

UNIVERSITY OF CALIFORNIA

Los Angeles

**Multivariate Ordinal Data Analysis**  
**with Pairwise Likelihood and Its**  
**Extension to SEM**

A dissertation submitted in partial satisfaction  
of the requirements for the degree  
Doctor of Philosophy in Statistics

by

**Juanmei Liu**

2007

© Copyright by  
Juanmei Liu  
2007

The dissertation of Juanmei Liu is approved.

---

Weng Kee Wong

---

Yingnian Wu

---

Jan de Leeuw

---

Peter M. Bentler, Committee Chair

University of California, Los Angeles

2007

*To my grandma, my parents and my husband . . .*

## TABLE OF CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Log linear model	2
1.2	Limitations of log linear models	4
<b>2</b>	<b>GCM and CGCM</b>	<b>6</b>
2.1	Grouped Continuous Model	6
2.1.1	Polychoric correlation	6
2.1.2	Definition of GCM	8
2.2	CGCM	9
2.3	Parameter estimation with full likelihood	11
2.3.1	MLE with Fletcher-Powell	12
2.3.2	Bayesian MCMC	12
2.4	Parameter estimation with other approaches	14
2.4.1	GCM simultaneous estimation with two way marginals	14
2.4.2	GCM estimation with one and two way marginals	16
2.4.3	CGCM estimation	17
<b>3</b>	<b>Maximum Pairwise Likelihood</b>	<b>19</b>
3.1	Definition of pairwise likelihood	19
3.2	Maximum pairwise likelihood approach for GCM and CGCM	20
3.3	Parameter estimation of pairwise likelihood	21
3.4	Asymptotic theories on pairwise likelihood	21

3.5	MPLE in hypothesis testing . . . . .	24
3.5.1	Wald tests . . . . .	24
3.5.2	PLRT . . . . .	25
<b>4</b>	<b>SEM with ordinal data . . . . .</b>	<b>29</b>
4.1	SEM . . . . .	29
4.2	SEM with ordinal data . . . . .	30
4.3	SEM with MPL - GLS . . . . .	31
<b>5</b>	<b>Simulation studies . . . . .</b>	<b>34</b>
5.1	First stage simulation . . . . .	34
5.1.1	A small simulation of first stage estimation . . . . .	34
5.1.2	A larger first stage estimation . . . . .	36
5.2	Second stage estimation: structural parameter estimates . . . . .	38
5.2.1	A small simulation on second stage estimation . . . . .	38
5.2.2	Simulation I: parameter estimation . . . . .	39
5.2.3	Simulation II: test statistic . . . . .	42
5.2.4	Power simulation . . . . .	43
5.2.5	A simulation on CGCM . . . . .	44
<b>6</b>	<b>Extension to multiple group . . . . .</b>	<b>47</b>
6.1	Model identification . . . . .	49
6.2	Parameter specification in a two population model: an example . . . . .	52
6.2.1	Group 1(reference group) parameters $\alpha^1, \sigma^1$ . . . . .	52

6.2.2	Group 2 parameters: $\alpha^{(2)}, \sigma^{(2)}$	53
6.3	Simulations: comparison of group difference	54
6.3.1	Model 1	55
6.3.2	Model 2	57
6.3.3	Model 3	59
6.4	Simulation: multiple groups parameter estimates	59
<b>7</b>	<b>Application</b>	<b>63</b>
7.1	Application in biomedical studies: clinical trials	63
7.2	Social science: linear model and anova	64
<b>8</b>	<b>Summary and conclusion</b>	<b>66</b>
<b>9</b>	<b>Appendix</b>	<b>67</b>
9.1	GCM gradient and expected hessian	67
9.1.1	Objective function $l_p(\theta)$	67
9.1.2	Gradient: $\frac{\partial l_p(\theta)}{\partial \theta}$	68
9.1.3	Expected hessian matrix $H(\theta)$	69
9.1.4	An example for expected hessian matrix	70
9.1.5	Observed Hessian Matrix $H_o(\theta)$	71
9.1.6	Observed variance-covariance matrix $V$	73
9.1.7	Estimated variance-covariance matrix $V$	73
9.2	CGCM gradient and hessian	75
9.2.1	Standardize $\mathbf{X}$	75

9.2.2	Objective function $l_p(\theta^*)$ . . . . .	75
9.2.3	Gradient $\frac{\partial l_p(\theta^*)}{\partial \theta^*}$ : . . . . .	76
9.2.4	Estimated Hessian Matrix $H(\theta^*)$ . . . . .	77
9.2.5	Transformation . . . . .	79
9.2.6	Estimated variance-covariance matrix $V$ . . . . .	79
9.3	Multiple groups gradient calculation . . . . .	81
9.3.1	$\partial \sigma_{t0}^{nr} / \partial \theta'$ . . . . .	82
9.3.2	$\partial \sigma_t^{nr} / \partial \theta'$ . . . . .	82
<b>10</b>	<b>References</b> . . . . .	<b>83</b>

## LIST OF TABLES

5.1	Mean estimates of MLE, Partition estimates and MPLE . . . . .	36
5.2	Comparison of RMSE of MLE, Partition estimates and MPLE. . . . .	37
5.3	Comparison of MPL estimates with marginal method . . . . .	38
5.4	Comparison of parameter estimates, se, and empirical s.e's. . . . .	39
5.5	MPL - GLS parameter estimates in SEM . . . . .	41
5.6	K-S test of T and empirical tail probability . . . . .	43
5.7	Rejection rate on power assessment . . . . .	44
5.8	MPL - GLS Parameter Estimates in SEM . . . . .	46
6.1	Model 1: MPL - GLS test statistic . . . . .	57
6.2	Model 1: Comparison of MPL-GLS with Mplus . . . . .	58
6.3	Model 2: MPL - GLS test statistic . . . . .	59
6.4	Model 3: MPL - GLS test statistic . . . . .	60
6.5	Estimates of MPL-GLS with Mplus . . . . .	62

## ACKNOWLEDGMENTS

First and most, no appreciation is enough to express my gratitude to my advisor, mentor and supporter, Peter M. Bentler. An excellent advisor, he keeps enlightening me how to think it correctly, do it correctly and opening any professional opportunity for me. Without his three years of continuous support, I might not be able to come this far.

Many thanks to David Sookne for his valuable work in the implementation of my algorithms into EQS; and to Eric Wu, EQS specialist, for his patience, availability in helping me with my simulation study on EQS. They have been an integral part of the projects described in this dissertation.

Special thanks to Dr. Yingnian Wu, Dr. Weng Kee Wong and Dr. Jan de Leeuw for being my doctoral committee members and their time and efforts on reviewing this dissertation.

## VITA

1975            Born, Cangzhou, Hebei, P.R. China

1997            B.S. (Mathematics), Hebei Normal University, P.R.China

2002            M.S. (Mathematics), University of Nebraska, Lincoln.

2005            M.S. (Statistics), UCLA.

## PUBLICATIONS AND PRESENTATIONS

J. Liu and P.M. Bentler (2007) "A simultaneous estimation methodology for multivariate ordinal data in SEM". Proceedings of 2007 American Psychological Association Convention.

J. Liu and P. M. Bentler (2007) "Binary clinical trials data analysis using pairwise likelihood". Proceedings of 2007 Joint Statistical Meetings.

J. Liu and P.M. Bentler (2007) "SEM of ordinal data with pairwise likelihood". Proceedings of 2007 International Meeting of the Psychometric Society .

S. Shoptaw, A. Huber, J. Peck, X. Yang, J. Liu, J. Dang, J. Roll, B. Shapiro, E. Rotheram-Fuller and W. Ling (2006). "Randomized, placebo-controlled trial of

setraline and contingency management for the treatment of methamphetamine dependence". *Journal of Drug and Alcohol Dependence*, **85-1**: 12-18

Yang, X., Nie, K., Belin, T., Liu, J., and Shoptaw, S (2006) "Markov Transition Models for Binary Repeated Measures with Ignorable and Nonignorable Missing Values". *Journal of Statistical Methods in Medical Research*, **15**: 1-18

K.G. Heinzerling, S. Shoptaw, J. A Peck, X.Yang, J. Liu, J. Roll, W. Ling (2006) "Randomized, placebo-controlled trial of baclofen and gabapentin for the treatment of methamphetamine dependence". *Journal of Drug and Alcohol Dependence*, **85-3**: 177-184

Karlamangla, D. Black, E. Barrett-Connor, J. Liu, D. Kado, G. Greendale "Increases in Femoral Neck Strength Indices with Alendronate versus Placebo". Proceedings of 2004 Annual Meeting of the American Society of Bone and Mineral Research.

J. Liu and P. M. Bentler. 'A Robust Method Analyzing Ordinal Data in SEM'. Slide presentation at the 2007 Annual Meeting and Exhibition of American Educational Research Association.

J. Liu and P. M. Bentler "Maximum Pairwise Likelihood for Ordinal Data In SEM". Slide presentation at the 2006 Society of Multivariate Experimental Psychology Meeting.

ABSTRACT OF THE DISSERTATION

**Multivariate Ordinal Data Analysis  
with Pairwise Likelihood and Its  
Extension to SEM**

by

**Juanmei Liu**

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2007

Professor Peter M. Bentler, Chair

In many modern applications of statistical models, high-dimensional interdependencies may cause standard likelihood-based inference meets difficulties. High dimensional ordinal data, for instance, will encounter the problem of prohibitively large computational demands. This dissertation develops the statistical theory for a new multistage ordinal methodology in the context of structural equation modeling (SEM), based on a recently developed maximum pairwise likelihood method. Unlike earlier methods, the maximum pairwise likelihood approach maximizes an objective function based on the product of bivariate probabilities from any two different pairs of variables to estimate both thresholds, polychoric and polyserial correlations. The asymptotic distribution of these estimators is used to develop a second stage estimation and testing procedure for SEM based on generalized least squares, and a new goodness-of-fit statistic is obtained that is asymptotically chi-square distributed. Simulation studies to evaluate the performance of the proposed method are described and summarized.

This method is further extended to multiple groups. Additional identification conditions are presented and simulation studies on detection of group difference are also provided.

# CHAPTER 1

## Introduction

Categorical data consists of variables which have a finite number of values that can be nominal, ordinal, interval or ratio data. Examples of categorical variables are race, sex, age group, and educational level. The analysis of categorical data usually apply the use of contingency tables which present categorical data by counting the number of observations that fall into cells. The simplest contingency table is two-by-two table. However, the concepts in two-by-two table apply equally well to more general multi-way tables.

Very often, the categorical data are actually obtained by putting thresholds on continuous variables. For example, the final grade of a student in a statistics class is A if his/her grade lies between 90 and 100, and B if between 80 and 89, and so on. In social and medical science, data with ordered categories occur frequently, either with outcomes or covariates or both. Observations on an ordinal variable represent responses to a set of ordered categories, such as a five-category Likert scale. Scores 1, 2, 3, ... assigned to categories are commonly treated as if they have metric properties but this is wrong. The level of difference between the categories makes no sense. It is only assumed that a person who selected one category has more of a characteristic than if he/she had chosen a lower category, but how much more seems unknown to us. For example, a marketing research firm wants to investigate what factors influence the size of coffee (small, medium, large or extra large) that people order at a starbucks. These factors may include

the gender of the customer, whether or not cookies are also ordered, and age of the consumer. While the outcome variable, size of coffee, is obviously ordered, the difference between the various sizes is not consistent, which are 6, 10, 12 ounces, respectively.

Ordinal variables have finite values and thus not continuous variables and should not be treated as if they are. Means, variances, and covariances of ordinal variables have no meaning because they do not have origins or units of measurements. The only information we have are counts of cases in each cell of a multi-way contingency table. Until the late 1960s, contingency tables were typically applied and analyzed by calculating chi-square values to test the hypothesis of independence. In the case of more than two variables, the chi-squares for two-way tables were computed by researchers and then again for multiple sub-tables formed from them in order to determine if associations and/or interactions were taking place among the variables. The analysis of contingency table changed quite dramatically in the 1970's with the publication of a series of papers on log linear models. Main contribution include S. N. Roy (1956), the emergence of log-linear models in the 1960s and modern log-linear model era (1970s through present). A complementary historical overview was given by Agresti (2002).

## **1.1 Log linear model**

The log linear model is one popular approach that relies on odds ratios or Pearson's correlations as measure of ordinal data association. It is one of the specialized cases of generalized linear models for Poisson-distributed data and an extension of the two-way contingency table where the conditional relationship between two or more discrete, categorical variables is analyzed by taking the natural logarithm of the cell frequencies within a contingency table. Log linear

models are more commonly used to evaluate multiway contingency tables that involve three or more variables, although it can be used to analyze the relationship between two categorical variables. Therefore, log linear models only demonstrate association between variables since no distinction is made between independent and dependent variables.

In comparison, the logistic or logit regression model, which focus on the prediction of one response factor, is useful when the study is interested in the relationship between the categorical response variable and the categorical and/or continuous explanatory variables. Whereas log-linear models treat all variables symmetrically and attempt to model important associations among them. In this sense, the goal of log-linear models is to determine the patterns of dependence and independence among a set of variables. Moreover, log linear models provide a richer analysis of the structure of association among all factors, specially when there is no distinction between the response and explanatory variable.

The basic strategy in log linear modeling involves computing the observed and expected frequencies in the cross-tabulation of categoric variables. The models can then be assessed by evaluating the difference among the two types of frequencies. There are different types of model, such as multinomial model, poisson model or product binomial model, which depend on different data collection assumptions. The choice of a preferred model among different fitted models is typically based on a formal comparison of goodness-of-fit statistics, i.e.  $\chi^2$ , which is associated with models that are related hierarchically. Ideally, the preferred model should distinguish between the pattern of the variables in the data and sampling variability to provide a defensible interpretation.

## 1.2 Limitations of log linear models

Although log linear model is a popular model for the ordered categorical data, it has its limitations. First, when there are large number of variables, the inclusion of so many variables in log linear models often makes interpretation very difficult. Moreover, only a between subjects design may be analyzed. The frequency in each cell is independent of frequencies in all other cells. Even more, log linear model requires adequate sample size. You need to have at least 5 times the number of cases as cells in your data. For example, if you have a  $2 \times 2 \times 2$  table, then you need to have at least 40 cases. If you do not have the required amount of cases, then you need to increase the sample size or eliminate one or more of the variables. For all two-way associations, the expected cell frequencies should be greater than one, and no more than 20% should be less than five. Failing to meet this requirement will result in a reduced power to the point where analysis of the data is worthless, although the Type I error rate usually does not increase. The following strategies could be done in case of low expected frequencies:

- Add a constant, typically 0.5, to each cell. Although the type I error rate only improves minimally, power will drop.
- Collapse categories for variables with more than two levels.
- Delete variables that are least associated with other variables to reduce the number of cells.

However, as we can see, those strategies may change the raw data ( i.e., number of variables or categories is reduced) and thus associations between the variables can be lost. The power, as a result, will be reduced for testing those associations. Because of the limitations of log linear models, an alternative approach,

Grouped Continuous Model (GCM) and Conditional Grouped Continuous Model (CGCM) will be the focus of this research.

## CHAPTER 2

### GCM and CGCM

#### 2.1 Grouped Continuous Model

##### 2.1.1 Polychoric correlation

The Grouped Continuous Model (GCM) postulates the existence of an underlying multivariate normal distribution for the latent variables. It relies on polychoric correlations to model the correlation structure of the data. The tetrachoric correlation (Pearson, 1901), for binary data, and the polychoric correlation, for ordered-category data, are excellent ways to measure rater agreement. Polychoric correlation is an extension of tetrachoric correlation when the bivariate normal is polytomized. They estimate what the correlation between raters would be if ratings were made on a continuous scale; they are, theoretically, invariant over changes in the number or "width" of rating categories. The tetrachoric and polychoric correlations also provide a framework that allows testing of marginal homogeneity between raters. Thus, these statistics let one separately assess both components of rater agreement: agreement on trait definition and agreement on definitions of specific categories.

Another statistic measuring the strength of association or dependency between two categorical variables in a contingency table is Cramer's V (H. Cramer, 1999). Different from Pearson's correlation and Polychoric correlation, Cramer's

$V$  is a correlation coefficient that indicates the relationship among two categorical variables. It is similar to Pearson's correlation coefficient for two quantitative variables. Like Pearson's coefficient, Cramer's  $V$  ranges from -1 to 1, with 0 indicating no relationship and -1 or 1 indicating a perfect relationship. Also like Pearson's coefficient, the square of Cramer's  $V$  indicates the proportion of the total possible association (i.e., the maximum possible value of the chi-square statistic) that is present in the data.  $\phi$  coefficient is a special case of Cramer's  $V$ .

Despite these similarities, Cramer's  $V$  is more difficult to interpret for several reasons. First, the maximum possible association (chi-square) is related to the sample size and the number of levels of each categorical variable. Thus, just changing the definitions of the levels of a categorical variable can change Cramer's  $V$ . Also, small values of Cramer's  $V$  often correspond to quite large proportional differences between groups, so the proximity of  $V$  to 0 can be misleading. These problems suggest that it is more reasonable to compare proportions than it is to focus solely on Cramer's  $V$ . If  $V$  is used as a measure of a relationship, it should be a secondary index that is interpreted in conjunction with a discussion of proportional differences.

The benefit of applying polychoric correlations is that they do not restrict the correlation parameter space, since they are just the usual pairwise correlations between the continuous latent variables. Unlike Pearson's correlations and Cramer's  $V$ , the number of polychoric correlations does not increase with the number of levels of an ordinal variable, a common problem with using odds ratios.

### 2.1.2 Definition of GCM

The GCM was introduced by Anderson and Pemberton (1985). Let the ordinal vector  $\mathbf{Y}^T = (y_1, \dots, y_Q)$  be observed  $Q$  dimensional outcome vector which is of interest, where  $y_q$  indicates the  $q$ th variable, which has  $l_q + 1$  levels:  $\eta_q^1 < \eta_q^2 < \dots < \eta_q^{l_q+1}$ ,  $q = 1, \dots, Q$ , where  $\eta_q^m$  is a possible value which  $y_q$  may take,  $m = 1, \dots, l_q + 1$ . It is assumed that underlying  $\mathbf{Y}$  is the unobservable continuous latent vector  $\mathbf{Z}^T = (z_1, \dots, z_Q)$ , which has the same dimension as  $\mathbf{Y}$ . With categorical  $\mathbf{Y}$  variables, the scale of the latent response variable  $\mathbf{Z}$  is indeterminate (Muthen, 1984). Without loss of generality, let  $\mathbf{Z}$  follows a normal distribution with mean  $\mathbf{0}$  and correlation matrix  $\mathbf{R}$ . i.e.,  $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{R})$ , where  $\mathbf{R}$  is a  $Q \times Q$  correlation matrix. The relationship between  $y_q$  and  $z_q$  is:  $y_q = \eta_q^i \iff \alpha_q^{i-1} < z_q \leq \alpha_q^i$ , where  $\alpha_q^i$  is the  $i$ th threshold for the variable  $z_q$ . For easy notation, define  $\alpha_q^0 = -\infty$  and  $\alpha_q^{l_q+1} = +\infty$ ,  $i = 1, 2, \dots, l_q + 1$ ,  $q = 1, \dots, Q$ . Without loss of generality, it is assumed that  $\eta_q^i = i$ , for  $i = 1, 2, \dots, l_q + 1$ .

Let  $\mathbf{I}^T = (i_1, \dots, i_Q)$  be a possible value of  $\mathbf{Y}$ . Then

$$Pr(\mathbf{Y} = \mathbf{I}) = \int_{\alpha_1^{i_1-1}}^{\alpha_1^{i_1}} \dots \int_{\alpha_Q^{i_Q-1}}^{\alpha_Q^{i_Q}} f(\mathbf{Z}) d\mathbf{Z} \quad (2.1)$$

where  $f(\cdot)$  is the standard multivariate normal density of  $\mathbf{z}$ . This gives the grouped continuous multivariate ordinal probability for  $Pr(\mathbf{y}_i)$ . The parameters of this model include the thresholds and the upper (or lower) triangle elements  $\mathbf{r}$  of  $\mathbf{R}$ , with  $\mathbf{r} = \text{vech}(\mathbf{R})$ , which may be estimated by maximizing the log likelihood function:

$$l(\theta) = \sum_{\mathbf{i}} n(\mathbf{i}) \log \left( \int_{\alpha_1^{i_1-1}}^{\alpha_1^{i_1}} \dots \int_{\alpha_Q^{i_Q-1}}^{\alpha_Q^{i_Q}} f(\mathbf{Z}, \mathbf{R}) d\mathbf{Z} \right) \quad (2.2)$$

where  $n(\mathbf{i})$  is number of observations with  $\mathbf{Y} = \mathbf{I}$ .

