrix Σ, can it be decomposed as

$$\Sigma = \Lambda_k \Lambda_k' + \Psi_k, \tag{1}$$

where Λ_k is a loading matrix of rank k and Ψ_k is a diagonal matrix with positive diagonal elements, for assumed k $(< p)$?
(P2) Uniqueness of the decompositions
If the decomposition exists, is it unique?

Reviews of identifiability of the FA model are given by Shapiro([Sh85]) and Sato([S92]). In particular, for the confirmatory factor analysis model, Jöreskog([J69], [J79]) and Bullen and Jöreskog([BJ85]) discussed the identifiability condition.

2 Uniqueness of the decompositions

Throughout this section, we assume that Σ has a decomposition (1). The uniqueness problem for the number m of factors is as follows: Does there exist $G_m \neq \Psi_k$ such that

$$\Sigma = F_m F_m' + G_m,$$

where F_m is a $p \times m$ real matrix of rank m and G_m is a diagonal matrix with positive diagonal elements, for given m $(< p)$? For $m = k$, first, we will discuss sufficient conditions for uniqueness. The most famous theorem (Theorem 5.1 of Anderson and Rubin[AR56]) is as follows:

*The author expresses my thanks for Professor KANO Yutaka of the organizer for FA100 Symposium.
†Hiroshima Prefectural College of Health Sciences

Theorem 1 (Anderson and Rubin's condition) *A sufficient condition for uniqueness is that if any one row of Λ_k is deleted then there remain two disjoint submatrices of rank k.* ∎

Ihara and Kano([IK86]) gave another proof for this theorem and constructed a new estimator.

Theorem 2 (Williams' condition) *A sufficient condition for uniqueness is that there exists a symmetric ordering and partitioning of the rows and columns of Σ, such as $\Sigma = (\Sigma_{ij})$, $i, j = 1, 2, 3$ where*
$\{W1\}$ Σ_{12} is $k \times k$ and of rank k and
$\{W2\}$ no row of $\Sigma_{21}^{-1}\Sigma_{23}$ contains all zeros. ∎

Now, each condition is abbreviated to {AR} and {W} respectively. The following Lemma holds.

Lemma 1 *$\{W\}$ is equivalent to the following condition $\{W^*\}$:*
Λ can be partitioned into two submatrices $A(k \times k)$ and $B((p-k) \times k)$ such that
$\det|A| \neq 0$ and
if any row of B is deleted, the remaining rows of B form a matrix of rank k. ∎

Using the above proposition, we see that {AR} closely resembles {W}. We try to clarify the relation between two conditions.

Proposition 1 *$\{W\}$ yields $\{AR\}$.* ∎

Proposition 2 *For $k = 1$ and 2, $\{AR\}$ yields $\{W\}$.* ∎

Conjecture 1 *For $k = 3$, $\{AR\}$ yields $\{W\}$.* ∎

Remark 1 *When $k \geq 4$, $\{AR\}$ does <u>NOT</u> yield $\{W\}$.* ∎

In fact, for $k \geq 4$, the following $3(k-1) \times k$ matrix satisfies {AR}, however, never satisfies {W*}:

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{bmatrix}'$$

In this sense, {AR} is better than {W} for $k \geq 4$.

3 An application of estimating air pollution source profiles and amounts

Multivariate receptor models aim to identify the pollution sources based on multivariate air pollution data. Park et al. ([PSH02]) concern with estimation of the source profiles (pollution recipes) and their contributions (amounts of pollution). They used the Proposition given by Anderson(pp.576–577 in [A84] or p.593 of [A03]).

Even if a condition for unknown parameters is presented, we cannot examine whether the condition is satisfied or not. The condition proposed in [PSH02] is not for a source composition matrix P but for a source contribution matrix A. Both P and A are unknown parameters, however, the authors point out that existence of missing values can be utilized for A.

