

New Developments of Latent Variable Models with Ordinal Data

Nuovi sviluppi dei modelli con variabili latenti per dati ordinali

Silvia Cagnone, Attilio Gardini, Stefania Mignani¹

Dipartimento di Scienze Statistiche

Università di Bologna

(cagnone@stat.unibo.it, gardini@stat.unibo.it, mignani@stat.unibo.it)

Riassunto: Nel contesto delle scienze sociali e comportamentali i modelli a variabili latenti rivestono un ruolo fondamentale nella valutazione degli atteggiamenti e delle abilità. In tale contesto, frequentemente i dati raccolti presentano modalità di tipo ordinale. Nel presente lavoro vengono illustrati due differenti approcci per la modellizzazione dei legami tra variabili latenti continue e variabili manifeste di natura ordinale, l'*Underlying Variable Approach* (UVA) e l'*Item Response Function* (IRF) approach. In prima istanza, le due soluzioni sono confrontate in termini teorici e successivamente con riferimento alle proprietà delle stime, attraverso uno studio di simulazione. Per la prima volta, viene proposto un algoritmo di stima per un modello di IRF con più di due variabili latenti. Infine tali modelli sono applicati ad un data set reale che raccoglie i risultati di un'indagine di Customer Satisfaction degli utenti dei mezzi di trasporto pubblico.

Keywords: Underlying Variable Approach, Item Response Function Approach, LISREL, Proportional Odds Model, Customer Satisfaction.

1. Introduction

Statistical models with latent variables occupy a prominent place in the analysis of multivariate data. The importance of these models is principally due to the meaningful and relevant role that latent constructs assume in many fields to which statistical methods can be applied. This is particularly true for evaluation studies such as satisfaction analysis in the context of social sciences and attitudes, personality characteristics and abilities in psychological and educational fields where the concepts analyzed are often not directly observable and hence not directly measurable. Nevertheless, a set of indicator variables that are assumed to be related to the underlying construct can be measured. Thus, the task of the researcher is to establish a theoretical framework within which the information contained in a given number of observed variables is conveyed in a much smaller set of non-observable variables. This means specifying a statistical model that makes the relationships between indicators and the correspondent latent construct explicit. Although latent variable models may take different forms, they are all characterized by the specific assumptions on the relationships between the latent variables and their observed indicators. These assumptions depend on the nature of both the observed and the latent variables.

In this paper we deal with latent variable models with observed ordinal variables and continuous latent variables. These specific models are suitable for survey analysis in

¹Indirizzo per corrispondenza: Stefania Mignani, Dip. Scienze Statistiche, via delle Belle Arti 41, 40126 Bologna

which data are coded into a number of ordered categories. Typical ordinal variables express attitudinal statements with response alternatives like "strongly disagree", "disagree", "strongly agree" or "very dissatisfied", "dissatisfied", "satisfied" and "very satisfied".

In the literature there are two main approaches for conducting latent variable analysis with ordinal observed data. The most popular one is the *Underlying Variable Approach* (UVA) which assumes that the manifest variables are indirect observations of multivariate normal variables. This approach is used in the general framework of structural equation modelling (LISREL). The developments are available in commercial software (that has the same name as the theoretical model, LISREL) that incorporates methods allowing to deal with a large variety of applied situations. The other main approach is the *Item Response Function* (IRF) approach according to which the manifest variables are treated as they are. The unit of analysis is the entire response pattern of a subject, so no loss of information occurs. Following this approach the model that formalizes the relationship between the manifest and the latent variables takes into account the nature of data. This type of models was first developed by Bartholomew (1980) and more recently by Moustaki and Knott (2000). They have proposed a generalized linear latent variable model (GLLVM) that allows to link the latent variables to observed ones of different types. In this paper we present the recent developments of GLLVM for ordinal data (Moustaki, 2000). In particular, we consider the Proportional Odds Model (POM) and investigate its performance through a comparison with LISREL in terms of both simulations and an application to real data. For the first time we implement a Proportional Odds Model with more than two latent variables.

The paper is organized as follows: Sections, 2 and 3, are dedicated to a brief review of the UVA and the IRF respectively. The comparison between the two approaches is illustrated in Section 4. In Section 5 a simulation study is carried out in order to investigate the statistical properties of the two approaches and compare them. Section 6 reports an application to a data set concerning a Customer Satisfaction survey.