Besides the fact that the number of polychoric correlations depends only on the dimension of ordinal vectors, but not on the number of levels of an ordinal variable which will increase the number of Pearson's correlation. Another advantage of GCM is that it can be extended very easily to mixed data with ordinal and continuous variables, whose latent variables are assumed to be jointly normal distributed. This is called Conditional Grouped Continuous Model (CGCM) because we see in the following section that the conditional distribution of latent variables given the continuous vector is being modeled.

## 2.2 CGCM

CGCM is a model which extends the GCM to include continuous outcome variables. Consider a vector  $\mathbf{S}^T = (s_1, \dots, s_C)$  of continuous variables in addition to  $\mathbf{Y}$ , such that  $\mathbf{S}$  and  $\mathbf{Z}$  are jointly normally distributed with  $E(\mathbf{S}) = \mu$ ,  $\text{var}(\mathbf{S}) = \Sigma_{\mathbf{SS}}$ , and  $\text{cov}(\mathbf{S}, \mathbf{Z}) = \Sigma_{\mathbf{SZ}}$ . That is,

$$\begin{pmatrix} \mathbf{S} \\ \mathbf{Z} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{SS}} & \Sigma_{\mathbf{SZ}} \\ \Sigma_{\mathbf{SZ}}^T & \mathbf{R} \end{pmatrix}\right)$$

Therefore, the conditional distribution of  $\mathbf{Z}$  given  $\mathbf{S}$ ,  $\mathbf{Z}|\mathbf{S}$ , is multivariate normal distributed with mean  $\mu_{\mathbf{Z}|\mathbf{S}} = \Sigma_{\mathbf{SZ}}^T \Sigma_{\mathbf{SS}}^{-1}(\mathbf{S} - \mu)$  and covariance matrix  $\Sigma_{\mathbf{Z}|\mathbf{S}} = \mathbf{R} - \Sigma_{\mathbf{SZ}}^T \Sigma_{\mathbf{SS}}^{-1} \Sigma_{\mathbf{SZ}}$ :

$$\mathbf{Z}|\mathbf{S} \sim N(\mu_{\mathbf{Z}|\mathbf{S}}, \Sigma_{\mathbf{Z}|\mathbf{S}})$$

Rewrite  $\Sigma_{\mathbf{Z}|\mathbf{S}} = \mathbf{D}\mathbf{R}^*\mathbf{D}$ , where  $\mathbf{D}$  is a diagonal matrix with diagonal elements be the inverse square root of the diagonal elements of  $\Sigma_{\mathbf{Z}|\mathbf{S}}$  and  $\mathbf{R}^*$  is the corresponding correlation matrix. Thus,  $[\mathbf{D}^{-1}\mathbf{Z} - \mathbf{B}(\mathbf{S} - \mu)]$  given  $\mathbf{S}$  is standard-

ized multivariate normal with mean  $\mathbf{0}$  and correlation matrix  $\mathbf{R}^*$ , where  $\mathbf{B} = \mathbf{D}^{-1}\Sigma_{\mathbf{SZ}}^T\Sigma_{ss}^{-1}$ .

$$[\mathbf{D}^{-1}\mathbf{Z} - \mathbf{B}(\mathbf{S} - \mu)]|\mathbf{S} \sim N(\mathbf{0}, \mathbf{R}^*)$$

Now we can apply equation (2.1) to calculate the conditional grouped continuous multivariate ordinal probability for  $Pr(\mathbf{Y}|\mathbf{S})$ :

$$Pr(\mathbf{Y} = \mathbf{I}|\mathbf{S}) = \int_{v_1^{i_1-1}}^{v_1^{i_1}} \cdots \int_{v_Q^{i_Q-1}}^{v_Q^{i_Q}} f(\mathbf{Z}^*, \mathbf{R}^*) d\mathbf{Z}^* \quad (2.3)$$

where  $\mathbf{Z}^* = [\mathbf{D}^{-1}\mathbf{Z} - \mathbf{B}(\mathbf{S} - \mu)]$  is the transformed variables from  $\mathbf{Z}$  and  $\mathbf{R}^*$  in this context is called the conditional polychoric correlation matrix of  $\mathbf{Y}$ .

The log likelihood function is:

$$l(\theta) = \log Pr(\mathbf{S}) + \log Pr(\mathbf{Y}|\mathbf{S}) = l_1(\theta_1) + l_2(\theta_2) \quad (2.4)$$

where  $\log Pr(\mathbf{S})$  is the log likelihood function for the continuous vector  $\mathbf{S}$ , and  $\log Pr(\mathbf{Y}|\mathbf{S})$  is the log likelihood function for observed ordinal vector of  $\mathbf{Y}$  given the observed continuous vector  $\mathbf{S}$ .  $\theta_1 = \{\mu, \Sigma_{ss}\}$  is the mean and covariance matrix of  $\mathbf{S}$ .  $\theta_2$  include also the mean and covariance matrix of  $\mathbf{S}$ , the thresholds for  $\mathbf{Y}$ ,  $\alpha$ , the polychoric correlations for  $\mathbf{Y}$ ,  $\mathbf{R}$  and the polyserial correlations between  $\mathbf{S}$  and  $\mathbf{Y}$ ,  $\Sigma_{\mathbf{yz}}$ . i.e.,  $\theta_2 = \{\mu^T, \text{vech}(\Sigma), \alpha_q^i, \text{vech}(\mathbf{R}), \text{vech}(\Sigma_{\mathbf{yz}})\}$ . Note,  $d_q$  is the qth diagonal element of  $\mathbf{D}$  and  $\beta_q^T$  is the qth row of  $\mathbf{B}$ , denoting the polyserial correlation between  $y_q$  and  $\mathbf{S}$ .

It is extremely complicated to do the maximization because  $\theta_2$  also include elements of  $\theta_1$  and we can't maximize  $l_1(\theta_1)$  and  $l_2(\theta_2)$  separately. Usually, and throughout literature, a transformation of parameters is required (Poon and Lee, 1987). Define  $\gamma_q^i = v_q^i + \beta_q^T \mathbf{x}$ , where  $v_q^i = \alpha_q^i/d_q + \beta_q^T \mu$ . Here  $d_q$  is the qth diagonal

element of  $\mathbf{D}$  and  $\beta_q^T$  is the  $q$ th row of  $\mathbf{B}$ , and  $\delta_q^0 = -\infty$ ,  $\delta_q^{l_q+1} = +\infty$ . Maximizing the likelihood in terms of the standardized parameters usually will significantly simplify the estimation problem. Then,  $\theta_2^* = \{\gamma_q^i, \text{vech}(\mathbf{R}^*), \text{vec}(\mathbf{B})\}$ .

After a transformation process, the new parameters in  $\theta_2^*$  will be in one to one correspondence with those in  $\theta_2$ , but do not include  $\theta_1$  any more. Therefore, the maximization becomes much easier as the two functions can be maximized separately:

$$\text{Max}_{\theta} l(\theta) \iff \text{Max}_{\theta_1} l(\theta_1) + \text{Max}_{\theta_2^*} l(\theta_2^*)$$

It is obvious that the sample mean  $\bar{\mathbf{X}}$  and sample covariance matrix  $\mathbf{S}_{\text{SS}}$  are the maximum likelihood estimator (MLE). The computation task concentrates on the second part,  $\text{Max}_{\theta_2^*} l(\theta_2^*)$ . Please note that when we standardize the conditional distribution of  $\mathbf{Y}|\mathbf{S}$ , the thresholds  $\alpha_q^i$  are also standardized as  $\delta_q^i = \alpha_q^i/d_q + \beta_q^T(\mu - \mathbf{Z})$ . Also standardized are the polychoric and polyserial correlations. Note here  $\beta_q$ ,  $q = 1, \dots, Q$ , represent the polyserial correlations between  $\mathbf{S}$  and  $\mathbf{Y}$ . From this we can construct the likelihood function to estimate the parameters

$$l(\theta_2^*) = \sum_{i=1}^N \log \int_{v_1^{i_1-1}}^{v_1^{i_1}} \cdots \int_{v_Q^{i_Q-1}}^{v_Q^{i_Q}} f(\mathbf{z}^*, \mathbf{R}^*) d\mathbf{z}^* \quad (2.5)$$

After  $\theta_2^*$  is estimated, it is necessary to transform it back to the original parameter  $\theta_2$ .

### 2.3 Parameter estimation with full likelihood

As we can see from above, to maximize the likelihood functions of GCM and CGCM for parameter estimation, the straightforward maximum likelihood method involves multiple integrals of the multivariate normal distribution functions. Ob-

viously, this will require a lot of computations (Lee, Poon and Bentler, 1990a). To avoid direct calculation of multiple integrals, there is currently a lot of ongoing work regarding different approximations, which include Gauss-Hermite quadrature, adaptive quadrature, Monte Carlo methods, Laplace approximation, and most popular approximation: Bayesian MCMC methods.

### 2.3.1 MLE with Fletcher-Powell

Since second derivative of the log likelihood is very complicated to derive, the Fletcher-Powell algorithm, which involves only the first derivative, was used to calculate the MLE (Lee and Poon, 1986):

$$\frac{\partial \Phi_Q(\alpha_1, \dots, \alpha_Q; \mathbf{R})}{\partial \alpha_i} = \phi(\alpha_i) \Phi_{Q-1}\left(\dots, \frac{\alpha_j - \rho_{ij} \alpha_i}{(1 - \rho_{ij}^2)^{1/2}}, \dots; \mathbf{R}_i\right) \quad (2.6)$$

But as we see from the equation above, the derivative of the log likelihood still involves Q - 1 dimension of multiple integrals. Therefore, when the dimension Q of the ordinal vector Y is large, a computation problem exists.

### 2.3.2 Bayesian MCMC

Markov Chain Monte Carlo (MCMC, dynamic Monte Carlo) was introduced by Metropolis et.al in 1950s in the field of statistical physics, and has been used in various field of physics such as studies in liquid, magnets, and lattice gauge theory. In 1980s, it was introduced in statistical science, and after 1990s the combination of MCMC and hierarchical Bayesian models become a standard tool for advanced data analysis. Now MCMC is also applied to the other fields such as computer graphics and becoming recognized as a universal method for sampling from multivariate distributions with unknown normalization constants.

In a standard Bayesian analysis, one must calculate the posterior distribution of the unknown parameters of interest. The posterior distribution is proportional to the product of prior information and the likelihood function

$$p(\theta|Y) \propto p(\theta)p(Y|\theta) \tag{2.7}$$

$p(Y|\theta)$  involves multi-dimensional integrals. Instead of working directly with  $p(Y|\theta)$ , the idea of data augmentation (Tanner and Wong, 1987) was applied here. These observed data matrices  $Y$  will be augmented with the latent data matrix  $Z$  in the posterior analysis. To analyze the joint posterior distribution  $p(\theta, Z|Y)$ , a sequence of  $\theta$ ,  $Z$  sampled from the posterior distribution will be generated via the Gibbs sampler (Geman and Geman, 1984) as follows: At the  $j$ th iteration with current values  $\theta(j)$ ,  $Z(j)$ :

- generate  $\theta(j + 1)$  from  $p[\theta|Z(j), Y]$
- generate  $Z(j + 1)$  from  $p[Z|Y, \theta(j + 1)]$

Under some mild regularity conditions, Geman and Geman (1984) showed that the joint density of the observation  $[\theta(j), Z(j)]$  geometrically converges in distribution to the posterior density of  $(\theta, Z|Y)$ . Hence, for sufficiently large  $j$ , say  $J$ ,  $[\theta(J), Z(J)]$  can be regarded as an observation from the joint posterior distribution  $(\theta, Z|Y)$ . The rate of convergence depends on many factors, such as starting values. For more details, please refer to Song and Lee (Song and Lee, 2002).

Yet, due to polytomous variables, it requires to simulate observations from a multivariate truncate normal distribution. This is a well-known difficult problem in statistical computing. Moreover, full likelihood analysis would involve the  $Q$ -dimensional normal integral and that quite apart from any computational

difficulties there might be fears about the robustness of the specification as it involves higher order integrals (Cox and Reid, 2004).

## 2.4 Parameter estimation with other approaches

Yet, instead of approximation of full likelihood, the widely used methods now separate the full likelihood into several parts as several objective functions. e.g., Lee, Poon and Bentler (1990b), Muthen (1984). By doing this, the problem of integration of high dimensional normal integrals at the heart of the iterative procedure is avoided. There are mainly two approaches.

### 2.4.1 GCM simultaneous estimation with two way marginals

For ordinal vector  $Y$  with  $Q$  dimensions, this approach will first divide  $Q$  ordinal variables into  $\frac{Q(Q-1)}{2}$  different pairs. Then for each pair, which can be seen as in the form of two way contingency table, apply Olsson (1979) approach.

Olsson (1979) estimated the polychoric correlation in a two way contingency table, which became the basis for most of the later work on more general models. For ordered variables  $y_1$  and  $y_2$ , which has  $s$  and  $r$  categories respectively, let  $n_{ij}$  and  $\pi_{ij}$  be number of observations and probability with  $y_1 = i$  and  $y_2 = j$ . Also let  $\rho$  be the polychoric correlation between  $y_1$  and  $y_2$ . Then the log-likelihood is:

$$l = \sum_{i=1}^r \sum_{j=1}^s n_{ij} \ln \pi_{ij} \tag{2.8}$$

One approach to estimate the parameters is to simultaneously estimate the parameters  $\rho$ , the polychoric correlation coefficient, and  $a_1, \dots, a_{r-1}, b_1, \dots, b_{s-1}$ , thresholds for  $y_1, y_2$ . Parameter estimates are obtained via the following three

equations from partial derivatives with respect to the parameters:

$$\frac{\partial l}{\partial \rho} = \sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}}{\pi_{ij}} \{ \phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1}) \} \quad (2.9)$$

$$\frac{\partial l}{\partial a_k} = \sum_{j=1}^s \left( \frac{n_{kj}}{\pi_{kj}} - \frac{n_{(k+1)j}}{\pi_{(k+1)j}} \right) \phi_1(a_k) \left\{ \Phi_1 \frac{(b_j - \rho a_k)}{(1 - \rho^2)^{1/2}} - \Phi_1 \frac{(b_{j-1} - \rho a_k)}{(1 - \rho^2)^{1/2}} \right\} \quad (2.10)$$

$$\frac{\partial l}{\partial b_m} = \sum_{i=1}^r \left( \frac{n_{im}}{\pi_{im}} - \frac{n_{i,m+1}}{\pi_{i,m+1}} \right) \phi_1(b_m) \left\{ \Phi_1 \frac{(a_i - \rho b_m)}{(1 - \rho^2)^{1/2}} - \Phi_1 \frac{(a_{i-1} - \rho b_m)}{(1 - \rho^2)^{1/2}} \right\} \quad (2.11)$$

Further, the asymptotic covariance matrix  $V$  of the parameter estimates  $\hat{\theta}$  are obtained from  $V = I_{\hat{\theta}}^{-1}$ , where  $\hat{\theta}$  is the vector of parameter estimates and matrix  $I_{\hat{\theta}}$  is expected second derivatives with respect to  $\theta$  which is obtained from

$$[I_{(\hat{\theta})}]_{m,n} = N \sum_{i=1}^r \sum_{j=1}^s \frac{1}{\pi_{ij}} \left( \frac{\partial \pi_{ij}}{\partial \theta_m} \right) \left( \frac{\partial \pi_{ij}}{\partial \theta_n} \right) \quad (2.12)$$

Olsson's work involves only 2 ordinal variables. For more general data with large dimension  $Q$  ( $Q > 2$ ) of polytomous variable, a suboptimal estimation procedure based on the observed two-way marginal totals was developed by Anderson and Pemberton (1985) in dealing with a set of ordinal variables. This procedure divides the several ordinal variables into any two different pairs. For each different pair of total  $\frac{Q(Q-1)}{2}$  pairs, the above two way contingency table estimation is applied to get one polychoric correlation and thresholds for the two ordinal variables.

### 2.4.2 GCM estimation with one and two way marginals

To further relieve the computational labor in the two way table, a second approach proposed by Olsson (Olsson, 1979) was to first estimate the thresholds from marginals, i.e.,  $a_i = \Phi_1^{-1}(P_{i.})$  and  $b_j = \Phi_1^{-1}(P_{.j})$ , where  $P_{i.}$  and  $P_{.j}$  are observed cumulative marginal proportions of the table. Subsequently, these estimates of thresholds, considered as known, are then plugged into equation (2.9). The asymptotic covariance matrix of the parameter estimates  $\hat{\theta}$  is also of the form  $(\frac{\partial F}{\partial \theta})^{-1}(\frac{\partial F}{\partial P})\Sigma(\frac{\partial F}{\partial P})'(\frac{\partial F}{\partial \theta})^{-1}$ , where F is system of equations to get the parameter estimates, P is the vector of cell probabilities, the covariance matrix of which is  $\Sigma$ .

A Newton-Raphson algorithm was utilized by Olsson to solve the equations. He also compared the results obtained from the above two methods, full maximum likelihood and the two-step method, and the difference seems to be small.

For a large number Q ( $Q > 2$ ) of ordinal variables, Anderson and Pemberton generalized Olsson's work based on the observed one- and two-way marginal totals. In this approach, through the separation of all Q ordinal variables into many different two-way contingency tables, threshold estimation is based on one-way univariate marginal totals, while the polychoric correlation estimates are based on two-way bivariate marginal totals treating threshold estimates as known.

Consider a sample of N observations taken from the joint probability  $Pr(\mathbf{y} = \mathbf{i})$ . Let  $n(\mathbf{i})$  be number of observations with  $\mathbf{y} = \mathbf{i}$ . Define  $n(i|r)$  the one way marginal totals for  $y_r$ . Since the marginal distribution of  $Z_r$  is  $N(0,1)$ , therefore the marginal probabilities for  $y_r$  are given by

$$P(i|r) = Pr(y_r = i) = \Phi[\alpha_r^i] - \Phi[\alpha_r^{i-1}], i = 1, \dots, l_r + 1 \quad (2.13)$$

where  $\Phi[\cdot]$  is the standard normal cumulative distribution function.

Thus, the marginal log-likelihood for the  $r$ th categorical variable  $y_r$ , is

$$l_r[\alpha_r] = \sum_{i=1}^{l_r+1} n(i|r) \log[P(i|r)] \quad (2.14)$$

Maximizing above equation with respect to  $\alpha_r = \alpha_r^1, \dots, \alpha_r^{l_r}$ , this gives marginal maximum likelihood estimates in form of a probit function,

$$\hat{\alpha}_r^m = \Phi^{-1}[(1/N) \sum_{i=1}^m n(i|r)], m = 1, \dots, l_r. \quad (2.15)$$

Let  $\rho_{rs}$  be the  $r$ th row and  $s$ th column element of the correlation matrix  $R$ . To estimate  $\rho_{rs}$ , define  $n(i, j|r, s)$  the two way marginal totals for  $y_r$  and  $y_s$ . The joint marginal probability for  $y_r$  and  $y_s$  is:

$$P(i, j|r, s) = Pr(y_r = i, y_s = j) \int_{\alpha_r^{i-1}}^{\alpha_r^i} \int_{\alpha_s^{j-1}}^{\alpha_s^j} f(x_1, x_2) dx_1 dx_2$$

where  $f(.,.)$  is the standardized bivariate normal density function. Hence the marginal log-likelihood for the  $(r,s)$  marginal table is

$$l_{rs}[\hat{\alpha}_r, \hat{\alpha}_s, \rho_{rs}] = \sum_{i=1}^{l_r+1} \sum_{j=1}^{l_s+1} n(i, j|r, s) \log[P(i, j|r, s)] \quad (2.16)$$

This is a function of  $\rho_{rs}$  and estimated  $\alpha_r$  and  $\alpha_s$ . Maximizing above equation with respect to  $\rho_{rs}$  after plugging in  $\hat{\alpha}_r, \hat{\alpha}_s$ , we can get the polychoric correlation estimate from this two-way marginal total.

### 2.4.3 CGCM estimation

For CGCM, which contains continuous vector  $S$ , the polyserial correlations should be estimated additionally. Currently there are mainly two approaches for CGCM.

In the first approach (Muthen, 1984), all thresholds are estimated from one way marginals. Treating the estimated thresholds known, the polychoric correlations are then estimated from the two way contingency table. This was exactly

the same as above simultaneous estimation with one and two way marginals. Each polyserial correlation is then estimated from the conditional distribution of ordinal variable on the continuous variable. This is approached by the standardization the continuous variables first.

In the second approach, a partition method (Poon and Lee, 1987) is applied.

