



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 66*

Some Aspects on Confirmatory Factor Analysis of Ordinal Variables and Generating Non- normal Data

HAO LUO



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2011

ISSN 1652-9030
ISBN 978-91-554-8035-6
urn:nbn:se:uu:diva-149423

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, May 6, 2011 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Luo, H. 2011. Some Aspects on Confirmatory Factor Analysis of Ordinal Variables and Generating Non-normal Data. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 66. 21 pp. Uppsala. ISBN 978-91-554-8035-6.

This thesis, which consists of five papers, is concerned with various aspects of confirmatory factor analysis (CFA) of ordinal variables and the generation of non-normal data.

The first paper studies the performances of different estimation methods used in CFA when ordinal data are encountered. To take ordinality into account the four estimation methods, *i.e.*, maximum likelihood (ML), unweighted least squares, diagonally weighted least squares, and weighted least squares (WLS), are used in combination with polychoric correlations. The effect of model sizes and number of categories on the parameter estimates, their standard errors, and the common chi-square measure of fit when the models are both correct and misspecified are examined.

The second paper focuses on the appropriate estimator of the polychoric correlation when fitting a CFA model. A non-parametric polychoric correlation coefficient based on the discrete version of Spearman's rank correlation is proposed to contend with the situation of non-normal underlying distributions. The simulation study shows the benefits of using the non-parametric polychoric correlation under conditions of non-normality.

The third paper raises the issue of simultaneous factor analysis. We study the effect of pooling multi-group data on the estimation of factor loadings. Given the same factor loadings but different factor means and correlations, we investigate how much information is lost by pooling the groups together and only estimating the combined data set using the WLS method. The parameter estimates and their standard errors are compared with results obtained by multi-group analysis using ML.

The fourth paper uses a Monte Carlo simulation to assess the reliability of the Fleishman's power method under various conditions of skewness, kurtosis, and sample size. Based on the generated non-normal samples, the power of D'Agostino's (1986) normality test is studied.

The fifth paper extends the evaluation of algorithms to the generation of *multivariate* non-normal data. Apart from the requirement of generating reliable skewness and kurtosis, the generated data also need to possess the desired correlation matrices. Four algorithms are investigated in terms of simplicity, generality, and reliability of the technique.

Keywords: confirmatory factor analysis, ordinal variables, maximum likelihood, weighted least squares, polychoric correlation, non-normal data, Fleishman's method, Monte Carlo simulation

Hao Luo, Department of Statistics, Uppsala University, SE-75120 Uppsala, Sweden.

© Hao Luo 2011

ISSN 1652-9030

ISBN 978-91-554-8035-6

urn:nbn:se:uu:diva-149423 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-149423>)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I Yang-Wallentin, F. , Jöreskog, K. G., and Luo, H. (2010) Confirmatory Factor Analysis of Ordinal Variables with Misspecified Models. *Structural Equation Modeling*, 17(3):392-423.
- II Luo, H., Lyhagen, J., and Yang-Wallentin, F. (2010) Analysis of Ordinal Variables Using Rank-Based Polychoric Correlation.
- III Luo, H. (2011) The Effect of Pooling Multi-Group Data on the Estimation of Factor Loadings.
- IV Luo, H. (2010) Generation of Non-normal Data - A Study of Fleishman's Power Method.
- V Luo, H. (2010) An Evaluation of Algorithms on Generating Multivariate Non-normal Data.

Reprints were made with permission from the publishers.

Contents

Acknowledgments	7
1 Introduction	9
1.1 Ordinal Data	9
1.2 Polychoric Correlations	10
1.3 Confirmatory Factor Analysis Model	11
1.4 Estimation Methods	12
1.4.1 Estimation Methods for Continuous Data	13
1.4.2 Estimation Methods for Ordinal Data	14
1.4.3 Multi-group Analysis	15
1.5 Algorithms for Generating Non-normal Data	16
1.6 The Contribution of this Thesis	17
2 Summary of Papers	19
2.1 Paper I: Confirmatory factor analysis of ordinal variables with misspecified models	19
2.2 Paper II: Analysis of ordinal variables using rank-based polychoric correlation	20
2.3 Paper III: The effect of pooling multi-group data on the estimation of factor loadings	21
2.4 Paper IV: Generation of non-normal data – A study of Fleishman’s power method	22
2.5 Paper V: An evaluation of algorithms on generating multi-variate non-normal data	23
3 References	29

Acknowledgments

This thesis would not have been possible without the guidance and help of several people who in one way or another contributed in the completion of this work. I would like to take this special opportunity to express my sincere gratitude to some of them.

First and foremost, I wish to send my utmost gratitude to my supervisor Fan Yang-Wallentin, who brought the subject structural equation modeling into my life and inspired me to become a PhD in the first place. To Fan, thank you for always backing me up, being available, and tolerating my one-after-the-other careless mistakes and scatterbrain. I will always remember the face-to-face sentence by sentence paper correction time. The effort and care that you have laid down on me goes far beyond and above all duty. I have occupied so much of your time and all the things you taught me are gifts enough for a lifetime. I feel extremely lucky and blessed of having you as my supervisor.

I want to thank my assistant supervisor Johan Lyhagen, who has enlightened me in many different ways. You always cheer me up with enthusiasm and a positive attitude, which reminds me the "Don't worry be happy" song all the time. A special acknowledgement goes to my assistant supervisor Karl Jöreskog, who has been very generous with helps of all sorts of matters—research ideas, programming, insightful comments on drafts, and presentation tips. It has been a privilege for me to work with such a qualified supervisor.

Also, I would like to thank Professor Rolf Larsson for being a good teacher and impressively patient for commenting on all my papers. Except for the constructive critiques you wrote on my papers, the delightful music that sometimes came out of your office had made my mornings more enjoyable. I wish to thank also Professor Anders Christofferson, who told me "our doors are always open" four years ago and brought valuable research ideas for me to give a try.

My thanks also goes to all my colleagues from the department. Thanks to Bo Wallentin for being so nice and supportive all the time. Thanks to Lisbeth Hansson, who quietly but noticeably showing considerations with warm smiles. I wish to thank Thommy Perlinger, Lennart Norell, Anders Ågren, Adam Taube, Anna Gunsjö and Dag Sörbom for being amiable and encouraging. Big thanks to Lars Forsberg, for showing me how to bake gingerbread, make marshmallow sandwich, and find the most frequently used sample sizes in applied research. Thanks also to Daniel Preve for arranging helpful seminars and the comforting words when I got panic after reading your excellent thesis. Thanks all of you for providing such a pleasant working environment and making my four years in the department a wonderful stay.