2. The Underlying Variable Approach in LISREL

One of the most widespread latent variable models for analyzing behavioural data is LISREL, since it allows to deal with a large variety of applied problems. According to this approach, the relationship between manifest and latent variables is linear, being based on the classical factor analysis, and it can be represented as follows:

$$\mathbf{y} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} \quad (1)$$

where $\mathbf{y}$ is the vectors of the observed variables, that are assumed to be continuous and normally distributed, and $\boldsymbol{\eta}$ is the vector of the latent ones. $\mathbf{\Lambda}$ is the matrix of the coefficients (factor loadings) of the latent variables on the observed variables and $\boldsymbol{\epsilon}$ is the vector of the measurement errors in $\mathbf{y}$. When the observed variables are categorical, binary or ordinal, it is not possible to specify a linear relation like (1). Nevertheless, in this case we can refer to the Underlying Variable Approach (UVA) by which each manifest variable is considered indicator of an underlying continuous variable distributed as standardized normal variable. If the y variables are binary or ordered polytomous, each of them with c_i ($i = 1, \dots, p$) categories, a vector of p continuous underlying variables, $\mathbf{y}^*$, one for each observed variable is considered, such that (Muthén, 1984):

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq \tau_{i1} \\ 2 & \text{if } \tau_{i1} < y_i^* \leq \tau_{i2} \\ \vdots & \\ c_i - 1 & \text{if } \tau_{ic_i-2} < y_i^* \leq \tau_{ic_i-1} \\ c_i & \text{if } \tau_{ic_i-1} < y_i^* \end{cases} \quad (2)$$

where the parameters τ_{is} ($s = 1, \dots, c_i - 1$) with $\tau_{i0} = -\infty$ and $\tau_{ic_i} = +\infty$ are interpreted as thresholds associated to y_i^* and allow to transform categorical variables into continuous ones. It follows that the expression (1) becomes:

$$y^* = \Lambda \eta + \epsilon \quad (3)$$

Model estimation is based on the correlation matrix of the underlying variables defined in (2). This matrix is called polychoric correlation matrix. Different methods of estimation have been discussed in the literature (Muthén, 1984; Jöreskog, 1990; Lee *et al.*, 1990a,b). They differ because of the different ways in which the parameters of the model (thresholds, polychoric correlations and coefficients) are estimated and because of the stages required for the estimation procedure. One of the most used procedure (implemented in the software LISREL, Jöreskog and Sörbom (1988)) is a three stage estimation method based on the following steps:

1. Thresholds estimation:

from the assumptions on the underlying variable y_i^* we get:

$$\pi_{s_i}^{(i)} = P(y_i = s_i) = \int_{\tau_{s_i-1}^{(i)}}^{\tau_{s_i}^{(i)}} \phi(y_i^*) dy_i^* = \Phi(\tau_{s_i}^{(i)}) - \Phi(\tau_{s_i-1}^{(i)}) \quad (4)$$

where $\phi(y_i^*)$ is the standard normal density function and $\pi_{s_i}^{(i)}$ can be interpreted as the unknown probability of a response to category s of the variable i . By maximizing the loglikelihood:

$$\ln L = \sum_{s_i=1}^{c_i} n_{s_i}^{(i)} \ln \pi_{s_i}^{(i)}$$

we get the maximum likelihood estimates of the thresholds:

$$\hat{\tau}_{s_i}^{(i)} = \Phi^{-1}(p_1^{(i)} + p_2^{(i)} + \dots + p_{s_i}^{(i)}), \quad s_i = 1, \dots, c_i - 1$$

2. Polychoric correlation estimation:

from the bivariate normality of the underlying variables y_i^* and y_j^* with correlation ρ (the polychoric correlation) we get:

$$\pi_{s_i s_j}^{(ij)} = P(y_i = s_i, y_j = s_j) = \int_{\tau_{s_i-1}^{(i)}}^{\tau_{s_i}^{(i)}} \int_{\tau_{s_j-1}^{(j)}}^{\tau_{s_j}^{(j)}} \phi_2(y_i^*, y_j^*, \rho_{ij}) dy_i^* dy_j^*$$