- Based on the sample mean and covariance, for each ordinal component variable,  $y_i$ ,  $i=1, \dots, Q$ , the polyserial correlations between  $\mathbf{x}$  and  $y_i$  are estimated based on the observed random sample corresponding to  $(\mathbf{S}', y_i)$ . Since the dimension of  $y_i$  is one, the Newton-Raphson algorithm developed in Lee and Poon (1986) can be employed to get the maximum likelihood estimates. This gives the so-called partition maximum likelihood estimates of polyserial correlations and a set of thresholds estimates.
- For  $i, j = 1, \dots, Q$ ,  $i < j$ , the polychoric correlations  $\rho_{ij}$  between the observed frequencies corresponding to  $y_i$  and  $y_j$  are estimated based on above simultaneous 2-way marginals.

The asymptotic normality properties of the partition maximum estimates were developed in 1995 (Lee, Poon and Bentler, 1995).

The above mentioned non-full likelihood methods involve calculation of up to bivariate integrations, thus have the advantage of requiring much less computer time and effort than the maximum likelihood method. However, estimation of the parameters is done separately for several models with common parameters, which causes a loss of information because these partitions are not independent (de Leeuw, 1983). And since they are not simultaneous, it remains a problem on how to combine the multiple sets of threshold estimates to obtain the final estimates.

## CHAPTER 3

### Maximum Pairwise Likelihood

#### 3.1 Definition of pairwise likelihood

The pseudo likelihood, pairwise likelihood, has recently received a lot of attention. It originated from the composite likelihood approach (Lindsay, 1988) and has become popular in many fields. Kuk and Nott (2000) applied the pairwise likelihood analyzing clustered or longitudinal binary data. Fu et al (2000) applied it for the multivariate ordinal probit model to estimate the parameters including regression coefficients, thresholds, and polychoric correlations. They proceeded via a non-linear weighted least squares method. Heagerty and Lele (2000) proposed a pairwise distribution on binary spatial data. Cox and Reid (2004) further investigated the asymptotic properties and the conditions under which consistent estimators of parameters may obtain. Varin and Battisti (2006) showed how this methodology worked successfully on ordinal time series data within the class of autoregressive ordered probit models. They also emphasized its potential usefulness for methodology inference and model selection within more general classes of models.

The pairwise likelihood is a product of bivariate likelihoods for within cluster pairs of observations, and its maximizer is the Maximum Pairwise Likelihood Estimator (MPLE).

Let  $l_{ijk} = \log P_{ijk}(y_{ij}, y_{ik})$  denote the bivariate log-likelihood based on obser-

vations  $j \neq k$  in cluster  $i$ . By summing over all distinct pairs  $(j, k)$ ,  $j > k$  in cluster  $i$ , we obtain the pairwise log-likelihood based on cluster  $i$ ,  $l_i = \sum_{j=2}^{n_i} \sum_{k=1}^{j-1} l_{ijk}$ , where  $n_i$  is cluster size for the  $i$ th cluster. By summing over clusters, we obtain an over-all pairwise log-likelihood,  $l_p = \sum_{i=1}^M l_i$ , where  $M$  is the number of clusters.

The assumptions behind the pairwise likelihood approach are not strong. Only bivariate normality between each pair of variables is assumed.

### 3.2 Maximum pairwise likelihood approach for GCM and CGCM

De Leon (2005) proposed the maximum pairwise likelihood (MPL) approach to GCM and CGCM, and he did a small simulation to show the performance of the parameter estimator. This is a special case of the pairwise likelihood with equal cluster size. Let  $\theta$  be vector of parameters of this model, which include the thresholds and polychoric correlations, i.e,  $\theta = (\vec{\alpha}, \vec{r})'$ . The pairwise log-likelihood function for  $\theta$  is defined as:

$$l_p(\theta) = \sum_{i=1}^N \sum_{q < q'} l_{iqq'} \quad (3.1)$$

where  $l_{iqq'}$  is the bivariate pairwise log-likelihood function for  $y_q$  and  $y_{q'}$ .

Similarly, the pairwise log-likelihood function for  $\theta$  is defined as:

$$l(\theta) = \log Pr(\mathbf{S}) + \log Pr(\mathbf{Y}|\mathbf{S}) = l_1(\theta_1) + l_2(\theta_2) \quad (3.2)$$

and

$$l_p(\theta_2) = \sum_{i=1}^N \log \int_{v_1^{i_1-1}}^{v_1^{i_1}} \cdots \int_{v_Q^{i_Q-1}}^{v_Q^{i_Q}} f(\mathbf{z}^*, \mathbf{R}^*) d\mathbf{z}^* \quad (3.3)$$

### 3.3 Parameter estimation of pairwise likelihood

The parameter estimates can be obtained by maximizing  $l_p(\theta)$ . And the maximizer,  $\hat{\theta}_p$ , is called the Maximum Pairwise Likelihood Estimator (MPLE). Define the pairwise score equation  $S(\theta)$  as

$$S(\theta) = \dot{l}_p(\theta) = \sum_{i=1}^N \sum_{q < q'} \partial l_{iqq'} / \partial \theta = 0, \quad (3.4)$$

This equation can be solved via a modified Fisher scoring algorithm. Specially, the first and second derivative of  $l_p(\theta)$  are required for estimation. These are given in the appendix 9.1 (for GCM) and 9.2 (for CGCM).

### 3.4 Asymptotic theories on pairwise likelihood

Assuming that the partial derivative with respect to  $\theta$  can be passed under the integral sign, the pairwise likelihood function shares a common character with full likelihood:

**Theorem 1:**  $E[\dot{l}_p(\theta)] = 0$

Proof: 
$$E_\theta[\dot{l}_p(\theta)] = \sum_{j < k} E\left[\frac{\partial}{\partial \theta} \log Pr(X_j, X_k | \theta)\right]$$

$$= \sum_{j < k} \int \left[\frac{(\partial / \partial \theta) Pr(X_j, X_k | \theta)}{Pr(X_j, X_k | \theta)}\right] Pr(X_j, X_k | \theta) d\nu(jk)$$

Let  $\hat{\theta}_p$  be the maximum pairwise likelihood estimator (MPLE) of a data with sample size  $N$  and  $\theta_0$  be the true value of  $\theta$ . It follows from the standard theory of estimating equations that, under certain regularity conditions,  $\hat{\theta}_p$  is consistent and asymptotically normal.

**Theorem 2:**  $\sqrt{N}(\hat{\theta}_p - \theta_0) \longrightarrow N(0, \Delta)$ , where

$$\Delta = N[E(-\ddot{l}_p(\theta_0))]^{-1}[\sum_{i=1}^N E\dot{l}_{p_i}(\theta_0)\dot{l}_{p_i}(\theta_0)^T][E(-\ddot{l}_p(\theta_0))^{-1}]^T$$

Proof: First, taylor expansion around  $\theta_0$ :

$0 = \dot{l}_p(\hat{\theta}_p)\dot{l}_p(\theta_0) + \ddot{l}_p(\tilde{\theta})(\hat{\theta}_p - \theta_0)$ , where  $\tilde{\theta}$  is within neighborhood of  $\theta_0$ . Specially,  $\tilde{\theta} \rightarrow \theta_0$  when  $N \rightarrow \infty$ . Therefore,

$$\hat{\theta}_p - \theta_0 = (-\ddot{l}_p(\tilde{\theta}))^{-1}\dot{l}_p(\theta_0) \quad (3.5)$$

From Central Limit Theorem, and because  $E_{\theta_0}[\dot{l}_p(\theta_0)] = 0$ ,

$$\frac{1}{\sqrt{N}}\dot{l}_p(\theta_0) = \sqrt{N}\left(\frac{1}{n}\sum_{i=1}^N \dot{l}_{p_i}(\theta_0)\right) \rightarrow Z \in N(0, Var(\dot{l}_{p_i}(\theta_0))) \quad (3.6)$$

with  $Var(\dot{l}_{p_i}(\theta_0)) = E[\dot{l}_{p_i}(\theta_0)\dot{l}_{p_i}(\theta_0)^T] - E[\dot{l}_{p_i}(\theta_0)]E[\dot{l}_{p_i}(\theta_0)]^T = E[\dot{l}_{p_i}(\theta_0)\dot{l}_{p_i}(\theta_0)^T]$ , equality holds from Theorem 1.

Therefore,

$$\sqrt{N}(\hat{\theta}_p - \theta_0) = \left(-\frac{1}{N}\ddot{l}_p(\tilde{\theta})\right)^{-1}\frac{1}{\sqrt{N}}\dot{l}_p(\theta_0) \rightarrow \left(-\frac{1}{N}\ddot{l}_p(\tilde{\theta})\right)^{-1}Z \quad (3.7)$$

From Law of Large Numbers

$$\left[\frac{\ddot{l}_p(\tilde{\theta})}{N}\right] \rightarrow E[\ddot{l}_{p_i}(\theta_0)] \quad (3.8)$$

Combining above three equations,

$$\sqrt{N}(\hat{\theta}_p - \theta_0) = \left[-\frac{\ddot{l}_p(\tilde{\theta})}{N}\right]^{-1} \frac{1}{\sqrt{N}} \dot{l}_p(\theta_0) \longrightarrow N(0, \Delta)$$

$$\begin{aligned} \Delta &= [E(-\ddot{l}_{p_i}(\theta_0))]^{-1} E[\dot{l}_{p_i}(\theta_0) \dot{l}_{p_i}(\theta_0)^T] [E(-\ddot{l}_{p_i}(\theta_0))^{-1}]^T \\ &= N * [E(-\ddot{l}_p(\theta_0))]^{-1} [\sum_{i=1}^N E \dot{l}_{p_i}(\theta_0) \dot{l}_{p_i}(\theta_0)^T] [E(-\ddot{l}_p(\theta_0))^{-1}]^T \end{aligned}$$

Therefore,  $\hat{\theta}_p$  is consistent and asymptotically normal with asymptotic covariance matrix  $[E(-\ddot{l}_p(\theta_0))]^{-1} [\sum_{i=1}^n E \dot{l}_{p_i}(\theta_0) \dot{l}_{p_i}(\theta_0)^T] [E(-\ddot{l}_p(\theta_0))^{-1}]^T$ .

The pairwise likelihood approach for GCM and CGCM is more conceptually appealing because it entails maximizing a single objective function, the log-likelihood function, to obtain a single set of parameter estimates. Unlike the marginal estimation method, no extra work is needed to combine the several sets of parameter estimates. Moreover, the well developed statistical properties of parameter estimates are useful for further inference. Specially, the asymptotic normality and consistency property of MPLE allows it to do estimation under restricted parameters. This is an unique features of MPLE which make it distinct from other estimators introduced above.

## 3.5 MPLE in hypothesis testing

In a variety of modeling situations, it is hypothesized that parameters are restricted. However, this is not the case with MPLE. We will see below how the MPL method can do estimation under restricted parameters, to allow for hypothesis testing purpose.

### 3.5.1 Wald tests

For example, a psychological test theory application may inquire whether the thresholds for several variables are equal, or the intraclass correlation model requires that all correlations in  $R$  are equal. Because of normality of the MPLE distribution, Wald statistic can be applied here to test for different forms of restricted maximum pairwise likelihood models.

First, The simplest form of hypothesis states that the parameter are equal to specific values. That is,  $H_0: \theta = \theta_0$ , with  $\dim(\theta) = d$ . To test this hypothesis, we will construct the test hypothesis  $Z_{wald}$  as:

$$Z_{wald} = (\hat{\theta}_p - \theta_0)^T \Sigma^{-1} (\hat{\theta}_p - \theta_0) \quad (3.9)$$

where  $\hat{\theta}_p$  is MPLE and  $\Sigma$  is its covariance matrix. Under  $H_0$ ,  $Z_{wald}$  follows a  $\chi_d^2$  distribution. We reject  $H_0$  if it is too large.

Second, for the hypothesis which puts restriction only on part of the parameter. Let  $\theta_{p \times 1} = ((\theta_1)_{q \times 1}, (\theta_2)_{(p-q) \times 1})$ . The hypothesis is on  $\theta_2$ ,  $H_0: \theta_2 = \theta_{20}$ . Now define

$$Z_{wald} = (\hat{\theta}_{2p} - \theta_{20})^T \hat{\Sigma}_{22}^{-1} (\hat{\theta}_{2p} - \theta_{20}) \quad (3.10)$$

where  $\Sigma_{22}$  is its covariance matrix of  $\hat{\theta}_{2p}$ . Under  $H_0$ ,  $Z_{wald}$  follows a  $\chi_{p-q}^2$  distri-

bution. We reject  $H_0$  if it is too large.

Finally, for the generalized form of hypothesis,  $H_0: \eta(\theta_0) = \eta_0$ , where  $\eta: \Xi_{p \times 1} \rightarrow \Theta_{k \times 1}$ . Then

$$Z_{wald} = (\eta(\hat{\theta}_p) - \eta_0)^T \mathbf{V}^{-1} (\eta(\hat{\theta}_p) - \eta_0) \quad (3.11)$$

where  $\mathbf{V} = \zeta'(\theta_0)^T \Sigma \zeta'(\theta_0)$ , with  $\zeta'(\theta_0) = \frac{\partial \eta(\theta_0)}{\partial \theta}$ . Under  $H_0$ ,  $Z_{wald}$  follows a  $\chi_k^2$  distribution. We reject  $H_0$  if it is too large.

### 3.5.2 PLRT

We may also develop a parallel likelihood ratio test procedure, called a Pairwise likelihood ratio test (PLRT):

Let  $X_1, \dots, X_n$  be a sample from density  $f(x|\theta)$ , where  $\theta \in \Theta \subset R^k$ . To test  $H_0: \theta \in \Theta_0$  versus  $H_1: \theta \in \Theta - \Theta_0$  for a given subset  $\Theta_0$  of  $\Theta$ . Let  $\theta_p^*$  be the MPLE over  $\Theta_0$ , and  $\hat{\theta}_p$  be the MPLE over  $\Theta$ .

**Theorem 3:** Suppose  $H_0: \theta_1 = \theta_2 \dots = \theta_r$ , where  $1 \leq r \leq k$ . Suppose that the true value  $\theta_0$  satisfies  $H_0$ . Then

$$\lambda = -2[l_p(\theta_p^*) - l_p(\hat{\theta}_p)] \longrightarrow \sum_{i=1}^r \lambda_i \chi_{1i}^2 .$$

Proof: Expand  $l_p(\theta_p^*)$  about  $\hat{\theta}_p$ :

$$l_p(\theta_p^*) = l_p(\hat{\theta}_p) + \dot{l}_p(\hat{\theta}_p)(\theta_p^* - \hat{\theta}_p) - (\theta_p^* - \hat{\theta}_p)' \frac{\ddot{l}_p(\tilde{\theta})}{2} (\theta_p^* - \hat{\theta}_p)$$

Therefore,

$$\lambda = -2[l_p(\theta_p^*) - l_p(\hat{\theta}_p)] = (\theta_p^* - \hat{\theta}_p)^T \ddot{l}_p(\tilde{\theta})(\theta_p^* - \hat{\theta}_p) \quad (3.12)$$

Under simple  $H_0 : \theta = \theta_0$ , then  $\theta_p^* = \theta_0$  and

$$\lambda = \sqrt{n}(\hat{\theta}_p - \theta_0)^T \left[ \frac{\ddot{l}_p(\tilde{\theta})}{n} \right] \sqrt{n}(\hat{\theta}_p - \theta_0) \quad (3.13)$$

From Theorem 2,  $\sqrt{n}(\hat{\theta}_p - \theta_0) \longrightarrow N(0, \Lambda)$

From LLN,  $\frac{\ddot{l}_p(\tilde{\theta})}{n} \longrightarrow E[\ddot{l}_p(\theta_0)]$

Therefore,  $\lambda \longrightarrow \sum_{i=1}^q \lambda_i \chi_{1i}^2$ , where  $\lambda_i, i = 1, \dots, q$  are nonzero eigenvalues of  $\Lambda^{1/2} H \Lambda^{1/2}$ . It's easy to show that  $q = r$ , and  $H = E[\ddot{l}_p(\theta_0)]$ .

To find the asymptotic distribution of  $\sqrt{n}(\theta_p^* - \hat{\theta}_p)$  in general, expand  $\dot{l}_p(\theta_p^*)$  around  $\hat{\theta}_p$ :

$$\begin{aligned} \frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*) &= \frac{1}{\sqrt{n}} \dot{l}_p(\hat{\theta}_p) + \frac{1}{\sqrt{n}} \ddot{l}_p(\tilde{\theta})(\theta_p^* - \hat{\theta}_p) = \frac{\ddot{l}_p(\tilde{\theta})}{n} \sqrt{n}(\theta_p^* - \hat{\theta}_p) \\ &\sim E(\ddot{l}_p(\theta_0)) \sqrt{n}(\theta_p^* - \hat{\theta}_p) \end{aligned} \quad (3.14)$$

Thus,

$$\sqrt{n}(\theta_p^* - \hat{\theta}_p) \sim E(\ddot{l}_p(\theta_0))^{-1} \frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*) \quad (3.15)$$

and plug into (3.12):

$$\lambda \sim \frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*)^T E(\ddot{l}_p(\theta_0))^{-1} \frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*) \quad (3.16)$$

To find the asymptotic distribution of  $\dot{l}_p(\theta_p^*)$ , expand about  $\theta_0$ :

$$\frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*) = \frac{1}{\sqrt{n}} \dot{l}_p(\theta_0) + E(\ddot{l}_p(\theta_0)) \sqrt{n}(\theta_p^* - \theta_0) \quad (3.17)$$

Partition  $E(\ddot{l}_p(\theta_0))$  into four matrices,

$$E(\ddot{l}_p(\theta_0)) = \begin{pmatrix} r \times r & r \times (k-r) \\ G_1 & G_2 \\ (k-r) \times r & (k-r) \times (k-r) \\ G_2^T & G_3 \end{pmatrix}$$

and let

$$H = \begin{pmatrix} 0 & 0 \\ 0 & G_3^{-1} \end{pmatrix}$$

According to the property of  $\theta_p^* = (0, \dots, 0, \theta_p^{(*)r+1}, \dots, \theta_p^{(*)k}) = \operatorname{argmax}_{\theta \in \Theta_0} l_p(\theta)$ , the last  $k-r$  components of  $\dot{l}_p(\theta_p^*)$  are zero, so that  $H\dot{l}_p(\theta_p^*) = 0$ .

Multiply  $H$  on both sides of equation (3.17) and

$$H \frac{1}{\sqrt{n}} \dot{l}_p(\theta_0) \sim HE(\ddot{l}_p(\theta_0))\sqrt{n}(\theta_p^* - \theta_0) \quad (3.18)$$

since the first  $r$  components of  $\theta_p^*$  and  $\theta_0$  are zero.

Plug into equation (3.17), we find

$$\frac{1}{\sqrt{n}} \dot{l}_p(\theta_p^*) \sim [I - E(\ddot{l}_p(\theta_0))H] \frac{1}{\sqrt{n}} \dot{l}_p(\theta_0) \quad (3.19)$$

From central limit theorem,

$$\frac{1}{\sqrt{n}} \dot{l}_p(\theta_0) = \sqrt{n} \frac{1}{n} \dot{l}_p(\theta_0) \rightarrow Z \in N\left(0, \frac{\operatorname{Var}(\dot{l}_p(\theta_0))}{n}\right)$$

Hence,

$$\frac{1}{\sqrt{n}}\dot{l}_p(\theta_p^*) \rightarrow [I - E(\ddot{l}_p(\theta_0))H]Z$$

So that from Eq. (3.16),

$$\begin{aligned}\lambda &\rightarrow Z^T[I - E(\ddot{l}_p(\theta_0))H]^T E(\ddot{l}_p(\theta_0))^{-1}[I - E(\ddot{l}_p(\theta_0))H]Z \\ &= Z^T[E(\ddot{l}_p(\theta_0))^{-1} - H]Z,\end{aligned}$$

because  $HE(\ddot{l}_p(\theta_0))H = H$ .

Therefore,  $\lambda \rightarrow \sum_{i=1}^r \lambda_i \chi_{1i}^2$ , with  $\lambda_i, i = 1, \dots, r$  are nonzero eigenvalues of  $\frac{E[\dot{l}_p(\theta_0)\dot{l}_p(\theta_0)^T]}{n}[E(\ddot{l}_p(\theta_0))^{-1} - H]\frac{E[\dot{l}_p(\theta_0)\dot{l}_p(\theta_0)^T]}{n}$ .

As we mentioned, marginal estimators could not do the testing because it has no way to estimate the parameters while the parameters are restricted. This is a special feature of MPLE.

## CHAPTER 4

### SEM with ordinal data

#### 4.1 SEM

Structural Equation Modeling (SEM), which is also called mean and covariance analysis, is a very powerful multivariate analysis technique that integrates regression, factor analysis and path analysis.

