



UPPSALA  
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations  
from the Faculty of Social Sciences 86*

# Composite Likelihood Estimation for Latent Variable Models with Ordinal and Continuous, or Ranking Variables

MYRSINI KATSIKATSOU



ACTA  
UNIVERSITATIS  
UPSALIENSIS  
UPPSALA  
2013

ISSN 1652-9030  
ISBN 978-91-554-8571-9  
urn:nbn:se:uu:diva-188342

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, February 15, 2013 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

### **Abstract**

Katsikatsou, M. 2013. Composite Likelihood Estimation for Latent Variable Models with Ordinal and Continuous, or Ranking Variables. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 86. 31 pp. Uppsala. ISBN 978-91-554-8571-9.

The estimation of latent variable models with ordinal and continuous, or ranking variables is the research focus of this thesis. The existing estimation methods are discussed and a composite likelihood approach is developed. The main advantages of the new method are its low computational complexity which remains unchanged regardless of the model size, and that it yields an asymptotically unbiased, consistent, and normally distributed estimator.

The thesis consists of four papers. The first one investigates the two main formulations of the unrestricted Thurstonian model for ranking data along with the corresponding identification constraints. It is found that the extra identifications constraints required in one of them lead to unreliable estimates unless the constraints coincide with the true values of the fixed parameters.

In the second paper, a pairwise likelihood (PL) estimation is developed for factor analysis models with ordinal variables. The performance of PL is studied in terms of bias and mean squared error (MSE) and compared with that of the conventional estimation methods via a simulation study and through some real data examples. It is found that the PL estimates and standard errors have very small bias and MSE both decreasing with the sample size, and that the method is competitive to the conventional ones.

The results of the first two papers lead to the next one where PL estimation is adjusted to the unrestricted Thurstonian ranking model. As before, the performance of the proposed approach is studied through a simulation study with respect to relative bias and relative MSE and in comparison with the conventional estimation methods. The conclusions are similar to those of the second paper.

The last paper extends the PL estimation to the whole structural equation modeling framework where data may include both ordinal and continuous variables as well as covariates. The approach is demonstrated through an example run in R software. The code used has been incorporated in the R package lavaan (version 0.5-11).

*Keywords:* latent variable models, factor analysis, structural equation models, Thurstonian model, item response theory, composite likelihood estimation, pairwise likelihood estimation, maximum likelihood, weighted least squares, ordinal variables, ranking variables, lavaan

*Myrsini Katsikatsou, Uppsala University, Department of Statistics, SE-751 20 Uppsala, Sweden.*

© Myrsini Katsikatsou 2013

ISSN 1652-9030

ISBN 978-91-554-8571-9

urn:nbn:se:uu:diva-188342 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-188342>)

# List of papers

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

- I Katsikatsou, M., and Yang-Wallentin, F. (2012) On the identification of the unrestricted Thurstonian model for ranking data.
- II Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and Jöreskog, K. G. (2012) Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, 56, p. 4243-4258.
- III Katsikatsou, M. (2012) Composite likelihood estimation for Thurstonian models with ranking data.
- IV Katsikatsou, M. (2012) Pairwise likelihood estimation for structural equation modeling with ordinal and continuous variables.

Reprints were made with permission from the publishers.



# Contents

1	Research Goal .....	7
2	Framework .....	9
2.1	Ordinal variables .....	9
2.2	Ranking variables .....	9
2.3	Relationship between ordinal and ranking variables .....	10
2.4	Latent variable modeling .....	11
2.4.1	Structural Equation Modeling .....	11
2.4.2	Item Response Theory approach .....	13
2.4.3	The Thurstonian model .....	14
3	Estimation methods overview .....	17
3.1	Estimation in Structural Equation Modeling .....	17
3.2	Estimation in Item Response Theory approach .....	18
3.3	Estimation in Thurstonian modeling .....	18
3.4	Composite likelihood estimation methods .....	19
4	Summary of Papers .....	21
4.1	Paper I .....	21
4.2	Paper II .....	22
4.3	Paper III .....	23
4.4	Paper IV .....	25
5	Contribution of the thesis .....	27
	Acknowledgments .....	28
	References .....	30



# 1. Research Goal

The main research goal of this thesis is to study the current estimation methods employed within latent variable modeling with data sets of ordinal and continuous variables, or data sets of ranking variables, and develop an alternative, hopefully better, estimation method. The criteria of comparison among the methods are their practical feasibility and the statistical properties of the provided estimators such as unbiasedness and consistency. The motivation for this research interest lies on the fact that maximum likelihood (ML) estimation is computationally infeasible for large latent models with the aforementioned type of data sets. On the other hand, the conventional step-wise limited information estimation approaches (Muthén, 1984) require the estimation of a weight matrix to compute correct standard errors, the dimension of which grows rapidly with the number of observed variables. Besides, relatively large sample sizes are needed to get a reliable estimate of the matrix. For this, a composite likelihood estimation method is developed for the type of models and data in question. The merits of the proposed approach are that it is computationally feasible regardless of the model size, it does not require the estimation of a weight matrix to provide correct standard errors, and it yields an estimator which is asymptotically unbiased, consistent and normally distributed (Lindsay, 1988; Varin, 2008; Varin et al., 2011).

To achieve the main goal the research project was split into four major parts resulting into the four papers composing this thesis. Firstly, before proceeding to any studies on the estimation of latent variable models with ranking variables, it was necessary to investigate the special nature of ranking data and how they can be analyzed within the framework of latent variable modeling. The latter can be done by applying the Thurstonian model (Thurstone, 1927). To get the model identified two approaches are suggested in the literature. The one is more well established but inference on a basic question ranking data aim to answer becomes tricky. The other, which has been recently suggested (Maydeu-Olivares & Böckenholt, 2005), makes the inference on all basic questions straightforward. This second identification strategy is investigated in Paper I by mainly checking how it affects the parameter estimates.

Ranking variables set up a more challenging framework than ordinal variables do, mainly due to their comparative and discrete nature. For this, the study of estimation methods starts off with a somehow simpler theoretical structure. Paper II focuses on factor analysis models with ordinal variables. The existing estimation methods are discussed and a composite, namely a pairwise likelihood (PL) estimation is developed. The performance of the latter is

studied in terms of bias and mean squared error (MSE) and compared with the performance of the conventional estimation methods via a simulation study and through some real data examples.

The positive results of the second paper combined with the conclusions of the first paper lead to Paper III. That firstly presents an overview of the existing estimation methods for Thurstonian models with ranking data followed by the development of a composite likelihood estimation method. In that paper as well, a simulation study investigates the performance of the suggested approach with respect to bias and MSE and in comparison with the current estimation approaches.

Paper IV extends the method presented in the second paper to the whole structural equation model where covariates may be included and the data may consist of both ordinal and continuous variables. The proposed method is demonstrated with an example of empirical data run in R software (R Development Team, 2008) and using the R package lavaan (Rosseel, 2012; Rosseel et al., 2012) which our self-written R code has been incorporated into. The R commands used in the example are provided.

Hopefully, this way, the thesis accomplishes the main research purpose and adds to the knowledge of composite maximum likelihood methods; in particular, how they can be applied within latent variable models with ranking variables or with ordinal and continuous variables, and how they perform with respect to bias and MSE and in comparison with the mainstream estimation methods.