where $\phi_2(y_i^*, y_j^*, \rho_{ij})$ is the density function of the standardized bivariate normal distribution with correlation ρ_{ij} , $\tau_{s_i}^{(i)}$ ($s_i = 1, \dots, c_i - 1$) are the thresholds associated

to the variable y_i^* and $\tau_{s_j}^{(j)}$ ($s_j = 1, \dots, c_j - 1$) are the thresholds associated to y_j^* ; $\pi_{s_i s_j}^{(ij)}$ represents the unknown probability that an observation falls into cell (s_i, s_j) of the $c_i \times c_j$ bivariate contingency table referred to the two observed variables. As for the thresholds, the estimation of ρ is obtained through the maximum likelihood method by maximizing the loglikelihood of the multinomial distribution for a sample of size n (Olsson, 1979):

$$\ln L = \sum_{s_i=1}^{c_i} \sum_{s_j=1}^{c_j} n_{s_i s_j}^{(ij)} \ln \pi_{s_i s_j}^{(ij)} \quad (5)$$

3. Coefficient estimation:

in the third stage the parameters of the model (3) are estimated by a WLS approach. To this aim we consider a fitting function given by:

$$F_{WLS}(\boldsymbol{\lambda}) = (\hat{\boldsymbol{\rho}} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\hat{\boldsymbol{\rho}} - \boldsymbol{\sigma}) \quad (6)$$

where $\hat{\boldsymbol{\rho}}$ is a vector containing the polychoric correlations between all pairs of underlying variables. Any positive definite matrix $\mathbf{W}$ may be used to obtain consistent estimates. However, the asymptotic properties of test and standard errors are guaranteed by choosing a weight matrix $\mathbf{W}$ which is a consistent estimate of the asymptotic covariance matrix of the polychoric correlations (Jöreskog, 1994).

Parameter estimation is performed by minimizing the fitting function (6); since we get non linear equations in the parameters numerical minimization methods are necessary; Jöreskog (1969) refers to a modification of Fletcher and Powell's minimization procedure which is still used in LISREL; other suggested methods are Fisher's scoring method, Newton-Raphson procedure or a modified Gauss-Newton procedure (Jöreskog, 1978). These methods are based on iterative algorithms by which, once initial values for parameters have been selected, $\boldsymbol{\lambda}^{(0)}$, a sequence of successive points are generated, $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \dots, \boldsymbol{\lambda}^{(l)}$ so that $F(\boldsymbol{\lambda}^{(l+1)}) < F(\boldsymbol{\lambda}^{(l)})$ where $l = 0, 1, 2, \dots$. The iterative procedure continues until convergence is achieved.

3. The Item Response Function approach in GLLVM

A different approach to deal with latent variable models is the so called Generalized Linear Latent Variable Model (GLLVM). According to the GLLVM both the manifest and the latent variables are considered random variables so that the relationships between them is expressed in terms of probability distributions. In particular, as only $\mathbf{y}$ can be observed, we can refer to its joint distribution given by:

$$f(\mathbf{y}) = \int_{R_{\boldsymbol{\eta}}} g(\mathbf{y} | \boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} \quad (7)$$

where $g(\mathbf{y} | \boldsymbol{\eta})$ is the conditional distribution of $\mathbf{y}$ given $\boldsymbol{\eta}$, $h(\boldsymbol{\eta})$ is the prior distribution of $\boldsymbol{\eta}$ and $R_{\boldsymbol{\eta}}$ is the range space of $\boldsymbol{\eta}$. Precise assumptions are made on these distributions for the specification of the complete models. In more detail we assume:

- the conditional independence of $g(\mathbf{y} | \boldsymbol{\eta})$ by which:

$$g(\mathbf{y} | \boldsymbol{\eta}) = \prod_{i=1}^p g_i(y_i | \boldsymbol{\eta}) \quad (8)$$

- $g(\mathbf{y} \mid \boldsymbol{\eta})$ belongs to the exponential family within the generalized linear model framework (Bartholomew and Knott, 1999).
- Independence and standard normal distribution of the latent variables:

$$h(\boldsymbol{\eta}) = \prod_{i=1}^q h(\eta_i) \quad (9)$$

where q is the number of latent variables.