It is usually approached by strictly confirmatory, model development and alternative models approach. For the strictly confirmatory approach, a model is tested using SEM goodness-of-fit tests to determine whether the pattern of variances and covariances in the data is consistent with the specified structural model. However, an accepted model is only one of the comfortable models out of which unexamined models may exist that fit the data as well or better; In alternative models approach one may test and compare two or more causal models to determine which has the better fit; For the model development approach, a first model is proposed and tested using SEM procedures, when found not fitted well, a second model based on changes suggested by SEM previous output are then fitted, and so on until a model is fitted. By doing this, over fitting problem may exist because the fitted model over fit the uniqueness of the data and thus unstable. A cross-validation strategy may be applied to overcome this problem by developing model using a calibration data sample and then using an independent validation sample to confirm.

Major applications of structural equation modeling include:

- Covariance and correlation structure models, which hypothesize that a covariance or correlation matrix has a particular form;
- Confirmatory factor analysis (CFA), which extends factor analysis so that the structure of the factor loading and intercorrelations can be hypothesized and tested
- Second order factor analysis in which the correlation matrix of the common factors is itself factor analyzed to provide second order factors
- Regression models which extend linear regression analysis to allow for constrained weights
- Causal modeling which hypothesizes causal relationships among variables, either manifest variables or latent variables or both, and tests the causal models with a linear equation system.

## 4.2 SEM with ordinal data

To use ordinal variables in structural equation models requires other techniques than those that are traditionally employed with continuous variables. Currently, most ordinal data analysis in SEM applied multiple stage estimation (Muthen, 1984; Lee, Poon and Bentler, 1990). The multiple stage estimation approached as following.

1. Obtain the thresholds, polychoric and polyserial estimates.
2. Obtain the asymptotic distribution of the above estimates.

3. Estimate the structural parameters via generalized least squares.

We already learned from section 2.4 how the first step was accomplished. That became the basis for ordinal data in SEM. Lee, Poon and Bentler (1995) extended the above mentioned partition method to estimation in structural equation models. It was proved (and has to satisfy) that the joint distribution of partition maximum estimates is asymptotical multivariate normal. And the asymptotic covariance matrix of the partition maximum estimates were obtained for structure parameter estimation. This method (LPB) has become an integral part of the EQS program (Bentler, 1995). Similarly, Mplus is (Muthen and Muthen, 2004) three stage limited information estimator that is proceeded by first obtaining the estimator in first step (see section 2.4). The explicit asymptotic covariance matrix was given in 1995 (Muthen and Satorra, 1995).

As we can see, the asymptotic consistency and normality of the first step estimator is very important and required for further extension into SEM. The speed of convergence to normal and consistent will affect the performance of SEM estimator. We have already seen in Chapter three that the MPLE has asymptotic properties which satisfy the requirement in SEM. Therefore we are interested in seeing how it works in SEM.

### 4.3 SEM with MPL - GLS

We extend the MPL model to include latent structures on the correlations, such as those that can arise from factor analysis on linear relations models (e.g., de Leeuw, 1983).

In SEM, let  $\gamma$  be a vector of more basic parameters such that  $\alpha = \alpha(\gamma)$ ,  $\mathbf{R} = \mathbf{R}(\gamma)$ , and define  $\mathbf{r} = \text{vech}(\mathbf{R})$ , where  $\dim(\gamma) < \dim(\alpha, \mathbf{r})$ . To test a correlation

structure hypothesis, the following MPL - GLS approach is proposed.

Stage I. Thresholds, polychoric and polyserial correlations are estimated simultaneously using the Newton Raphson's algorithm. The gradient and the second derivatives were given in Appendix. A weight matrix  $\mathbf{W}$  which converges in probability to  $\Lambda_C$ , the asymptotic covariance matrix of  $\hat{\mathbf{r}}_p$ , is also computed at this stage.

Stage II. The structural parameter vector  $\gamma$  in the structure  $\mathbf{r}(\gamma)$  is estimated by minimizing the generalized least squares function

$$Q(\gamma) = (\mathbf{r} - \mathbf{r}(\gamma))^T \mathbf{W}^{-1} (\mathbf{r} - \mathbf{r}(\gamma)) \quad (4.1)$$

Based on the above specifications, consider a sample of  $Q$  ordinal variables with sample size  $N$ , we have:

**Theorem 4.**

(i)  $\hat{\gamma}$  is asymptotically consistent and normally distributed with mean  $\gamma_0$  and covariance matrix  $[(\frac{\partial \mathbf{r}}{\partial \gamma})^T \mathbf{W}^{-1} \frac{\partial \mathbf{r}}{\partial \gamma}]^{-1} |_{\gamma=\gamma_0}$ .

(ii) The asymptotic distribution of  $T = Q(\hat{\gamma})$  is chi-squared with degrees of freedom  $Q(Q-1)/2 - q$ .

(iii) If  $\mathbf{W}$  is chosen other than the one which converges in probability to the asymptotic covariance matrix of  $\hat{\gamma}$ , then  $T$  is a mixture of chi-square variates with one degree of freedom.

(iv) If in (iii) all the eigenvalues are equal, then a rescaling of  $T$  gives an asymptotic chi-squared variate.

**Proof:** Since  $W$  was chosen to converge in probability to  $\Lambda_C$ , and  $\hat{\gamma}$  is the minimum  $\chi^2$  estimate, based on theory of Ferguson (Ferguson, 1996), (i) holds

automatically.

In practice, if  $\Lambda_C$  is nonsingular,  $\mathbf{W}$  can be chosen as  $\Lambda_C$ . In the case that  $\Lambda_C$  is singular,  $\mathbf{W}$  may be chosen so that  $\mathbf{W}^{-1}$  is a generalized inverse of  $\Lambda_C$ , i.e.,  $\Lambda_C \mathbf{W}^{-1} \Lambda_C = \Lambda_C$ .

In this case,  $T = Q(\hat{\gamma})$  follows a chi-square distribution with degrees of freedom  $Q(Q-1)/2 - q$ .

If  $\mathbf{W}$  is a positive definite matrix which does not converge to  $\Lambda_C$  in probability, then  $T \rightarrow \sum_{i=1}^q \lambda_i \chi_{1i}^2$ , where  $\lambda_i$ 's are nonzero eigenvalues of the  $\Lambda_c^{-1/2} W \Lambda_c^{-1/2}$ .

A special case of (iii) is when all the eigenvalues of  $\Lambda_c^{-1/2} W \Lambda_c^{-1/2}$  are equal, then  $T \rightarrow \sum_{i=1}^q \lambda_i \chi_1^2 = \lambda \sum_{i=1}^q \chi_1^2 \rightarrow \lambda \chi_q^2$ . Therefore, a rescaling of  $T$ ,  $\frac{T}{\lambda}$  is a chi-squared variate.

## CHAPTER 5

### Simulation studies

Compared with other multiple stage approaches, the MPL-GLS procedure may be more robust since it maximizes a single objective function in stage I. To evaluate this, several simulation studies were conducted using the MPL-GLS procedure. We began with the first stage parameter estimates and compared them with current methods. Later on, second stage model parameter estimates and standard errors were evaluated and the test statistic,  $T$  and its distribution were assessed. We proceed from small simulations to simulations including more and more variables. At the same time, the power of the test statistic was also considered.

#### 5.1 First stage simulation

##### 5.1.1 A small simulation of first stage estimation

We first conducted a small simulation study to assess the performance of maximum pairwise likelihood estimates. The aim of this small study is to get a general view of the MPLE performance, and find out if further a larger simulation is needed.

In this study, random samples with sample sizes  $N=100$  and  $N=500$  were generated from a 3-dimensional multivariate normal latent distribution with cor-

relation matrix  $R$ , hence  $Q = 3$ . The data  $z_1, \dots, z_N$  were then transformed into  $y_1, \dots, y_N$  using a set of pre-assigned thresholds with  $\alpha_{11} = -0.5$ ,  $\alpha_{12} = 0.5$ ,  $\alpha_{21} = -0.8$ ,  $\alpha_{22} = 0.6$ ,  $\alpha_{31} = 0$ ,  $\alpha_{32} = 1.2$  and

$$R = \begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.5 \\ 0.9 & 0.5 & 1 \end{pmatrix}$$

For each sample with sample sizes 50 and 100, the thresholds  $\alpha_{11}$ ,  $\alpha_{12}$ ,  $\alpha_{21}$ ,  $\alpha_{22}$ ,  $\alpha_{31}$ ,  $\alpha_{32}$  and polychoric correlations  $r_{12}$ ,  $r_{13}$  and  $r_{23}$  were estimated using the pairwise, partition and maximum likelihood method respectively. This was replicated for a total of  $R = 20$  times, and the average estimates were calculated for each estimation methods. The results are listed in Table 5.1. Table 5.2 reports the root mean-squared error  $RMSE = \sqrt{\sum_{i=1}^R (\hat{\theta}_i - \theta_i)^2 / R}$  to measure the accuracy of estimates.

As we can see from Table 1, even when the sample is as small as 100, all three methods are not far away from the true value. From Table 2, for all three methods the RMSE for both threshold and polychoric estimates decrease as the sample size increases from 100 to 500. Furthermore, the polychoric correlation estimates do not seem to be affected by extremeness of the true parameter values.

When sample size is 100, most MPLEs are slightly better than partition method, for those that are not, the difference is very small. When sample size is 500, all the estimates from the two approaches are exactly the same, except for a very slight difference on the estimates of polychoric correlations. Overall, from this small simulation, the MPL and partition estimates perform as well as the MLE in terms of both average bias and RMSE. This small simulation provides us

True Parameter	$N = 100$			$N = 500$		
	MLE	Partition	MPLE	MLE	Partition	MPLE
<i>Thresholds</i>						
$\alpha_{11}=-0.500$	-0.405	-0.465	-0.485	-0.473	-0.493	-0.493
$\alpha_{12}=0.500$	0.574	0.513	0.501	0.518	0.497	0.497
$\alpha_{21}=-0.800$	-0.775	-0.810	-0.808	-0.788	-0.801	-0.801
$\alpha_{22}=0.600$	0.669	0.627	0.632	0.622	0.611	0.611
$\alpha_{31} = 0$	0.082	0.010	-0.007	0.018	0.001	0.001
$\alpha_{32}=1.200$	1.289	1.229	1.221	1.234	1.205	1.205
<i>Polychoric correlations</i>						
$\rho_{12}=0.100$	0.120	0.099	0.084	0.100	0.101	0.100
$\rho_{13}=0.900$	0.911	0.916	0.919	0.898	0.896	0.918
$\rho_{23}=0.500$	0.495	0.484	0.475	0.500	0.501	0.501

Table 5.1: Mean estimates of MLE, Partition estimates and MPLE

a basic information on how the estimates work. However, to make any reasonable conclusion, we need larger dimension data simulation.

### 5.1.2 A larger first stage estimation

In this simulation, we generated  $Q = 10$  dimensional normal vector  $\mathbf{Z}$ , with mean vector  $\mathbf{0}$  and correlation matrix  $\mathbf{R}$ , all the off diagonal elements of which are 0.8.  $\mathbf{Z}$  was then truncated into ordinal vector  $\mathbf{Y}$ , with equal thresholds -0.3 and 0.8.

With sample size 100, 200 and 500, each with 100 replications, R 2.6.1 was used to generate the Mplus estimation which applied the two way marginals. EQS 6.1 was applied to create the MPL estimates. We may compare the consistency of MPLE with the two way contingency table estimation.

To save space, only selected the estimates are displayed here. The standard

True Parameter	$N = 100$			$N = 500$		
	MLE	Partition	MPLE	MLE	Partition	MPLE
<i>Thresholds</i>						
$\alpha_{11}$	0.184	0.165	0.149	0.078	0.066	0.066
$\alpha_{12}$	0.132	0.133	0.122	0.078	0.071	0.071
$\alpha_{21}$	0.100	0.095	0.099	0.062	0.059	0.059
$\alpha_{22}$	0.126	0.100	0.100	0.052	0.048	0.048
$\alpha_{31}$	0.149	0.133	0.106	0.069	0.051	0.051
$\alpha_{32}$	0.174	0.159	0.159	0.100	0.091	0.091
<i>Polychoric correlations</i>						
$\rho_{12}$	0.127	0.139	0.120	0.048	0.040	0.040
$\rho_{13}$	0.036	0.044	0.050	0.017	0.018	0.047
$\rho_{23}$	0.099	0.119	0.117	0.043	0.043	0.042

Table 5.2: Comparison of RMSE of MLE, Partition estimates and MPLE.

error estimates are also omitted because they are also close to each other. The mean and Root Mean Square Error (RMSE) obtained from the 100 replications are provided in Table 5.3. The two estimators are close to each other. This pattern is not affected by the sample size. In common, both estimator get more and more closer to the true value 0.8 as sample size gets larger.

The first stage estimator is consistent and asymptotically normal. And this is a requirement for this method to be extended to SEM. Now we look at the performance of second stage estimator, the structural parameter estimator.

<i>Sample size</i>	<i>MPLE</i>					<i>Marginal</i>				
	r12	r23	r34	r46	r58	r12	r23	r34	r46	r58
n =100										
mean	.797	.804	.800	.804	.795	.797	.803	.799	.804	.795
rmse	.051	.054	.059	.051	.055	.051	.054	.060	.051	.055
n =200										
mean	.800	.796	.797	.801	.797	.799	.796	.797	.800	.797
rmse	.036	.038	.037	.036	.041	.036	.038	.037	.036	.041
n =500										
mean	.803	.800	.801	.802	.803	.802	.800	.801	.802	.803
rmse	.026	.027	.025	.028	.025	.026	.027	.026	.028	.025

Table 5.3: Comparison of MPL estimates with marginal method

## 5.2 Second stage estimation: structural parameter estimates

### 5.2.1 A small simulation on second stage estimation

Again, we will conduct a small simulation to look at the parameter estimates and the test statistic. Because if it doesn't work in this small simulation, there seems no meaning to see its larger dimension performance. A one factor CFA is used as this small simulation.

Each of the 100 replications has sample size of 100. There are 4 ordinal variables in each sample. Each ordinal variable has 3 categories. Therefore, the parameters of this one factor model include 4 factor loadings. Results using MPL-GLS method are compared with those using the partition method. Both the formulae based standard errors of the estimates and the empirical standard deviations are listed in table 5.4. We first notice that the parameter estimates

are mutually close to each other and to the true values. We also note that the estimated standard errors are close to the empirical ones. Finally, the test statistic formed by the estimates is  $\chi^2$  distributed, as shown by the Kolmogorov-Smirnov (K-S) test statistic.

True Parameter:	MPL-GLS			LPB		
	est.	s.e est	emp s.e	est.	s.e est.	emp s.e
<i>Factor loading</i>						
$\gamma_1 = 0.8$	0.814	0.070	0.074	0.816	0.067	0.075
$\gamma_2 = 0.8$	0.822	0.069	0.076	0.824	0.067	0.076
$\gamma_3 = 0.8$	0.816	0.069	0.076	0.816	0.067	0.077
$\gamma_4 = 0.8$	0.802	0.072	0.074	0.803	0.069	0.074
K-S test p-value	0.866			0.490		

Table 5.4: Comparison of parameter estimates, se, and empirical s.e's.

As the small simulation worked, it stimulated us to go further.

### 5.2.2 Simulation I: parameter estimation

In this simulation study, 12 dimensional multivariate normal random numbers  $\mathbf{z} = (z_1, \dots, z_{12})$  are generated from  $N_{12}(0, \mathbf{R})$ , with sample size 100, 300, 500 respectively. The normal data  $\mathbf{y}$  were then transformed into ordinal variables  $\mathbf{y} = (y_1, \dots, y_{12})$  using a set of pre-assigned thresholds with  $\alpha_1 = -2.0$ ,  $\alpha_2 = -0.5$ ,  $\alpha_3 = 0.4$ . So, each ordinal variable has 4 categories and the thresholds here are symmetric and also extreme, -2.0. This procedure is repeated 100 times.

The correlation matrix  $\mathbf{R}$  is obtained from a two-factor confirmatory factor

analysis model

$$\mathbf{R} = \Lambda\Phi\Lambda' + \Psi, \quad (5.1)$$

where  $\Lambda$  is the factor loading matrix,  $\Phi$  is the correlation matrix of the factors and  $\Psi$  is diagonal covariance matrix of error variances with  $\Psi = \mathbf{I} - \text{diag}(\Lambda\Phi\Lambda')$ . The particular population values chosen are

$$\Lambda = \begin{pmatrix} 0.95 & 0.95 & 0.95 & 0.95 & 0.95 & 0.95 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.95 & 0.95 & 0.95 & 0.95 & 0.95 & 0.95 \end{pmatrix}$$

$$\Phi = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

The parameters in this model include the 12 factor loadings, and the correlation between the factors. Table 5.5 gives results on the parameter estimates and standard errors from the simulation. For comparison to the formula standard error estimates, the empirical standard deviations are also reported. The  $n$  stands for sample size, for here, it is 100, 300 and 500 respectively, mean is the mean of the parameter estimates for the 100 samples; rmse is the root mean squared error; s.e gives estimated standard errors; while emp.se refers to the means the empirical standard deviation of the estimates across 100 replications.

It's easy to see that as sample size  $n$  gets larger, the more accurate are the parameter estimates. This is seen from the closeness of the estimates to the true parameter and the smaller rmse values. Furthermore, the average formula-based standard errors become closer to the empirical ones. The Lee, Poon and Bentler (1990) method mentioned above is also applied in this simulation. The results are omitted from Table 5.5 because they are quite close.

There is an important issue for the choice of extreme parameter value, namely

True	.80	.95	.95	.95	.95	.95	.95	.95	.95	.95	.95	.95	.95
n =100													
mean	.94	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
rmse	.145	.043	.043	.043	.042	.043	.042	.043	.043	.043	.041	.043	.043
se	.021	.007	.008	.008	.008	.008	.008	.008	.008	.008	.008	.008	.008
emp. se	.037	.012	.009	.010	.012	.013	.012	.012	.011	.011	.010	.012	.014
n =300													
mean	.86	.97	.97	.97	.97	.97	.97	.97	.97	.97	.97	.97	.97
rmse	.068	.022	.022	.022	.022	.022	.022	.022	.022	.022	.022	.022	.022
se	.021	.009	.009	.009	.009	.009	.009	.009	.009	.009	.009	.009	.009
emp. se	.025	.010	.009	.009	.010	.009	.009	.010	.009	.008	.010	.009	.011
n =500													
mean	.83	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96
rmse	.043	.014	.014	.015	.015	.014	.016	.016	.015	.015	.016	.015	.015
se	.018	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007
emp. se	.019	.008	.007	.008	.008	.007	.008	.008	.007	.008	.008	.008	.008

Table 5.5: MPL - GLS parameter estimates in SEM

the factor loading of 0.95. It is known that LPB method and Muthen's three stage method could break down in this set up. We expect that the MPL-GLS, which at the first stage has only one objective function, may be more stable compared with above two methods. There is, however, no broken data set for all methods. A comparison of number of non-sensible factor loadings, with number of data sets that broke down (Heywood Cases), are then calculated. By saying non-sensible factor loading, we say the factor loading is greater or equal than 1, such that the error variances would be non-positive.

For the 100 data sets with sample size at 100, the LPB gives output of 371 non-positive error variances out of 1200, and 97 data sets with Heywood cases. While MPL-GLS gives 281 non-positive error variances and 92 heywood cases. With sample size of 200, LPB decreases to 10 non-positive error variances and 8 Heywood cases. MPL-GLS reduces to 4 non-positive error variances and 4 Heywood cases. In summary, MPL-GLS performs slightly more robust than LPB from this simulation.

### 5.2.3 Simulation II: test statistic

We already know that under  $H_0$ , the test statistic  $T$  should follow a  $\chi^2$  distribution. We will see from simulations how this will work. With exactly the same design, same thresholds, dimension  $Q$  and same two factor model for  $R$ , we generated one large sample with sample size 200000 under  $H_0$  to see whether the test statistic  $T$  follows a  $\chi_{53}^2$  asymptotically. This same big sample is then divided into many different small samples with size  $n = 100, 200, 300, 500, 700$  respectively. For each small sample, MPL-GLS is applied and a  $T$  is calculated. The Kolmogorov-Smirnov test and the tail probabilities are calculated. In table 5.6 we present the results, giving the empirical tail probabilities at the theoretical .01, .05, and .10 points in the  $\chi_{53}^2$  distribution. We also give the K-S test results in the last column. LPB means the two stage estimation proposed by Lee, et al (1990).

From Table 5.6, we can observe that  $T$  asymptotically follows a  $\chi^2$  distribution. For sample size as large as 300, the value of the one sample K-S test is greater than .05 for MPL-GLS, allowing us to accept that  $T$  follow a  $\chi^2$  at .05 significance level. This is not true for the LPB method. For  $n$  as large as 500, both methods generate a well distributed  $T$ . Also, the tail probabilities from the sample are comparable with the true values, showing that the behavior of  $T$  in the accept/reject region becomes acceptable at around  $N = 700$ . This simulation provides basic evidence that the proposed method is a reliable alternative to a current approach.