Matrix Methods and its Relationships with Factor Analysis

National Center for University Entrance Examinations Haruo Yanai

1 Introduction

Since the introduction of the Spearman's two factor model in 1904, a number of books and articles on factor analysis theories have been published over the last 100 years. In accordance with the development of theories, a number of matrix methods have been developed as well. In this paper, we attempt to relate the matrix methods to some important topics of factor analysis such as an identifiability condition(Anderson & Rubin(1956), Ihara & Kano(1976)), communality problem with special references to squared multiple correlation (SMC)(Roff(1936), Yanai & Ichikawa(1990)), analyses of image and anti-image variables (Yanai & Mukherjee(1987), factor rotation, and estimation of factor scores, and thus we attempt to extend some of the earlier theories of these methods.

2 Main Result

Let A, B and (A, B) be matrices of orders $n \times p$, $n \times q$ and $n \times (p+q)$. Let $S(A)$, $S(B)$ and $S(A, B)$ be subspaces spanned by column vectors of A, B and (A, B), respectively. Further let $P(A)$ and $P(B)$ be orthogonal projector onto subspaces $S(A)$ and $S(B)$,respectively. Then we have

Property 1. (Rao & Yanai, 1979) Let $P(A, B)$ be the orthogonal projectors onto $S(A, B)$. Then

$$P(A, B) = P(A) + P(Q(A)B) = P(B) + P(Q(B)A).$$

Using this property, we can establish

Property 2. Let $X_{(j)} = (x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p)$ be an n by (p-1) column centered matrix. Then

$$h_j^2 \geq \mathrm{SMC}(x_j) \text{ for } j = 1, \cdots, p \tag{1}$$

and the equality holds when $(x_j)'Q(X_{(j)})F = 0$ where h_j^2 is the communality of variable x_j and $\mathrm{SMC}(x_j)$ is the squared multiple correlation of x_j on $X_{(j)}$.

Proof: The proof follows from Property 1, since we have

$$P(F, X_{(j)}) = P(F) + P(Q(F)X_{(j)}) = P(X_{(j)}) + P(Q(X_{(j)})F). \tag{2}$$

Further, we have

$$h_j^2 = \mathrm{SMC}(j) + (x_j)'P(Q(X_{(j)})F)x_j \tag{3}$$

which implies (1), since $(x_j)'P(Q(X_{(j)})F)x_j$ is nonnegative. (Q.E.D.)

Observe that the equality of (1) holds in the following two cases.
Case1: SMC(j)=1, Case2: SMC(j) $\neq 1$ and $\phi_i = 0$ for any i except for j.

Example. Given four variables x_1, x_2, x_3, x_4, assume that the number of common factors is two and the factor loading matrix of these variables is given by the elements of the first and second column vectors of A given below. We computed the communalities and SMC's of these variables in the third and fourth columns of the matrix A. Then the communalities of x_3 and x_4 are exactly the same as the SMC's of these variables.

$$A = \begin{array}{c} \begin{matrix} (x_1 & x_2 & com & smc) \end{matrix} \\ \begin{pmatrix} 1 & 0 & 1 & 2a^2 \\ 0 & 1 & 1 & 2a^2 \\ a & a & 2a^2 & 2a^2 \\ a & -a & 2a^2 & 2a^2 \end{pmatrix} \end{array} \text{ and } R = \begin{pmatrix} 1 & 0 & a & a \\ 0 & 1 & a & -a \\ a & a & 1 & 0 \\ a & -a & 0 & 1 \end{pmatrix}.$$

This result covers the case which was given as the third Theorem of Roff(1936),in which he mentions that the SMC(j) equals the communality of test j if the group of tests contains r $(r < p)$ statistically independent tests with a communality of unity.

Reference

Andesrson, T. W. and Rubin, H. (1956). Statistical inferences in factor analysis. *Proceedings of third Berkley Symposium on Math. Statist. Prob.*, **5**, 11-150, Univ. of California Press.

Ihara, M. and Kano, Y. (1986). A new estimator of the uniqueness in factor analysis. *Psychometrika*, **51**, 563-566.

Rao, C. R. and Yanai, H. (1979). General definition and decomposition of projectors and some applications to statistical problems. *J. of Statistical Planning and Inferences*, **3**, 1, 1-17.

