A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables

Irini Moustaki*

*Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE. email:i.moustaki@lse.ac.uk

Abstract

Moustaki (2000a) discusses a general class of latent variable models for analyzing ordinal manifest variables. This work is extended here to allow for covariate effects on the manifest ordinal variables and the latent variables.

A full maximum likelihood estimation method is used for estimating simultaneously all the model parameters. Goodness-of-fit statistics are discussed.

Two examples from the 1996 British Social Attitudes Survey are used to illustrate the methodology .

keywords ordinal items, full-information ML estimation, generalized latent variable models, covariate effects.

1 Introduction

McCullagh (1980) shows how to fit regression models where there is a single dependent observed ordinal variable and a set of observed explanatory variables. That class of models without covariates was used in Moustaki (2000a) to construct latent variable models where a number of latent (unobserved) variables account for the interrelationships among a set of ordinal observed variables.

In this paper we extend that work to allow for covariate effects both on the manifest variables and on the latent variables. The part of the model that shows the effect of the latent variables on the manifest variables is called the measurement model and the part of the model that links the covariates with the latent variables is called the structural part of the model. Covariates are allowed to affect the manifest variables indirectly through the latent variables or directly. However, there might be situations where we would like to model the effect of a set of covariates on the latent variables and the effect of a different set of covariates directly on the manifest variables. In the applications section we discuss an example in which we are interested in measuring overall satisfaction (latent variable) with the National Health system in respondents' area from five ordinal indicators controlling for the respondents' political affiliation (observed covariate). In addition we allow for covariates age and gender to effect the latent construct satisfaction.

Covariate effects on the latent variables can be estimated in one stage or

two stages. In the one stage approach the parameters of the measurement and the structural part of the model are estimated simultaneously. In the two stage approach the measurement model is fitted first, then factor scores (Moustaki and Knott 2000) are computed and used as dependent variables on further analysis. Croon and Bolck (1997) mention that in the one-stage approach it is more difficult to identify any mispecifications in either the measurement or the structural part of the model. Also due to the higher model complexity it might be possible that a local rather than a global solution will be found. However, they found that the two-stage approach based on the use of factor scores as observed variables regressed on a set of explanatory variables leads into biased estimates.

Jöreskog and Goldberger (1975) discussed a multiple indicators and multiple causes (MIMIC) model for normal manifest variables with a single latent variables that allows for direct and indirect effects of covariates on the latent and manifest variables respectively. In their results it is apparent that parameter estimates of the measurement and the structural models differ from the one to the two-stage method. They also found that the one-stage method gives more efficient parameter estimates. Muthén (1989) discusses the MIMIC model for other types of manifest variables such as binary and ordinal for capturing heterogeneity across groups (groups are defined through the covariates). He argues that the MIMIC model is a good alternative to multi-group analysis when not enough data are available to

estimate a model in each group.

The MIMIC model has been developed within the structural equation modeling framework. By that we mean that the approach used in the ordinal case for estimating the parameters of the measurement model is based on polychoric correlations estimated by maximum likelihood. In the structural equation modeling framework the ordinal variables are taken to be manifestations of some underlying unobserved variables. Packages such as LISREL (Jöreskog and Sörbom 1996) and Mplus (Muthén and Muthén 2000) fit the MIMIC model to ordinal manifest variables. A comparison between the LISREL type models for ordinal variables and the models presented here without the covariate effects can be found in Jöreskog and Moustaki (2001) and Moustaki (2000b).

In this paper we discuss a model for ordinal manifest variables that allows for covariate effects both on the latent and manifest variables using full maximum likelihood. This approach is based on an extension of the models for ordinal variables discussed by Samejima (1969), Muraki and Carlson (1995) and Moustaki (2000a) to allow for covariate effects.

2 Model and estimation

Let $y_1, y_2, \ldots, y_p$ be the ordinal observed variables. Small letters are used to denote both the variables and the values that these variables take. Let m_i denote the number of categories for the ith variable.

The m_i ordered categories have probabilities $\pi_{i1}(\mathbf{z}, \mathbf{x}), \pi_{i2}(\mathbf{z}, \mathbf{x}), \cdots, \pi_{im_i}(\mathbf{z}, \mathbf{x})$, which are functions of the vector of latent variables $\mathbf{z}$ and the vector of observed covariates $\mathbf{x}$. The vector of covariates $\mathbf{x}$ affects directly the manifest ordinal variables. In addition, we allow for covariates $w_1, w_2, \ldots, w_k$ affecting only the vector of latent variables $\mathbf{z}$.

2.1 Measurement model

The general form given in Moustaki (2000a) for the latent variable model with ordinal variables is extended here to allow for covariate effects:

$$\text{link}[\gamma_{is}(\mathbf{z}, \mathbf{x})] = \alpha_{is} - \sum_{j=1}^{q} \alpha_{ij} z_j + \sum_{l=1}^{r} \beta_{il} x_l,$$
$$i = 1, \ldots, p; s = 1, \ldots, m_i \qquad (1)$$

where $\gamma_{is}(\mathbf{z}, \mathbf{x})$ is the cumulative probability of a response in category s or lower of item y_i, written as:

$$\gamma_{is}(\mathbf{z}, \mathbf{x}) = \pi_{i1}(\mathbf{z}, \mathbf{x}) + \pi_{i2}(\mathbf{z}, \mathbf{x}) + \cdots + \pi_{i,s}(\mathbf{z}, \mathbf{x})$$

The $\gamma_{is}(\mathbf{z}, \mathbf{x})$ is a function of the latent variables $\mathbf{z}$ and the observed covariates $\mathbf{x}$. To simplify notation we just write γ_{is}.