## 2. Framework

### 2.1 Ordinal variables

In social sciences, the variables that are mainly encountered are categorical, ordinal and/or nominal. These are used to measure, among others, attitudes, beliefs, and abstract characteristics. The ordinal scale mostly employed is the Likert scale typically consisted of four, five, or seven points. Most of the time the points are labeled with numbers such as 1, 2, 3, etc., and with verbal text such as "Strongly disagree", "Disagree", "Agree", etc. Examples of ordinal variables are given in Table 2.1. Possible observations are, for example, (1,2,4), (3,4,5), (1,2,2), (2,5,5), etc. Within each vector, the numbers denote the response category chosen for each variable. In a statistical analysis with such variables, there are two important things that one should take into account. Firstly, the numbers of the response categories do not have metric properties; they just label ordered categories simply indicating different levels of a characteristic without giving information on the degree that the levels differ. Hence, addition (or subtraction) and multiplication (or division) are meaningless for ordinal variables. Secondly, whenever ordinal variables are used, an assumption is always implied, that the respondents understand the questions/ statements and use the response scale in the same way. Violation of this assumption leads to interpersonally incomparable responses which, in turn, are highly probable to lead to incorrect statistical results and inference (e.g. Brady, 1989).

### 2.2 Ranking variables

Ranking variables can also be used to measure attitudes, beliefs, and abstract characteristics. They are discrete as ordinal variables are but free of scale. In a ranking experiment, a set of  $m$  objects is presented to the respondents who are asked to assign a rank, from 1 to  $m$ , to all objects (complete ranking) or to a subset of them (partial ranking) according to their personal preference or a prespecified criterion. Usually 1 is defined as the rank to be assigned to the most preferred object and  $m$  to the least preferred one. An example of a ranking variable is given in Table 2.2. Possible responses are, for example, (1,2,3), (2,1,3), and (1,2,2) if ties are allowed. The main questions ranking variables seek to answer are a) which preference patterns are dominant, and b) what the inter-relationships among the ranked objects are. Adjacent

**Table 2.1.** *Example of ordinal variables*

Policy A is appropriate in order to deal with problem X.				
1	2	3	4	5
Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Policy B is appropriate in order to deal with problem X.				
1	2	3	4	5
Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Policy C is appropriate in order to deal with problem X.				
1	2	3	4	5
Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

ranks systematically assigned to certain objects indicate that these objects are perceived similarly and can possibly substitute each other. A detailed introduction to ranking variables and how they can be modeled and analyzed can be found in Marden (1995).

### 2.3 Relationship between ordinal and ranking variables

The relationship between a set of ordinal variables and a ranking variable can be illustrated through the examples discussed in sections 2.1 and 2.2. The first thing to note is that in the ordinal variable set-up the response pattern of an individual is a vector of dimension equal to the number of ordinal variables, while in the ranking set-up for each ranking variable a vector of size equal to the number of ranked objects is observed. The observations (1, 2, 4) and (3, 4, 5) for the three ordinal variables correspond to the observation (1, 2, 3) for the ranking variable and the observations (1, 2, 2) and (2, 5, 5) correspond to the ranking (1, 2, 2). Hence, ordinal data can be transformed into ranking data but not the other way round. A ranking response does not convey any information about the level an object is preferred. Therefore, the inference based on ranking data can only be made in relative terms and is valid only for the set of objects included in the ranking experiment. On the other hand, ranking designs impose less restrictions on the response mechanism than ordinal ones. They do not restrict participants to use a certain scale to provide their answer. This way, they avoid the assumption that all participants understand and use the scale in the same way. Because of these advantages and disadvantages

of ordinal and ranking variables, some researchers prefer designs where both types of variables are combined appropriately.

**Table 2.2.** *Example of a ranking variable*

Policy A, Policy B, and Policy C are three alternatives that can be applied in order to deal with problem X. Rank the three policies assigning rank 1 to the policy you consider as the most suitable, 2 to the next most suitable, and 3 to the least suitable.			
Policy	A	B	C
Rank	–	–	–

## 2.4 Latent variable modeling

Latent variable modeling is a multivariate type of modeling that can be applied for any type of observed variables. The observed variables are also referred to as indicators or items in the literature. The core idea of the analysis is that the latent variables account for the dependencies among the observed variables in the sense that if the former are held fixed, the latter are independent. The latent variables are also referred to as factors or constructs. Theoretically, latent variable model analysis can be distinguished into exploratory and confirmatory, but practice usually lies between the two. In exploratory analysis the goal is to summarize the information observed in a large data set of  $p$  variables into a smaller set of  $q$  latent variables, where  $q$  is much smaller than  $p$ . A typical example of such situation is questionnaires, which usually include at least 15 questions. In confirmatory analysis, the objective is to verify a theory where the variables of interest are abstract constructs such as those often faced in social and behavioral sciences. Examples of such variables are political beliefs, intelligence, emotional conditions, attitudes, etc. Hence, a latent variable model is specified in advance, the latent variables are measured through indicating variables, and the fit of the model to the empirical data is tested. Within latent variable modeling, there are two approaches, the Structural Equation Modeling (SEM) approach (e.g. Bollen, 1989; Jöreskog, 1990, 1994, 2002; Lee et al., 1990, 1992; Muthén, 1984) and the Item Response Theory (IRT) approach (e.g. Bartholomew et al., 2011; Muraki, 1990; Muraki & Carlson, 1995; Samejima, 1969). However, Bartholomew et al. (2011) explain how SEM, where both observed and latent variables are assumed normally distributed, can be seen as a special case of the more general IRT framework. A more detailed description of the two approaches is provided below.

### 2.4.1 Structural Equation Modeling

SEM mainly aims to test the hypothesis that the covariance matrix of the observed variables is equal to the covariance matrix implied by a hypothesized

latent variable model. SEM was developed first for continuous variables and then extended to ordinal ones by adopting the underlying response variable (URV) approach (Jöreskog, 1990, 1994; Muthén, 1984; Olsson, 1979).

Let  $\mathbf{x}$  be an observed  $p$ -dimensional vector of continuous variables assumed to follow a multivariate normal distribution. The SEM in its general form is written as follows:

$$\mathbf{x} = \mathbf{v} + \Lambda \xi + \mathbf{K}\mathbf{w} + \delta \quad (2.1)$$

$$\xi = \alpha + \mathbf{B}\xi + \Gamma\mathbf{z} + \zeta, \quad (2.2)$$

where  $\xi$  is a  $q$ -dimensional vector of latent variables,  $\mathbf{w}$  and  $\mathbf{z}$  are vectors of covariates,  $\delta$  and  $\zeta$  are vectors of error, and  $\mathbf{v}$  and  $\alpha$  are vectors of intercepts. The standard assumptions of the model are that: a)  $\mathbf{x}|\mathbf{w}, \mathbf{z} \sim N_p(\mu, \Sigma)$ , b)  $\xi$  follows a multivariate normal distribution, c)  $\delta \sim N_p(\mathbf{0}, \Theta)$  with  $\Theta$  being diagonal, d)  $\zeta \sim N_q(\mathbf{0}, \Psi)$ , e)  $Cov(\xi, \delta) = Cov(\xi, \zeta) = Cov(\delta, \zeta) = \mathbf{0}$ , and f)  $I - \mathbf{B}$  is not singular, where  $I$  is the identity matrix. Equation (2.1) is referred to as the measurement model and links the observed variable vector  $\mathbf{x}$  with the latent variable vector  $\xi$ . The effect of  $\xi$  on  $\mathbf{x}$  is given by the matrix of loadings  $\Lambda$  and the measurement model also reads as  $\xi$  is measured by  $\mathbf{x}$ . Equation (2.2) is referred to as the structural model and shows the relationships among the latent variables. The regression coefficients are included in matrix  $\mathbf{B}$ . In both equations covariates may be included. The only restriction is that  $\mathbf{w}$  and  $\mathbf{z}$  should contain different covariates for identification reasons.