Roff, M. (1936). Some properties of the communality in multiple factor theory. *Psychometrika*, **1**, 1-16.

Yanai, H. and Mukherjee, B. N. (1987). Generalized method of image analysis from an intercorrelation matrix which may be singular. *Psychomerika*, **42**, 555-564.

Yanai, H. and Ichikawa, M. (1990). New lower and upper bounds for communality in factor analysis. *Psychometrika*, **55**, 405-409.

Mixed Factors Analysis for Finding Groups in Gene Expression Pattern

Ryo Yoshida [1] yoshidar@ism.ac.jp
Tomoyuki Higuchi [1] higuchi@ism.ac.jp
Seiya Imoto [2] imoto@ims.u-tokyo.ac.jp

[1] Research Organization of Information and Systems, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 103-8569, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

1 Introduction

In microarray experiments, each array contains the expression levels of several thousands of genes on a tissue sample. Our objective is to find groups in expression patterns of tissue samples on the basis of genes. Major difficulty in this problem is that the number of tissues to be grouped is much smaller than the dimension of data which corresponds to the number of genes. In such a case, the use of conventional model-based clustering using finite mixture models, e.g. Gaussian mixture, leads to overfitting during the density estimation process. To overcome such difficulty, we consider an parametric model so referred to as the mixed factors model. As a result of this modeling, we can avoid the occurrence of overfitting during the density estimation process even when the dimension of data is more than several thousands and the number of sample is lesser than one hundred.

2 Modeling

Let $\boldsymbol{x}$ be an observed variable distributed over $\mathbf{R}^d$. In microarray studies, the dimension of data d is determined by the number of genes used in the analysis, and thus, typically ranging from 10^2 to 10^4. The basic idea underlying the mixed factor analysis [1] is to relate $\boldsymbol{x}$ to the factor variable $\boldsymbol{f} \in \mathbf{R}^q$ as

$$\boldsymbol{x} = \Xi \boldsymbol{f} + \boldsymbol{\epsilon}. \tag{1}$$

Here $q < d$, and the $\boldsymbol{\epsilon}$ is an observational noise to be Gaussian, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \lambda \boldsymbol{I})$. The matrix of order $d \times q$, Ξ, contains the factor loadings and is referred to as the factor loading matrix. Our intention is to describe the group structure of $\boldsymbol{x} \in \mathbf{R}^d$ by using the factor variable being of the lower-dimensional random variable. Consider now that the total population of $\boldsymbol{x}$ and $\boldsymbol{f}$ consists of G subpopulations, $\mathcal{P}_1, \cdots, \mathcal{P}_G$. Let $\boldsymbol{l}^T = (l_1, \cdots, l_G)$ be a vector of unknown class labels to indicate the subpopulations such that l_g takes value one if $(\boldsymbol{x}, \boldsymbol{f}) \in \mathcal{P}_g$, otherwise zero. Moreover, the $\boldsymbol{l}$ is assumed to be multinomial distribution, $\boldsymbol{l} \sim M_G(\boldsymbol{\alpha})$ in which the probabilities of entity are defined by $\boldsymbol{\alpha}^T = (\alpha_1, \cdots, \alpha_G)$. Beside, given $l_g = 1$, the factor $\boldsymbol{f}$ is assumed to be Gaussian $\boldsymbol{f}|l_g = 1 \sim N(\boldsymbol{\mu}_g, \Sigma_g)$ for $g \in \{1, \cdots, G\}$. Then the unconditional distribution of $\boldsymbol{f}$ results in the G-components Gaussian mixture with density taken in the form,

$$p(\boldsymbol{f}) = \sum_{g=1}^{G} \alpha_g \phi(\boldsymbol{f}; \boldsymbol{\mu}_g, \Sigma_g). \tag{2}$$

Here $\phi(\boldsymbol{f}; \boldsymbol{\mu}_g, \Sigma_g)$ denotes the Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix Σ_g. We refer to the observed system consisting of (1) and (2), the mixed factors model.