The link function can be the logit, the complementary log-log function, the inverse normal function, the inverse Cauchy, or the log-log function.

All those link functions are monotonically increasing functions that map $(0,1)$ onto $(-\infty,\infty)$. The parameters α_{is} are referred as 'cut-points' on the logistic, probit or other scale where $\alpha_{i1} < \alpha_{i2} < \cdots < \alpha_{i,m_i} = +\infty$. The α_{ij} parameters can be considered as factor loadings since they measure the effect of the latent variables $\mathbf{z}$ on some function of the cumulative probability of responding up to a category of the ith item. In the one latent variable case the negative sign in front of the slope parameter is used to indicate that as $\mathbf{z}$ increases the response on the observed item y_i is more likely to fall at the high end of the scale. The β_{il} are regression coefficients.

If we put the model into the generalized linear model framework then the p random response variables, $y_1,\ldots,y_p$ have a distribution from the exponential family. The systematic component is the one in which the latent variables $\mathbf{z}$ and the set of covariates $\mathbf{x}$ produce a linear predictor η_i corresponding to each y_i:

$$\eta_i = \alpha_{is} - \sum_{j=1}^{q} \alpha_{ij} z_j + \sum_{l=1}^{r} \beta_{il} x_l \,, \qquad i = 1,\ldots,p.$$

And finally the link between the systematic component and the conditional means of the random component distributions: $\eta_i = v_i(\mu_i)$ where $\mu_i = E(y_i \mid \mathbf{z},\mathbf{x})$ and $v_i(.)$ is the link function which can be any monotonic differentiable function.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ represent the whole response pattern for a ran-

domly selected individual. The density function $f(\mathbf{y})$ of the manifest variables $\mathbf{y}$ is:

$$f(\mathbf{y}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) h(\mathbf{z}, \boldsymbol{\lambda}) d\mathbf{z} \tag{2}$$

where $g(\mathbf{y} \mid \mathbf{z}, \mathbf{x})$ is the conditional density function of $\mathbf{y}$ given $\mathbf{z}$ and $\mathbf{x}$ and $h(\mathbf{z}, \boldsymbol{\lambda})$ is the density function of $\mathbf{z}$ conditional on a vector of covariates $\mathbf{w}$. The latent variables are assumed to be independent with normal distributions.

Under the assumption of conditional independence the vector of latent variables and the vector of observed covariates $\mathbf{x}$ account for the interrelationships among the observed ordinal variables so that when the latent variables are held fixed the responses to the p observed variables are independent:

$$g(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) = \prod_{i=1}^{p} g(y_i \mid \mathbf{z}, \mathbf{x}). \tag{3}$$

For a manifest item y_i the conditional probability of $(y_i \mid \mathbf{z}, \mathbf{x})$ is given by:

$$\begin{aligned} g(y_i \mid \mathbf{z}, \mathbf{x}) &= \prod_{s=1}^{m_i} \pi_{is}(\mathbf{z}, \mathbf{x})^{y_{i,s}} \\ &= \prod_{s=1}^{m_i} (\gamma_{i,s} - \gamma_{i,s-1})^{y_{i,s}} \ , \end{aligned} \tag{4}$$

where $y_{i,s} = 1$ if the response y_i is in category s and $y_{i,s} = 0$ otherwise. Equation (4) can be also written in the following form:

$$g(y_i \mid \mathbf{z}, \mathbf{x}) = \prod_{s=1}^{m_i-1} \left(\frac{\gamma_{i,s}}{\gamma_{i,s+1}}\right)^{y^*_{i,s}} \left(\frac{\gamma_{i,s+1} - \gamma_{i,s}}{\gamma_{i,s+1}}\right)^{y^*_{i,s+1} - y^*_{i,s}} \tag{5}$$

where $y^*_{i,s} = 1$ if a randomly selected individual responds into category s or a lower one of the ith item and $y^*_{i,s} = 0$ otherwise. If we take the log of (5) we have:

$$\begin{aligned} \log g(y_i \mid \mathbf{z}, \mathbf{x}) &= \sum_{s=1}^{m_i-1} [y^*_{i,s} \log \frac{\gamma_{i,s}}{\gamma_{i,s+1} - \gamma_{i,s}} - y^*_{i,s+1} \log \frac{\gamma_{i,s+1}}{\gamma_{i,s+1} - \gamma_{i,s}}] \\ &= \sum_{s=1}^{m_i-1} [y^*_{i,s} \theta_{i,s}(\mathbf{z}, \mathbf{x}) - y^*_{i,s+1} b(\theta_{i,s}(\mathbf{z}, \mathbf{x}))] \end{aligned} \tag{6}$$