In particular, I greatly enjoyed the accompany of my fellow colleagues and friends. To Petra Ornstein, I have benefited a lot from the valuable discussions we had and the careful proofreading you offered me. Your frankness and optimistic mood made me feel so relaxed and comfortable to share the office with you. To James Blevins, thank you for the books you lend to me. We might have different political views, but your nice words actually mean a lot to me. Thanks to Katrin Kraus for lending me the mug in my first day and bringing me to my first coffee break. To Joakim Ekström, thank you for inviting me to your conferment ceremony and letting me

try your wreath. To Martin Solberger, thank you very much for the jumpy moments and high-pitch screams you gave to me by hiding behind my door and coming out from nowhere. To Nicklas Pihlström, thank you for teasing me now and then, which helped me a lot by keeping my brain vigorous and active just in order to fight back. Big hugs to Myrsini Katsikatsou and Ronnie Pingel. You two are always the nice and encouraging ones, which made me feel really safe. Jianxin Wei, Xijia Liu and Xingwu Zhou, my friends from home, your diligence has inspired me so much. Having you guys around made me believe that I always have some one to count on. To all of you who made my life more colorful, the beers, the awkward dance in the faculty club, the singstar game, the crazy parties and the tons of fun we had are safely locked in my memory.

Many other people made my life in Sweden exciting and entertaining. I want to thank Haishan Yu, my dearest friend from the neighbor department. Just by seeing your yellow light on, I knew my "comrade-in-arms" is being through the same struggle. A tremendously big thank-you goes to my sweet friends Ran He, Yang Song, Zhaoguo Ding, Jia Zhou, Qiao Wei and Shaobo Jin. Thank you guys for sending me post cards during my hardship, cheering me up by your stupid laughs, and cooking me delicious food. I do feel I have a big family here just because of you. Last but not the least, to Camille Madec, I truly enjoyed our movie times together (despite the fact that you used my points to become a golden member). Your scarily deep knowledge of statistics (as an ecology PhD student) had given me so much pressure which urged me to study harder just to make sure that I can get a job as a real statistician.

Finally, I have to share my joy with my incredible parents who made me who I am today and always truly believe in me no matter what. Thanks to my mother for setting such a good example to me. It is your independence, your diligence, your dedication to research, and your warmth that motivated me to become a strong and kindhearted woman like you. Thanks to my sweet father, who stands by me all the time and takes care of all my tear bursting moments. It is your spirit of freedom that influenced me to always follow my heart and become a big-hearted girl (but forgive me for wasting so much money on the ballet lessons, the violin and the piano). Thank you both for putting up with my five years absence. Without your love, I will never find my way back home.

Hao Luo
Uppsala, March 2011

1. Introduction

1.1 Ordinal Data

Ordinal data are common in many empirical investigations in the social and behavioral sciences. Observations on an ordinal variable are assumed to have a logical ordering to categories. This logical ordering is typical when data are collected from questionnaires. A good example is the Likert Scale that is frequently used in survey research: 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly agree. Although a question is designed to measure a theoretical concept, the observed responses are only a discrete realization of a small number of categories and the distances between categories are unknown. For example, political philosophy can be categorized as liberal, moderate, and conservative. Although a person categorized as moderate is more liberal than a person categorized as conservative, no numerical values describe how much more liberal that person is.

Following Muthén (1984), Lee, Poon, & Bentler (1990), Jöreskog (1990), and others, it is assumed that there is a continuous variable x_i^* underlying the ordinal variable x_i , $i = 1, 2, \dots, p$. This continuous variable x_i^* represents the attitude underlying the ordered responses to x_i and is assumed to have a range from $-\infty$ to $+\infty$.

The underlying variable x_i^* is unobservable. Only the ordinal variable x_i is observed. For an ordinal variable x_i with m_i categories, the connection between the ordinal variable x_i and the underlying variable x_i^* is

$$x_i = c \iff \tau_{c-1}^{(i)} < x_i^* < \tau_c^{(i)}, \quad c = 1, 2, \dots, m_i, \quad (1.1)$$

where

$$\tau_0^{(i)} = -\infty, \quad \tau_1^{(i)} < \tau_2^{(i)} < \dots < \tau_{m_i-1}^{(i)}, \quad \tau_{m_i}^{(i)} = +\infty, \quad (1.2)$$

are threshold parameters. For variable x_i with m_i categories, there are $m_i - 1$ strictly increasing threshold parameters $\tau_1^{(i)}, \tau_2^{(i)}, \dots, \tau_{m_i-1}^{(i)}$.

Because only ordinal information is available about x_i , the distribution of x_i^* is determined only up to a monotonic transformation. It is convenient to let x_i^* have the standard normal distribution with density function $\phi(\cdot)$ and distribution function $\Phi(\cdot)$. Then the probability $\pi_c^{(i)}$ of

a response in category c on variable x_i , is

$$\pi_c^{(i)} = Pr[x_i = c] = Pr[\tau_{c-1}^{(i)} < x_i^* < \tau_c^{(i)}] = \int_{\tau_{c-1}^{(i)}}^{\tau_c^{(i)}} \phi(u) du = \Phi(\tau_c^{(i)}) - \Phi(\tau_{c-1}^{(i)}), \quad (1.3)$$

for $c = 1, 2, \dots, m_i - 1$, so that

$$\tau_c^{(i)} = \Phi^{-1}(\pi_1^{(i)} + \pi_2^{(i)} + \dots + \pi_c^{(i)}), \quad (1.4)$$

where Φ^{-1} is the inverse of the standard normal distribution function. The quantity $(\pi_1^{(i)} + \pi_2^{(i)} + \dots + \pi_c^{(i)})$ is the probability of a response in category c or lower.

The probabilities $\pi_c^{(i)}$ are unknown population quantities. In practice, $\pi_c^{(i)}$ can be estimated consistently by the corresponding percentage $p_c^{(i)}$ of responses in category c on variable x_i . Then, estimates of the thresholds can be obtained as

$$\hat{\tau}_c^{(i)} = \Phi^{-1}(p_1^{(i)} + p_2^{(i)} + \dots + p_c^{(i)}), \quad c = 1, \dots, m - 1. \quad (1.5)$$

The quantity $(p_1^{(i)} + p_2^{(i)} + \dots + p_c^{(i)})$ is the proportion of cases in the sample responding in category c or lower on variable x_i .