3 Mixed Factors Analysis

Given a set of observations, the mixed factors model can be fitted based on the maximum likelihood. This can be achieved by the EM algorithm. Once the model has been fitted to data, our method offers the applications to clustering, dimension reduction. These can be addressed by evaluating the empirical Bayes estimators of the hidden variables, $\boldsymbol{f}$, $\boldsymbol{l}$, respectively. The analysis also covers the method of extracting some sets of genes to be relevant to explain the presence of groups on a dataset. Let $\boldsymbol{\rho}(\boldsymbol{x}, f)$ be the canonical correlations between $x_1, \cdots, x_d$ and $f_1, \cdots, f_q$. By investigating all values in the correlation matrix, each of q-coordinates can be understood. For instance, if the hth gene, i.e. the hth element of $\boldsymbol{x}$, is highly correlated with $f_1, \cdots, f_q$, then it is judged to be relevant to explain the grouping. In practice, it will be useful to list some genes to give the highest positive correlation with f_k at Ω_+^k and to give the highest negative correlation with f_k at Ω_-^k for each $k \in \{1, \cdots, q\}$. As is demonstrated in Figure 3, for gene expression profiles, these $2q$ sets can be helpful to find the biologically meaningful groups of genes to be co-expressed and also to explain the presence of groups on a given data.

The leukemia data are available at `http://www.broad.mit.edu/cancer/`. Originally it was reported that this leukemia data contains three types of acute leukemias: acute myeloid leukemia (AML, 47 cases) and acute lymphoblastic leukemia (B-cell ALL, 38 cases, T-cell ALL, 9 cases). The 10 sets of genes in Figure 3 are selected by the mixed factors model with $G = 3$, $q = 5$. The left panel shows the expression patterns of 5 sets of genes, Ω_+^k for $k = 1, \cdots, 5$. The right panel shows the expression patterns of 5 sets of genes, Ω_-^k for $k = 1, \cdots, 5$. The first 25 columns refer to the AML cases, and the last columns, the ALL 26-72 cases (B-cell ALL, 26-63 and T-cell ALL, 64-72). From this plot, it can be seen that the grouping corresponds to the molecular subtypes, AML, B-cell ALL and T-cell ALL, is related to the genes in Ω_+^k, Ω_-^k for $k = 1, 2$. All genes in a set are co-expressed. In addition, notice that each pairs of Ω_+^k and Ω_-^k shows the opposite expression pattern for $k \in \{1, \cdots, 5\}$. This implies that all genes in Ω_+^k are expressed in combination with ones in Ω_-^k.

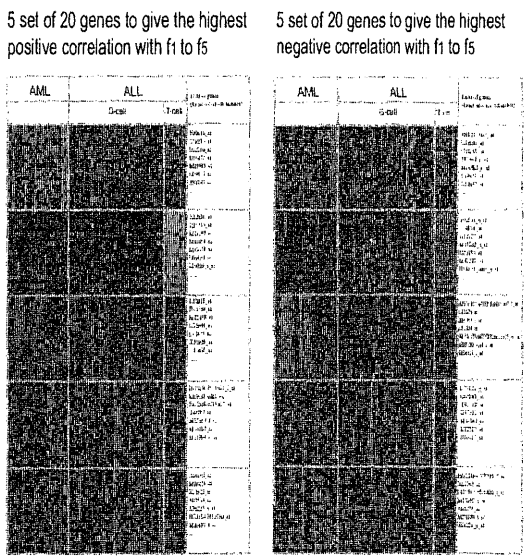

Figure 1: Expression pattern of genes judged to be relevant to the grouping.

References

[1] Yoshida, R., Higuchi, T., and Imoto, S., A mixed factors model for dimension reduction and extraction of a group structure in gene expression data, *Proc. 3rd Computational Systems Bioinformatics*, 161-172, 2004.