From (6) we see that each component is in the form of the general expression of the exponential family distribution. More specifically:

$$\theta_{i,s}(\mathbf{z}, \mathbf{x}) = \log \frac{\gamma_{i,s}}{\gamma_{i,s+1} - \gamma_{i,s}}, \qquad s = 1, \ldots, m_i - 1 \tag{7}$$

and

$$b(\theta_{i,s}(\mathbf{z}, \mathbf{x})) = \log \frac{\gamma_{i,s+1}}{\gamma_{i,s+1} - \gamma_{i,s}} = \log\{1 + \exp(\theta_{i,s}(\mathbf{z}, \mathbf{x}))\}, \qquad s = 1, \ldots, m_i - 1 \tag{8}$$

To simplify the notation we write $\theta_{i,s}$ and $b(\theta_{i,s})$. The canonical parameter $\theta_{i,s}$ is not a linear function of the latent variable. Expression (6) brings our

model to generalized linear models form and so this will be used to derive the maximum likelihood results.

2.2 Structural model

Let us now assume that the latent variables $\mathbf{z}_h$ are related to a set of observed covariates $\mathbf{w}_h$ in a simple linear form:

$$\mathbf{z}_h = \boldsymbol{\Lambda}\mathbf{w}_h + \boldsymbol{\delta}_h \qquad h = 1, \ldots, n \tag{9}$$

where $\mathbf{z}_h$ is $q \times 1$ vector, $\boldsymbol{\Lambda}$ is a $q \times k$ matrix of regression coefficients and the $\boldsymbol{\delta}_h$ is a $q \times 1$ vector of independent standard normal variables.

2.3 Model estimation

The parameters to be estimated are $\boldsymbol{\Lambda}, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. If $\mathbf{z}$ was directly observable we could maximize the complete data likelihood:

$$\log f(\mathbf{y}, \mathbf{z} \mid \mathbf{x}, \mathbf{w}) = \sum_{h=1}^{n} \left[\sum_{i=1}^{p} \log g(y_{ih} \mid \mathbf{z}, \mathbf{x}_h) + \log h(\mathbf{z}, \boldsymbol{\Lambda}) \right] \tag{10}$$

Because $\mathbf{z}$ is unknown $\log f(\mathbf{y}, \mathbf{z} \mid \mathbf{x}, \mathbf{w})$ is maximized using an EM algorithm that computes the expected score function of the model parameters where the expectation is with respect to the posterior distribution of $\mathbf{z}$ given the observations.

2.3.1 Estimation of $\mathbf{\Lambda}$

From (10) we see that the estimation of the parameters contained in matrix $\mathbf{\Lambda}$ does not depend on the first component of the complete loglikelihood. Therefore its estimation can be done separately from the rest of the parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). In addition the latent variables are assumed to be independent so that $h(\mathbf{z}, \mathbf{\Lambda}) = h(z_1, \boldsymbol{\lambda}_1) \times \cdots \times h(z_q, \boldsymbol{\lambda}_q)$

The expected score function with respect to the parameter vector $\boldsymbol{\lambda}_j; j = 1 \ldots, q$ takes the form:

$$ES_h(\boldsymbol{\lambda}_j) = \int \cdots \int S_h(\boldsymbol{\lambda}_j) h(\mathbf{z} \mid \mathbf{y}_h, \mathbf{x}_h) d\mathbf{z} \tag{11}$$

where

$$S_h(\boldsymbol{\lambda}_j) = \frac{\partial \log h(z_j, \boldsymbol{\lambda}_j)}{\partial \boldsymbol{\lambda}_j}; \qquad j = 1, \ldots, q$$

Equation (11) becomes:

$$ES_h(\boldsymbol{\lambda}_j) = \int \cdots \int \mathbf{w}_h (z_j - \mathbf{w}_h' \boldsymbol{\lambda}_j) h(\mathbf{z} \mid \mathbf{y}_h, \mathbf{x}_h) d\mathbf{z} \tag{12}$$

Solving $\sum_{h=1}^{n} ES_h(\boldsymbol{\lambda}_j) = 0$ and approximating the integral with Gauss-Hermite quadrature points we get an explicit solution for the maximum likelihood estimator of $\boldsymbol{\lambda}_j$:

$$\hat{\boldsymbol{\lambda}_j} = \frac{\sum_{h=1}^{n} \mathbf{w}_h \sum_{t_1=1}^{\nu_1} \cdots \sum_{t_q=1}^{\nu_q} z_{t_j} h(z_{t_1}, \cdots, z_{t_q} \mid \mathbf{y}_h, \mathbf{x}_h)}{\sum_{h=1}^{n} \mathbf{w}_h \mathbf{w}_h'} \tag{13}$$

where

$$h(z_{t_1}, \cdots, z_{t_q} \mid \mathbf{y}_h, \mathbf{x}_h) = \frac{g(\mathbf{y}_h \mid z_{t_1} \cdots z_{t_q}, , \mathbf{x}_h) h(z_{t_1}, \boldsymbol{\lambda}_1) \cdots h(z_{t_q}, \boldsymbol{\lambda}_q)}{f(\mathbf{y}_h, \mathbf{x}_h)}$$

This equation is updated at each step of the EM algorithm described in section 2.3.3.

2.3.2 Estimation of the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

The estimation of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ depends on the first component of (10).