Based on the model, it follows that:

$$\mu = E(\mathbf{x}|\mathbf{w}, \mathbf{z}) = \mathbf{v} + \Lambda(I - \mathbf{B})^{-1}(\alpha + \Gamma\mathbf{z}) + \mathbf{K}\mathbf{w},$$

$$\Sigma = Cov(\mathbf{x}|\mathbf{w}, \mathbf{z}) = \Lambda(I - \mathbf{B})^{-1}\Psi\left[(I - \mathbf{B})^{-1}\right]'\Lambda' + \Theta,$$

and

$$Cov(\mathbf{x}|\mathbf{w}, \mathbf{z}, \xi) = \Theta.$$

The last two equations along with the fact that  $\Theta$  is typically assumed to be diagonal implies that, given the covariates, SEM imposes a certain structure on the covariance matrix of  $\mathbf{x}$ , and accounts for all correlations among the observed variables once the values of latent variables are also given. For a detailed and instructive introduction to SEM with continuous variables see Bollen (1989).

To include ordinal observed variables into the model the (URV) approach is adopted which assumes that the ordinal variables are generated by underlying continuous variables. The connection between an ordinal variable  $x_i^*$  and its underlying continuous counterpart  $x_i$  is

$$x_i^* = c_{i,j} \iff \tau_{i,j-1} < x_i < \tau_{i,j}, \quad (2.3)$$

where  $c_{i,j}$  is the  $j$ -th response category of variable  $x_i^*$ ,  $j = 1, \dots, C_i$ ,  $\tau_{i,j}$  is the  $j$ -th threshold of variable  $x_i$ , and  $-\infty = \tau_{i,0} < \tau_{i,1} < \dots < \tau_{i,C_i-1} < \tau_{i,C_i} = +\infty$ .

Since only ordinal information is available in the data, the distribution of the underlying variable  $x_i$  is determined only up to a monotonic transformation. In practice it is often assumed that  $x_i \sim N(0, 1)$  and the thresholds are free to be estimated. An alternative parametrization where it is assumed  $x_i \sim N(\mu_i, \sigma_{ii})$  is also possible (e.g. Jöreskog, 2002).

Based on the URV approach, a  $p$ -dimensional vector of observed ordinal variables  $\mathbf{x}^*$  is matched to its underlying continuous counterpart  $\mathbf{x}$ . It is the latter that is involved in SEM, particularly in Equation (2.1). The probability of a response pattern for variable  $\mathbf{x}^*$  is written in terms of the distribution of  $\mathbf{x}$  as follows:

$$\pi(\mathbf{x}^*) = \pi(x_1^* = c_{1,j}, \dots, x_p^* = c_{p,j}) = \int_{\tau_{1,j-1}}^{\tau_{1,j}} \dots \int_{\tau_{p,j-1}}^{\tau_{p,j}} f(\mathbf{x}) d\mathbf{x}, \quad (2.4)$$

where  $f(\mathbf{x})$  is the assumed  $p$ -dimensional normal distribution of vector  $\mathbf{x}$ . If the observed variable vector is  $\mathbf{x}^* = \begin{pmatrix} \mathbf{x}_1^* \\ \mathbf{x}_2 \end{pmatrix}$ , where  $\mathbf{x}_1^*$  is a vector of  $p_1$  observed ordinal variables and  $\mathbf{x}_2$  is a vector of  $p_2$  observed continuous variables,  $p_1 + p_2 = p$ , then the vector  $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$  is defined, where  $\mathbf{x}_1$  is the vector of the  $p_1$  underlying continuous variables. Again, it is the latter which is involved in the measurement model. For an introduction of SEM with ordinal variables see for example Jöreskog (2002).

A special case of SEM is when only the measurement model is considered. Then, the model is usually called factor analysis model. The standard assumptions of SEM as well as the way that ordinal variables can be incorporated into the model remain the same.

### 2.4.2 Item Response Theory approach

The core assumption of IRT is that of conditional independence or otherwise local independence. It is assumed that the latent variables account for all the dependencies among the observed variables. This way, the conditional distribution of  $\mathbf{x}$  given  $\xi$ ,  $g(\mathbf{x}|\xi)$ , can be written as the product of the univariate conditional distributions of  $x_i$  given  $\xi$ ,  $g(x_i|\xi)$ , i.e.

$$g(\mathbf{x}|\xi) = \prod_{i=1}^p g(x_i|\xi).$$

This assumption cannot be tested but rather has the role of an axiom. IRT assumes the model:

$$f(\mathbf{x}) = \int_{R_\xi} g(\mathbf{x}|\xi) h(\xi) d\xi$$

which, based on the assumption of local independence, is simplified to

$$f(\mathbf{x}) = \int_{R_\xi} \prod_{i=1}^p g(x_i|\xi) h(\xi) d\xi, \quad (2.5)$$

where  $f(\mathbf{x})$  is the joint distribution of variable vector  $\mathbf{x}$ ,  $h(\xi)$  is the joint distribution of latent variables, and  $R_\xi$  is the latent variable area. The model is very general and it accommodates any type and combination of variables. Each observed variable can be of different type, since only its univariate conditional distribution  $g(x_i|\xi)$  needs to be determined.

In the case of an ordinal observed variable  $x_i^*$  the conditional distribution  $g(x_i^*|\xi)$  is a multinomial one, i.e.

$$g(x_i^*|\xi) = \prod_{j=1}^{C_i} \pi(x_i^* = c_{i,j}|\xi)^{I(x_i^* = c_{i,j})},$$

where  $I(x_i^* = c_{i,j})$  is the indicator variable whether  $x_i^*$  falls into the response category  $c_{i,j}$ . A measurement model is applied to the cumulative probabilities  $\gamma(x_i^* \leq c_{i,j}|\xi)$ , where

$$\pi(x_i^* = c_{i,j}|\xi) = \gamma(x_i^* \leq c_{i,j}|\xi) - \gamma(x_i^* \leq c_{i,j-1}|\xi).$$

A typical model for  $\gamma(x_i^* \leq c_{i,j}|\xi)$  in IRT is

$$\gamma(x_i^* \leq c_{i,j}|\xi) = F\left(\alpha_{i,j} - \sum_{k=1}^q \beta_{ik} \xi_k\right), \quad (2.6)$$

where the  $\alpha_{i,j}$ 's are thresholds ( $-\infty = \alpha_{i,0} < \alpha_{i,1} < \dots < \alpha_{i,C_i-1} < \alpha_{i,C_i} = +\infty$ ), the  $\beta_{ik}$ 's are loadings, and  $F$  is a link function (e.g. logit, probit, etc.). A very rigorous introduction to IRT and its links to SEM can be found in Bartholomew et al. (2011).

### 2.4.3 The Thurstonian model

In an experiment of complete ranking of  $m$  objects,  $\{O_1, \dots, O_m\}$ , where no ties are allowed the sampling distribution is a multinomial one with  $m!$  categories. The log-likelihood function of a random sample of  $n$  ranking vectors is of the form:

$$\ln L(\theta; (\mathbf{r}_1, \dots, \mathbf{r}_n)) = \sum_{c=1}^{m!} n_c \ln \pi_c(\theta), \quad (2.7)$$

where  $\mathbf{r}_i$  is the  $i$ -th observed ranking which is an  $m$ -dimensional vector of permuted integers from 1 to  $m$ ,  $\theta$  is a parameter vector,  $n_c$  is the observed frequency of ranking pattern  $c$ ,  $\sum_{c=1}^{m!} n_c = n$ , and  $\pi_c(\theta)$  is the corresponding probability under the model,  $\pi_c(\theta) > 0$ , and  $\sum_{c=1}^{m!} \pi_c(\theta) = 1$ .