1.2 Polychoric Correlations

Among many alternatives for estimating polychoric correlations, the estimator proposed by Olsson (1979), which has been widely used for decades, is the most popular. Let x_i and x_j be two ordinal variables with m_i and m_j categories, respectively. Their marginal distribution in the sample is represented by a contingency table

$$\begin{pmatrix} n_{11}^{(ij)} & n_{12}^{(ij)} & \dots & n_{1m_j}^{(ij)} \\ n_{21}^{(ij)} & n_{22}^{(ij)} & \dots & n_{2m_j}^{(ij)} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_i 1}^{(ij)} & n_{m_i 2}^{(ij)} & \dots & n_{m_i m_j}^{(ij)} \end{pmatrix}, \quad (1.6)$$

where $n_{ab}^{(ij)}$ is the number of cases in the sample in category a on variable x_i and in category b on variable x_j . The underlying variables x_i^* and x_j^* are assumed to be bivariate normal with zero means, unit variances, and with correlation ρ_{ij} , the polychoric correlation.

The polychoric correlation can be estimated by maximizing the log-likelihood of the multinomial distribution (see Olsson, 1979).

$$\ln L = \sum_{a=1}^{m_i} \sum_{b=1}^{m_j} n_{ab}^{(ij)} \log \pi_{ab}^{(ij)}, \quad (1.7)$$

where

$$\pi_{ab}^{(ij)} = Pr[x_i = a, x_j = b] = \int_{\tau_{a-1}^{(i)}}^{\tau_a^{(i)}} \int_{\tau_{b-1}^{(j)}}^{\tau_b^{(j)}} \phi_2(u, v) dudv, \quad (1.8)$$

and

$$\phi_2(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)}, \quad (1.9)$$

is the standard bivariate normal density with correlation ρ_{ij} . Maximizing $\ln L$ gives the sample polychoric correlation denoted r_{ij} .

The polychoric correlation can be estimated by a two-step procedure. In the first step, the thresholds are estimated from the univariate marginal distributions by Equation 1.5. In the second step, the polychoric correlations are estimated from the bivariate marginal distributions by maximizing $\ln L$ for given thresholds.

1.3 Confirmatory Factor Analysis Model

Structural Equation Modeling (SEM) has become the preeminent multivariate technique in the social and behavioral sciences. Within the area of SEM, confirmatory factor analysis is the most common type of analysis. The basic idea of factor analysis is to explain the correlation between a large set of manifest variables in terms of a small number of latent factors. Let \mathbf{x} be the vector of order $p \times 1$ observed variables. Let the vectors $\boldsymbol{\xi}$ of order $k \times 1$ and $\boldsymbol{\delta}$ of order $p \times 1$ represent the factors and the unique variables which are assumed to be uncorrelated. A typical CFA model has the form

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \quad (1.10)$$

where the matrix $\mathbf{\Lambda}$ of order $p \times k$ contains the factor loadings λ_{ij} . To make sure the model is identified, some elements of $\mathbf{\Lambda}$ may be fixed at zero.

Let $\boldsymbol{\Phi}$ of order $k \times k$ and $\boldsymbol{\Theta}$ of order $p \times p$ be the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$, respectively. We assume that the unique factors are uncorrelated so that $\boldsymbol{\Theta}$ is a diagonal matrix. The covariance matrix of \mathbf{x} is

$$\boldsymbol{\Sigma}(\mathbf{\Lambda}, \boldsymbol{\Phi}) = \mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Theta}. \quad (1.11)$$

We write $\boldsymbol{\Sigma}(\mathbf{\Lambda}, \boldsymbol{\Phi})$ to emphasize that $\boldsymbol{\Sigma}$ is a function of $\mathbf{\Lambda}$ and $\boldsymbol{\Phi}$.

The path diagram of a typical CFA model with two factors and six indicators is shown in Figure 1.1. In matrix form, the model is:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{pmatrix}. \quad (1.12)$$

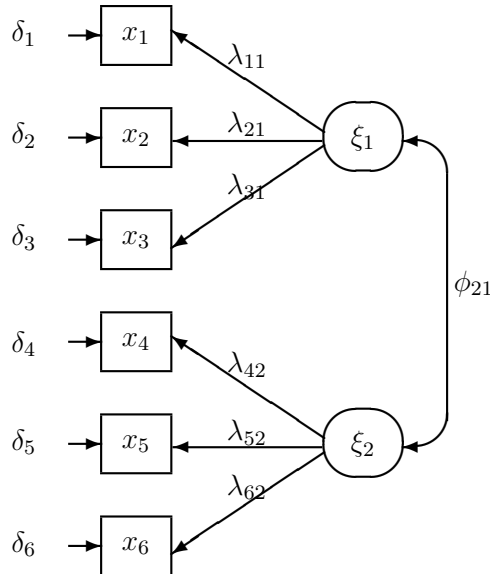


Figure 1.1: Path Diagram of a typical CFA model with two factors and six indicators

1.4 Estimation Methods

The fundamental hypothesis for the structural equation procedures is that the covariance matrix of the observed variables is a function of a set of parameters. If the model were correct and the parameters are known, the population covariance matrix would be exactly reproduced. As a result, to estimate the model we minimize the difference between the sample covariances and the predicted covariances implied by the model. Four alternative methods are considered and compared in this thesis, namely *maximum likelihood* (ML), *unweighted least squares* (ULS), *diagonally*

weighted least-squares (DWLS), and *weighted least squares* (WLS). The ML and the WLS methods are applied when the observed variables are continuous. All four methods are studied when ordinal data are encountered and the fit functions have been slightly adjusted.

1.4.1 Estimation Methods for Continuous Data

Maximum Likelihood (ML)

The ML estimator maximizes the likelihood of the parameter given the data. Let \mathbf{S} be the sample covariance matrix and Σ be defined as in Equation 1.11. The ML fit function is

$$F_{ML} = \log |\Sigma(\mathbf{\Lambda}, \mathbf{\Phi})| + \text{tr}(\mathbf{S}\Sigma(\mathbf{\Lambda}, \mathbf{\Phi})^{-1}) - \log |\mathbf{S}| - p, \quad (1.13)$$

which is to be minimized with respect to the free elements of $\mathbf{\Lambda}$ and $\mathbf{\Phi}$. To date, this fitting function is still the most widely used function for general structural equation models. It is derived based on the assumption that the observed variables have a multinormal distribution or that \mathbf{S} has a Wishart distribution.