When the observed variables are ordinal we refer to the Item Response Function (IRF) approach. The c_i ($i = 1, \dots, p$) ordered categories of these variables have associated the probabilities $\pi_1^{(i)}(\boldsymbol{\eta}), \pi_2^{(i)}(\boldsymbol{\eta}), \dots, \pi_{c_i}^{(i)}(\boldsymbol{\eta})$ which are functions of the vector of the latent variables $\boldsymbol{\eta}$. Within this framework, the unit of analysis is the response pattern of an individual; for the r -th individual it is defined as $\mathbf{y}_r = (y_1 = s_1, y_2 = s_2, \dots, y_p = s_p)$. The expression (7) becomes (Moustaki, 2000):

$$\pi_r = \int_{R_\zeta} \pi_r(\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} \quad (10)$$

where π_r is the probability of $\mathbf{y}_r$ and $\pi_r(\boldsymbol{\eta})$ is the multinomial probability function:

$$\pi_r(\boldsymbol{\eta}) = g(y_i \mid \boldsymbol{\eta}) = \prod_{s=1}^{c_i} \pi_{s_i}^{(i)}(\boldsymbol{\eta})^{y_{is_i}} = \prod_{s_i=1}^{c_i} (\gamma_{s_i}^{(i)} - \gamma_{s_i-1}^{(i)})^{y_{is_i}} \quad (11)$$

y_{is} is defined in the same way as for polytomous data that is $y_{is} = 1$ if the response falls in category s of the item i and $y_{is} = 0$ otherwise and $\gamma_{s_i}^{(i)} = \pi_1^{(i)}(\boldsymbol{\eta}) + \pi_2^{(i)}(\boldsymbol{\eta}) + \dots + \pi_{s_i}^{(i)}(\boldsymbol{\eta})$ that is the probability of a response in category s or lower on the variable i . As in the classical generalized linear model, the relation between the observed and the latent variables is expressed through a link function that in the case of ordinal variables is a logit:

$$\log \left[\frac{\gamma_{s_i}^{(i)}}{(1 - \gamma_{s_i}^{(i)})} \right] = \alpha_{i0}(s_i) - \sum_{j=1}^q \alpha_{ij} \eta_j, \quad s_i = 1, \dots, c_i - 1 \quad (12)$$

$\alpha_{i0}(s_i)$ and α_{ij} can be interpreted as the thresholds and the factor loadings of the model. The model so defined is called Proportional Odds Model (POM). Model estimation is performed by applying the maximum likelihood method through the E-M algorithm. For a sample of size n , the loglikelihood is defined as (Bartholomew and Knott, 1999):

$$\begin{aligned} L &= \sum_{h=1}^n \log f(\mathbf{y}_h) = \sum_{h=1}^n \log \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(\mathbf{y}_h \mid \boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} \\ &= \sum_{h=1}^n \log \sum_{t_1=1}^{k_1} \dots \sum_{t_q=1}^{k_q} g(\mathbf{y}_h \mid \eta_{t_1}, \dots, \eta_{t_q}) h(\eta_{t_1}, \dots, \eta_{t_q}) \end{aligned} \quad (13)$$

In the last equality the integrals are approximated by Gauss Hermite quadrature points. The expressions of the partial derivatives of L with respect to α_{i0} 's and α_{ij} 's are given in Moustaki (2000). Since we get non linear equations in the parameters a Newton-Raphson iterative procedure is used within the E-M algorithm.

4. Comparison between UVA and IRF

Although the UVA and the IRF approaches seem different because of the different assumptions they need, Bartholomew and Knott (1999) have noticed an equivalence between the two methods. In particular we have that:

$$\alpha_{i0}(s_i) = \tau_{s_i}^{(i)} / (1 - \sum_{j=1}^m \lambda_{ij}^2)^{1/2} \quad i = 1, \dots, p \quad (14)$$

and:

$$\alpha_{ij} = \lambda_{ij} / (1 - \sum_{j=1}^m \lambda_{ij}^2)^{1/2} \quad i = 1, \dots, p \quad (15)$$

The parameters $\tau_{s_i}^{(i)}$ and λ_{ij} are assumed to be standardized parameters (Jöreskog and Moustaki, 2001), whereas $\alpha_{i0}(s_i)$ and α_{ij} are assumed to be unstandardized. The equivalence between UVA and IRF is not exact because the former is based on a probit model and the latter on a logit model. It follows that if we want to make the parameters comparable we have to multiply the LISREL loadings by 1.7 (Lord and Novick, 1968).