Methods	1 % tail	5% tail	10% tail	K-S test P-value
n=100				
MPL-GLS	5	14	22	<.00001
LPB	10	17	22	< .00001
n= 200				
MPL-GLS	1.3	3.7	7.3	0.0077
LPB	0.4	2.6	5.4	< 0.0001
n=300				
MPL-GLS	0	3.0	6.0	0.056
LPB	0	2.0	5.1	0.003
n=500				
MPL-GLS	0	3.5	8.0	0.16
LPB	0	3.0	7.0	0.37
n=700				
MPL-GLS	0.7	4.9	9.5	0.20
LPB	0.7	2.8	8.8	0.38

Table 5.6: K-S test of T and empirical tail probability

#### 5.2.4 Power simulation

We have seen the performance of type I errors in Table 5.6. Now we interested in the power of the test statistic. The power is the probability of correct rejection. We will compare the power of using MPL-GLS method with that from LPB method. We used the same data generated from model 5.1. However, a two factor model but with some incorrect loadings is fitted. i.e., variable 2, 4, 6, 8 were loaded onto  $F_2$ , while remaining variables loaded onto  $F_1$ . In Table 5.7, the first column is the sample size, while the second column gives four significance levels.

The two right columns give the number of rejections out of 100 replications from partition method and MPL - GLS method respectively.

Sample size	Significance Level	LPB	MPL-GLS
50	0.005	36	36
	0.01	44	45
	0.05	65	66
	0.1	77	77
100	0.005	48	53
	0.01	57	63
	0.05	83	83
	0.1	92	93
200	0.005	94	98
	0.01	96	99
	0.05	99	100
	0.1	100	100

Table 5.7: Rejection rate on power assessment

Rejection rates are similar across the two approaches, with MPL performed slightly better for  $n = 100$  with significance level of 0.005, 0.001.

### 5.2.5 A simulation on CGCM

In this simulation study, 8 dimensional multivariate normal random numbers  $\mathbf{Z} = (z_1, \dots, z_8)$  are generated from  $N_8(0, \mathbf{R})$ , with sample size 50,100, 200, respectively. The normal data  $\mathbf{Z}$  were then transformed into mixed data containing 4 ordinal variables  $\mathbf{Y} = (y_1, y_2, y_3, y_4, z_5, z_6, z_7, z_8)$  using a set of pre-assigned thresh-

olds with  $\alpha_1 = -1.0$ ,  $\alpha_2 = 1.0$ . So, here we have  $C = 4$  continuous variables and  $S = 4$  ordinal variables each of which has 3 categories with symmetric thresholds. This procedure is repeated 100 times.

The correlation matrix  $\mathbf{R}$  is obtained from a two-factor confirmatory factor analysis model as given in (5.1).

The particular population values chosen are

$$\Lambda' = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.8 & 0.8 & 0.8 \end{pmatrix}$$

$$\Phi = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

The parameters in this model include the eight factor loadings, and the correlation between the factors. Table 5.8 gives results on the parameter estimates and standard errors from the simulation. The residual variances are given as  $(\mathbf{I} - \text{diag}(\Lambda\Phi\Lambda'))$ . To compare the formula based standard error estimates with those the empirical standard deviations, the latter are also reported using the same notation as in Table 5.1. For each replication, MPL-GLS is applied and a T is calculated. We also test if the T follows a  $\chi_{df}^2$  distribution as expected, where  $df = \frac{56}{2} - 9 = 19$ . Finally, the probabilities associated with the Kolmogorov-Smirnov test whose hypothesis is that T follows a  $\chi_{19}^2$  are listed in the last column.

From Table 5.8, it is easy to see that as sample size  $n$  gets larger, the more accurate are the parameter estimates. This is seen from the closeness of the estimates to the true parameter and the smaller RMSE values. Furthermore, the average formula-based standard errors become closer to the empirical ones, although the formula-based estimates underestimate variability in the parameters at all sample sizes, especially at  $n = 50$  and  $n = 100$ . This might further con-

True parameter	.80	.80	.80	.80	.80	.80	.80	.80	.80	T (p-value)
n =50										
mean	.83	.84	.83	.81	.84	.84	.85	.85	.85	< .0001
rmse	.12	.13	.12	.11	.16	.12	.11	.10	.11	
se	.04	.02	.02	.03	.05	.04	.03	.03	.03	
emp. se	.11	.13	.11	.12	.16	.11	.09	.09	.10	
n =100										
mean	.82	.81	.81	.82	.80	.83	.82	.82	.82	0.0007
rmse	.08	.08	.08	.07	.08	.07	.07	.06	.08	
se	.04	.02	.02	.03	.05	.04	.03	.03	.03	
emp. se	.07	.08	.06	.07	.08	.06	.06	.05	.06	
n =200										
mean	.81	.80	.81	.80	.79	.82	.81	.80	.81	0.1625
rmse	.06	.06	.06	.06	.06	.04	.04	.04	.04	
se	.03	.02	.02	.02	.04	.03	.03	.03	.03	
emp. se	.05	.05	.05	.06	.05	.03	.03	.03	.04	

Table 5.8: MPL - GLS Parameter Estimates in SEM

firmed Hoogland’s comments: ”It is not recommended to use the ADF standard error estimator, unless the sample size is extremely large” relative to number of variables (Hoogland, 1999. p.142). We can also observe that T asymptotically follows a  $\chi^2$  distribution. For sample size as large as 200, the value of the one sample K-S test is greater than .05 for MPL-GLS, allowing us to accept that T follow a  $\chi^2$  at .05 significance level. This simulation provides basic evidence that the proposed method is a reliable alternative to a current approach.

## CHAPTER 6

### Extension to multiple group

Many times, the data may come from many groups instead of a single population. SEM analysis with multiple group or multi-sample has been widely used for cross validation, experimental design, longitudinal analysis and cross-sectional data. For example, compare model calibration sample with a model validation sample for cross-validation, compare treatment group with control group in experimental research and compare an earlier sample with that at a later time, as well as simply to compare the female group with male group in a cross-sectional data. The interest of multiple group SEM lies in the difference and similarities between groups, which may be done with respect to structural parameters, namely the factor means, factor variances and covariances, and error variances.

Although differences and similarities between groups can be studied also in a multi-way contingency table via log linear models, only relationships between observed variables are considered. Multiple population extension of the MPL-GLS must be considered if we wish to study factor invariance.

For group  $g$ ,  $g = 1, 2, \dots, G$ , below is an  $r$  factor model for the latent response vector  $Z^g(Q \times 1)$ , which is underlying observed ordinal vector  $Y^g(Q \times 1)$  :

$$\mathbf{Z}^g = \tau^g + \mathbf{\Lambda}^g \mathbf{F}^g + \varepsilon^g \quad (6.1)$$

$\tau^g$  is a  $Q \times 1$  latent intercept parameter vector,  $\mathbf{\Lambda}^g$  is  $Q \times r$  factor loadings matrix,  $\varepsilon^g$  is random error term with  $V(\varepsilon^g) = \mathbf{\Psi}^g$ , a  $Q \times Q$  diagonal matrix.

$\mathbf{F}^g$  is the  $r \times 1$  vector of factor scores, with mean  $E(\mathbf{F}^g) = \nu^g$  and covariance matrix  $V(\mathbf{F}^g) = \Phi^g$ . Therefore, the mean and variance of the latent vector are

$$E(\mathbf{Z}^g) = \tau^g + \Lambda^g \nu^g, \quad V(\mathbf{Z}^g) = \Lambda^g \Phi^g \Lambda^{g'} + \Psi^g = \Sigma^g \quad (6.2)$$

For continuous vector  $\mathbf{Z}^g$ ,  $\tau^g$  represents the mean of  $\mathbf{Z}^g$  when  $E(\mathbf{F}^g) = 0$ . While in two groups the difference between  $\tau^i$  and  $\tau^j$  indicates an intercept difference between the group  $i$  and  $j$ , which has practical meaning. Because of the character of ordinal vector, during the estimation procedure,  $\tau^g$  is usually fixed at 0 for all  $g$ . Thus, the model used for ordinal data estimation is:

$$\mathbf{Z}^g = \Lambda^g \mathbf{F}^g + \varepsilon^g \quad (6.3)$$

This constraint does not change the factor difference among groups. Different from the standard GCM an CGCM model in chapter two, where  $\mathbf{Z}$  is assumed to follow a standardized multivariate normal distribution with mean vector  $\mathbf{0}$  and correlation matrix  $\mathbf{R}$ , the value of  $\mathbf{Z}$  depends on the factor mean, factor variance and error variance, therefore the value of  $\mathbf{Y}$ .

The relationship between latent variable and observed variable is:

$$y_q = m \text{ iff } \alpha_q^{m-1} \leq \lambda * f + \varepsilon < \alpha_q^m$$

Therefore, multiple groups study on the difference of is directly related to the factor mean. Thus if  $\mathbf{Y}$  from a given group  $g$  has more higher categories than another group  $f$ , then  $\mathbf{Y}$  should have bigger thresholds and therefore group  $g$  has bigger factor mean than group  $f$ .

The study of differences and similarities among groups may include the hypothesis test of equality of measurement instrument across groups. For example, the thresholds are the same for all groups,  $\alpha^1 = \alpha^2 = \dots = \alpha^G$ ; or invariant factor loadings:  $\Lambda^1 = \Lambda^2 = \dots = \Lambda^G$ , where  $G$  is total number of groups, or invariant

error variances across groups. We may extend the MPL - GLS procedure to multiple groups in the following two steps.

step I: For each group  $g$ ,  $g = 1, 2, \dots, G$ , get the estimates of the thresholds vector  $s_1^g$  and polychoric correlations vector  $s_2^g$  using MPL. A weight matrix  $W^g$  is also obtained just as before. So we have a vector of  $(s^1, s^2, \dots, s^G)$ , where  $s^g = (s_1^g, s_2^g)^t = (\alpha_1^g, \alpha_2^g, \dots, \alpha_Q^g, r_{21}^g, r_{31}^g, \dots, r_{Q(Q-1)}^g)^t$ .  $\alpha_m^g$  denotes the threshold vector for the  $m$ th variable in  $g$ th group,  $r_{st}^g$  denotes the polychoric correlation between  $s$ th and  $t$ th variable in  $g$ th group.  $g=1, 2, \dots, G$ .

step II: The model parameter vector  $\theta$  estimate is obtained by minimizing  $F$ , the discrepancy between sample entity  $s$  and its corresponding population entity  $\sigma(\theta)$ :

$$Q(\theta) = \sum_{g=1}^G (s^g - \sigma^g(\theta))' W^{g-1} (s^g - \sigma^g(\theta)) \quad (6.4)$$

Under  $H_0$ ,  $F = Q(\hat{\theta}) \sim \chi^2$  asymptotically and the asymptotic covariance matrix of  $\hat{\theta}$  is  $acov(\hat{\theta}) = [\sum_{g=1}^G C^{gT} W^{g-1} C^g]^{-1}$ , where  $C^g = \partial \sigma^g(\theta) / \partial \theta'$ . When  $G = 1$ , a one group case,  $F = T$  mentioned in chapter 4.

## 6.1 Model identification

For the purpose of comparing outcomes from different groups, it is important to note a fact that the GCM and CGCM are not identifiable. That is, there may exist more than one set of parameter estimate which can produce the same value for the likelihood function. This is because the observed ordinal categorical data, unlike ratio scaled data, do not have an inherent origin or unit of measurement. Some constraints on the model parameters are required for the model to be identified. That is, an origin and a unit of measurement must be selected as the reference. The minimum identification conditions are difficult to find. However, there are

currently many different but mathematically equivalent forms of constraints used to identify the model. Different choices of fixed values only result in changes of scale, see reasonings given in Lee, Poon and Bentler (1990).

Some identification methods have already been implemented in softwares, such as Mplus and Lisrel. Song and Lee (2003) used a MC method for multiple samples with missing data.

Millsap and Tein (2004) illustrated and compared in details the different constraints put on Mplus and Lisrel. They state "On balance, Mplus appears at present to offer a more flexible system for invariance modeling in ordered-categorical data" (p. 498) than LISREL. Please refer to Millsap and Tein (2004) for more details.

Basically, what kind of constraints should be imposed in baseline model depends mainly on the ordinal data type and the factor loading matrix,  $\Lambda$ . Specially, conditions depend on if the ordinal variables are dichotomous, or if the factor loading matrix has only one column(single factor model), or if each variable has only one nonzero factor loading. These will be explained in more details below.

- $\nu$  is free in all groups except the reference group, where it is 0
- $\Phi$  is free in all groups
- Every column in  $\Lambda$  has one fixed at 1.0, and different for all variables, and enough identification conditions as in standard confirmatory factor analysis.
- For variables that have three or more categories:
  1. free but hold equal two thresholds each reference variable and one threshold for each non-reference variable.  $\Psi$  is free in all groups except

the reference group, where  $\Psi = \mathbf{I} - \text{diag}(\Lambda\Phi\Lambda')$ .

2. In the model allows any variable to have more than one free parameters, free and hold equal a second threshold for each non-reference variable.

This is corresponding to the fact that any row of factor loading matrix has more than one free parameters. (Thus exclude any one factor model).

- For dichotomous variables that has only two categories:
  1. free but hold equal for the threshold.  $\Psi = \mathbf{I} - \text{diag}(\Lambda\Phi\Lambda')$  in reference group. In remaining groups,  $\Psi$  is free in remaining groups except the element of  $\Psi$  that is corresponding to the reference binary variable is  $1 - \Psi_{qq}$ .
  2. In the model allows any variable to have more than one free parameters, constrain other dichotomous variables other than the reference variable to be  $1 - \Psi_{qq}$ .

By saying reference variable, we mean the variable whose factor loading is prefixed at zero.

In summary, all factor covariance matrix elements are free. Factor means in the reference group are fixed to zero, and free in other groups. In the reference group, the error variance matrix is fixed as the identity matrix, a while it is free in other groups.

## 6.2 Parameter specification in a two population model: an example

We illustrate the idea of identification conditions using an example. Consider two data sets representing two independent population. Each data contains 4 ordinal variables each has four categories. A one factor model is fitted to both data.

First, thresholds, polychoric correlations and the weight matrix are obtained via MPL for each group separately. Second, Obtain the corresponding population entity including the population thresholds and polychoric correlations in both groups. The population entity corresponding to this model for this data are obtained as follows:

### 6.2.1 Group 1(reference group) parameters $\alpha^1, \sigma^1$

For group 1 (which is defined as the reference group), the corresponding population thresholds  $\alpha^1$  are:

$$\alpha^1 = (\alpha_{11}^1, \alpha_{12}^1, \alpha_{13}^1, \alpha_{21}^1, \alpha_{22}^1, \alpha_{23}^1, \alpha_{31}^1, \alpha_{32}^1, \alpha_{33}^1, \alpha_{41}^1, \alpha_{42}^1, \alpha_{43}^1)^T$$

The corresponding population polychoric correlations,  $r^1$ , is a vector built by stacking the lower off-diagonal matrix of  $R^1$ , whose diagonal elements are fixed at 1.

$$R^1 = \Lambda^1 \Phi^1 \Lambda^{1'} + \Psi^1$$

- $\Lambda^1 = (1, \lambda_1^1, \lambda_2^1, \lambda_3^1)^T$  is factor loading vector for group 1. Note the first loading of 1 indicates that the first ordinal variable is the reference variate.

- $\Phi^1$  is the variance for the factor in group 1
- $\Psi^1$  is the diagonal error variance matrix of group 1, which does not contain free parameters because  $\text{Diag}(\Sigma^1) = \mathbf{I}$

### 6.2.2 Group 2 parameters: $\alpha^{(2)}, \sigma^{(2)}$

The group 2 corresponding population  $r^{(2)}$  are the lower off-diagonal elements of  $\Sigma^{(2)}$  :

$$\Sigma^{(2)} = \Lambda^{(2)}\Phi^{(2)}\Lambda^{(2)'} + \Psi^{(2)} \quad (6.5)$$

- $\Lambda^{(2)} = (1, \lambda_1^{(2)}, \lambda_2^{(2)}, \lambda_3^{(2)})^T$  is the factor loading vector of group 2
- $\Phi^{(2)}$  is group 2 factor variance.
- $\Psi^{(2)}$  is group 2 diagonal error variance, a free parameter vector.
- $\nu^{(2)}$  is the factor mean in group 2 (factor mean is fixed at 0 in group 1)
- $D^{(2)} = [\text{diag}(\Sigma^{(2)})^{-1/2}]$  ( $D^{(1)}$  is fixed as identity matrix in group 1)

Based on the model, the elements,  $\alpha_{ij}^{(2)}$ , of second group thresholds  $\alpha^2$  is

$$\alpha_{ij}^{(2)} = (D_{ii}^{(2)} * (\alpha_{ij}^* - \Lambda_i^{(2)}\nu^{(2)})), \quad (6.6)$$

$i = 1, \dots, 4; j = 1, 2, 3$ .  $D_{ii}^{(2)}$  is the  $i$ th diagonal element of  $D^{(2)}$  and  $\Lambda_i^{(2)}$  is the  $i$ th element of  $\Lambda^{(2)}$ .  $\alpha_{ij}^*$  denotes the element of  $\alpha^*$  corresponds to the  $j$ th threshold parameter for  $i$ th variable with  $\alpha^* = (\alpha_{11}^1, \alpha_{12}^1, \alpha_{13}^2, \alpha_{21}^1, \alpha_{22}^2, \alpha_{23}^2, \alpha_{31}^1, \alpha_{32}^2, \alpha_{33}^2, \alpha_{41}^1, \alpha_{42}^2, \alpha_{43}^2)^T$ . Note that  $\alpha_{ij}^g$  denotes the  $j$ th threshold for  $i$ th variable for group  $g$ ,  $g = 1, 2$ . Some elements of  $\alpha^{(2)}$  come from group 1, such as  $\alpha_{11}^1$ , this is because the invariant thresholds condition for identification purpose we illustrated earlier.

Therefore, the elements of  $\alpha^2$  can be written as

$$\begin{aligned} \alpha^{(2)} = & (D_{11}^{(2)} * (\alpha_{11}^1 - \Lambda_1^{(2)} \nu^{(2)}), D_{11}^{(2)} * (\alpha_{12}^1 - \Lambda_1^{(2)} \nu^{(2)}), D_{11}^{(2)} * (\alpha_{13}^2 - \Lambda_1^{(2)} \nu^{(2)}), D_{22}^{(2)} * (\alpha_{21}^1 - \\ & \Lambda_2^{(2)} \nu^{(2)}), D_{22}^{(2)} * (\alpha_{22}^2 - \Lambda_2^{(2)} \nu^{(2)}), D_{22}^{(2)} * (\alpha_{23}^2 - \Lambda_2^{(2)} \nu^{(2)}), D_{33}^{(2)} * (\alpha_{31}^1 - \Lambda_3^{(2)} \nu^{(2)}), D_{33}^{(2)} * \\ & (\alpha_{32}^2 - \Lambda_3^{(2)} \nu^{(2)}), D_{33}^{(2)} * (\alpha_{33}^2 - \Lambda_3^{(2)} \nu^{(2)}), D_{44}^{(2)} * (\alpha_{41}^1 - \Lambda_4^{(2)} \nu^{(2)}), D_{44}^{(2)} * (\alpha_{42}^2 - \Lambda_4^{(2)} \nu^{(2)}), D_{44}^{(2)} * \\ & (\alpha_{43}^2 - \Lambda_4^{(2)} \nu^{(2)}))' \end{aligned}$$

The model parameters for this data set include:  $\alpha^1, \Lambda^1, \Phi^1$ , elements without  $*$  in  $\alpha^{2*}, \Lambda^{(2)}, \Phi^{(2)}, \Psi^{(2)}, \nu^{(2)}$ .

Finally, minimize  $F(\theta)$  to get parameter estimates, where

$$F(\theta) = \sum_{g=1}^2 (s^g - \sigma^g(\theta))' W^{g-1} (s^g - \sigma^g(\theta)) \quad (6.7)$$

$F(\hat{\theta}) \sim \chi_{df}^2$  asymptotically. Further, the asymptotic covariance matrix  $acov(\hat{\theta})$  of  $\hat{\theta}$  is

$$acov(\hat{\theta}) = \left[ \sum_{g=1}^2 C^{gT} W^{g-1} C^g \right]^{-1}, \quad (6.8)$$

where  $C^g = \partial \sigma^g(\theta) / \partial \theta'$ .