Weighted Least Squares (WLS)

The WLS method is an asymptotically distribution free method that has the fitting function

$$F_{WLS} = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \quad (1.14)$$

$$= \sum_{g=1}^p \sum_{h=1}^g \sum_{i=1}^p \sum_{j=1}^i w^{gh,ij} (s_{gh} - \sigma_{gh})(s_{ij} - \sigma_{ij}), \quad (1.15)$$

where

$$\mathbf{s}' = (s_{11}, s_{21}, s_{22}, s_{31}, \dots, s_{pp}) \quad (1.16)$$

is a vector of the elements in the lower half, including the diagonal elements, of the sample covariance matrix \mathbf{S} ;

$$\boldsymbol{\sigma}' = (\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \dots, \sigma_{pp}) \quad (1.17)$$

is the vector of corresponding elements of $\Sigma(\mathbf{\Lambda}, \mathbf{\Phi})$; and \mathbf{W}^{-1} is a positive definite matrix of order $s \times s$, where $s = p(p+1)/2$. In most case the elements of \mathbf{W}^{-1} are obtained by inverting a matrix \mathbf{W} whose typical element is denoted as $w_{gh,ij}$. The usual way of choosing \mathbf{W} is to let $w_{gh,ij}$ be a consistent estimate of the asymptotic covariance between s_{gh} and s_{ij} , *i.e.*, \mathbf{W} is the asymptotic covariance matrix of the elements of \mathbf{S} .

1.4.2 Estimation Methods for Ordinal Data

For the special case of ordinal variables, the model to be estimated is

$$\mathbf{x}^* = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} , \quad (1.18)$$

where \mathbf{x}^* is a vector of order $p \times 1$ of *underlying* variables corresponding to the $p \times 1$ vector of the observed ordinal variables \mathbf{x} , as defined in Section 1.1. For convenience, we assume that Φ is a correlation matrix with ones in the diagonal. Because the underlying variables x_i^* have variances equal to 1, it follows from Equation 1.11 that

$$\Theta = \mathbf{I} - \text{diag}(\mathbf{\Lambda}\Phi\mathbf{\Lambda}') , \quad (1.19)$$

so that

$$\Sigma(\mathbf{\Lambda}, \Phi) = \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \mathbf{I} - \text{diag}(\mathbf{\Lambda}\Phi\mathbf{\Lambda}') . \quad (1.20)$$

This is the correlation matrix implied by the model to be fitted to the matrix of polychoric correlations.

Maximum Likelihood (ML) for Ordinal Variables

The method of ML has no theoretical justification for use with ordinal variables. Nevertheless, it works if used as follows. Let \mathbf{R} be the matrix of polychoric correlations with ones in the diagonal and let Σ be defined as in Equation 1.20. The ML fit function is

$$F_{ML}(\mathbf{R}, \mathbf{\Lambda}, \Phi) = \log |\Sigma| + \text{tr}(\mathbf{R}\Sigma^{-1}) - \log |\mathbf{R}| - p . \quad (1.21)$$

Three Least Squares Methods for Ordinal Variables

The three least squares methods are two-step methods. In the first step, the polychoric correlations \mathbf{r} and their asymptotic covariance matrix \mathbf{W} are estimated as described earlier. Note that $\mathbf{r} = (r_{21}, r_{31}, r_{32}, \dots, r_{p,p-1})'$ is a vector of polychoric correlations below the diagonal of the polychoric correlation matrix \mathbf{R} . The 1s in the diagonal are not included in the vector \mathbf{r} . Both \mathbf{r} and \mathbf{W} are estimated from the sample data without the use of the model. Let $s = p(p-1)/2$. The vector \mathbf{r} is of order $s \times 1$ and the matrix \mathbf{W} is of order $s \times s$.

In the second step, $\mathbf{\Lambda}$ and Φ are fitted to \mathbf{r} by minimizing the fit function

$$F(\mathbf{r}, \mathbf{\Lambda}, \Phi) = [\mathbf{r} - \boldsymbol{\rho}(\mathbf{\Lambda}, \Phi)]' \mathbf{V} [\mathbf{r} - \boldsymbol{\rho}(\mathbf{\Lambda}, \Phi)] , \quad (1.22)$$

where \mathbf{V} is a positive matrix and $\boldsymbol{\rho}(\mathbf{\Lambda}, \Phi)$ is a vector of the elements of $\mathbf{\Lambda}\Phi\mathbf{\Lambda}'$ below the diagonal. The three least squares methods differ in the choice of weight matrix \mathbf{V} :

$$\text{ULS} : \mathbf{V} = \mathbf{I} , \quad (1.23)$$

$$\text{DWLS} : \mathbf{V} = (\text{diag} \mathbf{W})^{-1} , \quad (1.24)$$

$$\text{WLS} : \mathbf{V} = \mathbf{W}^{-1} . \quad (1.25)$$

The main difference between these weight matrices is that the weight matrix is diagonal for ULS and DWLS, whereas for WLS the weight matrix is the inverse of the full matrix \mathbf{W} . For DWLS, only the diagonal elements of \mathbf{W} are used. In scalar form, these fit functions can be written as

$$\text{ULS} : F(\mathbf{r}, \mathbf{\Lambda}, \mathbf{\Phi}) = \sum_i (r_i - \rho_i)^2 , \quad (1.26)$$

$$\text{DWLS} : F(\mathbf{r}, \mathbf{\Lambda}, \mathbf{\Phi}) = \sum_i (r_i - \rho_i)^2 / w_{ii} , \quad (1.27)$$

$$\text{WLS} : F(\mathbf{r}, \mathbf{\Lambda}, \mathbf{\Phi}) = \sum_i \sum_j (r_i - \rho_i)(r_j - \rho_j) w^{ij} , \quad (1.28)$$

where w^{ij} is an element of \mathbf{W}^{-1} .