One of the most influential models for ranking data in the literature is that suggested by Thurstone (1927). Maydeu-Olivares & Böckenholt (2005) show how a Thurstonian model can be incorporated in the more general framework of SEM. The basic idea of analysis is similar to that of the URV approach. Each ranked object is assumed to have an underlying continuous utility, i.e. the observed ranks assigned to the objects are assumed to be the result of the underlying object utilities. Furthermore, it is assumed that the differences in object utility assessments follow a multinormal distribution.

Let  $u_j$  be the underlying utility of object  $O_j$ ,  $j = 1, \dots, m$ , and  $\mathbf{u}' = (u_1, u_2, \dots, u_m)$  be the  $m$ -dimensional random vector of underlying utilities. As said, it is assumed that

$$\mathbf{u} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Let also:

- $O_{(h)}^c$  denote the object which has been assigned the rank  $h$  given the complete ranking pattern  $c$ ,  $h = 1, \dots, m$ ,  $c = 1, \dots, m!$ ,
- $\tilde{u}_i^c = u_{O_{(i)}^c} - u_{O_{(i+1)}^c}$  be the utility difference between objects with adjacent ranks within the ranking pattern  $c$ ,  $i = 1, \dots, m - 1$ ,
- $\tilde{\mathbf{u}}_c$  be the  $(m - 1)$ -dimensional vector containing all the above utility differences,
- $C_c$  be an  $(m - 1) \times m$  contrast matrix transforming vector  $\mathbf{u}$  into  $\tilde{\mathbf{u}}_c$ , i.e. its exact form depends on the ranking pattern  $c$  in question, and
- $D_c = [\text{diag}(C_c \boldsymbol{\Sigma} C_c')]^{-1/2}$ , where  $\text{diag}$  is the function which takes as an argument a square matrix and returns a diagonal matrix with the same main diagonal elements.

The probability of the ranking pattern  $c$ ,  $\pi_c$ , is modeled as follows:

$$\pi_c(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \Phi_{m-1}(D_c C_c \boldsymbol{\mu}; D_c C_c \boldsymbol{\Sigma} C_c' D_c), \quad (2.8)$$

where  $\Phi_{m-1}(D_c C_c \boldsymbol{\mu}; D_c C_c \boldsymbol{\Sigma} C_c' D_c)$  is the  $(m - 1)$ -dimensional cumulative normal distribution with correlation matrix  $D_c C_c \boldsymbol{\Sigma} C_c' D_c$  evaluated at the point  $D_c C_c \boldsymbol{\mu}$ . In the unrestricted Thurstonian model where no structure is assumed for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , the above model faces serious identification issues. For this reason, the  $(m - 1)$ -dimensional random vector  $\tilde{\mathbf{u}}$  is considered instead, where  $\tilde{\mathbf{u}}$  contains all the object utility differences with respect to the utility of a reference object (e.g. Chan & Bentler, 1998; Yao & Böckenholt, 1999). Choosing object  $O_1$  as a reference object, the random vector  $\tilde{\mathbf{u}}$  is of the form

$$\tilde{\mathbf{u}}' = (\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_{m-1}) = (u_1 - u_2, u_1 - u_3, \dots, u_1 - u_m).$$

It holds that:

$$\tilde{\mathbf{u}} = B\mathbf{u}, \quad (2.9)$$

where  $B$  is a  $(m-1) \times m$  contrast matrix of the form:

$$B = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ \vdots & & & \ddots & & \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Hence,  $\tilde{\mathbf{u}} \sim N_{m-1}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$  with  $\tilde{\boldsymbol{\mu}} = B\boldsymbol{\mu}$  and  $\tilde{\boldsymbol{\Sigma}} = B\Sigma B'$ . The ranking probability  $\pi_c$  is modified to:

$$\pi_c(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) = \Phi_{m-1}(\tilde{D}_c C_c \tilde{\boldsymbol{\mu}}; \tilde{D}_c C_c \tilde{\boldsymbol{\Sigma}} C_c' \tilde{D}_c), \quad (2.10)$$

where  $\tilde{D}_c = [\text{diag}(C_c \tilde{\boldsymbol{\Sigma}} C_c')]^{-1/2}$  and  $C_c$  is now an  $(m-1) \times (m-1)$  contrast matrix transforming vector  $\tilde{\mathbf{u}}$  into  $\tilde{\mathbf{u}}_c$ .

### 3. Estimation methods overview

The standard method of maximum likelihood (ML) estimation becomes computationally demanding or even impractical for relatively large latent models where some of the observed variables are ordinal or ranking and the latent variables are assumed normally distributed. This has motivated the development of limited information estimation methods, especially within SEM with ordinal variables and Thurstonian ranking models. A brief discussion about estimation under the three types of models mentioned in Section 2 is given below.

#### 3.1 Estimation in Structural Equation Modeling

Full maximum likelihood estimation is not practical under SEM with ordinal variables even when the number of ordinal variables is small (Lee et al., 1992; Liu, 2007; Poon & Lee, 1987). To demonstrate this, consider the variable vector  $\mathbf{x}^* = \begin{pmatrix} \mathbf{x}_1^* \\ \mathbf{x}_2 \end{pmatrix}$  and the SEM defined in Section 2.4.1. Then, the log-likelihood function for one observation is:

$$\ln L(\boldsymbol{\theta}; \mathbf{x}^*) = \ln \pi(\mathbf{x}_1^* | \mathbf{x}_2; \boldsymbol{\theta}) + \ln f(\mathbf{x}_2; \boldsymbol{\theta}), \quad (3.1)$$

where the parameter vector  $\boldsymbol{\theta}$  includes the thresholds and the SEM parameters,  $f(\mathbf{x}_2; \boldsymbol{\theta})$  is the multinormal distribution of the observed continuous variables  $\mathbf{x}_2$ , and  $\pi(\mathbf{x}_1^* | \mathbf{x}_2; \boldsymbol{\theta})$  is the probability of the response pattern for  $\mathbf{x}_1^*$  given  $\mathbf{x}_2$ . The latter is written as in Equation (2.4) with the difference that the distribution to be integrated is the conditional multinormal distribution of the underlying continuous variables  $\mathbf{x}_1$  given the observed continuous variables  $\mathbf{x}_2$ ,  $f(\mathbf{x}_1 | \mathbf{x}_2; \boldsymbol{\theta})$ . Hence, ML requires the evaluation of  $p_1$ -dimensional normal probabilities the computation of which is possible only for a very small number  $p_1$ . In practice though, this number is very rarely small.

The impractical nature of ML has initiated the development of limited information estimation methods that usually require the evaluation of up to bi-variate normal probabilities. Among them the three-stage estimation method proposed by Muthén (1984) is widely employed as well as implemented in commercial software such as LISREL (Jöreskog & Sörbom, 1996) and Mplus (Muthén & Muthén, 2010). In the first stage of the method, the thresholds of the underlying variables and the product-moment correlations, if continuous

observed variables are present, are estimated. The thresholds are estimated by maximizing one by one all the univariate likelihood functions of ordinal variables. In the second stage, given the threshold estimates, the polychoric correlations of all pairs of ordinal variables and the polyserial correlations of all pairs of one ordinal and one observed continuous variable are estimated by maximizing one by one the corresponding bivariate likelihood functions. These computations involve the evaluation of one- and two-dimensional normal probabilities only. Finally, in the third stage, given the estimates of all types of correlations, the SEM parameters are estimated by applying a type of least squares, i.e. weighted least squares (WLS), diagonally weighted least squares (DWLS), or unweighted least squares (ULS) (for comparisons see e.g. Forero et al., 2009; Yang-Wallentin et al., 2010). The weight matrix is an estimate of the asymptotic covariance matrix of the estimated correlations. The full weight matrix is needed in all three versions of least squares to compute correct standard errors. Then, DWLS and ULS are called robust DWLS (RDWLS) and robust ULS (RULS). A point to consider with respect to this three-stage estimation is that the size of the weight matrix grows very rapidly with the number of observed variables, e.g. for 15 observed variables, the matrix is  $105 \times 105$ . This way, the computational simplicity of the approach achieved in the first two stages is somehow canceled out by the size of the weight matrix. Besides, for a reliable estimate of the matrix a fairly large sample size is desired.