Let denote $\mathbf{a}'_i = (\alpha_{i1}, \ldots, \alpha_{i,m_i-1}, \alpha_i, \beta_{i1}, \ldots, \beta_{ir})$, $i = 1, \ldots, p$.

The expected score function of the parameter vector $\mathbf{a}_i$ where the expectation is taken with respect to $h(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$ is:

$$ES_h(\mathbf{a}_i) = \int \cdots \int S_h(\mathbf{a}_i) h(\mathbf{z} \mid \mathbf{y}_h, \mathbf{x}_h) d\mathbf{z}, \qquad h = 1, \ldots, n \qquad (14)$$

where

$$S_h(\mathbf{a}_i) = \frac{\partial \log g(\mathbf{y}_h \mid \mathbf{z}, \mathbf{x}_h)}{\partial \mathbf{a}_i}, \qquad i = 1, \ldots, p$$

Now,

$$\frac{\partial \log g(\mathbf{y}_h \mid \mathbf{z}, \mathbf{x}_h)}{\partial \mathbf{a}_i} = \sum_{s=1}^{m_i-1} [y^*_{i,s,h}\theta'_{i,s,h} - y^*_{i,s+1,h}b'(\theta_{i,s,h})] \tag{15}$$

Replace (15) into (14):

$$ES_h(\mathbf{a}_i) = \int \cdots \int \sum_{s=1}^{m_i-1} [y^*_{i,s,h}\theta'_{i,s,h} - y^*_{i,s+1,h}b'(\theta_{i,s,h})]h(\mathbf{z} \mid \mathbf{y}_h, \mathbf{x}_h)d\mathbf{z} \tag{16}$$

Solving $\sum_{h=1}^{n} ES_h(\mathbf{a}_i) = 0$ and approximating the integral with Gauss-Hermite quadrature points we get non-explicit solutions for the parameter vector $\mathbf{a}_i$:

$$\sum_{t_1=1}^{\nu_1} \cdots \sum_{t_q=1}^{\nu_q} \sum_{s=1}^{m_i-1} \left[\sum_{h=1}^{n} y^*_{i,s,h} \frac{\partial \theta_{i,s,h}}{\partial \mathbf{a}_i} - \sum_{h=1}^{n} y^*_{i,s+1,h} \frac{\partial b(\theta_{i,s,h})}{\partial \mathbf{a}_i} \right] h(z_{t_1} \cdots z_{t_q} \mid \mathbf{y}_h, \mathbf{x}_h) \tag{17}$$

Equation (17) is written as:

$$\sum_{t_1=1}^{\nu_1} \cdots \sum_{t_q=1}^{\nu_q} \sum_{s=1}^{m_i-1} \left[r_{i,s,t_1,\ldots,t_q} - r_{i,s+1,t_1,\ldots,t_q} \right] \tag{18}$$

where

$$r_{i,s,t_1,\ldots,t_q} = \sum_{h=1}^{n} h(z_{t_1},\ldots,z_{t_q} \mid \mathbf{y}_h, \mathbf{x}_h) y^*_{i,s,h} \frac{\partial \theta_{i,s,h}}{\partial \mathbf{a}_i} \tag{19}$$

$$r_{i,s+1,t_1,\ldots,t_q} = \sum_{h=1}^{n} h(z_{t_1},\ldots,z_{t_q} \mid \mathbf{y}_h, \mathbf{x}_h) y^*_{i,s+1,h} \frac{\partial b(\theta_{i,s,h})}{\partial \mathbf{a}_i} \tag{20}$$

From the above results we can see that to compute the derivatives with respect to the model parameters for any link function we need to find the first derivatives of the functions $\theta_{i,s,h}$ and $b(\theta_{i,s,h})$ with respect to the model parameters. The maximization of the loglikelihood is done by an E-M algorithm. The model without covariate effects has $\theta_{i,s,h}$ and $b(\theta_{i,s,h})$ functions not depending on the individual h.

2.3.3 E-M algorithm

The steps of the E-M algorithm are defined as follows:

step1 Choose initial estimates for the model parameters α_{is}, α_i, β_{il} and $\lambda_{j\nu}$ where $i = 1,\ldots,p$; $s = 1,\ldots,m_i - 1$; $l = 1,\ldots,r$; $j = 1,\ldots,q$; $\nu = 1,\ldots,k$.

step2 Compute the values $r_{i,s,t_1,\ldots,t_q}$ and $r_{i,s+1,t_1,\ldots,t_q}$ (E-step).

step3 Obtain improved estimates for the parameters by solving the non-

linear maximum likelihood equations for the parameters $\alpha_{is}, \alpha_i, \beta_{il}$ and explicit solutions for the parameters $\lambda_{j\nu}$ of the latent distribution. (M-step)

step4 Return to step 2 and continue until convergence is attained.

At the M-step a one-step Fisher scoring algorithm, is used to solve the non-linear maximum likelihood equations.