1.4.3 Multi-group Analysis

So far, we have discussed the methodology that can be used to analyze data from a single sample. In many studies, one needs to analyze data from several samples simultaneously. Consider a set of G populations. The samples may come from different countries, industries, strategic groups, corporations, or culturally or socioeconomically different groups. Let $N^{(g)}$ be the sample size of the g th group and N be the total sample size. Let $\mathbf{S}^{(g)}$ and $\mathbf{\Sigma}^{(g)}$ be the sample and model implied covariance matrices for the g th group, respectively. To estimate all the models simultaneously, the fit function to be minimized is

$$F = \sum_{g=1}^G (N^{(g)} / N) F_{ML}^{(g)} , \quad (1.29)$$

where

$$F_{ML}^{(g)} = \log |\mathbf{\Sigma}^{(g)}| + \text{tr}(\mathbf{S}^{(g)}(\mathbf{\Sigma}^{(g)})^{-1}) - \log |\mathbf{S}^{(g)}| - p , \quad (1.30)$$

is the fit function of the maximum likelihood method. Under different circumstances, the fit function $F_{ML}^{(g)}$ can be substituted by the fit functions of any other estimation methods, such as ULS, WLS, and DWLS.

1.5 Algorithms for Generating Non-normal Data

Monte Carlo simulations are used to compare the robustness of various estimation methods and the performance of fit statistics under different degrees of non-normality. The value of a simulation study is closely related to the generation of non-normal data. Therefore, the reliability and efficiency of the data generation method are of crucial importance.

Fleishman (1978) proposed that a polynomial transformation

$$Y = a + bX + cX^2 + dX^3, \quad (1.31)$$

where X is normally distributed with zero mean and unit variance ($N(0,1)$), may be used to obtain non-normal distributions. The constants a , b , c , and d may be chosen such that Y has a distribution with specified moments of the first four orders, *i.e.*, the mean, variance, skewness, and kurtosis. Suppose Y should have the first four moments $E(Y) = 0$, $E(Y^2) = 1$, $E(Y^3) = \gamma_1$, and $E(Y^4) = \gamma_2 + 3$, where γ_1 and γ_2 are the specified values of skewness and kurtosis, respectively. Then $a = -c$ and b , c , and d must satisfy the three equations, corresponding to Fleishman's Equation 11, 17, and 18:

$$F_1(b, c, d) = b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0, \quad (1.32)$$

$$F_2(b, c, d) = 2c(b^2 + 24bd + 105d^2 + 2) - \gamma_1 = 0, \quad (1.33)$$

$$F_3(b, c, d) = 24(bd + c^2[1 + b^2 + 28bd] + d^2[12 + 48bd + 141c^2 + 225d^2]) - \gamma_2 = 0. \quad (1.34)$$

Tadikamalla (1980) criticizes this method because the exact distribution of Y is not known and certain combinations of skewness and kurtosis are not feasible. Despite these drawbacks, there are two main advantages that made Fleishman's polynomial transformation still a widely used method for generating non-normal data: Its procedure is the easiest to implement and executed most quickly and it can easily be extended to generate multivariate random numbers with specified intercorrelations and univariate means, variances, skewness, and kurtosis.

Vale & Maurelli (1983) extended Fleishman's technique to the multivariate case. In addition to the use of coefficients for each variable that yield the desired moments, the key step in this process is to consider the effect of the non-normalizing transformations on the variable intercorrelations.

Headrick & Sawilowsky (1999) proposed another extension of Fleishman's procedure by combining a generalization of Theorems 1 and 2 from Knapp & Swoyer (1967) with the Fleishman procedure. It is claimed to be simpler than Vale & Maurelli's approach since the initial step of principal components decomposition on the population correlation matrix is avoided.

To increase the precision in the approximations of non-normal distribution and lower the boundary of the possible skewness and kurtosis combinations, Headrick (2002) derived a new method for simulating univariate and multivariate non-normal distributions using polynomial transformations of order five.

A recent technique, the *sample* and *iterate* (SI) technique, was proposed by Ruscio & Kaczetow (2008). The basic idea of the SI approach is to sample each variable directly from a specified population distribution and determine the intermediate correlation matrix through an iterative, trial-and-error process.

1.6 The Contribution of this Thesis

This thesis, which consists of five papers, deals with various aspects of confirmatory factor analysis with ordinal variables, as well as the evaluation of the algorithms used in generating univariate and multivariate non-normal distributions. The thesis consists of two main parts. The first two papers study the approaches used in confirmatory factor analysis when ordinal variables are encountered. In Paper I, the polychoric correlations of the ordinal variables are estimated and the models are fitted using ML, ULS, WLS, and DWLS methods. We carry out a simulation evaluation to study the behavior of different estimation methods in combination with polychoric correlations when the models are misspecified. The second paper focuses on the situation when the underlying distributions of the ordinal variables are not normal. Instead of the polychoric correlation proposed by Olsson (1979), a non-parametric polychoric correlation coefficient based on the discrete version of Spearman's rank correlation is proposed. The simulation study shows the benefits of the proposed estimator. Paper III is an independent study about the effect of pooling multi-group data on the estimation of factor loadings. Using the same model as the first two papers, this simulation study shows that, given the same factor structures among groups, pooling the groups and fitting the model by the WLS method have a negligible effect on the estimation of factor loadings.

The second part of this thesis is the evaluation of the algorithms on generating non-normal data. Paper IV uses a Monte Carlo simulation to assess the reliability of Fleishman's power method of generating univariate non-normal distributions. The results suggest that Fleishman's method has difficulties on generating non-normal samples with higher levels of skewness and kurtosis. In addition, Fleishman's (1978) parabola, which indicates the bottom boundary of the possible combination of skewness and kurtosis, is shown to be incorrect. Extending the univariate study to the multivariate case, the last paper examines four algorithms for gener-

ating multivariate non-normal distributions. Apart from the requirement of generating distributions with the reliable skewness and kurtosis as the pre-specified values, the ability of generating correct correlation matrices among variables is also a desirable property. Algorithms are compared in terms of simplicity, generality, and reliability of the technique. Recommendations regarding the application schemes are provided to applied researchers. In addition, the size and power of some well-known normality tests are investigated based on the generated non-normal samples.

2. Summary of Papers

2.1 Paper I: Confirmatory factor analysis of ordinal variables with misspecified models

Paper I assesses the effect of model size, sample size, and number of categories on the performance of the four estimation methods when the models are both correct and misspecified. The four alternative methods are RML, RULS, RDWLS, and WLS. Two models, referred to as Model 1 and Model 2, are used. Model 1, as shown in Section 1.3, is a small model with $p = 6$ observed variables and $k = 2$ factors and Model 2 is a large model with $p = 16$ observed variables and $k = 4$ factors. The five sample sizes (N) used in this simulation study are selected as 100, 200, 400, 800, and 1600. They represent sample sizes commonly encountered in applied research, ranging from fairly small ($N = 100$) to fairly large ($N = 1600$). Each ordinal variable x_i is assigned to have 2, 5, or 7 categories with and without symmetric distributions. The category probabilities are the same for all variables.