Multi-Trait Multi-Method Models with Data Incomplete by Design

Albert Satorra (Universitat Pompeu Fabra. Barcelona)
Willem E. Saris (University of Amsterdam)

In 1959 Campbell and Fiske suggested the multitrait-multimethod (MTMM) design for evaluating the validity of measurement instruments. Since then, several *factor analysis* models have been proposed for MTMM designs. A severe drawback of the MTMM design for assessing reliability and validity, is that respondents need to answer repeated questions of the same trait. In recent work, Saris, Satorra and Coenders (2004)(SSC) propose a multiple-group design for MTMM models (the split-ballot MTMM design, SB-MTMM) which reduces the burden of respondents (reducing the number of repeated questions to be answered). Here we extend the SSC work to a general model where the parameters are directly interpretable as reliability and validity coefficients.

We consider the following 9-indicators (3-traits and 3-methods) model: $z_{ij} = \lambda_{ij}F_{ij}$, $F_{ij} = \beta_{ij}G_{ij} + \epsilon_{ij}$ and $G_{ij} = \tau_{ij}T_i + \gamma_{ij}M_j$, where $i, j = 1-3$, the F's, G's, T_i's and the M's are latent variables, i indicates the trait, and j the method. The traits are freely correlated among them with covariances ϕ_{12}, ϕ_{13} and ϕ_{23}. The error terms ϵ_{ij}'s are allowed to have unconstrained variances ψ_{ij}'s. For the parameters β's and τ's to have direct interpretation as the square root of reliability and validity, the latent variables M's, F's and T's are constrained to have variance 1; so, we impose the following non-linear constraints: $\psi_{ij} = 1 - \beta_{ij}^2$ and $\tau_{ij}^2 = 1 - \gamma_{ij}^2$, for all i, j. Clearly, this model implies a specific covariance structure $\Sigma = \Sigma(\theta)$ for the covariance matrix of the 9 variables z's. Here θ is the parameter vector collecting the scaling parameters (the λ's), the β's and τ's, as well as the three correlations among the traits. We consider two- and three-group SB-MTMM design as discussed in SSC; for example, in the three group design: we observe variables $1-6$ in group 1, variables $1-3$ and $6-9$ in group 2, and variables $4-9$ in group 3.

Parameter estimates will be obtained minimizing the ML fitting function $F_{ML} = F_{ML}(\theta)$, where $F_{ML} := \sum_{g=1}^{G} \frac{n_g}{n} \left[\log \mid \Sigma_g \mid + \mathrm{tr}\left\{S_g \Sigma_g^{-1}\right\} - \log \mid S_g \mid - p_g\right]$, G is the number of groups, S_g and Σ_g $(= \Sigma_g(\theta))$ are the sample and population matrices for group gth, and p_g the number of observed variables in group gth. Since the same model applies for all the groups, we have

$\Sigma_g = L_g \Sigma L_g'$, where the L_g's are (known) fix selection matrices,

$$\Delta_g = \frac{\partial}{\partial\theta}\sigma_g(\theta) = U_g \frac{\partial}{\partial\theta}\operatorname{vec}\Sigma(\theta) = U_g\Delta$$

where $\sigma_g := \operatorname{vec}\Sigma_g(\theta)$. $\Delta := \frac{\partial}{\partial\theta}\operatorname{vec}\Sigma(\theta)$ and $U_g := L_g \otimes L_g$. Thus, $J := \sum_{g=1}^{G}\frac{n_g}{n}\Delta_g' W_g \Delta_g = \Delta'\left(\sum_{g=1}^{G}\frac{n_g}{n}U_g' W_g U_g\right)\Delta$, with $W_g := \frac{1}{2}\Sigma_g^{-1}\otimes\Sigma_g^{-1}$. So, the normal theory standard errors (se) will be given by the square root of the diagonal of J^{-1}/n. The asymptotic robust se (rse) will be obtained from $J^{-1}\Delta'\left(\sum_{g=1}^{G}\frac{n_g}{n}\,U_g' W_g \Gamma_g W_g U_g\right)\Delta J^{-1}/n$. Here, Γ_g is substituted by its consistent estimate $\hat{\Gamma}_{s_g} := \frac{1}{n_g-1}\sum_{i=1}^{n_g}(d_{gi}-s_g)(d_{gi}-s_g)'$, where $d_{gi} := \operatorname{vec}(z_{gi}z_{gi}')$, and s_g is the sample mean of the d_{gi}'s.