### 3.2 Estimation in Item Response Theory approach

Maximum likelihood estimation is the standard approach in IRT modeling. The E-M or Newton-Raphson algorithm are usually employed to carry out the maximization of the objective function. In both cases evaluations of multiple integrals the dimension of which is equal to the number of latent variables are required. These computations are performed numerically using methods such as Gauss-Hermite quadrature, adaptive quadrature, and Monte Carlo (for a discussion of different algorithms see e.g. Feddag & Bacci, 2009; Schilling & Bock, 2005). However, for all these algorithms, for a fixed level of computational accuracy, the computation time and burden grows fast with the dimension of the integration. Besides, the convergence of the EM algorithm slows down as the number of latent variables increases. As a result, ML becomes infeasible beyond a certain number of latent variables.

### 3.3 Estimation in Thurstonian modeling

The maximization of the log-likelihood in (2.7) requires the evaluation of  $(m - 1)$ -dimensional normal probabilities as defined in (2.10). To evalu-

ate such probabilities, the algorithm suggested by Schervish (1984) can be employed. However, the computational time increases very rapidly with the number of objects  $m$  so that ML is practically infeasible for even a moderate  $m$ . As a consequence, limited information estimation methods employing mainly generalized least squares (GLS) have been proposed (Brady, 1989; Chan & Bentler, 1998; Maydeu-Olivares & Böckenholt, 2005). In these approaches, the estimates are obtained by minimizing a fit function of the type:

$$F(\tilde{\theta}) = (\mathbf{p} - \pi(\tilde{\theta})) W^{-1} (\mathbf{p} - \pi(\tilde{\theta}))',$$

where  $\tilde{\theta}' = (\tilde{\mu}', [\text{vech}(\tilde{\Sigma})]')$ ,  $\text{vech}$  is the function transforming a symmetric matrix into a vector by stacking the elements of the lower triangular of the matrix, including its main diagonal, columnwise,  $\tilde{\mu}$  and  $\tilde{\Sigma}$  are as defined in Section 2.4.3,  $\mathbf{p}$  is the vector of observed low-order ranking probabilities,  $\pi(\tilde{\theta})$  is the vector of the corresponding probabilities under the model,  $W$  is a sample estimate of the covariance matrix of the random vector  $\mathbf{p}$ , and  $W^{-1}$  is the generalized inverse of  $W$ . The generalized inverse is computed as  $W$  is singular due to the inter-relationships among low-order rankings. Although GLS aims to be an estimation method of low computational complexity, the size of  $W$  grows extremely fast with the number of objects  $m$  rendering the inversion of the matrix and subsequently, the method computationally demanding. Maydeu-Olivares & Böckenholt (2005) show how this GLS approach can be carried out by the conventional three-stage estimation approach applied in SEM and described in Section 3.1.

### 3.4 Composite likelihood estimation methods

Composite likelihood estimation methods, already applied to a range of models, are gaining more research attention recently because they can substantially simplify the computations involved in estimation and at the same time yield asymptotically unbiased, consistent, and normally distributed estimators (Lindsay, 1988). Varin (2008) and Varin et al. (2011) give an extensive overview of these methods and their application areas.

In situations where the likelihood function either cannot be specified or is impractical to work with due to high computational complexity, one could consider instead a composite likelihood function. The latter is defined as follows (Lindsay, 1988; Varin, 2008; Varin et al., 2011): let  $\mathbf{x}$  be a  $p$ -dimensional random vector with probability density  $f(\mathbf{x}; \omega)$  for some unknown vector parameter  $\omega \in \Omega$ . Let  $\{\mathcal{A}_1, \dots, \mathcal{A}_K\}$  be a set of measurable marginal or conditional events with associated likelihoods  $\mathcal{L}_k(\omega; \mathbf{x}) \propto f(\mathbf{x} \in \mathcal{A}_k; \omega)$ . A composite likelihood (CL) function is the weighted product of the likelihoods cor-

responding to each single event,

$$CL(\boldsymbol{\omega}; \mathbf{x}) = \prod_{k=1}^K \mathcal{L}_k(\boldsymbol{\omega}; \mathbf{x})^{w_k},$$

where  $w_k$  are non-negative weights to be chosen. The maximum composite likelihood estimator  $\hat{\boldsymbol{\omega}}_{MCL}$  is obtained by maximizing the function  $CL(\boldsymbol{\omega}; \mathbf{x})$  over the parameter  $\boldsymbol{\omega}$ . Under regularity conditions on the component likelihoods, the central limit theorem for the composite likelihood score statistic can be applied leading to the result

$$\sqrt{n}(\hat{\boldsymbol{\omega}}_{MCL} - \boldsymbol{\omega}) \xrightarrow{d} N(0, G^{-1}(\boldsymbol{\omega})),$$

where  $G(\boldsymbol{\omega})$  is the Godambe (also called sandwich) information matrix of a single observation. In particular,

$$G(\boldsymbol{\omega}) = H(\boldsymbol{\omega})J^{-1}(\boldsymbol{\omega})H(\boldsymbol{\omega}),$$

where  $H(\boldsymbol{\omega}) = E\{-\nabla^2 \ln CL(\boldsymbol{\omega}; \mathbf{x})\}$ , and  $J(\boldsymbol{\omega}) = Var\{\nabla \ln CL(\boldsymbol{\omega}; \mathbf{x})\}$ . In general, the identity  $H(\boldsymbol{\omega}) = -J(\boldsymbol{\omega})$  does not hold because the assumed independence among the likelihood terms forming the composite function is usually not valid when the full likelihood is considered.

Varin et al. (2011) discuss some further qualities of the composite likelihood approach. It can be seen as a robust alternative in terms of modeling. It is easier and more straightforward to model lower order dimensional distributions while modeling uncertainty increases with dimensionality. By applying composite likelihood, possible misspecification of the higher order dimensional distributions can be avoided. In addition, a model assumed for lower order distributions can be compatible with more than one possible modeling options available for higher dimensional distributions.

## 4. Summary of Papers

### 4.1 Paper I

The framework of Paper I is that described in Section 2.4.3. The research interest lies on the identification of the model written with respect to  $\mu$  and  $\Sigma$  (see Equation (2.8)) and the research question is what impact the identification strategy proposed by Maydeu-Olivares & Böckenholt (2005) has on the parameter estimates.

More specifically, due to the discrete and comparative nature of ranking data, the utility random vector  $\mathbf{u}$  is unique only up to a linear transformation. The origin and the unit of the utility scale should be defined. This is usually done by fixing the utility mean and variance of one object equal to 0 and 1, respectively. Let  $\mu_1 = 0$  and  $\sigma_{11} = 1$ . As a consequence, the elements of  $\mu$  and  $\Sigma$  should be interpreted in relative terms and the inference should be based on the estimates of the standardized parameters, i.e. the correlations among object utilities  $\rho_{ij}$ ,  $i \neq j = 1, \dots, m$ , the ratios of object utilities variances, e.g.  $\sigma_{jj}/\sigma_{11}$ , and the standardized mean utility differences, e.g.  $(\mu_1 - \mu_j)/\sqrt{\sigma_{11} - 2\sigma_{1j} + \sigma_{jj}}$ ,  $j = 2, \dots, m$ .

