Factor Analysis and Latent Structure of Categorical Data

Irini Moustaki
Athens University of Economics and Business

Outline

- Objectives
- Factor analysis model
- Literature Approaches
- Item Response Theory
- Models for categorical responses (nominal, ordinal and survival)
- Applications

Objectives of Latent Variable Analysis

1. Identify the factors that explain the relationships among the observed items.

2. Design, construction and evaluation of educational and psychological tests.

   - Fitting a factor analysis model with one or more factors
   - Fitting a latent class model with two or more latent classes

3. Scale individuals on the identified latent dimensions

   - Estimate the posterior mean of the distribution of the latent variable given the response pattern.
   - Component Scores (Bartholomew and Knott, 1999).

Notation and general formulation

Manifest variables are denoted by: $y_1, y_2, \ldots, y_p$.

Latent variables are denoted by: $z_1, z_2, \ldots, z_q$.

One wants to find a set of latent factors $z_1, \ldots, z_q$, fewer in number than the observed variables ($q < p$), that contain essentially the same information

If both the response variables and the latent factors are normally distributed with zero means and unit variances, this leads to the model (see Jöreskog, 1979)

$$E(y_i \mid z_1, z_2, \ldots, z_q) = \lambda_{i1} z_1 + \lambda_{i2} z_2 + \cdots + \lambda_{iq} z_q$$

$$E(y_i y_j \mid z_1, z_2, \ldots, z_q) = 0 \ , i \neq j$$

In case of nominal or ordinal responses we must instead specify the probability of each response pattern as a function of $z_1, z_2, \ldots, z_q$:

$$Pr(y_1 = a_1, y_2 = a_2, \ldots, y_p = a_p \mid z_1, z_2, \ldots, z_q) = f(z_1, z_2, \ldots, z_q)$$

where $a_1, a_2, \ldots, a_p$ represent the different response categories of $y_1, y_2, \ldots, y_p$, respectively.

If the factors are independent, it follows that the correlation ρ_{ij} between y_i and y_j is

$$\rho_{ij} = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl} \ .$$

Literature Approaches

Item Response Theory Approach: specifies the conditional distribution of the complete p-dimensional response pattern as a function of the latent variables and/or explanatory variables.

Underlying Variable Approach: supposes that the categorical observed variables y are generated by underlying unobserved continuous variables assumed to be normally distributed.

Note

Model parameters equivalence exist between the two approaches (Takane and De Leeuw, 1987 Psychometrika).

Underlying Variable Approach

All the variables are treated as metric through assumed underlying and normal variables and by using ML, GLS or WLS as the estimation method. Estimation is done in two- or three-stages.

- Jöreskog and Sörbom (LISREL)
- Muthén and Muthén (M-Plus)
- Bentler (EQS)
- Arminger and Küsters (MECOSA)

Their work covers a wide range of models that allows relationships among the latent variables, inclusion of exogenous (explanatory) variables, multilevel analysis, analysis of panel data.

UVA for ordinal responses, multivariate normality

$$y_i^* = \lambda_{i1} z_1 + \lambda_{i2} z_2 + \cdots + \lambda_{ik} z_q + u_i \ , i = 1, 2, \ldots, p$$

The connection between the ordinal variable y_i and the underlying variable y_i^* is

$$y_i = a \quad \Longleftrightarrow \quad \tau_{a-1}^{(i)} < y_i^* \leq \tau_a^{(i)}, \ a = 1, 2, \ldots, m_i$$

where

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

- The mean and variance of y_i^* are set equal to zero and one respectively.
- $z_1, \ldots, z_q, u_1, \ldots, u_p$ are independent and normally distributed with

- $z_j \sim N(0,1)$
- $u_i \sim N(0, \psi_i^2)$.
- It follows that $y_1^*, \ldots, y_p^*$ has a multivariate normal distribution with zero means, unit variances and correlation matrix $\mathbf{P} = (\rho_{ij})$, where $\rho_{ij} = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}$.

$$\pi_r(\boldsymbol{\theta}) = Pr(y_1 = a_1, y_2 = a_2, \ldots, y_p = a_p) =$$
$$\int_{\tau_{a_1-1}^{(1)}}^{\tau_{a_1}^{(1)}} \int_{\tau_{a_2-1}^{(2)}}^{\tau_{a_2}^{(2)}} \cdots \int_{\tau_{ap-1}^{(p)}}^{\tau_{ap}^{(p)}} \phi_p(u_1, u_2, \ldots u_p|\mathbf{P}) du_1 du_2 \cdots du_p$$

UVA, bivariate normality, UBN

The UBN approach makes the same assumptions as the previous approach but uses only the data in the univariate and bivariate margins to estimate the model. (Jöreskog and Moustaki (2001, MBR)).

$$\pi_a^{(g)}(\boldsymbol{\theta}) = \int_{\tau_{a-1}^{(g)}}^{\tau_a^{(g)}} \phi(u)du \ ,$$

where $\phi(u)$ is the standard normal density function, and that the probability $\pi_{ab}^{(gh)}$ of a response in category a on variable g and a response in category b on variable h is

$$\pi_{ab}^{(gh)}(\boldsymbol{\theta}) = \int_{\tau_{a-1}^{(g)}}^{\tau_a^{(g)}} \int_{\tau_{b-1}^{(h)}}^{\tau_b^{(h)}} \phi_2(u, v|\rho_{gh})dudv$$

where $\phi_2(u, v|\rho)$ is the density function of the standardized bivariate normal distribution with correlation ρ.