As stated previously, of central interest is the behavior of different estimators when models are misspecified. For Model 1, although four values of λ_{41} (0, 0.1, 0.3, 0.5) are used for generating data, the model is consistently estimated using the same model by assuming $\lambda_{41} = 0$. Except for the condition of generating data using $\lambda_{41} = 0$, the population models used to generate the data differ from the model actually estimated for the other three conditions. Similar to Model 1, four values are used for both λ_{14} and $\lambda_{16,1}$. Thus, for each model, there are four population models. Altogether, there are 24 experimental cells for each model (4 population models, 3 number of categories with and without symmetric distributions).

To evaluate estimators under various conditions, we first generate \mathbf{x}^* according to different model settings. If $\tau_{c-1}^{(i)} < x_i^* < \tau_c^{(i)}$, set $x_i = c$, for $c = 1, 2, \dots, m_i$ and $i = 1, 2, \dots, p$. This procedure gives a single observation of \mathbf{x} . Repeat this procedure N times gives a random sample represented by a data matrix \mathbf{X} of order $N \times p$, where each element is an integer of $1, 2, \dots, m_i$ representing the observed category. From \mathbf{X} , the matrix of polychoric correlations \mathbf{R} and the asymptotic covariance matrix \mathbf{W} are estimated using PRELIS. \mathbf{R} and \mathbf{W} are used to estimate the model parameters and their standard errors with ULS, DWLS, ML, and WLS.

The outcome variables of interest are mainly bias and root mean square error for both parameter estimates and their standard errors.

Our results show that, in comparison with ULS, DWLS, and ML, WLS performs poorly under all conditions, although WLS performs better for the small than for the large model. Concerning the other three methods, it is difficult to claim that anyone of the methods (ULS, DWLS, and ML) is better than the others. Their performance on estimates, standard errors, and chi-squares is very similar. Even with $R = 2000$ replicates, we have not found any significant differences. In addition, the number of categories and shape of the distribution do not seem to matter. All methods behave similarly with respect to the different number of categories.

Our recommendation would be to use the ULS estimator in that it is computationally parsimonious and functions equally well as DWLS and ML. The \mathbf{W} is not needed for estimation but to obtain correct standard errors and chi-squares it is required.

2.2 Paper II: Analysis of ordinal variables using rank-based polychoric correlation

The estimators that have been studied in Paper I are all derived under the assumption that the underlying variables are bivariate normally distributed, which in practice is usually violated. We propose a *non-parametric polychoric correlation* coefficient based on the discrete version of Spearman's rho proposed by Nešlehová (2004).

Let $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$ be independent random vectors with common margins F (for X_1 and Y_1) and G (for X_2 and Y_2), which are both discrete with finite support. Let a_1, a_2, \dots, a_m and b_1, b_2, \dots, b_n be the finite supports of the margins F and G , respectively. The marginal probabilities are denoted as $p_i = P(X_1 = a_i)$ and $q_j = P(X_2 = b_j)$, where $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$. The *discrete version of Spearman's rho* is defined as

$$\rho_s^* = \frac{\tilde{\rho}_s}{\sqrt{(1 - \sum_{i=1}^m p_i^3)(1 - \sum_{j=1}^n q_j^3)}}, \quad (2.1)$$

where

$$\tilde{\rho}_s = 3 \{P[(X_1 - Y_1)(X_2 - Y_2) > 0] - P[(X_1 - Y_1)(X_2 - Y_2) < 0]\} \quad (2.2)$$

is the Spearman's rho corresponding to the standard extension copula. The non-parametric polychoric correlation is given by

$$\rho_{np} = 2 \sin(\rho_s^* \pi / 6) .$$

We carry out a Monte Carlo simulation to evaluate the performance of the non-parametric polychoric correlation in confirmatory factor analysis. For various experimental conditions, both the non-parametric polychoric correlations and the polychoric correlation proposed by Olsson (1979) are estimated and used to fit the CFA model using the ML method. To study the effect of skewness and kurtosis four underlying distributions, including the multivariate normal distribution, are generated. The four distributions are: (1) the multivariate normal, (2) the multivariate skew-normal (see Azzalini and Valle, 1996), (3) the multivariate t and (4) the multivariate skew- t (see Azzalini and Capitanò, 2003). The simulation result shows the superiority of using the non-parametric polychoric correlation to fit a CFA model for ordinal data when the underlying distributions of the ordinal variables are not normal.

2.3 Paper III: The effect of pooling multi-group data on the estimation of factor loadings

Structural equation modeling is widely used in the social, economic, and behavioral sciences. In particular, the simultaneous factor analysis approach to study the similarities and differences between groups is commonly applied to many areas of research. Applying the ML method to a multi-group analysis is optimal when data are normally distributed within each group. Often, there is a factor pattern that is invariant over groups, *i.e.*, the factor loadings for each group are the same. It may be argued that the populations were selected from a parent population for which the same common model holds. As a result, we can pool the groups together and estimate the model at one time. Given the situation of the same factor loadings but different factor means and factor correlations between groups, when pooling multi-group data together, the distribution of the pooled data will no longer be normally distributed. Instead, mixture distributions of the observed variables are created. The WLS method will be used to estimate the model, which should be the optimal method under such circumstances.

Given the same factor structure across groups, it is of interest to see how much information is lost by pooling the groups together and estimating the model using the combined data set. In this study, assuming the same factor structures for different groups, we compare the traditional multi-group analysis approach using the ML method with a procedure for estimating the factor loadings by the WLS method using the aggregated data over groups. The differences between groups are achieved by various settings of factor means and factor covariance matrices. We examine the effects of the level of differences between factor means, factor

covariance matrices, sample sizes, and group weights on the estimation of factor loadings.