2.4 Proportional Odds Models

The proportional odds model is a special case of model 1 with the logit as a link function.

$$\log\left[\frac{\gamma_{i,s}(\mathbf{z},\mathbf{x})}{1-\gamma_{i,s}(\mathbf{z},\mathbf{x})}\right] = \alpha_{is} - \sum_{j=1}^{q}\alpha_{ij}z_j + \sum_{l=1}^{r}\beta_{il}x_l, \qquad s = 1,\ldots,m_i-1; l = 1\ldots,r, \tag{21}$$

From (21) we get that:

$$\gamma_{i,s} = P(y_i \leq s \mid \mathbf{z},\mathbf{x}) = \frac{\exp(\alpha_{is} - \sum_{j=1}^{q}\alpha_{ij}z_j + \sum_{l=1}^{r}\beta_{il}x_l)}{1+\exp(\alpha_{is} - \sum_{j=1}^{q}\alpha_{ij}z_j + \sum_{l=1}^{r}\beta_{il}x_l)}\ , s = 1,2,\ldots,m_i-1\ . \tag{22}$$

Let us denote with $\mathbf{a}_i' = (\alpha_{i1},\ldots,\alpha_{iq},\beta_{i1},\ldots,\beta_{ir})$ and $\mathbf{v}' = (\mathbf{z},\mathbf{x})$ then for two individuals with values $\mathbf{v}_1$ and $\mathbf{v}_2$ the difference between two corresponding logits is $\mathbf{a}'(\mathbf{v}_2 - \mathbf{v}_1)$ and it does not depend on the category involved.

The derivatives required in (17) for the proportional odds model are:

$$\frac{\partial \theta_{i,s,h}}{\partial \alpha_{is}} = \frac{(1-\gamma_{i,s,h})\gamma_{i,s+1,h}}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})} \qquad s = 1,\ldots,m_i-1,$$

$$\frac{\partial \theta_{i,s-1,h}}{\partial \alpha_{is}} = -\frac{(1-\gamma_{i,s,h})\gamma_{i,s,h}}{(\gamma_{i,s,h}-\gamma_{i,s-1,h})}\,, \qquad s = 2,\ldots,m_i-1,$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \alpha_{is}} = \frac{\gamma_{i,s,h})(1-\gamma_{i,s,h})}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}\,, \qquad s = 1,\ldots,m_i-1$$

$$\frac{\partial b(\theta_{i,s-1,h})}{\partial \alpha_{is}} = -\frac{\gamma_{i,s-1,h}(1-\gamma_{i,s,h})}{(\gamma_{i,s,h}-\gamma_{i,s-1,h})}\,, s = 2,\ldots,m_i-1$$

$$\frac{\partial \theta_{i,s,h}}{\partial \alpha_{ij}} = -z_j\gamma_{i,s+1,h} \qquad j = 1,\ldots,q$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \alpha_{ij}} = -z_j\gamma_{i,s,h} \qquad j = 1,\ldots,q$$

$$\frac{\partial \theta_{i,s,h}}{\partial \beta_{il}} = x_{il}\gamma_{i,s+1,h}\,, \qquad l = 1,\ldots,r$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \beta_{il}} = x_{il}\gamma_{i,s,h}\,, l = 1,\ldots,r$$

2.5 Proportional Hazards Model

The proportional hazards model is used in the analysis of survival data and it is also a special case of the general framework presented above. The hazard function or risk function is defined to be the failure probability at time t conditional on survival up to time t and it is denoted by $\lambda_i(t;\mathbf{z},\mathbf{x})$. The proportional hazards model is written (see Cox 1972):

$$\lambda_i(t;\mathbf{z},\mathbf{x}) = \lambda_{i0}(t)\exp(-\alpha_{is} - \sum_{j=1}^{q}\alpha_{ij}z_j + \sum_{l=1}^{r}\beta_{il}x_l),$$

$$i = 1, \ldots, p; \quad l = 1, \ldots, r; \quad s = 1, \ldots, m_i$$

The function $\lambda_{i0}(t)$ changes with time and it is the value of the risk function at $\mathbf{z} = \mathbf{0}$ and $\mathbf{x} = \mathbf{0}$. The part in the exponent shows how the risk changes as a function of the latent variable and the set of observed covariates. In that model as well the ratio of two hazard functions for two individuals is independent of the category involved.

For discrete data the proportional hazards model is written as:

$$-\log[1 - \gamma_{is}(\mathbf{z}, \mathbf{x})] = \exp(\alpha_{is} - \sum_{j=1}^{q} \alpha_{ij} z_j + \sum_{l=1}^{r} \beta_{il} x_l)$$

where $(1 - \gamma_{is}(\mathbf{z}, \mathbf{x}))$ is the survival probability beyond category s.

The model is written in its linear form as:

$$\log[-\log[1 - \gamma_{is}(\mathbf{z}, \mathbf{x})]] = \alpha_{is} - \sum_{j=1}^{q} \alpha_{ij} z_j + \sum_{l=1}^{r} \beta_{il} x_l$$

This is called the log-log transformation. For that model also the difference between two log-logs is constant and does not depend on the category involved.