The choice of one of the two approaches in practical applications depends both on the properties of the estimates and on the efficiency of the estimation procedures. At this regard, we have showed before that the basic assumption of the UVA is that the observed variables are generated by normally distributed underlying variables allowing the definition of the so called polychoric correlations fundamental for parameters estimation. Nevertheless, in many instances the requirement of the normality of the underlying variables is not satisfied, especially in the cases of variables with few categories. The IRF approach within the GLLVM framework allows to overcome this limit by treating the observed variables as they are. In fact we model the conditional probability of a p -complete response pattern as a function of the latent variables through a generalized linear model where the link function is the logit.

Furthermore, this approach is based on a full information maximum likelihood method of estimation in the sense that the full distribution over all score patterns is considered in the estimation procedure and all the parameters of the model are estimated simultaneously. On the contrary the UVA is based on a limited information method (the parameters are estimated in three different steps). It follows that the former is more efficiency since there is no loss of information. Nevertheless, from a computational point of view, full information methods are much heavier than limited information one. This problem is overburdened by the limits of the E-M algorithm that in general becomes very slow when a large number of latent variables is considered.

At last, although from a theoretical point of view the GLLVM allows an unlimited number of latent variables, so far the application of the GLLVM has been limited to a maximum of two latent variables (Bartholomew and Knott, 1999; Moustaki, 2000).

In the next section we perform a simulation study in order to compare the two approaches highlighting the properties of the parameter estimates. For the first time, an IRF model including three latent factors will be considered in the analysis and will be compared with the LISREL results. This study allows to highlight which model is to be preferred.

5. A simulation study

In this section we compare the two approaches through a simulation study. In this way we can evaluate the parameter estimates in terms of bias and MSE. We have generated 400 samples from POM with known parameters in different experimental conditions and then analyzed each sample by POM and LISREL. The details of the generation procedure are reported in Jöreskog and Moustaki (2001). The experimental design is described

Table 1: Design of study

<i>N. of factors</i>	1,2,3
<i>N. of variables</i>	6
<i>N. of categories</i>	3,4
<i>Sample size</i>	200,1000

in Table 1. As for the POM estimation, Moustaki (2002) has developed the software GENLAT 1.1 written in Fortran 77 to estimate the GLLVM for various types of observed variables and at most two latent variables.

Starting from this computer program, we have implemented a new program in Fortran

Table 2: Estimates of $\hat{\lambda}_1$ for the one-factor model, 3 categories and 4 categories.

<i>True</i>	POM		LISREL		<i>True</i>	POM		LISREL	
	200	1000	200	1000		200	1000	200	1000
Mean					Mean				
0.31	0.38	0.40	0.75	0.46	0.31	0.38	0.41	0.60	0.46
0.29	0.30	0.34	0.31	0.36	0.29	0.31	0.34	0.31	0.36
0.31	0.35	0.41	0.34	0.44	0.31	0.36	0.40	0.43	0.44
0.31	0.38	0.40	0.37	0.45	0.31	0.38	0.39	0.40	0.44
0.32	0.41	0.43	0.46	0.47	0.32	0.43	0.44	0.43	0.49
0.32	0.36	0.41	0.40	0.46	0.32	0.39	0.43	0.38	0.48
Bias					Bias				
0.31	0.06	0.09	0.44	0.15	0.31	0.07	0.10	0.29	0.15
0.29	0.01	0.05	0.03	0.07	0.29	0.02	0.05	0.03	0.07
0.31	0.04	0.10	0.03	0.13	0.31	0.05	0.09	0.12	0.13
0.31	0.07	0.10	0.06	0.14	0.31	0.07	0.08	0.09	0.13
0.32	0.09	0.11	0.14	0.15	0.32	0.11	0.12	0.11	0.17
0.32	0.05	0.10	0.09	0.14	0.32	0.07	0.11	0.07	0.16
Mse					Mse				
0.31	0.09	0.02	0.38	0.03	0.31	0.08	0.02	0.19	0.01
0.29	0.07	0.02	0.13	0.02	0.29	0.06	0.01	0.12	0.01
0.31	0.08	0.02	0.13	0.01	0.31	0.07	0.02	0.17	0.01
0.31	0.09	0.02	0.21	0.01	0.31	0.07	0.02	0.16	0.01
0.32	0.09	0.03	0.17	0.01	0.32	0.08	0.02	0.14	0.01
0.32	0.08	0.02	0.19	0.01	0.32	0.08	0.02	0.18	0.01

95 that allows to handle models with binary and ordinal data and an unlimited number of latent variables (Cagnone, 2003). The main advantage of this new program is its flexibility with respect to the extensions and the great improvement in terms of computing time. As for the UVA estimation, we have considered the softwares PRELIS 2.0 and LISREL 8.52.