Assume that the data have been collected on a set of variables $\mathbf{x} = x_1, \dots, x_p$ for $g = 1, \dots, G$ separate groups. Let the mean of latent variables be $\kappa^{(g)}$ for the g th group, the pooled covariance matrix of the latent variables $\Phi^{(p)}$ is derived as

$$\Phi^{(p)} = \frac{1}{N-1} \left\{ \sum_{g=1}^G (N^{(g)} - 1) \Phi^{(g)} \right\} + \frac{1}{N-1} \left\{ \sum_{g=1}^G N^{(g)} \left(\kappa^{(g)} - \frac{1}{N} \sum_{g=1}^G N^{(g)} \kappa^{(g)} \right) \left(\kappa^{(g)'} - \frac{1}{N} \sum_{g=1}^G N^{(g)} \kappa^{(g)'} \right) \right\}.$$

The simulation results show that the effect of pooling multi-group data on the estimation of factor loadings is negligible. In addition, an example about customer-supplier relationship supports the results from the simulation study.

2.4 Paper IV: Generation of non-normal data – A study of Fleishman’s power method

Paper IV examines the effect of sample size as well as levels of skewness and kurtosis in relation to Fleishman’s method for generating reliable non-normal samples. Their effect on the power of the normality test proposed by D’Agostino is also a point at issue.

The basic idea of Fleishman’s method is that by constraining values of both skewness and kurtosis of Y , values of the four constants a , b , c , and d can be calculated correspondingly. We first compute the exact values of the four constants given pre-specified values of skewness and kurtosis of Y . The desired non-normal samples are obtained by generating a random sample from a standard distribution and transform it by a , b , c , and d according to formula (1.31). A question might be to what extent the generated non-normal samples can mimic the properties of the desired ones. To address this problem, we estimate the skewness and kurtosis of the generated samples under different conditions and compare them with the pre-specified values of skewness and kurtosis.

To examine the effect of sample size, the experimental sample sizes are chosen to be 10, 25, 50, 100, 200, 1000, and 2000, ranging from extremely small to fairly large. Regarding levels of skewness and kurtosis, there are 29 combinations with skewness ranging from 0 to 1.25 and kurtosis from -1 to 4. Each of these 203 experimental cells is repeated 2000 times. The evaluation criteria are set to be bias of the estimated skewness/kurtosis and their associated standard deviations across the 2000 replicates.

The normality test proposed by D'Agostino (1986) includes three aspects. It consists of two independent tests of zero skewness and kurtosis separately as well as an omnibus test to examine both skewness and kurtosis simultaneously. The test statistics for the independent tests are different z -scores justified by different conditions of sample size. Under normality, the z -scores are approximately normally distributed with mean zero and variance one. The omnibus test simply sums up the squares of the z -scores for skewness and kurtosis. Under normality, this K^2 statistic has approximately a chi-square distribution with two degrees of freedom. In the current study, we assess this normality test by comparing the power of the tests under different conditions of sample size and levels of skewness and kurtosis.

The results reveal that significant effects are found for both sample size and levels of skewness and kurtosis on generating the expected non-normal data. A sample size of 1000 can guarantee an accurate enough generating procedure. Fleishman's method has difficulty in properly generating non-normal samples characterized by higher levels of skewness and kurtosis. In point of the power of normality test, sample size has a positive effect on obtaining a trustworthy test decision. The null hypothesis of zero skewness and kurtosis is more often rejected for higher levels of skewness and kurtosis.

In addition, Fleishman (1978) has described a parabola as the bottom boundary for the possible combinations of skewness and kurtosis. Our result shows that this parabola is incorrect and therefore another parabola is provided to describe such a bottom boundary, which is seen to be reasonable.

2.5 Paper V: An evaluation of algorithms on generating multivariate non-normal data

Monte Carlo simulations requiring correlated data from non-normal populations are frequently used to investigate the small sample properties of competing statistics or the robustness of estimation methods. To guarantee the quality of the Monte Carlo simulation, a persuasive technique is required to generate multivariate non-normal distributions that can precisely mimic the properties of the desired distributions.

Only a few methods of generating multivariate non-normal distributions have been proposed. In this study, four algorithms to generate multivariate non-normal distributions are investigated, *i.e.*, Vale & Maurelli's extension of Fleishman's polynomial transformation to multivariate applications, Headrick & Sawilowsky's (1999) refinement for calculating the intermediate correlations based on Vale & Maurelli (1983), Headrick's

(2002) fifth-order polynomial transformation, and the iterative algorithm proposed by Ruscio & Kacetow (2008).

Vale & Maurelli's multivariate extension of Fleishman's method (*VM*)

Vale & Maurelli (1983) extended Fleishman's technique (see Fleishman, 1978) of generating univariate non-normal distributions to the multivariate case. In addition to the use of coefficients for each variable that yield the desired moments, the key step in this process is to take into account the effect of the non-normalizing transformations on the variable inter-correlations. Define the vector \mathbf{x} as

$$\mathbf{x}' = [1, X, X^2, X^3] . \quad (2.3)$$

The weight vector \mathbf{w}' contains the power function weights $a, b, c,$ and d is:

$$\mathbf{w}' = [a, b, c, d] , \quad (2.4)$$

then we have

$$Y = \mathbf{w}'\mathbf{x} . \quad (2.5)$$

Let $r_{Y_1Y_2}$ be the correlation between two non-normal variables Y_1 and Y_2 , and $\rho_{X_1X_2}$ be the correlation between two normal variables X_1 and X_2 , then

$$r_{Y_1Y_2} = E(Y_1Y_2) = E(\mathbf{w}'_1\mathbf{x}_1\mathbf{x}'_2\mathbf{w}_2) = \mathbf{w}'_1E(\mathbf{x}_1\mathbf{x}'_2)\mathbf{w}_2 = \mathbf{w}'_1\mathbf{R}\mathbf{w}_2 , \quad (2.6)$$

where

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & \rho_{X_1X_2} & 0 & 3\rho_{X_1X_2} \\ 1 & 0 & 2\rho_{X_1X_2}^2 + 1 & 0 \\ 0 & 3\rho_{X_1X_2} & 0 & 6\rho_{X_1X_2}^3 + 9\rho_{X_1X_2} \end{pmatrix} \quad (2.7)$$

is the expected matrix product of \mathbf{x}_1 and \mathbf{x}'_2 . Returning to scalar algebra, the relationship between $\rho_{X_1X_2}$ and $r_{Y_1Y_2}$ is:

$$r_{Y_1Y_2} = \rho_{X_1X_2}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) + \rho_{X_1X_2}^2(2c_1c_2) + \rho_{X_1X_2}^3(6d_1d_2) . \quad (2.8)$$

To generalize, we have

$$r_{Y_iY_j} = \rho_{X_iX_j}(b_ib_j + 3b_id_j + 3d_ib_j + 9d_id_j) + \rho_{X_iX_j}^2(2c_ic_j) + \rho_{X_iX_j}^3(6d_id_j) . \quad (2.9)$$

By setting the desired post-transformation correlation $r_{Y_iY_j}$, solving this polynomial for $\rho_{X_iX_j}$ provides the intermediate correlations.