### 6.3 Simulations: comparison of group difference

To illustrate the performance of MPL-GLS in testing group difference in ordered-categorical data, we generated three pairs of data sets from each of three true models based on different invariance conditions. Each pair of data sets contains two independent groups. Because of the fact that weighted least square (WLS) estimator will need a large sample size to perform adequately well, the sample size in each group is 3000. A single-factor model was used for four measured variables all the time and each of 4 measured variables has 4 ordered categories. Thus this is a congeneric case. Parameter estimates, standard errors and test

statistics were obtained from each simulated data set. We compared some of the results with those from Mplus 4.21. The procedure is described as follows.

1. For group 1, generate random factor score  $F^1$  from a  $N(\nu^1, \Phi^1)$  density. Then independently generate random vector  $\varepsilon$  from multivariate normal  $N(\mathbf{0}, \Psi^1)$ .
2. Calculate  $\mathbf{Z}^1 = \tau^1 + \Lambda^1 \mathbf{F}^1 + \varepsilon^1$
3. Create  $y_q^1 = m \Leftrightarrow \alpha_{q,m-1}^1 \leq z_q^g < \alpha_{q,m}^g$ ,  $q = 1, \dots, 4$ ,  $m = 1, \dots, 4$
4. Repeat steps 1-3 for  $n = 3000$  times, creating a  $3000 \times 4$  data matrix  $Y^1$
5. Repeat 1-4 using parameters for group 2 to generate  $3000 \times 4$  data matrix  $Y^2$ .

For each pair of data sets, the chi-square values for 4 different invariance hypothesis models were calculated below. The first underlying model is full factor invariant which means that the thresholds, factor loadings, and error variances are the same for the two groups. In the second underlying model, only the thresholds are not invariant. In the third model, only the error variances are invariant between the two groups. A sequences of hypotheses with increasing invariance are tested sequentially on each data pair.

### 6.3.1 Model 1

Model 1 is full factorial invariant. This implies threshold  $\alpha$ , factor loading  $\Lambda$ , and error variance  $\Psi$  are all equal for both groups. The parameter values are given below:

- $\tau^1 = \tau^2 = (0.25, 0.25, 0.50, 0.50)^T$

- $\nu^1 = 0, \Phi^1 = 1$  in group 1;  $\nu^2 = 0.25, \Phi^2 = 1.44$  in group 2
- $\Psi^1 = \Psi^2 = \mathbf{I}$
- $\Lambda^1 = \Lambda^2 = (0.4, 0.5, 0.6, 0.4)^T$
- $\alpha^1 = \alpha^2 = \begin{pmatrix} -0.45 & 0.25 & 0.95 \\ -0.45 & 0.25 & 0.95 \\ -0.3 & 0.5 & 1.3 \\ -0.2 & 0.5 & 1.2 \end{pmatrix}$

Table 6.1 gives the fit results from MPL-GLS for a sequence of four hypotheses model fit to the data generated under Model 1. The first hypothesis is the baseline model described earlier in which some thresholds are fixed at values that are invariant over groups. No other invariance constraints are imposed except the loading of the first variable is fixed to one. The second hypothesis adds invariance of the factor loadings to the baseline model. The third hypothesis adds invariance of thresholds to the second model. The fourth hypothesis adds invariance of error variances to the third hypothesis. For comparison purpose, the fit results from Mplus on hypothesis model three are provided under the table, which applied delta parameterizations with WLS estimator.

The fit results in Table 6.1 demonstrate that the four hypothesis models can not be rejected, as expected. The chi-square value, 8.748 from hypothesis model 3 based on MPL-GLS is close to that from Mplus, 9.150 with 14 degrees of freedom. In fact, MPL-GLS will always have the same degrees of freedom as Mplus with delta parameterizations.

Table 6.2 provided the parameter estimates and standard errors from hypothesis model 3 applying MPL-GLS and Mplus separately. Note  $\alpha_i^j$  stands for the  $j$ th threshold for variable  $i$ ;  $\lambda_j$  stands for the factor loading on the  $j$ th measured

Model	Chi-square	def	<i>p</i> -value
Baseline	2.267	4	0.687
Invariant $\Lambda$	4.595	7	0.709
Invariant $\Lambda, \alpha$	8.748	14	0.847
Invariant $\Lambda, \alpha, \Psi$	16.231	18	0.576

Mplus:  $\chi^2_{14} = 9.150$ ; Value = 0.821

Table 6.1: Model 1: MPL - GLS test statistic

variable( $\lambda_1 = 1$  in this setup). First of all, the most important information contained in this table is the factor mean. Since the factor mean in the reference group is fixed at 0,  $\nu^2$  is actually the group difference. Obviously,  $\nu^2 > 0$  significantly. We therefore conclude that group 2 has significantly larger factor mean. Also note that the last four rows of parameters are not comparable.  $\Psi_{11}$  -  $\Psi_{44}$  are error variances in MPL-GLS procedure, while  $d_1$  -  $d_4$  are scale factors from Mplus, which are the inverse of the standard deviations for the measured variables. Except slight difference, Table 6.2 demonstrates that both parameter estimates and standard errors provided by Mplus and MPL-GLS are close.

### 6.3.2 Model 2

Model 2 released the invariance of error variances. This model is identical to Model 1 except that error variances in group one were  $0.8*\mathbf{I}$ , the error variances used in group two were as in model 1. Table 6.3 was the fit results from data generated from Model 2. Only hypothesis model 4 yields a significant chi-square value, consistent with the theory. Again, chi-square values from MPL-GLS are close to that from Mplus.

<i>Parameter</i>	<i>MPL-GLS</i>		<i>Mplus</i>	
	Estimates	S.E.	Estimates	S.E.
$\alpha_1^1$	-0.630	0.017	-0.643	0.022
$\alpha_1^2$	-0.007	0.021	-0.009	0.018
$\alpha_1^3$	0.642	0.024	0.656	0.021
$\alpha_2^1$	-0.650	0.022	-0.669	0.022
$\alpha_2^2$	-0.025	0.019	-0.009	0.019
$\alpha_2^3$	0.611	0.021	0.612	0.021
$\alpha_3^1$	-0.714	0.023	-0.682	0.023
$\alpha_3^2$	-0.026	0.019	-0.034	0.019
$\alpha_3^3$	0.655	0.022	0.627	0.021
$\alpha_4^1$	-0.674	0.022	-0.664	0.022
$\alpha_4^2$	0.011	0.018	0.012	0.018
$\alpha_4^3$	0.628	0.021	0.665	0.021
$\lambda_2$	1.157	0.074	1.171	0.086
$\lambda_3$	1.251	0.078	1.261	0.093
$\lambda_4$	0.967	0.065	1.058	0.079
$\Phi^1$	0.152	0.015	0.142	0.016
$\nu^2$	0.070	0.016	0.094	0.016
$\Phi^2$	0.228	0.022	0.189	0.022
$\Psi_{11}(d_1)$	0.724	0.048	0.962	0.028
$\Psi_{22}(d_2)$	0.829	0.059	0.981	0.029
$\Psi_{33}(d_3)$	0.699	0.052	1.018	0.030
$\Psi_{44}(d_4)$	0.908	0.063	0.995	0.029

Table 6.2: Model 1: Comparison of MPL-GLS with Mplus

Model	Chi-square	def	<i>p</i> -value
Baseline	5.150	4	0.272
Invariant $\Lambda$	5.834	7	0.559
Invariant $\Lambda, \alpha$	8.222	14	0.877
Invariant $\Lambda, \alpha, \Psi$	55.120	18	< 0.001

Mplus:  $\chi^2_{14} = 8.413$ ; Value = 0.867

Table 6.3: Model 2: MPL - GLS test statistic

### 6.3.3 Model 3

Model 3 is identical to Model 1, except the thresholds in group 2 are not equal to those in group 1, but defined as

$$\alpha^2 = \begin{pmatrix} -0.45 & 0.25 & 0.75 \\ -0.45 & 0.55 & 1.10 \\ -0.3 & 0.7 & 1.3 \\ -0.2 & 0.5 & 1.1 \end{pmatrix}$$

Table 6.4 is the fit results from data generated from model 3. The last two hypothesis models which carry the invariance of thresholds were rejected. This again agrees with the theory. Even the rejected chi-square value, 154.88, from MPL-GLS are still very close to 154.61, the rejected chi-square value from Mplus.

## 6.4 Simulation: multiple groups parameter estimates

From last section, we already know that MPL-GLS works very well in hypothesis testing on group invariances or differences. Also, the parameter estimates from MPL-GLS are consistent with those from Mplus. However, because of the nature

Model	Chi-square	def	p-value
Baseline	3.612	4	0.461
Invariant $\Lambda$	7.148	7	0.414
Invariant $\Lambda, \alpha$	154.88	14	< 0.001
Invariant $\Lambda, \alpha, \Psi$	191.00	18	< 0.001

Mplus:  $\chi^2_{14} = 154.61$ ; Value < 0.001

Table 6.4: Model 3: MPL - GLS test statistic

of that simulation, we have no way to prove if the results in table 6.2 are close to their true parameter values.

In order to make up this part, we conduct another simulation study in which the true parameters satisfy the baseline model with identification conditions. There are 4 ordinal variables each has 4 categories, and the sample size is 3000. A similar data generation procedure as in last section is applied.

- $\tau^1 = \tau^2 = 0$
- $\nu^1 = 0, \Phi^1 = 0.36$  in group 1;  $\nu^2 = 0.5, \Phi^2 = 1.41$  in group 2
- $\Psi^1 = \mathbf{I}$
- $\Psi^2 = \text{Diag}(1, 2, 3, 4)$
- $\Lambda^1 = \Lambda^2 = (1, 0.8, 1.2, 1.5)^T$
- $\alpha^1 = \alpha^2 = \begin{pmatrix} -0.45 & 0.25 & 0.95 \\ -0.45 & 0.25 & 0.95 \\ -0.3 & 0.5 & 1.3 \\ -0.2 & 0.5 & 1.2 \end{pmatrix}$

Table 6.5 gives results from this simulation. It is easy to see that all true values, estimates from Mplus and from MPL-GLS are very close. The standard errors from the latter two are close to each other too. Mplus gives test statistic with  $\chi^2_{14} = 22.55$ , the corresponding pvalue is 0.069. For MPL - GLS, the test statistic is  $\chi^2_{14} = 22.86$  with pvalue 0.063.

<i>True parameter</i>	<i>Mplus</i>		<i>MPL-GLS</i>	
	Estimates	S.E.	Estimates	S.E.
$\alpha_1^1 = -0.45$	-0.45	0.022	-0.45	0.024
$\alpha_1^2 = 0.25$	0.24	0.021	0.24	0.021
$\alpha_1^3 = 0.95$	0.95	0.024	0.95	0.022
$\alpha_2^1 = -0.45$	-0.46	0.022	-0.45	0.023
$\alpha_2^2 = 0.25$	0.26	0.021	0.24	0.019
$\alpha_2^3 = 0.95$	0.97	0.024	0.95	0.025
$\alpha_3^1 = -0.30$	-0.28	0.022	-0.27	0.022
$\alpha_3^2 = 0.50$	0.51	0.022	0.51	0.020
$\alpha_3^3 = 1.30$	1.31	0.029	1.30	0.027
$\alpha_4^1 = -0.20$	-0.16	0.023	-0.16	0.022
$\alpha_4^2 = 0.50$	0.54	0.023	0.53	0.023
$\alpha_4^3 = 1.20$	1.21	0.029	1.21	0.026
$\lambda_2 = 0.80$	0.81	0.028	0.84	0.026
$\lambda_3 = 1.20$	1.20	0.032	1.21	0.033
$\lambda_4 = 1.50$	1.50	0.043	1.51	0.049
$\Phi^1 = 0.36$	0.37	0.017	0.36	0.018
$\nu^2 = 0.50$	0.53	0.031	0.53	0.036
$\Phi^2 = 1.41$	1.34	0.084	1.28	0.078
$\Psi_{11} = 1.0$	1.07	0.018	1.06	0.021
$\Psi_{22} = 2.0$	2.00	0.017	1.95	0.017
$\Psi_{33} = 3.0$	2.92	0.013	2.83	0.015
$\Psi_{44} = 4.0$	3.57	0.013	3.47	0.012

Table 6.5: Estimates of MPL-GLS with Mplus

# CHAPTER 7

## Application

The main application areas of this dissertation would include the analysis of multivariate longitudinal and event-history data, spatial statistics, bioinformatics social network analysis, and so on. We illustrate in some details on how it would be applied in biomedical studies and social science.

### 7.1 Application in biomedical studies: clinical trials

In clinical trials, the interest lies in comparison of outcomes among different groups. Most clinical trials are superiority trials with the aim to show a better performance of the new drug compared to the control drug. When the control drug is not placebo but a standard active drug, and it is conceived to be difficult to improve upon the efficacy of that standard drug, one might consider showing that the new drug has comparable efficacy. When the new drug is believed to have comparable efficacy and has other advantages, for example, a much cheaper cost, a noninferiority trial is an option. For a noninferiority trial, the aim is to show that the new medication is not (much) worse than the standard treatment. An equivalence trial attempts to show that the responses to two drugs/treatments differ by an amount that is clinically insignificant (Tang and Poon, 2007). Currently, noninferiority trials are becoming quite frequent due to the difficulty to improve upon existing therapies.

Outcomes from noninferiority clinical trials are often seen in ordinal forms. For instance, in the randomized controlled study on the effects of Assalix and Rofecoxib by Chrubasik et al.(2001), outcomes were described by 228 patients with acute back pain as 'poor', 'moderate', 'good' and 'very good'.

Biomedical studies with longitudinal designs frequently also collect data on ordered categorical repeated measures that indicate the degree of symptoms. For example, for binary data, it indicates the presence or absence of clinical or biological states. Binary repeated measures can be conveniently explained as "use" or "no use" of a certain drug in drug abuse treatment research. This strategy brings intuitive statistical interpretation to the study of dynamic changes in responses to treatment through time and across subjects. This has clear clinical interpretations and usefulness.

For the purpose of comparing ordinal outcomes from different groups, we may use the multiple group version of MPL -GLS. We can assume various factors relating to the drug effect and we can test a structural hypothesis. In many studies, the factor mean in the control group can be prefixed at 0, and the emphasis of the test is on the factor mean of the group of interest. Factor loadings in multiple indicator studies can also give a clear clinical interpretation.

## **7.2 Social science: linear model and anova**

Owing to the nature of the research problems or the design of questionnaires, most data in behavioral, psychological and social research are in polytomous form. In analyzing this kind of data with a multivariate linear model, ignoring the polytomous nature of the data may lead to incorrect results. The variances of the observed data are crucial in analysis of variance and regression, and

these variances computed from the observed polytomous data and the underlying continuous measurements can be very different. We can expect the problem if treating ordinal data as continuous may be severe in analyzing linear models.

The statistical analysis relating to the covariance matrix of polytomous variables has received a lot of attention, for example Olsson (1979), Olsson, Drasgow and Dorans (1982), Lee and Poon (1986), and Poon and Lee (1987) for estimation of polychoric and polyserial correlations; Muthen (1984) and Lee, Poon and Bentler (1992, 1995) for analysis of structural equation models with polytomous and continuous variables. However, not much attention devoting to the analysis of polytomous variables in the context of multivariate linear models that include the commonly used analysis of variances and regression as special cases.

Consider the following multivariate linear model:

$$\mathbf{z}_i = \beta \mathbf{x}_i + \epsilon_i, i = 1, \dots, n \quad (7.1)$$

where  $\mathbf{z}_i$  is the corresponding latent continuous variable underlying the observed ordinal variable  $\mathbf{y}_i$ .  $x_i$  is observed continuous covariate. The multivariate linear model defined in above equation is rather general, it subsumes models such as analysis of variance, multivariate linear regression models, etc, as special cases.

For this type of multivariate linear model with polytomous variables, we may use the MPL approach but extend it to include continuous or categorical covariates. According to the nature of the GCM and CGCM, adding covariates will not affect the estimation procedure. The only thing change will be the additional regression parameter  $\beta$ . That is, instead of  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{R})$ , now we have  $\mathbf{z} - \beta \mathbf{x}_i \sim N(\mathbf{0}, \mathbf{R})$ .

A simple transformation will be enough to apply the MPL method.

## CHAPTER 8

### Summary and conclusion

This article develops a MPL-GLS two stage methodology for performing ordinal data analysis in SEM. Methods are developed for estimation and model comparison (hypothesis testing), in the context of a general model that is common in behavioral, sociological and psychological research. We show by simulation that this approach is asymptotically efficient and works reasonably well in small samples. Parameter estimates especially became acceptable at quite small sample size, although standard error formulae underestimate empirical variability up to  $n = 200$ . This result is not unexpected since the approach is based on a minimum chi-square method. It is possible that estimation by least squares, followed up with corrections to standard error and test statistics, may work better in very small samples.

This method can also be applied to simultaneous estimation and testing in a multi-sample context as above. Simulation studies showed that parameter estimates, standard error estimates and test statistics are acceptable for large sample size, and comparable to the results from Mplus.

It may be noted that in Table 5.5 the mean parameter estimates always seem larger than the the true value, thus there might exist some systematic bias. If this is the case, then a bias correction may be useful. Further, if data is only marginally bivariate normal distributed, it is interesting to see if MPLE are really MLE. These are topics for future work.

# CHAPTER 9

## Appendix

### 9.1 GCM gradient and expected hessian

#### 9.1.1 Objective function $l_p(\theta)$

$$l_p(\theta) = \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \sum_{i=1}^{l_{s+1}} \sum_{j=1}^{l_{t+1}} n(i, j|s, t) \log Pr(i, j|s, t) \quad (9.1)$$

where  $Pr(i, j|s, t) = [\int_{\alpha_{s-1}^i}^{\alpha_s^i} \int_{\alpha_{t-1}^j}^{\alpha_t^j} \phi_2(x, y, \rho_{st}) dx dy]$   
 $= [\Phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) + \Phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) - \Phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}) - \Phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st})]$

- $\theta = \{\alpha_s^m, \rho_{st}; s, t = 1, \dots, Q, s \neq t; m = 1, \dots, l_s; \}$
- $\alpha_s^m$  is the mth threshold for the sth ordinal variable  $y_s$
- $\rho_{st}$  is the polychoric correlation for  $y_s$  and  $y_t$ .
- Q: number of ordinal variables
- $l_s$ : number of thresholds for the sth ordinal variable,  $y_s$ .
- $l_s + 1$ : number of categorical levels for  $y_s$ .
- $\phi_2(x, y, \rho_{st}) = \frac{1}{2\pi\sqrt{1-\rho_{st}^2}} \exp(-\frac{x^2+y^2-2\rho_{st}xy}{2(1-\rho_{st}^2)})$ , the standard bivariate normal density function with correlation  $\rho_{st}$ .

- $\Phi_2(x, y, \rho_{st}) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-\rho_{st}^2}} \exp\left(-\frac{x^2+y^2-2\rho_{st}xy}{2(1-\rho_{st}^2)}\right) dx dy$ , the standard bivariate normal distribution function with correlation  $\rho_{st}$ .
- $n(i, j|s, t)$ : number of observations with  $y_s = i$  and  $y_t = j$ .
- $Pr(i, j|s, t)$ : bivariate joint probability,  $Pr(y_s = i, y_t = j) = \Phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) + \Phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) - \Phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}) - \Phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st})$ . Note that for  $w=0$  or  $m=0$  or both,  $\Phi_2(\alpha_s^w, \alpha_t^m, \rho_{st}) = 0$ .
- It is defined that  $\alpha_s^0 = -\infty$ .

### 9.1.2 Gradient: $\frac{\partial l_p(\theta)}{\partial \theta}$

$$\frac{\partial l_p(\theta)}{\partial \theta} = \left( \frac{\partial l_p(\theta)}{\partial \alpha_s^m}; \frac{\partial l_p(\theta)}{\partial \rho_{st}}; s, t = 1, \dots, Q, t \neq s; m = 1, \dots, l_s \right) \quad (9.2)$$

1.  $\frac{\partial l_p(\theta)}{\partial \alpha_s^m} = \sum_{t \neq s} \sum_{j=1}^{l_t+1} \phi_1(\alpha_s^m) \left( \frac{n(m, j|s, t)}{Pr(m, j|s, t)} - \frac{n(m+1, j|s, t)}{Pr(m+1, j|s, t)} \right) \left[ \Phi_1\left(\frac{\alpha_t^j - \rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right]$ 
    - when  $j=1$ ,  $\Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) = 0$ ;
    - when  $j = l_t + 1$ ,  $\Phi_1\left(\frac{\alpha_t^j - \rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) = 1$ .
  2.  $\frac{\partial l_p(\theta)}{\partial \rho_{st}} = \sum_{i=1}^{l_s+1} \sum_{j=1}^{l_t+1} \frac{n(i, j|s, t)}{Pr(i, j|s, t)} \left[ \phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) + \phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) - \phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}) - \phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st}) \right]$ 
    - $\phi_2(\alpha_s^m, \alpha_t^n, \rho_{st}) = 0$ , if either  $m = 0$  or  $n = 0$  or  $m = l_s + 1$  or  $n = l_t + 1$  or any combination of these.
- $\phi_2(\cdot)$  is standard bivariate normal density function defined above.
  - $\Phi_1(x) = \int_{-\infty}^x \frac{\exp(-\frac{z^2}{2})}{\sqrt{2\pi}} dz$  is the univariate standard normal distribution function.