In Table 2 we report the results of the simulation for the one-factor model and for, respectively, 3 categories (Table on the left) and 4 categories (Table on the right). We consider the standardized parameters. For the reasons listed in section 4 the LISREL estimates have been multiplied by 1.7. The first column of each Table contains the true parameters, whereas the subsequent columns report mean, bias and MSE, for sample sizes 200 and 1000, of POM estimates and of LISREL estimates. Notice that the bias and the MSE of POM estimates are almost always smaller than LISREL estimates, and this is particularly evident for small sample sizes. Furthermore, if POM estimates are not influenced by the number of categories, being the bias and the MSE similar for 3 and 4 categories, the results vary sensibly for LISREL estimates. In particular, they are less accurate for variables with 3 categories.

Table 3 and Table 4 show the results for the two-factor model and for the three-factor model. In order to restrict the number of Tables, we report only the results for the stan-

Table 3: Estimates of $\hat{\lambda}_1$ for the two-factor model, 3 categories and 4 categories.

True	POM		LISREL		True	POM		LISREL	
	200	1000	200	1000		200	1000	200	1000
	Mean					Mean			
0.42	0.62	0.66	1.62	1.37	0.42	0.56	0.49	1.40	1.37
0.35	0.62	0.59	0.67	0.74	0.35	0.30	0.28	0.54	0.74
0.40	0.64	0.68	0.85	0.79	0.40	0.33	0.36	0.64	0.80
0.34	0.50	0.47	0.49	0.64	0.34	0.32	0.28	0.51	0.64
0.36	0.47	0.46	0.48	0.67	0.36	0.31	0.31	0.53	0.67
0.38	0.42	0.41	0.32	0.66	0.38	0.31	0.30	0.54	0.66
	Bias					Bias			
0.42	0.20	0.24	1.20	0.95	0.42	0.14	0.07	0.98	0.95
0.35	0.27	0.24	0.32	0.39	0.35	-0.05	-0.07	0.19	0.39
0.40	0.24	0.28	0.45	0.39	0.40	-0.07	-0.04	0.24	0.40
0.34	0.16	0.13	0.15	0.30	0.34	-0.02	-0.06	0.17	0.30
0.36	0.11	0.10	0.12	0.31	0.36	-0.05	-0.05	0.17	0.32
0.38	0.04	0.03	-0.06	0.28	0.38	-0.07	-0.08	0.16	0.28
	Mse					Mse			
0.42	0.08	0.08	0.29	0.15	0.42	0.09	0.03	0.17	0.15
0.35	0.12	0.09	0.03	0.16	0.35	0.09	0.03	0.22	0.16
0.40	0.10	0.11	0.07	0.09	0.40	0.08	0.02	0.17	0.10
0.34	0.10	0.06	0.03	0.12	0.34	0.08	0.03	0.19	0.12
0.36	0.08	0.04	0.06	0.11	0.36	0.09	0.03	0.19	0.11
0.38	0.07	0.03	0.09	0.12	0.38	0.10	0.03	0.18	0.12

standardized loadings $\hat{\lambda}_1$ associated to the first latent variable (Note that for the other loadings we have found similar results). We can observe that LISREL estimates don't change sen-

Table 4: Estimates of $\hat{\lambda}_1$ for the three-factor model, 3 categories and 4 categories.