The UBN approach minimizes the sum of all univariate and bivariate fit functions (equivalent to maximizing the sum of all univariate and bivariate likelihoods):

$$F_{UBN}(\boldsymbol{\theta}) = \sum_{g=1}^{p}\sum_{a=1}^{m_g} p_a^{(g)} \ln[p_a^{(g)}/\pi_a^{(g)}(\boldsymbol{\theta})] + \sum_{g=2}^{p}\sum_{h=1}^{g-1}\sum_{a=1}^{m_g}\sum_{b=1}^{m_h} p_{ab}^{(gh)} \ln[p_{ab}^{(gh)}/\pi_{ab}^{(gh)}(\boldsymbol{\theta})]$$

Only data in the univariate and bivariate margins are used. This approach is quite feasible in that it can handle a large number of variables as well as a large number of factors.

- Further work is being done in that direction by Karl Jöreskog.

Other approaches

- PRELIS/LISREL and described by Jöreskog (1990 Q&Q, 1994 Psychometrika)

- MPLUS and described by Muthen (1984, Psychometrika) and Muthen & Satorra (1995, Psychometrika).

These are three-step procedures:

1. Thresholds are estimated from the univariate margins of the observed variables.

2. Polychoric correlations are estimated from the bivariate margins of the observed variables for given thresholds.

3. The factor model is estimated from the polychoric correlations by weighted least squares.

Item Response Theory: list of models for categorical responses

- Rasch model (Rasch 1960)

$$P(y_{ij} = 1 \mid \theta) = \frac{\theta_j}{\theta_j + \delta_i} = \frac{1}{1 + \exp(-(\theta - \delta_i))},$$

where θ_j is a person parameter and δ_i is an item parameter.

- Two-parameter logistic or probit model (Lord, 1952 and Birnbaum).

$$P(y_i = 1 \mid \theta) = \frac{1}{1 + \exp(-\alpha_i(\theta - b_i))}$$

- Three-parameter model (Birnbaum)

$$P(y_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{1}{1 + \exp(-\alpha_i(\theta - b_i))}$$

- Nominal categories model (Bock, 1972 Psychometrika)
- Partial Credit model (Masters 1982, Psychometrika)
- Generalized Partial Credit Model (Muraki 1992, APM)
- Graded Response Model (Samejima 1969, Psychometrika)

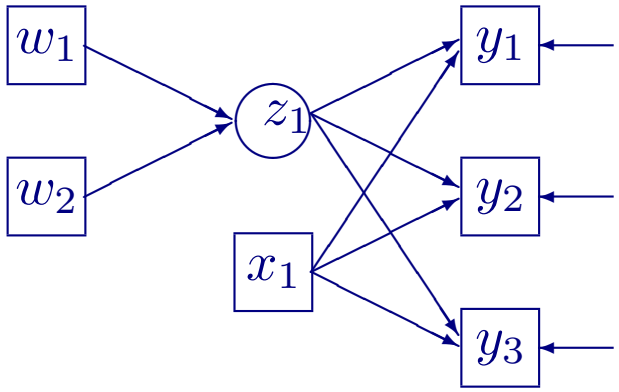
Handbook of Modern Item Response Theory edited by van der Linden, and Hambleton.

Non-parametric IRT models for polytomous responses:

Molenaar (1983), Agresti (1990), Mellenbergh (1995), van der Ark, Hemker and Sijtsma (2002), Tzamourani and Knott (2002).

- No distributional assumptions are made about the latent variables.
- Probability of a 'correct' response does not have a specific parametric form.

Path Diagram



Item Response Theory, Theoretical Framework

As only $\mathbf{y}$ can be observed any inference must be based on the joint distribution of $\mathbf{y}$:

$$f(\mathbf{y} \mid \mathbf{x}, \mathbf{w}) = \int_{R_\mathbf{z}} g(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) h(\mathbf{z} \mid \mathbf{w}) d\mathbf{z}$$

$h(\mathbf{z} \mid \mathbf{w})$: prior distribution of $\mathbf{z}$

$g(\mathbf{y} \mid \mathbf{z}, \mathbf{x})$: conditional distribution of $\mathbf{y}$ given $\mathbf{z}$ and $\mathbf{x}$.