The derivatives required in (17) for the proportional hazards model are:

$$\frac{\partial \theta_{i,s,h}}{\partial \alpha_{is}} = -\frac{\gamma_{i,s+1,h}(1 - \gamma_{i,s,h}) \log(1 - \gamma_{i,s,h})}{\gamma_{i,s,h}(\gamma_{i,s+1,h} - \gamma_{i,s,h})},$$

$$\frac{\partial \theta_{i,s-1,h}}{\partial \alpha_{is}} = \frac{(1-\gamma_{i,s+1,h})\log(1-\gamma_{i,s+1,h})}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \alpha_{i,s}} = -\frac{(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}\,,$$

$$\frac{\partial b(\theta_{i,s-1,h})}{\partial \alpha_{i,s}} = \frac{\gamma_{i,s-1,h}(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})}{(\gamma_{i,s,h}-\gamma_{i,s-1,h})}$$

$$\frac{\partial \theta_{i,s,h}}{\partial \alpha_{ij}} = \frac{z_j}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}[(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})\frac{\gamma_{i,s+1,h}}{\gamma_{i,s,h}}-(1-\gamma_{i,s+1,h})\log(1-\gamma_{i,s+1,h})]$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \alpha_{ij}} = -\frac{z_j}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}[\frac{\gamma_{i,s,h}}{\gamma_{i,s+1,h}}(1-\gamma_{i,s+1,h})\log(1-\gamma_{i,s+1,h})-(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})]$$

$$\frac{\partial \theta_{i,s,h}}{\partial \beta_{il}} = \frac{x_{il}}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}[(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})\frac{\gamma_{i,s+1,h}}{\gamma_{i,s,h}}-(1-\gamma_{i,s+1,h})\log(1-\gamma_{i,s+1,h})]$$

$$\frac{\partial b(\theta_{i,s,h})}{\partial \beta_{il}} = \frac{x_{il}}{(\gamma_{i,s+1,h}-\gamma_{i,s,h})}[(1-\gamma_{i,s+1,h})\log(1-\gamma_{i,s+1,h})\frac{\gamma_{i,s,h}}{\gamma_{i,s+1,h}}-(1-\gamma_{i,s,h})\log(1-\gamma_{i,s,h})]$$

The proportional hazards model is suggested here as an alternative link function for the ordinal manifest variables rather than a way of modeling manifest variables that measure survival time (duration of an event). Manifest variables that measure duration of time can be easily accommodated into the framework presented in this paper but we think it is outside of the scope of this paper.

All models under the framework presented are affected by an arbitrary permutation of the response categories. The proportional odds model is unaffected when only a reversal of category order occurs. Under those circumstances there is only a change in the sign of the regression and latent coefficients and a change in sign and order for the threshold parameters. This invariance does not hold in the proportional hazards model (log-log link function).

Standard errors

Asymptotically the standard errors of the maximum likelihood estimates are given by the diagonal elements of the inverse of the information matrix evaluated at the maximum likelihood solution. In the program LATENT the standard errors of the maximum likelihood estimates are based on an approximation of the information matrix given by :

$$I(\hat{\beta}) = \sum_{h=1}^{n} \frac{1}{f^2(\mathbf{y}_h)} \frac{\partial f(\mathbf{y}_h)}{\partial a_j} \frac{\partial f(\mathbf{y}_h)}{\partial a_k},$$

where a is the vector with the model parameters.

2.6 Goodness-of-Fit

The goodness-of-fit of the model can be theoretically checked by computing a Pearson chi-square or a likelihood ratio statistic. When the number of manifest ordinal variables is large it is expected that many response patterns will have expected frequency less than 5 and many will be so small that they will not occur at all. So from the practical point of view these tests cannot be used.

Alternatively we can compute the Pearson chi-square statistic or likelihood ratio statistic only for pairs and triplets of responses. Those values can be called residuals and provide information on how well the model predicts the two- and three-way margins. A detailed discussion on the use of those goodness-of-fit measures for ordinal variables can be found in Jöreskog and Moustaki (2001). However, for the model with covariate effects the Pearson chi-square statistic or likelihood ratio statistic for pairs and triplets of responses have to be computed for different values of the explanatory variables. That will eventually make the use of those residuals less informative with respect to goodness-of-fit.

Alternatively, instead of testing the goodness-of-fit of a specified model we could use a criterion for selecting among a set of different models. This procedure gives information about the goodness-of-fit for each model in com-

parison with other models. This can be useful for the determination of the number of factors required or for comparing the model with latent variables and covariate effects with the model with only latent variables. Sclove (1987) gives a review of some of the model selection criteria used in multivariate analysis such as those due to Akaike, Schwarz and Kashap. These criteria take into account the value of the likelihood at the maximum likelihood solution and the number of parameters estimated.

Akaike's criterion for the determination of the order of an autoregressive model in time series has also been used for the determination of the number of factors in factor analysis, see Akaike (1987).

$$AIC = -2[\max L] + 2m \tag{23}$$

where m is the number of model parameters. The model with the smallest AIC value is taken to be the best one.

In this paper we also use an information complexity criterion proposed by Bozdogan (2000). The criterion is defined as

$$ICOMP = -2[\max L] + 2\frac{p}{2}\log[\frac{tr(\Sigma)}{p}] - \frac{1}{2}\log\det\Sigma \tag{24}$$

where Σ is the covariance matrix of the parameter estimates and p is the number of parameters estimated.

3 Factor Scores

The effect of the covariates $\mathbf{w}$ on the vector of latent variables $\mathbf{z}$ can be also estimated in two stages. First the measurement model is fitted and latent scores are computed from the measurement model. Then the latent scores can be used as dependent variables on further analysis with the vector of covariates $\mathbf{w}$.