### 9.1.3 Expected hessian matrix $H(\theta)$

$$H(\theta) = \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \sum_{i=1}^{l_s+1} \sum_{j=1}^{l_t+1} n(i, j|s, t) \left( \frac{\partial \log Pr(i, j|s, t)}{\partial \theta} \right) \left( \frac{\partial \log Pr(i, j|s, t)}{\partial \theta} \right)^T \quad (9.3)$$

$$= \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \sum_{i=1}^{l_s+1} \sum_{j=1}^{l_t+1} \frac{n(i, j|s, t)}{Pr(i, j|s, t)^2} \left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) \left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right)^T$$

- a. For the position of  $\theta$  corresponding to  $\alpha_s^{i-1}$ ,  $2 \leq i \leq l_s + 1$ :

When  $i \geq 2$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) \phi_1(\alpha_s^{i-1}) [\Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^{i-1}}{\sqrt{1 - \rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^{i-1}}{\sqrt{1 - \rho_{st}^2}}\right)]$ , where  $\phi_1$  and  $\Phi_1$  are the standard univariate normal density and distribution function respectively;

Note for  $j = 1$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) - \phi_1(\alpha_s^{i-1}) \Phi_1\left(\frac{\alpha_t^1 - \rho_{st} \alpha_s^{i-1}}{\sqrt{1 - \rho_{st}^2}}\right)$ ;

for  $j = l_t + 1$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) = \phi_1(\alpha_s^{i-1}) \Phi_1\left(\frac{\alpha_t^{l_t} - \rho_{st} \alpha_s^{i-1}}{\sqrt{1 - \rho_{st}^2}}\right)$

- b. For the position of  $\theta$  corresponding to  $\alpha_s^i$ ,  $i = 1, 2, \dots, l_s$ :

When  $i = 1, 2, \dots, l_s$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) = \phi_1(\alpha_s^i) [\Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^i}{\sqrt{1 - \rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^i}{\sqrt{1 - \rho_{st}^2}}\right)]$  ;

Note for  $j = 1$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) \phi_1(\alpha_s^i) \Phi_1\left(\frac{\alpha_t^1 - \rho_{st} \alpha_s^i}{\sqrt{1 - \rho_{st}^2}}\right)$  ;

for  $j = l_t + 1$ ,  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) = -\phi_1(\alpha_s^i) \Phi_1\left(\frac{\alpha_t^{l_t} - \rho_{st} \alpha_s^i}{\sqrt{1 - \rho_{st}^2}}\right)$

- c.  $\left( \frac{\partial Pr(i, j|s, t)}{\partial \theta} \right) = \phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) + \phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) - \phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}) - \phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st})$ , for the position of  $\theta$  corresponding to  $\rho_{st}$ ;

Note:  $\phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) = 0$ , if either  $i = 1$  or  $j = 1$  or  $i = l_s + 1$  or  $j = l_t + 1$  or any combination of these.

- d. It is symmetric for  $\alpha_s$  and  $\alpha_t$ . Therefore, for the position of  $\theta$  corresponding to  $\alpha_t^{j-1}$ ,  $2 \leq j \leq l_t + 1$ , it has the same formulae as (a) except replacing  $s$  by  $t$  and  $i$  by  $j$  wherever they occur.

- e. Therefore, for the position of  $\theta$  corresponding to  $\alpha_t^j$ ,  $j = 1, \dots, l_t$ , it has the same formulae as (b) except replacing  $s$  by  $t$  and  $i$  by  $j$  wherever they occur.
- f.  $(\frac{\partial Pr(i,j|s,t)}{\partial \theta}) = 0$ , for other positions of  $\theta$ .

#### 9.1.4 An example for expected hessian matrix

Data: 2 ordinal variables  $y_1$  and  $y_2$ , each has 3 categories 1,2 and 3. Thus  $Q=2$ ,  $l_1 = 2$  and  $l_2 = 2$ . Parameter  $\theta = \{\alpha_1^1, \alpha_1^2, \alpha_2^1, \alpha_2^2, \rho_{12}\}$

$$\begin{aligned} H(\theta) &= \sum_{i=1}^{l_1+1} \sum_{j=1}^{l_2+1} \frac{n(i,j|s,t)}{Pr(i,j|s,t)^2} (\frac{\partial Pr(i,j|s,t)}{\partial \theta}) (\frac{\partial Pr(i,j|s,t)}{\partial \theta})^T \\ &= \sum_{i=1}^3 \sum_{j=1}^3 \frac{n(i,j|1,2)}{Pr(i,j|1,2)^2} (\frac{\partial Pr(i,j|1,2)}{\partial \theta}) (\frac{\partial Pr(i,j|1,2)}{\partial \theta})^T \end{aligned}$$

- $n(i, j|1, 2)$  is number of observations with  $y_1=i$  and  $y_2 = j$ ,  $i$  and  $j$  could take values from 1,2 and 3.
- $Pr(i, j|1, 2) = \Phi_2(\alpha_1^i, \alpha_2^j, \rho_{12}) + \Phi_2(\alpha_1^{i-1}, \alpha_2^{j-1}, \rho_{12}) - \Phi_2(\alpha_1^{i-1}, \alpha_2^j, \rho_{12}) - \Phi_2(\alpha_1^i, \alpha_2^{j-1}, \rho_{12})$
- An example:

$$\frac{\partial Pr(1,3|1,2)}{\partial \theta} = \begin{pmatrix} \phi_1(\alpha_1^1) [\Phi_1(\frac{\alpha_2^3 - \rho_{12} \alpha_1^1}{\sqrt{1 - \rho_{12}^2}}) - \Phi_1(\frac{\alpha_2^2 - \rho_{12} \alpha_1^1}{\sqrt{1 - \rho_{12}^2}})] \\ 0 \\ 0 \\ -\phi_1(\alpha_2^3) [\Phi_1(\frac{\alpha_1^1 - \rho_{12} \alpha_2^3}{\sqrt{1 - \rho_{12}^2}})] \\ \phi_2(\alpha_1^1, \alpha_2^3, \rho_{12}) - \phi_2(\alpha_1^1, \alpha_2^2, \rho_{12}) \end{pmatrix}$$

– Entry 3 came from  $\Phi_1(\frac{\alpha_1^0 - \rho_{12} \alpha_2^3}{\sqrt{1 - \rho_{12}^2}}) = 0$

– Entry 4 came from  $\phi_2(\alpha_1^0, \alpha_2^2, \rho_{12}) = \phi_2(\alpha_1^0, \alpha_2^3, \rho_{12}) = 0$

### 9.1.5 Observed Hessian Matrix $H_o(\theta)$

$$H_o(\theta) = \frac{\partial^2 l_p(\theta)}{\partial \theta \partial \theta'} \begin{pmatrix} \frac{\partial^2 l_p(\theta)}{\partial (\alpha_1^1)^2} & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^1 \partial \alpha_1^2} & \cdots & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^1 \partial \alpha_Q^1} & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^1 \partial r_{12}} & \cdots & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^1 \partial r_{(Q-1)Q}} \\ & \frac{\partial^2 l_p(\theta)}{\partial (\alpha_1^2)^2} & \cdots & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^2 \partial \alpha_Q^1} & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^2 \partial r_{12}} & \cdots & \frac{\partial^2 l_p(\theta)}{\partial \alpha_1^2 \partial r_{(Q-1)Q}} \\ & & \cdots & & & & \\ & & & & \frac{\partial^2 l_p(\theta)}{\partial (r_{12})^2} & \cdots & \frac{\partial^2 l_p(\theta)}{\partial r_{12} \partial r_{(Q-1)Q}} \\ & & & & \cdots & & \\ & & & & & & \frac{\partial^2 l_p(\theta)}{\partial (r_{(Q-1)Q})^2} \end{pmatrix}$$

Because of symmetry, only the upper triangle part is calculated:

A.  $\frac{\partial^2 l_p(\theta)}{\partial \alpha_s^m \partial \alpha_a^b}$ :

- If  $a=s$ , and  $b=m$ ;

$$\begin{aligned} \frac{\partial^2 l_p(\theta)}{\partial (\alpha_s^m)^2} &= \sum_{t \neq s} \sum_{j=1}^{t+1} \phi_1(\alpha_s^m) \left\{ \left[ \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right] \right. \\ &\quad \left[ -\alpha_s^m \left( \frac{n(m,j|s,t)}{Pr(m,j|s,t)} - \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)} \right) - \phi_1(\alpha_s^m) \left( \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right) \right. \\ &\quad \left. \left. \left( \frac{n(m,j|s,t)}{Pr(m,j|s,t)^2} + \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)^2} \right) \right] - \frac{\rho_{st}}{\sqrt{1-\rho_{st}^2}} \left( \frac{n(m,j|s,t)}{Pr(m,j|s,t)} - \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)} \right) \right. \\ &\quad \left. \left. \left( \phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right) \right\} \end{aligned}$$

- If  $a = s$  and  $b = m+1$ ,

$$\begin{aligned} \frac{\partial^2 l_p(\theta)}{\partial \alpha_s^m \partial \alpha_s^{m+1}} &= \sum_{t \neq s} \sum_{j=1}^{t+1} \phi_1(\alpha_s^m) \left[ \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right] \\ &\quad \left[ \phi_1(\alpha_s^{m+1}) \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)^2} \left( \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^{m+1}}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} - \rho_{st} \alpha_s^{m+1}}{\sqrt{1-\rho_{st}^2}}\right) \right) \right] \end{aligned}$$

- 0, if  $a = s$  and  $b > m+1$ . ( $b \geq m$  in this setup)

- If  $b = j$ ,  $a \neq s$ , say  $a = t$ :

$$\begin{aligned} \frac{\partial^2 l_p(\theta)}{\partial \alpha_s^m \partial \alpha_t^j} &= \phi_2(\alpha_s^m, \alpha_t^j, \rho_{st}) \left[ \frac{n(m,j|s,t)}{Pr(m,j|s,t)} - \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)} - \frac{n(m,j+1|s,t)}{Pr(m,j+1|s,t)} + \right. \\ &\quad \left. \frac{n(m+1,j+1|s,t)}{Pr(m+1,j+1|s,t)} \right] + \phi_1(\alpha_s^m) \phi_1(\alpha_t^j) \left\{ \left( \Phi_1\left(\frac{\alpha_t^{j+1} - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^j - \rho_{st} \alpha_s^m}{\sqrt{1-\rho_{st}^2}}\right) \right) \right\} \end{aligned}$$

$$\begin{aligned}
& \left[ \frac{-n(m,j+1|s,t)}{Pr(m,j+1|s,t)^2} (\Phi_1(\frac{\alpha_s^{m-1}-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}}) - \Phi_1(\frac{\alpha_s^m-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}})) + \frac{n(m+1,j+1|s,t)}{Pr(m+1,j+1|s,t)^2} (\Phi_1(\frac{\alpha_s^m-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}}) \right. \\
& - \Phi_1(\frac{\alpha_s^{m+1}-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}})) \left. \right] + (\Phi_1(\frac{\alpha_t^j-\rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}}) - \Phi_1(\frac{\alpha_t^{j-1}-\rho_{st}\alpha_s^m}{\sqrt{1-\rho_{st}^2}})) \left[ \frac{-n(m,j|s,t)}{Pr(m,j|s,t)^2} (\Phi_1(\frac{\alpha_s^m-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}}) \right. \\
& - \Phi_1(\frac{\alpha_s^{m-1}-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}})) + \frac{n(m+1,j|s,t)}{Pr(m+1,j|s,t)^2} (\Phi_1(\frac{\alpha_s^{m+1}-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}}) - \Phi_1(\frac{\alpha_s^m-\rho_{st}\alpha_t^j}{\sqrt{1-\rho_{st}^2}})) \left. \right] \}
\end{aligned}$$

B.  $\frac{\partial^2 l_p(\theta)}{\partial \alpha_s^m \partial \rho_{ab}}$ :

- If a = s or b = s, say, a = s

$$\begin{aligned}
\frac{\partial^2 l_p(\theta)}{\partial \alpha_s^m \partial \rho_{sb}} &= \sum_{j=1}^{l_b+1} \phi_1(\alpha_s^m) \{ [\Phi_1(\frac{\alpha_b^j-\rho_{sb}\alpha_s^m}{\sqrt{1-\rho_{sb}^2}}) - \Phi_1(\frac{\alpha_b^{j-1}-\rho_{sb}\alpha_s^m}{\sqrt{1-\rho_{sb}^2}})] [-\frac{n(m,j|s,b)}{Pr(m,j|s,b)^2} \\
& (\phi_2(\alpha_s^m, \alpha_b^j, \rho_{sb}) + \phi_2(\alpha_s^{m-1}, \alpha_b^{j-1}, \rho_{sb}) - \phi_2(\alpha_s^m, \alpha_b^{j-1}, \rho_{sb}) - \phi_2(\alpha_s^{m-1}, \alpha_b^j, \rho_{sb})) \\
& + \frac{n(m+1,j|s,b)}{Pr(m+1,j|s,b)^2} (\phi_2(\alpha_s^{m+1}, \alpha_b^j, \rho_{sb}) + \phi_2(\alpha_s^m, \alpha_b^{j-1}, \rho_{sb}) - \phi_2(\alpha_s^{m+1}, \alpha_b^{j-1}, \rho_{sb}) - \\
& \phi_2(\alpha_s^m, \alpha_b^j, \rho_{sb})) \left. \right] + (\frac{n(m,j|s,b)}{Pr(m,j|s,b)} - \frac{n(m+1,j|s,b)}{Pr(m+1,j|s,b)}) [\frac{\rho_{sb}\alpha_b^j-\alpha_s^m}{(1-\rho_{sb}^2)^{3/2}} \phi_1(\frac{\alpha_b^j-\rho_{sb}\alpha_s^m}{\sqrt{1-\rho_{sb}^2}}) \\
& - \frac{\rho_{sb}\alpha_b^{j-1}-\alpha_s^m}{(1-\rho_{sb}^2)^{3/2}} \phi_1(\frac{\alpha_b^{j-1}-\rho_{sb}\alpha_s^m}{\sqrt{1-\rho_{sb}^2}})] \}
\end{aligned}$$

- 0, else.

C.  $\frac{\partial^2 l_p(\theta)}{\partial \rho_{st} \partial \rho_{ab}}$ :

- If ab = st

$$\begin{aligned}
\frac{\partial^2 l_p(\theta)}{\partial (\rho_{st})^2} & \sum_{i=1}^{l_s+1} \sum_{j=1}^{l_t+1} \left\{ -\frac{n(i,j|s,t)}{Pr(i,j|s,t)^2} [(\phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) + \phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) \right. \\
& - \phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st}) - \phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}))^2 + \frac{n(i,j|s,t)}{Pr(i,j|s,t)} [\phi_2(\alpha_s^i, \alpha_t^j, \rho_{st}) (\frac{\rho}{1-\rho^2} \\
& - \frac{\rho(\alpha_s^i)^2 + \rho(\alpha_t^j)^2 - (1+\rho^2)\alpha_s^i\alpha_t^j}{(1-\rho^2)^2}) + \phi_2(\alpha_s^{i-1}, \alpha_t^{j-1}, \rho_{st}) (\frac{\rho}{1-\rho^2} \\
& - \frac{\rho(\alpha_s^{i-1})^2 + \rho(\alpha_t^{j-1})^2 - (1+\rho^2)\alpha_s^{i-1}\alpha_t^{j-1}}{(1-\rho^2)^2}) - \phi_2(\alpha_s^i, \alpha_t^{j-1}, \rho_{st}) (\frac{\rho}{1-\rho^2} \\
& - \frac{\rho(\alpha_s^i)^2 + \rho(\alpha_t^{j-1})^2 - (1+\rho^2)\alpha_s^i\alpha_t^{j-1}}{(1-\rho^2)^2}) - \phi_2(\alpha_s^{i-1}, \alpha_t^j, \rho_{st}) \\
& \left. (\frac{\rho}{1-\rho^2} - \frac{\rho(\alpha_s^{i-1})^2 + \rho(\alpha_t^j)^2 - (1+\rho^2)\alpha_s^{i-1}\alpha_t^j}{(1-\rho^2)^2}) \right] \}
\end{aligned}$$

- 0, else.

### 9.1.6 Observed variance-covariance matrix $V$

$$V = H_o^{-1}(\theta)C(\theta)H_o^{-1}(\theta)$$

$$C(\theta) = \sum_{i=1}^n \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta} \right) \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta} \right)^T$$

- All symbols, functions and index are same as those in gradient calculation
- $n$  is total number of observations for a data set
- $Q$  is total number of ordinal variables
- $Pr(y_s(i), y_t(i))$  is the bivariate probability for the  $s$ th and  $t$ th ordinal variables whose values taken as  $i$ th observations. May use the same formula as in calculation of  $Pr(i, j | s, t)$ , with  $i$  and  $j$ 's values are shown from the  $i$ th observation of a data set.

### 9.1.7 Estimated variance-covariance matrix $V$

$$V = H^{-1}(\theta)C(\theta)H^{-1}(\theta) \tag{9.4}$$

$$C(\theta) = \sum_{i=1}^n \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta} \right) \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta} \right)^T$$

- $n$  is total number of observations for a data set
- $Q$  is total number of ordinal variables
- $Pr(y_s(i), y_t(i))$  is the bivariate probability for the  $s$ th and  $t$ th ordinal variables whose values taken as  $i$ th observations. May use the same formula as

in calculation of  $\Pr(i, j | s, t)$ , with  $i$  and  $j$ 's values are shown from the  $i$ th observation of a data set.

- $\frac{\partial \log \Pr(y_s(i), y_t(i))}{\partial \theta}$  can be calculated the same way as in Hessian matrix calculation, see (a)-(f).

## 9.2 CGCM gradient and hessian

### 9.2.1 Standardize $\mathbf{X}$

1.  $\mu$ : sample mean of  $\mathbf{X}$ , the continuous part of data with dimension  $r$ .
2.  $\Sigma$ : sample variance of  $\mathbf{X}$ . Let  $\Sigma_{\mathbf{xx}}$  be correlation matrix of  $\mathbf{X}$ , which will be used below.
3. Standardize  $\mathbf{X}$  to  $(\mathbf{X} - \mu)\Sigma^{-\frac{1}{2}}$ , Replace  $\mathbf{X}$  by its standardized value.

### 9.2.2 Objective function $l_p(\theta^*)$

$$\begin{aligned}
 l_p(\theta^*) &= \sum_{i=1}^N \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \log Pr(y_s(i) = m, y_t(i) = j) \\
 &= \sum_{i=1}^N \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \log [\Phi_2(\alpha_s^m + \xi_s^T \mathbf{x}(i), \alpha_t^j + \xi_t^T \mathbf{x}(i), r_{st}) + \Phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}(i), r_{st}) - \Phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i), \alpha_t^j + \xi_t^T \mathbf{x}(i), r_{st}) - \Phi_2(\alpha_s^m + \xi_s^T \mathbf{x}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}(i), r_{st})]
 \end{aligned}$$

- $Q$ : number of ordinal variables
- $r$ : number(dimension) of continuous variable  $\mathbf{x}$ .
- $m = 1, \dots, l_s + 1$ ;  $n = 1, \dots, l_t + 1$  are possible values for  $y_s$  and  $y_t$ .
- $\theta^* = \{\alpha_s^m, \xi_s, r_{st}; s, t = 1, \dots, Q, t \neq s; m = 1, \dots, l_s; j = 1, \dots, l_t\}$
- $\alpha_s^m$  is the  $m$ th threshold for the  $s$ th ordinal variable  $y_s$
- $\xi_s = (\xi_s^1, \dots, \xi_s^r)$  is a  $r$  dimension vector;  $s = 1, \dots, Q$ .
- $\mathbf{x}(i)$  is the  $i$ th observed  $r$  dimension vector of  $\mathbf{x}$ ,  $i = 1, \dots, N$ .
- $r_{st}$  is the polychoric correlation for  $y_s$  and  $y_t$ .
- $l_s$ : number of thresholds for the  $s$ th ordinal variable,  $y_s$ .