For given values of the desired marginal skewness, kurtosis, and correlation matrix, we first determine the values of a_i , b_i , c_i , and d_i for each non-normal variable. The intermediate correlation matrix will be obtained by solving Equation 2.9 for each of the elements in the correlation matrix. Second, multivariate normal random numbers with a specified intermediate correlation matrix are generated and then univariately transformed to the desired shape.

Headrick & Sawilowsky's refinement for calculating the intermediate correlations (H99)

To generate multivariate *normal* distributions with the specified intermediate correlation matrix by the VM method, a preliminary step involves a principal components (or other factorization method) decomposition on the population correlation matrix. To avoid the factorization procedure, Headrick & Sawilowsky (1999) proposed a refined procedure to generate multivariate normal random numbers with the desired intermediate correlations.

Let Z_1 and V be normally distributed independent random variables with zero means and unit variances. Let E_1, \dots, E_p be a set of p normally distributed independent random variables with zero means and unit variances. Further, let

$$Z_{i+1} = r_0 Z_1 + \sqrt{1 - r_0^2} V, \quad (2.10)$$

where

$$i = \begin{cases} 0, & \text{if } r_0 = 1; \\ 1, & \text{if } r_0 < 1; \end{cases}$$

and r_0 is the correlation between Z_{i+1} and Z_1 . Let

$$X_j = r_j Z_1 + \sqrt{1 - r_j^2} E_j, \quad (2.11)$$

$$X_k = r_k Z_{i+1} + \sqrt{1 - r_k^2} E_k, \quad (2.12)$$

where r_j is the correlation between X_j and Z_1 , and r_k is the correlation between X_k and Z_{i+1} . Then the correlation between X_j and X_k is

$$\rho_{X_j X_k} = \begin{cases} r_j r_k, & \text{for } i = 0; \\ r_0 r_j r_k, & \text{for } i = 1. \end{cases}$$

As stated earlier, the only difference between the VM and the H99 approach is the procedure of generating a multivariate *normal* distribution with the specified intermediate correlation matrix. Hence, the relation-

ship between the post and intermediate correlation is still as listed in Equation 2.9.

Headrick's fifth-order polynomial transformation to achieve greater precision (*H02*)

To improve the precision of the approximation, Headrick (2002) proposed a method that extends Fleishman's polynomial transformation technique to the fifth order. The transformation is expressed as follows:

$$Y = c_0 + c_1X + c_2X^2 + c_3X^3 + c_4X^4 + c_5X^5, \quad (2.13)$$

where X is also normally distributed with zero mean and unit variance ($N(0, 1)$). The basic idea of this method is almost the same as we described for the H99 approach. However, to determine the the values of the six constants $c_{0i}, c_{1i}, c_{2i}, c_{3i}, c_{4i}$, and c_{5i} , six nonlinear equations are derived and the equation to calculate the intermediate correlations is much more complicated.

Ruscio & Kacetow's SI technique (*SI*)

Another alternative to the polynomial transformation methods is the technique proposed by Ruscio & Kacetow (2008). The main idea of this approach is to sample each variable directly from a non-normal population distribution and identify the intermediate correlation through an iterative, trial-and-error process.

The iterative nature of SI technique requires considerably more processing time than the other methods. However, several potential advantages of the SI technique can compensate for the disadvantage of the increase in processing time. First, it can handle distributions with undefined moments. Second, no limits are placed on the range of moments for the SI technique. It is also possible to generate data with a discrete distribution. Thus, many types of data commonly encountered in empirical research can be reproduced, such as binary and ordered categorical variables. Another important advantage of the SI technique is its capability of reproducing the characteristics observed in a unique sample of empirical data.

To evaluate the algorithms empirically, the data with population parameters γ_{1i}, γ_{2i} , and $r_{Y_iY_j}$ for 2, 3, and 4 variables multivariate sets are simulated by using different algorithms. The sample sizes are chosen to be 10, 20, 100, and 1000 in order to establish a range from extremely small to fairly large. The average values $\bar{\gamma}_{1i}, \bar{\gamma}_{2i}$, and $\bar{r}_{Y_iY_j}$ are computed based on $N \times 50000$ random deviates. To investigate the large sample properties of each algorithm additional sample sizes, $N = 10000$ and 100000 ,

are generated. We would like to see if there exists a uniformly better performed method on such criteria as bias and mean square errors of the sample correlations, estimated skewness, and kurtosis. In particular, the standard errors of the estimated marginal skewness, kurtosis, and correlations are of more interest in order to find a consistently well-performing method. Based on the generated data sets, the size and power of three well-known tests of multivariate normality are also compared.

According to the simulation results, we propose to use the H99 method for relatively small sample sizes ($N \leq 100$) and the SI technique for larger sample sizes ($N > 100$). The H99 method performs consistently well for all the experimental conditions. Furthermore, it is computationally parsimonious compared with the VM method since the factorization procedure is avoided and it is simpler to code in the programming language such as R. On the other hand, if a larger number of sample sizes with comparatively lower levels of correlation and dimensions are desired, the SI technique is highly recommended.

3. References

- Azzalini, A. and Capitano, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65:367–389.
- Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.
- D’Agostino, R. B. (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532.
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40(4):685 – 711.
- Headrick, T. C. and Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the fleishman power method. *Psychometrika*, 64(1):25–35.
- Jöreskog, K. G. (1990). New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24(4):387–404.
- Knapp, T. R. and Swoyer, V. H. (1967). Some empirical results concerning the power of bartlett’s test of the significance of a correlation matrix. *American Educational Research Journal*, 4(1):13–17.
- Lee, S., Poon, W., and Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, 9(1):91–97.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.
- Nešlehová, J. (2004). *Dependence of Non-Continuous Random Variables*. PhD thesis, Carl von Ossietzky Universität Oldenburg, Oldenburg.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460.

Ruscio, J. and Kacetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, 43(3):355–381.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, 45(2):273–279.

Vale, C. D. and Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3):465–471.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 66*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-149423



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2011