In the analysis of binary, nominal categorical and metric manifest variables the posterior distribution of the latent variables conditional on the manifest variables depend on the manifest variables through a set of q components called sufficient statistics where q is the number of latent variables and it is much less than p. Those components are a weighted sum of the observed responses where the weights are the coefficients of the latent variables. In contrast, in the model formulation for ordinal variables the canonical parameter is not a linear function of the latent variables and the set of observed covariates. As a result the posterior distribution of the latent variables conditional on the manifest variables does not depend on the manifest variables through any sufficient statistics such as the component scores. To score the individuals on the latent dimensions defined by the analysis one can use the mean of the posterior distribution of the latent variable z_j given the individual's response pattern $E(z_j \mid \mathbf{y_h}, \mathbf{x_h})$. In the qth factor model the

posterior mean is given by:

$$E(z_j \mid \mathbf{y_h}, \mathbf{x_h}) = \int_{R_{z1}} \cdots \int_{R_{zq}} z_j h(\mathbf{z} \mid \mathbf{y}_h, \mathbf{x}_h) d\mathbf{z} \tag{25}$$

where R_{zj} denotes the range of values for z_j.

4 Application

In that section we analyze two data sets from the 1996 British Social Attitudes Survey[1].

Both data sets are analyzed using the proportional odds model(POM). POM is a special case of the general model presented in section 1 with a logit link function.

4.1 Example 1

The first data set consists of five ordinal manifest variables given below.

On the whole do you think it should or not be the government's responsibility to..

- provide a job for everyone who wants one [JobEvery]
- keep prices under control [PriCon]
- provide a decent standard of living for the unemployed [LivUnem]

[1]Social and Community Planning Research, British Social Attitudes Survey,1996, [computer file] Colchester, Essex: The Data Archive [distributor], 2 December 1998. SN: 3921

- reduce income differences between the rich and the poor [IncDiff]
- provide decent housing for those who can't afford it [Housing]

y_1 to y_5 are the responses to these five items. The response alternatives given to the respondents are: definitely should be, probably should be, probably not be and definitely should not be. Item nonresponse vary between 2%-6%. After we excluded the missing values we were left with 822 respondents.

Model 1a

A covariable x that is constructed to measure left to right political identification, is used as a continuous explanatory variable for the manifest ordinal variables. The ML estimates of the threshold parameters are given in Table 1 and the discrimination and regression parameters are given in Table 2. The discrimination parameters are all positive and of similar magnitude indicating that the five ordinal items are all indicators of one latent variable with more or less the same discrimination power. Their positive sign indicate that the more an individual believes that the state should not be responsible for its citizens the less likely it is to be on the lower categories of the ordinal variables. The negative sign of the regression coefficients also show that the more right wing an individual is the lower the probability of being in the low level categories (i.e. yes) of the ordinal observed variables. The one- two- and three-way margins show no big discrepancies between

the observed and the expected under the model frequencies. Table 3 gives the AIC and ICOMP criteria for the model with and without the covariate. They both conclude that the model with the covariate effect is a better fit than the one without the covariate effect on the manifest variables.

Table 1: Threshold estimates

Item	category	α_{is}
JobEvery	1	-1.254
	2	1.178
	3	2.887
Pricon	1	-0.333
	2	2.114
	3	3.530
LivUnem	1	-1.635
	2	2.438
	3	4.613
IncDiff	1	-0.950
	2	1.254
	3	3.459
Housing	1	-0.971
	2	3.678
	3	6.270

Table 2: Factor loadings and regression parameters

Item	α_i	β_i
JobEvery	1.102	-2.108
Pricon	0.693	-1.425
LivUnem	2.185	-2.071
IncDiff	1.356	-2.632
Housing	2.250	-2.120

Model 1b

In the above example (Model 1a) we modeled the effect of the political identification variable on the manifest variables controlling for the latent

Table 3: Model selection criteria

	Model with no covariate	Model with covariate
AIC	8636.186	8244.148
ICOMP	8607.984	8199.568

Table 4: Factor loadings

Item	α_i
JobEvery	1.466
Pricon	0.954
LivUnem	1.420
IncDiff	1.803
Housing	1.453
Structural part of the model $\hat{\lambda} = 1.446$	

variable. Alternatively we could have modeled the effect of the political identification variable directly on the latent variable (model (9). Table 4 gives the parameter estimates when model (9) is fitted.

4.2 Example 2

The second application is also from the 1996 British Social Attitudes Survey. Five ordinal manifest variables were selected for the analysis. The items measure satisfaction with the National Health Service in respondents' area.

The items asked are whether the National Health Service in your area is, on the whole, satisfactory or in need of improvement.

- GP's appointment systems [Appointment]
- Amount of time GP gives to each patient [AmountTime]
- Being able to choose which GP to see [ChooseGP]
- Quality of medical treatment by GPs [Quality]

- Waiting areas at GP's surgeries [WaitingArea]

The response alternatives given to the respondents are: In need of a lot of improvement, in need of some improvement, satisfactory, and very good. Item nonresponse vary between 1.5%-2.5%. After we excluded the missing values we were left with 841 respondents. In the analysis we are interested in measuring overall satisfaction with GP's from the five ordinal manifest variables controlling for respondents' political identification (measured by an observed covariate with four categories: conservative, labour, liberal democrat and other). We also want to measure the effect of gender and age on the latent variable satisfaction. Age is given in four categories: 18-25, 26-44, 45-64, 65+.