The scaled and adjusted test statistics of Satorra and Bentler (1994) do also apply. For the case of the scaled statistic, let T denote the regular chi-square test statistic, then the scaled chi-square test is $sT := T/\alpha$ where α is given by

$$\alpha = \left[\sum_{1}^{G}\operatorname{tr} W_g\Gamma_g - \operatorname{tr} J^{-1}\Delta'\left(\sum_{g=1}^{G}\frac{n_g}{n}\,U_g' W_g \Gamma_g W_g U_g\right)\Delta\right]/r$$

with r being the degrees of freedom of the test. Clearly, in all the above formulae, sample values (consistent estimates) should replace population quantities.

For the sake of comparison, single group analysis of the sample covariance matrix S obtained using the pair-wise deletion option for missing data is also reported and discussed in this talk.

References

Campbell, D.T., and D.W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait Multimethod Matrices," *Psychological Bulletin 56*, 81 -105.

Saris, W.E., Satorra, A., and G. Coenders (2004, to appear), "A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design", *Sociological Methodology, 34*.

Satorra, A., and Bentler, P. M. (1994), "Corrections to test statistics and standard errors in covariance structure analysis", In A. von Eye and C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419), Thousand Oaks, CA: Sage.

Noniterative estimation and its properties in exploratory factor analysis

Osaka University Yutaka Kano

A factor analysis model for an observed p-vector $\boldsymbol{x}$ is characterized as

$$\mathrm{Cov}(\boldsymbol{x}) = \Lambda\Lambda^{\mathsf{T}} + \Psi \quad (= \Sigma, \quad \text{say}),$$

where Λ is a factor loading matrix of $p \times k$ and Ψ is a diagonal matrix of order p for unique variances. Here the k is the number of factors. In factor analysis, communality (= 1−unique-variance) estimation has been an issue since factor analysis was invented and generalized by Spearman(1904) and Thurston(1934). The conventional way of estimating communality is now the method of maximum likelihood assuming the Wishart distribution for a sample covariance matrix, by which all parameters including communality are jointly estimated. The method of maximum likelihood requires an iterative process for optimization. Historically there is a noniterative estimation called principal factor analysis. Nowadays, it has not been used anymore.

Ihara and Kano (1986) has proposed an alternative noniterative estimation in factor analysis. Let m be a positive integer that denotes the number of factors to be extracted. Let S be a sample covariance matrix based on the sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Partition

$$\Lambda = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \Lambda_2 \\ \Lambda_3 \\ \Lambda_4 \end{bmatrix}, \qquad \Psi = \begin{array}{c} \begin{bmatrix} \psi_1 & & & \\ & \Psi_2 & & \\ & & \Psi_3 & \\ & & & \Psi_4 \end{bmatrix} \end{array} \begin{array}{l} \}1 \\ \}m \\ \}m \\ \}p-(2m+1) \end{array},$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} & \boldsymbol{\sigma}_{13} & \boldsymbol{\sigma}_{14} \\ \boldsymbol{\sigma}_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \boldsymbol{\sigma}_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \boldsymbol{\sigma}_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{bmatrix}, \quad S = \begin{bmatrix} s_{11} & \boldsymbol{s}_{12} & \boldsymbol{s}_{13} & \boldsymbol{s}_{14} \\ \boldsymbol{s}_{21} & S_{22} & S_{23} & S_{24} \\ \boldsymbol{s}_{31} & S_{32} & S_{33} & S_{34} \\ \boldsymbol{s}_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \begin{array}{l} \}1 \\ \}m \\ \}m \\ \}p-(2m+1) \end{array}$$

When $m = k$, i.e., the number of extracted factors is correctly specified, Ihara and Kano (1986) derived

$$\psi_1 = \sigma_{11} - \boldsymbol{\sigma}_{12}\Sigma_{32}^{-1}\boldsymbol{\sigma}_{31}, \tag{1}$$

provided that Σ_{32} is nonsingular.