- $l_s + 1$ : number of categorical levels for  $y_s$ .
- $\phi_2(x, y, r_{st}) = \frac{1}{2\pi\sqrt{1-r_{st}^2}} \exp\left(-\frac{x^2+y^2-2r_{st}xy}{2(1-r_{st}^2)}\right)$ , the standard bivariate normal density function with correlation  $r_{st}$ .
- $\Phi_2(x, y, r_{st}) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-r_{st}^2}} \exp\left(-\frac{x^2+y^2-2r_{st}xy}{2(1-r_{st}^2)}\right) dx dy$ , the standard bivariate normal distribution function with correlation  $r_{st}$ .
- $Pr(m, j|s, t)$ : bivariate joint probability which is not fixed as the GCM, but now depends on value of  $\mathbf{X}$ . For the  $i$ th observed vector  $\mathbf{X}$  corresponding to  $y_s = m, y_t = j$ ,

$$Pr(y_s = m, y_t = j) = \Phi_2(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) + \Phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) - \Phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) - \Phi_2(\alpha_s^m, \alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}).$$

Note that for  $w=0$  or  $m=0$  or both,  $\Phi_2(\alpha_s^w + \xi_s^T \mathbf{x}_{m,w}(i), \alpha_t^m + \xi_t^T \mathbf{x}_{m,w}(i), r_{st}) = 0$ .

- It is defined that  $\alpha_s^0 = -\infty$  and  $\alpha_s^{l_s+1} = \infty$

### 9.2.3 Gradient $\frac{\partial l_p(\theta^*)}{\partial \theta^*}$ :

$$\frac{\partial l_p(\theta^*)}{\partial \theta^*} = \left( \frac{\partial l_p(\theta^*)}{\partial \alpha_s^m}; \frac{\partial l_p(\theta^*)}{\partial \xi_s}; \frac{\partial l_p(\theta^*)}{\partial r_{st}}; s, t = 1, \dots, Q, t \neq s; m = 1, \dots, l_s \right) \quad (9.5)$$

$$\begin{aligned} \bullet \frac{\partial l_p(\theta^*)}{\partial \alpha_s^m} = & \sum_{t \neq s} \sum_{j=1}^{l_t+1} \left\{ \sum_{i=1}^{n(m,j|s,t)} \frac{\phi_1(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i))}{Pr(m,j|s,t)} \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) \right. \right. \\ & - \left. \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) \right] - \sum_{i=1}^{n(m+1,j|s,t)} \frac{\phi_1(\alpha_s^m + \xi_s^T \mathbf{x}_{m+1,j}(i))}{Pr(m+1,j|s,t)} \\ & \left. \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}_{m+1,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m+1,j}(i))}{\sqrt{1-r_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m+1,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m+1,j}(i))}{\sqrt{1-r_{st}^2}}\right) \right] \right\} \end{aligned}$$

Note:  $\mathbf{x}_{m,j}(i)$  denotes the  $i$ th of  $n(m, j|s, t)$  observed vector  $\mathbf{x}$  corresponding to  $y_s = m$  and  $y_t = j$ .

- $\frac{\partial l_p(\theta^*)}{\partial \xi_s} = \sum_{t \neq s} \sum_{m=1}^{l_s+1} \sum_{j=1}^{l_t+1} \sum_{i=1}^{n(m,j|s,t)} \frac{\mathbf{x}_{m,j}(i)}{Pr(m,j|s,t)} \left\{ \phi_1(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i)) \right.$   
 $\left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) \right]$   
 $- \phi_1(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i)) \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) \right]$   
 $- \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i))}{\sqrt{1-r_{st}^2}}\right) \left. \right\}$
- $\frac{\partial l_p(\theta^*)}{\partial r_{st}} = \sum_{m=1}^{l_s+1} \sum_{j=1}^{l_t+1} \sum_{i=1}^{n(m,j|s,t)} \frac{1}{Pr(m,j|s,t)} \left[ \phi_2(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) \right.$   
 $+ \phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) - \phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^j +$   
 $\left. \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) - \phi_2(\alpha_s^m + \xi_s^T \mathbf{x}_{m,j}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}_{m,j}(i), r_{st}) \right]$
- $\phi_2(\cdot)$  is standard bivariate normal density function defined above.
- $\Phi_1(x) = \int_{-\infty}^x \frac{\exp(-\frac{z^2}{2})}{\sqrt{2\pi}} dz$  is the univariate standard normal distribution function.

## 9.2.4 Estimated Hessian Matrix $H(\theta^*)$

:

$$H(\theta^*) = \sum_{i=1}^N \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \left( \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right) \left( \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right)^T$$

$$= \sum_{i=1}^N \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))^2} \left( \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right) \left( \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right)^T$$

Suppose  $y_s(i) = m$  and  $y_t(i) = j$ , i.e., The  $i$ th observed values for  $y_s$  and  $y_t$  are  $m, j$  respectively.  $m=1, \dots, l_s + 1; j=1, \dots, l_t + 1$ ; Then

- a. For the position of  $\theta^*$  corresponding to  $\alpha_s^{m-1}$ , with  $2 \leq m \leq l_s + 1$ :  
 $\left( \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right) = \phi_1(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i)) \left[ \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) - \right.$

$\Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right)$ ], where  $\phi_1$  and  $\Phi_1$  are the standard univariate normal density and distribution function respectively;

- b. For the position of  $\theta^*$  corresponding to  $\alpha_s^m$ , with  $m = 1, 2, \dots, l_s$ :

$$\left(\frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*}\right) = \phi_1(\alpha_s^m + \xi_s^T \mathbf{x}(i)) \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) \right]$$

- Note: It is symmetric for  $\alpha_s$  and  $\alpha_t$ . Therefore, for the position of  $\theta^*$  corresponding to  $\alpha_t^{j-1}$ ,  $2 \leq j \leq l_t + 1$ , it has the same formulae as (a) except replacing  $s$  by  $t$  and  $m$  by  $j$  wherever they occur; for the position of  $\theta^*$  corresponding to  $\alpha_t^j$ ,  $1 \leq j \leq l_t$ , it has the same formulae as (b) except replacing  $s$  by  $t$  and  $m$  by  $j$  wherever they occur.

- c. For the position of  $\theta^*$  corresponding to  $\xi_s, \xi_t$ .

$$\left(\frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*}\right) \mathbf{x}(i) \left\{ \begin{aligned} & \phi_1(\alpha_s^m + \xi_s^T \mathbf{x}(i)) \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}_{m,j}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^m + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) \right] - \\ & \phi_1(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i)) \left[ \Phi_1\left(\frac{\alpha_t^j + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) - \Phi_1\left(\frac{\alpha_t^{j-1} + \xi_t^T \mathbf{x}(i) - r_{st}(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i))}{\sqrt{1-r_{st}^2}}\right) \right] \end{aligned} \right\}$$

- Note, symmetry applies to  $\xi_t$ . i.e. for the position of  $\theta^*$  corresponding to  $\xi_t$ , same formula as (c) except replacing  $s$  by  $t$ ,  $m$  by  $j$  wherever they occur.

- d. For the position of  $\theta^*$  corresponding to  $r_{st}$ ;

$$\left(\frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*}\right) \phi_2(\alpha_s^m + \xi_s^T \mathbf{x}(i), \alpha_t^j + \xi_t^T \mathbf{x}(i), r_{st}) + \phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}(i), r_{st}) - \phi_2(\alpha_s^{m-1} + \xi_s^T \mathbf{x}(i), \alpha_t^j + \xi_t^T \mathbf{x}(i), r_{st}) - \phi_2(\alpha_s^m + \xi_s^T \mathbf{x}(i), \alpha_t^{j-1} + \xi_t^T \mathbf{x}(i), r_{st})$$

- e.  $\left(\frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*}\right) = 0$ , for other positions of  $\theta^*$ .

### 9.2.5 Transformation

Transform  $\theta^*$  to  $\theta$ ,  $\theta = \{\eta_s^m, \mathbf{c}_s, \rho_{st}; s, t = 1, \dots, Q, t \neq s; m = 1, \dots, l_s\}$ .

- $\eta_s^m = \frac{\alpha_s^m}{(1 + \xi_s' \Sigma_{xx} \xi_s)^{\frac{1}{2}}}$
- $\mathbf{c}_s = \frac{-\Sigma_{xx} \xi_s}{(1 + \xi_s' \Sigma_{xx} \xi_s)^{\frac{1}{2}}}$
- $\rho_{st} = \frac{r_{st} + \xi_s' \Sigma_{xx} \xi_t}{[(1 + \xi_s' \Sigma_{xx} \xi_s)(1 + \xi_t' \Sigma_{xx} \xi_t)]^{\frac{1}{2}}}$

### 9.2.6 Estimated variance-covariance matrix V

$$V = \frac{\partial \theta}{\partial \theta^*} H^{-1}(\theta^*) C(\theta^*) H^{-1}(\theta^*) \left( \frac{\partial \theta}{\partial \theta^*} \right)^T \quad (9.6)$$

$$C(\theta^*) = \sum_{i=1}^N \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right) \\ \left( \sum_{s=1}^{Q-1} \sum_{t=s+1}^Q \frac{1}{Pr(y_s(i), y_t(i))} \frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta^*} \right)^T$$

where  $\frac{\partial Pr(y_s(i), y_t(i))}{\partial \theta}$  can be calculated the same way as in Hessian matrix calculation.

If  $\dim(\theta) = \dim(\theta^*) = M$ , Then  $\frac{\partial \theta}{\partial \theta^*}$  is  $M \times M$  matrix. i.e.,

$$\frac{\partial \theta}{\partial \theta^*} = \begin{pmatrix} \frac{\partial \theta_1}{\partial \theta_1^*} & \cdots & \frac{\partial \theta_1}{\partial \theta_M^*} \\ \cdots & \cdots & \cdots \\ \frac{\partial \theta_M}{\partial \theta_1^*} & \cdots & \frac{\partial \theta_M}{\partial \theta_M^*} \end{pmatrix}$$

- $\frac{\partial \eta_s^m}{\partial \alpha_s^m} = \frac{1}{(1 + \xi_s' \Sigma_{xx} \xi_s)^{\frac{1}{2}}}$

- $\frac{\partial \eta_s^m}{\partial \xi_s} = -\frac{1}{2} \xi_s' \alpha_s^m (1 + \xi_s' \Sigma_{xx} \xi_s)^{-\frac{3}{2}}$
- $\frac{\partial \eta_s^m}{\partial \theta^*} = 0$ , for other  $\theta^*$ .
- $\frac{\partial c_s}{\partial \xi_s} = \frac{-\Sigma_{xx} (1 + \xi_s' \Sigma_{xx} \xi_s)^{\frac{1}{2}} + \frac{1}{2} \xi_s (1 + \xi_s' \Sigma_{xx} \xi_s)^{-\frac{1}{2}} \Sigma_{xx} \xi_s'}{(1 + \xi_s' \Sigma_{xx} \xi_s)}$
- $\frac{\partial c_s}{\partial \theta^*} = 0$ , for other  $\theta^*$ .
- $\frac{\partial \rho_{st}}{\partial r_{st}} = \frac{r_{st}}{[(1 + \xi_s' \Sigma_{xx} \xi_s)(1 + \xi_t' \Sigma_{xx} \xi_t)]^{\frac{1}{2}}}$
- $\frac{\partial \rho_{st}}{\partial \xi_s} = \frac{\Sigma_{xx} \xi_t [(1 + \xi_s' \Sigma_{xx} \xi_s)(1 + \xi_t' \Sigma_{xx} \xi_t)]^{\frac{1}{2}} - \frac{1}{2} (1 + \xi_t' \Sigma_{xx} \xi_t)^{\frac{1}{2}} (1 + \xi_s' \Sigma_{xx} \xi_s)^{-\frac{1}{2}} \Sigma_{xx} \xi_s (r_{st} + \xi_s' \Sigma_{xx} \xi_t)}{[(1 + \xi_s' \Sigma_{xx} \xi_s)(1 + \xi_t' \Sigma_{xx} \xi_t)]}$
- $\frac{\partial \rho_{st}}{\partial \xi_t}$  is symmetric to  $\frac{\partial \rho_{st}}{\partial \xi_s}$ .
- $\frac{\partial \rho_{st}}{\partial \theta^*} = 0$  for other  $\theta^*$ .

### 9.3 Multiple groups gradient calculation

In general, the gradient of the minimization function  $F$  includes two parts, namely, the reference group part  $\partial\sigma^0(\theta)/\partial\theta'$  and the non-reference group part  $\partial\sigma^{nr}(\theta)/\partial\theta'$ . To simplify the number of parameters, we can generalize  $\theta$  as

$$\theta = (\alpha^0, \Lambda^0, \Phi^0, \alpha^{nr}, \Lambda^{nr}, \nu^{nr}, \Phi^{nr}, \Psi^{nr})^T \quad (9.7)$$

where  $\alpha^0$  and  $\alpha^{nr}$  are threshold parameters for reference and non-reference group separately;  $\Lambda^0$  and  $\Lambda^{nr}$  are factor loading parameters,  $\Phi^0$  and  $\Phi^{nr}$  are factor variance parameters for reference and non-reference group;  $\nu^{nr}$  and  $\Psi^{nr}$  are factor mean and error variance parameters for non-reference group.

For reference group,  $\partial\sigma^0(\theta)/\partial\theta'$  is easy to compute: the threshold part w.r.t itself is 1, and 0 else. We will put emphasis on non-reference group part  $\partial\sigma^{nr}(\theta)/\partial\theta'$ . The vector  $\sigma^{nr}(\theta)$  includes elements as standardized thresholds and polychoric correlations for the non-reference groups. To simplify its elements, define the following notations:

- $\sigma_{i0}^{nr}(\theta)$  is defined as the elements of  $\sigma^{nr}(\theta)$  that correspond to the thresholds whose ordinal variable has been fixed at 1 on factor loading.
- $\sigma_t^{nr}(\theta)$  is defined as the elements of  $\sigma^{nr}(\theta)$  that correspond to the thresholds whose ordinal variable has free parameter on factor loading, and the free parameter needs to be estimated.
- $\sigma_{corr}^{nr}(\theta)$  is defined as the elements of  $\sigma^{nr}(\theta)$  that correspond to the polychoric correlation. This part is already built in EQS and we will not do it here.

The gradient matrix can be obtained by stacking the rows of different types of elements in  $\sigma^{nr}(\theta)$  w.r.t different type of elements in  $\theta$ .

### 9.3.1 $\partial\sigma_{i0}^{nr}/\partial\theta'$

- $\partial\sigma_{i0}^{nr}/\partial\alpha^{nr} = (\Phi^{nr} + \Psi^{nr})^{-0.5}$ 
  - $\Phi^{nr}$  is the corresponding factor variance
  - $\Psi^{nr}$  is the corresponding error variance
  - By corresponding, we mean that each ordinal variable that the thresholds parameter come from has its unique error variance, and its underlying factor has its unique factor variance too.
- $\partial\sigma_{i0}^{nr}/\partial\nu^{nr} = -(\Phi^{nr} + \Psi^{nr})^{-0.5}$
- $\partial\sigma_{i0}^{nr}/\partial\Phi^{nr} = \partial\sigma_{i0}^{nr}/\partial\Psi^{nr} = -0.5^* \frac{\alpha^{nr} - \nu^{nr}}{(\Phi^{nr} + \Psi^{nr})^{1.5}}$
- $\partial\sigma_{i0}^{nr}/\partial\theta' = 0$  for all other  $\theta$ .

### 9.3.2 $\partial\sigma_t^{nr}/\partial\theta'$

- $\partial\sigma_t^{nr}/\partial\alpha^{nr} = ((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{-0.5}$ ;
- $\partial\sigma_t^{nr}/\partial\lambda^{nr} = \frac{-\nu^{nr}}{((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{0.5}} - \frac{\lambda^{nr}\Phi^{nr}(\alpha^{nr} - \lambda^{nr}\nu^{nr})}{((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{1.5}}$
- $\partial\sigma_t^{nr}/\partial\nu^{nr} = \frac{-\lambda^{nr}}{((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{0.5}}$
- $\partial\sigma_t^{nr}/\partial\Phi^{nr} = \frac{-0.5(\lambda^{nr})^2(\alpha^{nr} - \lambda^{nr}\nu^{nr})}{((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{1.5}}$
- $\partial\sigma_t^{nr}/\partial\Psi^{nr} = \frac{-0.5(\alpha^{nr} - \lambda^{nr}\nu^{nr})}{((\lambda^{nr})^2\Phi^{nr} + \Psi^{nr})^{1.5}}$
- $\partial\sigma_t^{nr}/\partial\theta' = 0$  for all other  $\theta$ .

## CHAPTER 10

### References

- [1] Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.
- [2] Anderson, J. A. and Pemberton, J. D., 1985. The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41**, 875-885.
- [3] Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.
- [4] Chrubasik, S., Kunzel, O., Model, A., Conradt, C. and Black, A. (2001), "Treatment of low back pain with a herbal or synthetic anti-rheumatic: a randomized controlled study." Willow bark extract for low back pain. *Rheumatology*, **40**, 1388C1393.
- [5] Cox, D. R. and Reid, N. 2004. Miscellanea A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.
- [6] Cramer, H. *Mathematical Methods of Statistics*, Princeton University Press, 1999.
- [7] de Leeuw, J., 1983. Models and methods for the analysis of correlation coefficients. *Journal of Econometrics* **22**, 113-137.
- [8] de Leon, A.R., 2005. Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters* **75**, 49-57.

- [9] Eickhoff, J.C., 2005. Quasi-Maximum likelihood estimation for latent variable models with mixed continuous and polytomous data. *Journal of Modern Applied Statistical Methods* **4**, 473-481.
- [10] Ferguson, T.S., 1996. *A course in large sample theory*. London: Chapman and Hall.
- [11] Fu, T.T., Li, L.A., Lin, Y.M., and Kan, K., 2000. A limited information estimator for the multivariate ordinal probit model. *Applied Economics* **32**, 1841-1851.
- [12] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.
- [13] Heagerty, P. J. and Lele, S. R., 2000. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099-1111.
- [14] Hoogland, J.J. (1999). The robustness of estimation methods for covariance structure analysis. Unpublished doctoral dissertation, University of Groningen, The Netherlands.
- [15] Kuk, A.Y.C., Nott, D.J., 2000. A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters* **47**, 329-335.
- [16] Lee, S. Y., Poon, W. Y. and Bentler, P. M., 1990a. Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability letters* **9**, 91-97.
- [17] Lee, S.Y., Poon, W.Y., and Bentler, P.M., 1990b. A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika* **55**, 45-51.

- [18] Lee, S.Y., Poon, W.Y., and Bentler, P.M., 1992. Structural equation models with continuous and polytomous variables. *Psychometrika* **57**, 89-105.
- [19] Lee, S.Y., Poon, W.Y., and Bentler, P.M., 1995. A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology* **48**, 339-358.
- [20] Lindsay, B.G., 1988. Composite likelihood methods. *Contemporary Mathematics* **80**, 221-239.
- [21] Millsap, R. E. and Tein, J. Y., 2004. Assessing factorial invariance in ordered-categorical Measures. *Multivariate Behavioral Research* **39**(3), 479-515
- [22] Muthen, B., 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115-132.
- [23] Muthen, B. and Satorra, A., 1995. Technical aspects of Muthen's Liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika* **60**, 489-503.
- [24] Olsson, U., 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443-460.
- [25] Olsson, U., Drasgow, F. and Dorans, N.J., 1982. The polyserial correlation coefficient. *Psychometrika* **47**, 337-347.
- [26] Pearson, K. I., 1901. Mathematical contribution to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. *Phil. Trans. Royal Society of London*, **195A**, 1-47.
- [27] Poon, W.Y., Lee, S.Y., 1987. Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* **52**, 409-430.

- [28] Roy, S. N. and Kastenbaum, M. A., 1956. On the hypothesis of no interaction in a multi-way contingency table. *Annals of Mathematical Statistics* **27(3)**, 749-757.
- [29] Roy, S. N. and S. K. Mitra (1956). An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika* **43**, 361-376.
- [30] Roy, S. N. (1957). Some Aspects of Multivariate Analysis. New York: Wiley.
- [31] Song, X.Y., Lee, S.Y., 2002, Bayesian estimation and model selection of multivariate linear model with polytomous variables. *Multivariate Behavioral Research* **37 (4)**, 453-47
- [32] Song, X.Y., Lee, S.Y., 2006, A maximum likelihood approach for multisample nonlinear structural equation models with missing continuous and dichotomous Data. *Structural Equation Modeling* **13(3)**, 325-351
- [33] Tang, M.L., Poon, W.Y., 2007, Statistical inference for equivalence trials with ordinal responses: A latent normal distribution approach. *Computational Statistics & Data Analysis* **51**, 5918-5926.
- [34] Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association* **86**, 79-86.
- [35] Varin, C. and Vidoni, P., 2006. Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis* **51**, 2365-2373.
- [36] Yule, G. U., 1900. On the association of attributes in statistics: with illustration from the material of the childhood society, *Philosophical Transaction of the Royal Society* **194**, 257-319.