However, the model written with respect to  $\mu$  and  $\Sigma$  is identified only if at least  $m$  extra constraints, additional to those defining the scale origin and unit, are set. Maydeu-Olivares & Böckenholt (2005) suggest that one could fix the covariances of the utility of the  $m$ -th object with all other object utilities equal to 0, i.e.  $\sigma_{mi} = 0$ ,  $i = 1, \dots, m - 1$ , and the variance of the last object  $\sigma_{mm}$  equal to 1. These  $m$  extra constraints imply that the correlations  $\rho_{mi}$ ,  $i = 1, \dots, m - 1$ , are equal to 0 and the variance ratio  $\sigma_{mm}/\sigma_{11}$  is equal to 1, something which may not hold in the population. Consequently, the model to be estimated may be misspecified and then, the estimates of the free correlations and variance ratios are expected to present bias and relatively high MSE. Note that the goodness-of-fit statistics are unable to detect possible model misspecifications coming from the extra identification constraints.

To investigate the size of the misspecification impact on the estimates a simulation study is conducted. 36 different experimental conditions derived by nine different misspecification situations, two model sizes ( $m = 4, 7$ ), and two sample sizes ( $n = 500, 1000$ ), are examined. The bias and MSE of the estimates obtained under no misspecification within each combination of model and sample size are used as benchmarks. The results indicate that the identification approach suggested in the literature leads to reliable estimates of all the standardized parameters as long as the extra identification constraints coincide

with the true values of the constrained parameters. When this does not hold, the estimates of almost all correlations and variance ratios are seriously biased and present relatively high MSE. The level of bias and MSE increases with the misspecification level and not in a uniform way for all parameters. An increase in the sample size seems to have very marginal effect in decreasing the bias and MSE. As a result, the approach should be used with great caution. Before adopting any extra constraints one should resort to already existing theory or previous empirical studies concerning the specific set of objects to confirm that the extra constraints are reasonable and can be justified for the population in question.

## 4.2 Paper II

Papers II focuses on latent variable model analysis, both confirmatory and exploratory, of ordinal variables. Factor analysis models both under the URV approach and the IRT approach, as described in sections 2.4.1 and 2.4.2, respectively, are considered. The research objective is to examine the conventional estimation methods employed within these two types of modeling, and propose a new one that is of low computational complexity regardless of the model size and performs equally well as the current methods.

Within factor analysis, where the URV approach is adopted, the model is

$$\mathbf{x} = \Lambda\xi + \delta, \quad (4.1)$$

where  $\mathbf{x}$  is the vector of underlying continuous variables corresponding to the vector of the observed ordinal variables  $\mathbf{x}^*$ ,  $\xi$  is the vector of latent variables, and  $\delta$  is the vector of errors. It is assumed that each underlying continuous variable follows standard normal distribution,  $\xi \sim N_q(\mathbf{0}, \Phi)$ , where  $\Phi$  has ones on its main diagonal,  $\delta \sim N_p(\mathbf{0}, \Theta)$  with  $\Theta = I - \text{diag}(\Lambda\Phi\Lambda')$ , and  $\text{Cov}(\xi, \delta) = \mathbf{0}$ . The parameter vector to be estimated is  $\theta' = (\lambda', \varphi', \tau')$ , where  $\lambda$  and  $\varphi$  are the vectors of free non-redundant parameters in matrices  $\Lambda$  and  $\Phi$ , respectively, and  $\tau$  is the vector of all free thresholds. As explained in Section 3.1, ML estimation of  $\theta$  is practically infeasible and the three-stage limited information estimation presented by Muthén (1984) is applied.

Under the IRT approach, the model in Equation (2.6) in Section 2.4.2 is considered. As said in Section 3.2, ML is the standard estimation method but it becomes impractical when the number of latent variables is large.

We propose a pairwise maximum likelihood (PML) method under the factor analysis model and the URV approach to estimate the parameter  $\theta$ . The PML estimator  $\hat{\theta}_{PML}$  is the value of  $\theta$  maximizing the pairwise log-likelihood

function. The form of the latter for one observation is:

$$pl(\theta; \mathbf{x}^*) = \sum_{i < i'} \ln L(\theta; (x_i^*, x_{i'}^*)) = \quad (4.2)$$

$$= \sum_{i < i'} \sum_{j=1}^{C_i} \sum_{j'=1}^{C_{i'}} I(x_i^* = c_{i,j}, x_{i'}^* = c_{i',j'}) \ln \pi_{c_{i,j}c_{i',j'}}(\theta), \quad (4.3)$$

where,  $I(x_i^* = c_{i,j}, x_{i'}^* = c_{i',j'})$  is the indicator variable taking the value 1 if the variables  $x_i^*$  and  $x_{i'}^*$  fall into the categories  $c_{i,j}$  and  $c_{i',j'}$ , respectively, and 0 otherwise; based on the Equation (2.4),

$$\begin{aligned} \pi_{c_{i,j}c_{i',j'}}(\theta) &= \pi(x_i^* = c_{i,j}, x_{i'}^* = c_{i',j'}; \theta) = \\ &= \Phi_2(\tau_{i,j}, \tau_{i',j'}; \rho_{x_i x_{i'}}) - \Phi_2(\tau_{i,j}, \tau_{i',j'-1}; \rho_{x_i x_{i'}}) - \\ &\quad - \Phi_2(\tau_{i,j-1}, \tau_{i',j'}; \rho_{x_i x_{i'}}) + \Phi_2(\tau_{i,j-1}, \tau_{i',j'-1}; \rho_{x_i x_{i'}}), \\ \rho_{x_i x_{i'}}(\theta) &= (\lambda_{i \cdot}) \Phi(\lambda_{i' \cdot})', \end{aligned}$$

and  $\lambda_{i \cdot}$  is a  $1 \times q$  row vector containing the elements of the  $i$ -th row of matrix  $\Lambda$ . Since PML belongs to the general family of composite likelihood methods, the general result reported in Section 3.4 can be applied. Hence, the estimator  $\hat{\theta}_{PML}$  is asymptotically unbiased, consistent, and normally distributed. The advantage of PML over ML is mainly computational since the former involves the evaluation of integrals of bivariate normal distributions only, regardless of the number of observed ordinal variables or factors. The main advantages of PML over the three-stage limited information estimators are that all model parameters are estimated in one single step and there is no need of estimating a weight matrix to obtain correct standard errors.

The performance of PML estimator in finite samples with respect to bias and MSE is studied via a simulation study under eight experimental conditions (four different sample sizes and two model sizes). It is also compared with the performance of the three-stage approaches RDWLS and RULS, and that of ML as implemented under the IRT approach. Moreover, PML is demonstrated and compared with RDWLS, RULS, and ML within some real data examples both in an exploratory and confirmatory analysis set-up. The general conclusions are that: a) PML estimates and their standard errors have bias and MSE very close to zero, both decreasing with the sample size, b) all the methods considered in the study provide very similar results, and c) there is a tendency for the PML and RDWLS estimates and standard errors to be slightly closer to those of ML than those of the RULS approach.