Ihara and Kano (1986) used the formula in (1) to suggest a closed form estimator for uniquness as

$$\widehat{\psi}_1^{(m)} \quad = \quad s_{11} - \boldsymbol{s}_{12}S_{32}^{-1}\boldsymbol{s}_{31}, \tag{2}$$

One can obtain noniterative estimators for the other uniquenesses by exchanging observed variables suitably. Let $\hat{\Psi}^{(m)} = \mathrm{diag}(\hat{\psi}_1^{(m)}, \ldots, \hat{\psi}_p^{(m)})$ be an estimator for Ψ, constructed as described above. A factor loading estimator is obtained by a spectrum decomposition of $S - \hat{\Psi}^{(m)}$. Kano (1990a) and Cudeck(1991) have given an efficient tabular method for calculating $\hat{\psi}_i^{(m)}$. A program of MS-Excel with Visual Basic for the noniterative estimator is distributed by Hori. The non-iterative estimator has some important properties.

Proposition 1. (Ihara and Kano (1986), Kano (1990b)) Assume that the factor analysis model holds with Σ_{32} nonsingular and normal assumption for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is met. Assume further that $m = k$, i.e., the number of factors is correctly specified. The noniterative estimator $\hat{\psi}_1^{(k)}$ is then scale invariant, consistent, and asymptotically normal with the asymptotic covariance matrix as

$$\psi_1^2 + (\psi_1 + \boldsymbol{\sigma}_{12}\Sigma_{32}^{-1}\Psi_3\Sigma_{23}^{-1}\boldsymbol{\sigma}_{21})(\psi_1 + \boldsymbol{\sigma}_{13}\Sigma_{23}^{-1}\Psi_2\Sigma_{32}^{-1}\boldsymbol{\sigma}_{31}).$$

Proposition 2. (Kano (1990a, 1991)) Assume that the factor analysis model holds with Σ_{32} nonsingular and normal assumption for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is satisfied. Assume that $k < m \quad (< (p-1)/2)$, i.e., the number of factors is overestimated. The noniterative estimator $\psi_1^{(m)}$ is then still scale invariant and consistent. The asymptotic distribution is not normal but instead we have

$$\sqrt{n}(\psi_1^m - \psi_1) \xrightarrow{L} \psi Z_1 + c(\psi_1^2/\sigma_{11})(Z_2 + \boldsymbol{z}_1^T \boldsymbol{z}_2/Z_3),$$

where Z_1, Z_2 and Z_3 are univariate normal variates, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ have $N_{m-k}(\mathbf{0}, I_{m-k})$ and these variates are independent, and where c is a constant.

An unexpected result is that the consistency still holds for the case where the number of factors is overestimated, because the factor analysis model is not identified for the overestimated case. The MLE does not have this property. The property gives a theoretical background for the fact that the non-iterative estimator seldom causes an improper spoluton (Heywood case).

The property will be able to be used to develop an alterntive statistic for choosing an appropriate number of factors. For this, however, we need to make a further research to derive the joint distributions of $\hat{\Psi}^{(m)}$ and $S - \hat{\Psi}^{(m)}$. Distribution theory for non-normal populations also need to be developed.

References

Cudeck, R. (1991). Noniterative factor analysis estimators, with algorithms for subset and instrumental variable selection. Journal of Educational Statistics, 16, 35-52.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. Psychometrika, 47, 209-222.

Ihara, M. and Kano, Y. (1986). A new estimator of the uniqueness in factor analysis. Psychometrika, 51, 563-566.

Kano, Y. (1990a). Noniterative estimation and the choice of the number of factors in exploratory factor analysis. Psychometrika, 55, 277-291.

Kano, Y. (1990b). Comparative studies of non-iterative estimators based on Ihara and Kano's method in exploratory factor analysis. Communications in Statistics, Part A, 19, 431-444.

Kano, Y. (1991). The asymptotic distribution of a noniterative estimator in exploratory factor analysis. The Annals of Statistics, 19, 272-282.

Spearman, C. (1904). General intelligence: Objectively determined and measured. American Journal of Psychology, 15, 201-293.

Thurstone, L. L. (1934). The vectors of mind. Psychological Review, 41, 1-32