First we fit the one factor model to the five ordinal manifest variables without allowing for any covariate effects. Table 5 gives the factor loadings (α_i). The factor loadings are all positive and of similar magnitude indicating that the five ordinal items are all indicators of one latent variable with more or less the same discrimination power. Their positive signs indicate that the more satisfied an individual is with the National Health Service in his area the less likely he/she is to be in the lower categories of the ordinal variables. The fit of the one-factor model is satisfactory judging from the one- two- and three-way margins. Most of those margins show small discrepancies between the observed and the expected under the model frequencies. Most of those values are between 0-4. Table 6 gives pairs of items and categories

Table 5: One factor model: factor loadings

Item	α_i
Appointment	1.915
AmountTime	3.252
ChooseGP	2.292
Quality	2.309
WaitingArea	1.443

Table 6: One factor model: $(O-E)^2/E$ for the margins

Item	2	3	4	5
1	(1,4), (3,4)	(3,4)	(1,4), (2,4), (3,4)	
2			(2,4), (3,4), (4,2),(4,3)	
3				(1,2), (1,4)
4				(2,4)

for which the chi-square value computed for those combinations of items and categories is greater than four.

Then we fitted the one factor model that allows for covariate effects. The ML estimates of the thresholds parameters are given in Table 7 and the factor loadings and regression parameters are given in Table 8. The effect of the covariates age and gender on the latent variables are given in Table 9.

The factor loadings (α_i) remain all positive and of similar magnitude. Also the values of the factor loadings have not changed much when the covariates were introduced into the model indicating factorial invariance within the groups defined by the covariates. The direct effects of the political party covariate on the manifest ordinal variables are similar with the exception of variable 3 [ChooseGP]. Respondents that tend to vote for the labour party

Table 7: Threshold estimates

Item	category	$\alpha_{i(s)}$
Appointment	1	-1.254
	2	1.178
	3	2.887
AmountTime	1	-0.333
	2	2.114
	3	3.530
ChooseGP	1	-1.635
	2	2.438
	3	4.613
Quality	1	-0.950
	2	1.254
	3	3.459
WaitingArea	1	-0.971
	2	3.678
	3	6.270

Table 8: Factor loadings and direct effect parameters

Item	α_i	labour	liberal	other
Appointment	1.786	0.910	0.703	0.551
AmountTime	2.914	1.309	0.613	0.714
ChooseGP	2.146	0.568	-0.202	0.313
Quality	2.129	0.906	0.613	1.090
WaitingArea	1.329	0.533	0.051	0.690

Table 9: Structural parameters

Female	26-44	45-64	65+
-0.070	0.165	0.483	0.698

are more likely to express dissatisfaction with each one of the five ordinal items than those respondents that tend to vote for the conservative party. Finally, from Table 9 we see that gender has no effect on overall satisfaction with the National Health Service but as respondents age increases so does their satisfaction with the Health Service.

The AIC criterion for the model without the covariates is 7966.5 and for the model with the covariates is 7936.9. We conclude that the model with the covariate effects is a better fit than the one without.

5 Conclusion

We propose a general framework for fitting latent variable models that allows for covariate effects both on the manifest and on the latent variables. The approach proposed is full maximum likelihood and it is distinct from the approaches used by researchers such as Jöreskog and Muthen. The drawback of the approach proposed here is the computational burden of the integrations required in the evaluation of the loglikelihood function.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika 52*, 317–332.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology 44*, 62–91.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B 34*, 187–220.

Croon, M. and A. Bolck (1997). On the use of factor scores in structural equations models. Technical Report 97.10.102/7, Work and Organization Research Centre, Tilburg University. WORC Paper.

Jöreskog, K. G. and A. S. Goldberger (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association 70*, 631–639.

Jöreskog, K. G. and I. Moustaki (2001). Factor analysis of ordinal variables: a comparison of three approaches. *To appear Multivariate Behavioural Research.*

Jöreskog, K. G. and D. Sörbom (1996). *LISREL 8: Structural Equation Modelling with the SIMPLIS Command Language Computer manual.* Hove and London: Scientific Software International.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B 42*, 109–142.

Moustaki, I. (2000a). A latent variable model for ordinal variables. *Applied Psychological Measurement 24*, 211–223.

Moustaki, I. (2000b). A review of exploratory factor analysis for ordinal categorical data. In R. Cudeck, S. Du Toit, and D. Sörbom (Eds.), *Structural equation modeling: present and future.* Scientific Software International.

Moustaki, I. and M. Knott (2000). Generalized latent trait models. *Psychometrika 65*, 391–411.

Muraki, E. and E. Carlson (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement 19*, 73–90.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika 54*(4), 557–585.

Muthén, B. O. and L. Muthén (2000). *Mplus: The comprehensive modeling program to applied researchers.* 11965 Venice Boulevard, Suite 407, Los Angeles, CA 90066.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17 34.*

Sclove, S. (1987). Application of model-selection criteria to some problems of multivariate analysis. *Psychometrika 52*, 333–343.