<i>True</i>	POM		LISREL		<i>True</i>	POM		LISREL	
	200	1000	200	1000		200	1000	200	1000
	Mean					Mean			
0.57	0.36	0.33	0.40	0.37	0.57	0.60	0.58	1.09	0.84
0.36	0.51	0.54	0.25	0.67	0.36	0.04	0.08	0.54	0.43
0.76	0.42	0.43	0.21	0.54	0.76	0.26	0.36	0.58	0.63
0.92	0.57	0.65	0.27	0.89	0.92	0.79	0.83	1.09	0.99
0.87	0.63	0.71	0.40	1.17	0.87	0.83	0.85	1.05	0.95
0.89	0.57	0.60	0.36	1.05	0.89	0.81	0.84	1.06	0.96
	Bias					Bias			
0.57	-0.21	-0.24	-0.17	-0.20	0.57	0.03	0.01	0.52	0.27
0.36	0.15	0.18	-0.11	0.31	0.36	-0.32	-0.28	0.18	0.07
0.76	-0.34	-0.33	-0.55	-0.22	0.76	-0.50	-0.40	-0.18	-0.13
0.92	-0.35	-0.27	-0.65	-0.03	0.92	-0.12	-0.08	0.17	0.07
0.87	-0.24	-0.16	-0.47	0.30	0.87	-0.04	-0.02	0.18	0.08
0.89	-0.32	-0.29	-0.52	0.16	0.89	-0.09	-0.05	0.17	0.07
	Mse					Mse			
0.57	0.10	0.07	0.23	0.14	0.57	0.02	0.01	0.16	0.24
0.36	0.07	0.05	0.24	0.03	0.36	0.14	0.12	0.14	0.15
0.76	0.18	0.13	0.55	0.22	0.76	0.32	0.24	0.32	0.35
0.92	0.21	0.11	0.85	0.22	0.92	0.03	0.01	0.22	0.42
0.87	0.15	0.05	0.81	0.12	0.87	0.01	0.01	0.19	0.39
0.89	0.21	0.11	0.83	0.18	0.89	0.02	0.01	0.21	0.40

sibly by introducing additional latent variables. As before, the results are influenced by the number of categories. On the contrary, bias and MSE of POM estimates increase as the number of factors increases. Nevertheless, the MSE of POM is always smaller than that of LISREL.

6. A real data analysis

In this section, we analyze a real data set and compare the results provided by LISREL and POM. The data are extracted from a survey conducted in the year 2000 in order to investigate user satisfaction with public transportation services in Rimini. The database consists in satisfaction judgements expressed by 453 bus passengers about several service attributes (item). Permitted responses to these judgements are: "not satisfied", "satisfied" "very satisfied". In particular, we have selected the following seven attributes resulted most important in a previous exploratory analysis (Cagnone *et al.*, 2003):

- Safety in vehicles (SAFE)
- Comfort (COMF)
- Trip Time (TIME)
- Frequency departure (FREQ)
- Time schedule (scheduling) (SCHED)
- Information (INFO)
- Channels for complaints (COMPL)

From the results of the exploratory analysis, we decided to fit a three-factor model to the data using LISREL and POM. The parameter estimates are presented in Table 5. In order to obtain a unique solution, we have imposed some constraints on the parameters. In particular, λ_{12} , λ_{13} , λ_{23} are set to zero (i.e. not estimated).

Table 5: Estimates of factor loadings for POM and LISREL

Item	POM			LISREL		
	COND	MANAG	RELAT	COND	MANAG	RELAT
	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}_{i3}$	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}_{i3}$
SAFE	0.87	-	-	0.76	-	-
COMF	0.67	-0.36	-	0.45	0.34	-
TIME	0.74	-0.08	-0.10	0.52	0.13	0.01
FREQ	0.12	-0.26	-0.93	0.13	0.34	0.71
SCHED	0.36	-0.18	-0.81	0.33	0.18	0.81
INFO	0.36	-0.44	-0.42	0.22	0.49	0.22
COMPL	0.18	-0.95	0.06	0.17	0.47	0.01

By examining the factor loadings we can notice that the interpretation of the latent variables is similar for the two models. More in detail, the first latent variable presents high factor loadings in correspondence of safety, comfort and trip time. We can interpret this factor as satisfaction for the "CONDITIONS" of the trip. As for the second factor, the most important loadings are the ones related to information and channels for complaints. Thus, we can refer to this latent variable as satisfaction for the "RELATIONSHIPS" management. The last factor is related to frequency and time schedule, so that we can call it satisfaction for the "MANAGEMENT" of the trip.

It is possible to highlight some differences between the estimated parameters of the two models. As regards the second factor in the POM model $\hat{\lambda}_{72}$ is higher with respect to $\hat{\lambda}_{62}$ whereas in the LISREL model these parameters are very similar. From a practical point of view the POM results allow to conclude that user satisfaction is mainly influenced

by channels for complaints. As for RELATIONSHIPS, in accordance with LISREL, time schedule is more outstanding than frequency.