### 4.3 Paper III

The promising results of the second paper lead to Paper III. The framework is the unrestricted Thurstonian model with ranking data as described in Section

2.4.3. Based on the results of the first paper we consider the estimation of the model written with respect to  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\boldsymbol{\Sigma}}$  (Equation (2.10)). This model is identified by only defining the unit of the scale of utility differences. An overview of the existing estimation methods is given in the paper, a summary of which is provided in Section 3.3.

The composite likelihood estimation proposed is based on the notion of trinary rankings. Trinary rankings are the relative rankings of triplets of objects implied by a given complete ranking of  $m$  objects. Chan & Bentler (1998) explain that it is enough to consider only the  $(m-1)(m-2)/2$  triplets which include the reference object. The trinary rankings of the rest of the object triplets contain redundant information. This way, we suggest a trinary composite likelihood (TCL) estimation where, for one observation, the log-likelihood function to be maximized is:

$$l_{tc}(\tilde{\boldsymbol{\theta}}; \mathbf{r}) = \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \ln L_{ij}(\tilde{\boldsymbol{\theta}}; \mathbf{r}_{(O_1, O_{i+1}, O_{j+1})}), \quad (4.4)$$

where  $\tilde{\boldsymbol{\theta}}' = \left( \tilde{\boldsymbol{\mu}}', [\text{vech}(\tilde{\boldsymbol{\Sigma}})]' \right)$ ,  $\mathbf{r}_{(O_1, O_{i+1}, O_{j+1})}$  is the observed trinary ranking of the triplet  $(O_1, O_{i+1}, O_{j+1})$  implied by the complete ranking of the  $m$  objects  $\mathbf{r}$ , and  $\ln L_{ij}(\tilde{\boldsymbol{\theta}}; \mathbf{r}_{(O_1, O_{i+1}, O_{j+1})})$  is the log-likelihood function for this triplet of objects. The specific form of the latter is:

$$\ln L_{ij}(\tilde{\boldsymbol{\theta}}; \mathbf{r}_{(O_1, O_{i+1}, O_{j+1})}) = \sum_{t=1}^6 I(\mathbf{r}_{(O_1, O_{i+1}, O_{j+1})} = t) \ln \pi_t^{(O_1, O_{i+1}, O_{j+1})}(\tilde{\boldsymbol{\theta}}), \quad (4.5)$$

where  $I(\mathbf{r}_{(O_1, O_{i+1}, O_{j+1})} = t)$  is the indicator variable which takes the value 1 if the observed trinary ranking for the triplet  $(O_1, O_{i+1}, O_{j+1})$  is equal to pattern  $t$  and 0 otherwise, and  $\pi_t^{(O_1, O_{i+1}, O_{j+1})}(\tilde{\boldsymbol{\theta}})$  is the corresponding probability under the model. Based on Equation (2.10), this probability is written as follows:

$$\pi_t^{(O_1, O_{i+1}, O_{j+1})}(\tilde{\boldsymbol{\theta}}) = \Phi_2(D_t^{(ij)} C_t \tilde{\boldsymbol{\mu}}^{(ij)}; D_t^{(ij)} C_t \tilde{\boldsymbol{\Sigma}}^{(ij)} C_t' D_t^{(ij)}),$$

$$\text{where } \tilde{\boldsymbol{\mu}}^{(ij)} = \begin{pmatrix} \tilde{\boldsymbol{\mu}}_i \\ \tilde{\boldsymbol{\mu}}_j \end{pmatrix}, \quad \tilde{\boldsymbol{\Sigma}}^{(ij)} = \begin{cases} \begin{pmatrix} c & \\ \tilde{\sigma}_{j1} & \tilde{\sigma}_{jj} \end{pmatrix} & \text{if } i = 1 \\ \begin{pmatrix} \tilde{\sigma}_{ii} & \\ \tilde{\sigma}_{ji} & \tilde{\sigma}_{jj} \end{pmatrix} & \text{otherwise} \end{cases},$$

$D_t^{(ij)} = [\text{diag}(C_t \tilde{\boldsymbol{\Sigma}}^{(ij)} C_t')]^{-1/2}$ ,  $C_t$  is a  $2 \times 2$  contrast matrix, and  $c$  is a known positive constant. Note that  $\tilde{\sigma}_{11}$  is fixed to a constant  $c$ ,  $c > 0$ , to define the scale unit.

The estimator  $\hat{\boldsymbol{\theta}}_{TCL}$  shares the asymptotic properties of the composite likelihood estimators detailed in Section 3.4. The method is computationally gen-

eral as it involves the evaluation of only bivariate normal probabilities regardless of the number of objects  $m$ . Compared to the three-stage SEM estimation methods, the TCL approach estimates all parameters simultaneously and it does not require the estimate a weight matrix to get correct standard errors. The performance of TCL estimation in finite samples under different model sizes and sample sizes is investigated with respect to relative Bias and relative MSE through a simulation study. It is also compared with the performance of RDWLS and RULS as implemented within SEM with ordinal variables. It is found that all three methods yield similar estimates and standard errors for all experimental conditions with TCL and RULS performing slightly better than RDWLS with respect to relative bias. Interestingly enough, the great deal of redundant information that is used within the three-stage RULS and RDWLS approaches does not affect the accuracy and efficiency of the methods as those are compared with the accuracy and efficiency of the TCL method.

## 4.4 Paper IV

Paper IV extends the pairwise estimation method proposed in the second paper to the whole SEM where covariates may be included and the indicators of the latent variables can be both ordinal and continuous. The variable vector  $\mathbf{x}^*$  and the SEM model introduced in Section 2.4.1 (the model consisting of equations (2.1) and (2.2)) are considered. Hence, the pairwise log-likelihood function for one observation of Equation (4.2) is modified to:

$$pl(\theta; \mathbf{x}^*) = \sum_{i < i'} \ln L(\theta; (x_{1i}^*, x_{1i'}^*)) + \sum_{k < k'} \ln L(\theta; (x_{2k}, x_{2k'})) + \sum_i \sum_k \ln L(\theta; (x_{1i}^*, x_{2k})),$$

where  $\ln L(\theta; (x_{1i}^*, x_{1i'}^*))$  is the bivariate log-likelihood function of a pair of ordinal variables,  $\ln L(\theta; (x_{2k}, x_{2k'}))$  is the bivariate log-likelihood of a pair of observed continuous variables, and  $\ln L(\theta; (x_{1i}^*, x_{2k}))$  is the bivariate log-likelihood function of a pair of one ordinal and one continuous observed variables. The form of  $\ln L(\theta; (x_{1i}^*, x_{1i'}^*))$  is as defined in Equation (4.3), the function  $\ln L(\theta; (x_{2k}, x_{2k'}))$  is a bivariate normal log-likelihood, and

$$\begin{aligned} \ln L(\theta; (x_{1i}^*, x_{2k})) &= \ln f(x_{1i}^* | x_{2k}; \theta) + \ln f(x_{2k}; \theta) = \\ &= \sum_{j=1}^{C_i} I(x_{1i}^* = c_{i,j} | x_{2k}) \ln \pi(x_{1i}^* = c_{i,j} | x_{2k}; \theta) + \ln f(x_{2k}; \theta). \end{aligned}$$

The function  $\ln f(x_{2k}; \theta)$  is the log-likelihood of a univariate normal distribution,  $I(x_{1i}^* = c_{i,j} | x_{2k})$  is the indicator variable taking the value 1 if the ordinal variable  $x_{1i}^*$  falls into the category  $c_{i,j}$  given the value of  $x_{2k}$  and 0 otherwise, and  $\pi(x_{1i}^* = c_{i,j} | x_{2k}; \theta)$  is the probability that  $I(x_{1i}^* = c_{i,j} | x_{2k})$  takes the value 1. The latter is written in terms of univariate normal probabilities.