The Customer Satisfaction analysis is strictly related to the estimation of the latent scores within the structural equation model approach, each factor score representing an estimate of the individual satisfaction. Unfortunately, these scores are not uniquely identified and precise conditions have to be imposed in order to get consistent solutions. Since this problem is not discussed in this work, the interested reader is referred to the dedicated literature (Vittadini, 1999).

7. Conclusions

In this paper we have discussed new developments for the application of latent variable models with ordinal data, considering two different approaches, IRF and UVA. Previous studies (Jöreskog and Moustaki, 2001) have shown that IRF seems to perform better than UVA but so far from a practical point of view the former has been implemented only up to a two-factor model. We have extended the estimation of IRF to a three-factor model and then we have compared the two approaches by means of a large simulation study and an application to a real data set in the context of Customer Satisfaction.

The results have highlighted that in general the IRF gives more accurate estimates, particularly for manifest variables with few categories. This behavior is due to the fact that IRF takes into account the nature of the data, whereas LISREL forces them into a multivariate normal framework. Furthermore, the IRF is based on a full information method that uses all information in the data. On the contrary, the UVA is a limited information method that uses information from the marginal distributions. Nevertheless, unlike LISREL, the IRF is computationally heavy and doesn't allow to include a structural part that specifies a relationship between latent variables. We are presently working on both these extensions (Cagnone, 2003).

References

- Bartholomew D. (1980) Factor analysis for categorical data, *Journal of the Royal Statistical Society, Series B*, 42, 3, 293–321.
- Bartholomew D. and Knott M. (1999) *Latent Variable Models and Factor Analysis*, Kendall's Library of statistics, London, second edition.
- Cagnone S. (2003) *Latent Variable for Ordinal Data*, Ph.D. thesis, University of Bologna.
- Cagnone S., Lubisco A. and Mignani S. (2003) Modelli a variabili latenti per l'analisi della customer satisfaction degli utenti dei servizi di trasporto pubblico, Technical Report 2, Appunti e Ricerche, Università di Bologna.
- Jöreskog K. (1969) A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika*, 34, 183–202.
- Jöreskog K. (1978) Structure analysis of covariance and correlation matrices, *Psychometrika*, 43, 443–477.
- Jöreskog K. (1990) New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares, *Quality and Quantity*, 24, 387–404.
- Jöreskog K. (1994) On the estimation of polychoric correlations and their asymptotic covariance matrix, *Psychometrika*, 59, 381–389.
- Jöreskog K. and Moustaki I. (2001) Factor analysis of ordinal variables: A comparison of three approaches, *Multivariate Behavioral Research*, 36, 347–387.
- Jöreskog K. and Sörbom D. (1988) *LISREL 8: Users' Reference Guide*, SSI, Chicago, second edition.
- Lee S., Poon W. and Bentler P. (1990a) Full maximum likelihood analysis of structural equation models with polytomous variables, *Statistics and probability letters*, 9, 91–97.
- Lee S., Poon W. and Bentler P. (1990b) A three-stage estimation procedure for structural equation models with polytomous variables, *Psychometrika*, 55, 45–51.
- Lord F. and Novick M. (1968) *Statistical theories of mental test scores*, Addison-Wesley Publishing Co.
- Moustaki I. (2000) A latent variable model for ordinal data, *Applied psychological measurement*, 24, 211–223.
- Moustaki I. (2002) Genlat 1.1: A computer program for fitting a one- or two- factor latent variables model to categorical, metric and mixed observed items with missing values, Technical report, London school of Economics and Political Science, Statistics Department.
- Moustaki I. and Knott M. (2000) Generalized latent trait models, *Psychometrika*, 65, 391–411.
- Muthén B. (1984) A general structural equation model with dichotomous ordered categorical and continuous latent variable indicators, *Psychometrika*, 49, 115–132.
- Olsson U. (1979) Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika*, 4, 443–460.
- Vittadini G. (1999) Analysis of qualitative variables in structural models with unique solutions, in: *Classification and Data Analysis - Theory and Application*, e O. Opitz M.V., ed., Springer, 203–210.