It is worthy to note that the proposed method does not apply only to SEM but to any kind of model that assumes a parametric structure for  $\mu$  and  $\Sigma$ , where  $\mu = E(\mathbf{x}|\mathbf{w}, \mathbf{z})$ ,  $\Sigma = Cov(\mathbf{x}|\mathbf{w}, \mathbf{z})$ ,  $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ , and  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are the vectors of underlying and observed continuous variables, respectively. Observe also that PL estimation assumes only bivariate normality among the elements of the vector  $\mathbf{x}$  given  $\mathbf{w}$  and  $\mathbf{z}$  and not joint normality, i.e.  $\mathbf{x}|\mathbf{w}, \mathbf{z} \sim N(\mu, \Sigma)$ , as SEM does. That it is an important difference since bivariate normality is an assumption that can be tested (see e.g. Jöreskog, 2002).

The suggested method is demonstrated using an example with empirical data. To run the model an R code has been written and later on incorporated into the R package lavaan (version 0.5-11). The R code used for the example is provided. Interestingly enough, ML as implemented in Mplus (version 5.21) and LISREL (version 9.10) is not feasible for the demonstrated example. On the other hand, RDWLS gives very similar estimates and slightly smaller standard errors than those of PL.

## 5. Contribution of the thesis

In this thesis a composite, namely a pairwise likelihood estimation method is developed for SEM with ordinal and continuous variables where covariates may also be included, and is appropriately adjusted for Thurstonian models with ranking variables. The performance of the method with respect to bias and MSE in finite samples is studied within factor analysis models with ordinal variables, and within Thurstonian models with ranking variables. Moreover, in both types of models, the performance of PL is compared with that of RDWLS and RULS. ML as implemented under the IRT approach is also considered in the performance comparisons in the case of factor analysis. The main conclusion of the simulation studies is that PL shows a close to zero bias and MSE decreasing with the sample size and it is competitive to the other estimation methods. The proposed approach is also demonstrated with examples of empirical ordinal data in the case of exploratory and confirmatory factor analysis, and in the case of SEM. To run all these models a code in R has been written and later on a part of it has been incorporated in the R package `lavaan` (version 0.5-11). This way, PL for SEM with ordinal variables is accessible to researchers at the time this thesis is written. A secondary result of the thesis is that one should be careful with the formulation of the unrestricted Thurstonian ranking model. A certain parametrization requires extra identification constraints than the typical ones which are highly probable to affect the quality of the estimates in terms of bias and MSE negatively.

The thesis adds knowledge on composite likelihood estimation methods applied to latent variable models with ordinal and continuous variables, and latent variables models with ranking variables. Optimistically, it could be used as the starting point for further research on the performance of PL under more experimental conditions, especially under large models with small sample size; on the development of chi-square test statistics and model selection criteria under PL for SEM and Thurstonian models; and on the treatment of missing values when PL is applied.

# Acknowledgments

I am enormously thankful to my supervisors, Prof. Fan Yang-Wallentin, Prof. Irini Moustaki, and Prof. Karl Jöreskog, who taught me and guided me scientifically as well as supported me psychologically. It is self-evident that without their tremendous help I would not have been able to accomplish the demanding task of PhD. Fan, thank you so much for being always available for advice, solving various annoying administrative issues for me, and taking the effort to ensure more than enough financial support for our project. Irini, thank you so much for "bringing me up" in the world of statistics from my day one up to now with understanding and patience. The knowledge and encouragement you offered me during my master thesis were decisive in applying for a PhD. Thank you for actively supervising my PhD without being paid! Karl, it was an honor and privilege to have such an outstanding researcher of the field like you as my supervisor. Your expertise and experience were more than valuable.

I am very thankful to Prof. Yves Rosseel of Department of Data Analysis of Ghent University in Belgium who enthusiastically accepted to cooperate with me in order to incorporate my amateurish R code in his amazing R package `lavan`. Apart from the fact that my programming skills have been improved thanks to this cooperation, I find that my PhD gets more value as part of my research becomes easily accessible to other researchers through his extremely user-friendly package.

I want to express my appreciation to all professors, teachers, PhD students, and staff of the department. I could not wish a better working environment. The fruitful statistical discussions, the substantial help with various tasks during the PhD, the encouragement, the financial support, and of course the refreshing coffee breaks are also important to carry on with and carry out a PhD successfully.

Sincere thanks go to all the teaching staff of the Department of Statistics of Athens University of Economics and Business in Greece who worked during 2004-2006. They offered a high quality and inspiring education. I got a solid background in statistics, indispensable prerequisite for continuing further my studies.

My dearest friends, many many thanks for being supportive and encouraging but mostly for the great fun we have together! Dino, thank you so much for being my co-traveler in the world of statistics from my very first days, for all the statistical jokes, the crazy fun, and the great deal of traveling!

Finally, I am grateful to my family for the endless love and the unconditional support, psychological and financial. Thanks mum and dad for what I

am today. You also played a very small role in this thesis! Thanks mum for teaching me how to compose a piece of text properly! Thanks dad for helping me with my math homework! These were the very first steps! Giorgo, thank you for the good pieces of advice on how to deal with the challenging PhD life abroad. *Σας ευχαριστώ πολύ!*

The current PhD was partly funded by the Swedish Research Council (projects: "Structural Equation Models with Ordinal Ipsative Variables" and "Structural Equation Modeling with Ordinal Variables").

# References

- Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley series in Probability and Statistics, 3rd ed.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.
- Brady, H. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, *54*, 181–202.
- Chan, W., & Bentler, P. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63*, 369–399.
- Feddag, M.-L., & Bacci, S. (2009). Pairwise likelihood for the longitudinal mixed Rasch model. *Computational Statistics and Data Analysis*, *53*, 1027–1037.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 625–641.
- Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*, 387–404.
- Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.
- Jöreskog, K. G. (2002). Structural equation modeling with ordinal variables using LISREL. <http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf>.
- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Chicago, IL: Scientific Software International.
- Lee, S., Poon, W., & Bentler, P. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability Letters*, *9*, 91–97.
- Lee, S., Poon, W., & Bentler, P. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*, 89–105.
- Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.
- Liu, J. (2007). *Multivariate ordinal data analysis with pairwise likelihood and its extension to SEM*. Ph.D. thesis, University of California, Los Angeles, <http://theses.stat.ucla.edu/72/Thesis>
- Marden, J. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.
- Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*, 285–304.
- Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, *14*, 59–71.
- Muraki, E., & Carlson, E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

- Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variables indicators. *Psychometrika*, *49*, 115–132.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus 6 [Computer Software]*. Muthén and Muthén, Los Angeles.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.
- Poon, W. Y., & Lee, S. Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, *52*, 409–430.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48* (2), 1–36, <http://www.jstatsoft.org/v48/i02/paper>.
- Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., & Merkle, E. (2012). *Package lavaan*. <http://cran.r-project.org/web/packages/lavaan/lavaan.pdf>.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.
- Schervish, M. (1984). Algorithm AS 195: Multivariate normal probabilities with errors bound. *Applied Statistics*, *3*, 81–94.
- Schilling, S., & Bock, R. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.
- Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.org>.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.
- Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*, 1–28.
- Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 1–41.
- Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 392–423.
- Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*, 79–92.

# Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations  
from the Faculty of Social Sciences 86*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences.



ACTA  
UNIVERSITATIS  
UPSALIENSIS  
UPPSALA  
2013

Distribution: [publications.uu.se](http://publications.uu.se)  
urn:nbn:se:uu:diva-188342