$h(\mathbf{z} \mid \mathbf{w})$ and $g(\mathbf{y} \mid \mathbf{z}, \mathbf{x})$ are not uniquely determined.

What we want to know: $h(\mathbf{z} \mid \mathbf{y})$

If correlations among the y's can be explained by a set of latent variables and a set of explanatory variables then when all z's and the x's are accounted for the y's will be independent (local independence).

q must be chosen so that:

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

The question is whether $f(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$ admit the presentation:

$$f(\mathbf{y} \mid \mathbf{x}, \mathbf{w}) = \int_{R_{\mathbf{z}}} \prod_{i=1}^{p} g(y_i \mid \mathbf{z}, \mathbf{x}) h(\mathbf{z} \mid \mathbf{w}) d\mathbf{z}$$

for some small value of q.

Model Estimation

For a random sample of size n the loglikelihood is written as:

$$L = \sum_{h=1}^{n} \log f(\mathbf{y}_h \mid \mathbf{x}, \mathbf{w}) = \sum_{h=1}^{n} \log \int_{R_\mathbf{z}} \prod_{i=1}^{p} g(y_{ih} \mid \mathbf{z}, \mathbf{x}) h(\mathbf{z} \mid \mathbf{w}) d\mathbf{z}$$

The integrals can be approximated with Gauss Hermite quadrature, adaptive quadrature points, Monte Carlo, Laplace approximation.

Maximization of the log-likelihood is done with the E-M (Bock and Aitkin 1981, Psychometrika) or the Newthon-Raphson algorithm (Rabe-Hesketh, Pickles and Skrondal, GLLAMM).

Bayesian estimation methods have been applied too. (A review in two papers by Patz and Junker 1999, JEBS).

Logistic model for Binary Data

With p variables, each having two outcomes, there are 2^p different response patterns which are possible.

y_i: independent Bernoulli variables taking values 0 and 1.

The individuals in the sample are denoted with h where $h = 1, \ldots, n$.

The logit/probit model:

$$\text{logit}\pi_i(\mathbf{z}, \mathbf{x}) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} z_+ \sum_{l=1}^{k} \beta_{il} x_l$$

where

$$\pi_i(\mathbf{z}, \mathbf{x}) = P(y_i = 1 \mid \mathbf{z}, \mathbf{x})$$

Interpretation of Parameters

The coefficient α_{i0} is the value of logit$\pi_i(\mathbf{z}, \mathbf{x})$ at $\mathbf{z} = \mathbf{x} = \mathbf{0}$. The probability of a positive response from the median individual. The α_{i0} is called **'difficulty'** parameter.

$$\pi_i(\mathbf{0}) = P(y_i = 1 \mid \mathbf{0}) = \frac{\exp(\alpha_{i0})}{1 + \exp(\alpha_{i0})}$$

The coefficients $\alpha_{ij}, \quad j = 1, \ldots, q$ are called **'discrimination'** coefficients.

They also measure the extent to which the latent variable z_j discriminates between individuals.

Polytomous items, nominal

$$y_{i(s)} = \begin{cases} 1, & \text{if the response falls in category } s, \\ & s = 1, \ldots, c_i \\ 0, & \text{otherwise} \end{cases}$$

where c_i denotes the number of categories of variable i. The observed variable y_i is replaced by a vector $\mathbf{y}_i$ of c_i elements, where $\sum_s y_i(s) = 1$.

$$g(\mathbf{y}_i \mid \mathbf{z}, \mathbf{x}) = \prod_{s=1}^{c_i} (\pi_{i(s)}(\mathbf{z}, \mathbf{x}))^{y_{i(s)}}$$

$$\pi_{i(s)}(\mathbf{z}, \mathbf{x}) = P(y_{i(s)} = 1 \mid \mathbf{z}, \mathbf{x})$$

$$\text{logit}\pi_{i(s)}(\mathbf{z}, \mathbf{x}) = \alpha_{i0(s)} + \sum_{j=1}^{q} \alpha_{ij(s)} z_j + \sum_{l=1}^{k} \beta_{il(s)} x_l$$

Ordinal observed variables

To take into account the ordinality property of the items we model the cumulative probabilities, $\gamma_{i,s}(\mathbf{z}, \mathbf{x}) = P(y_i \leq s \mid \mathbf{z}, \mathbf{x})$.

The response category probabilities are denoted by

$$\pi_{i,s}(\mathbf{z}, \mathbf{x}) = \gamma_{i,s}(\mathbf{z}, \mathbf{x}) - \gamma_{i,s-1}(\mathbf{z}, \mathbf{x}), \quad s = 1, \ldots, m_i$$

m_i the number of categories for the ith item.

The model used is the proportional odds model:

$$\ln \left[\frac{\gamma_{i,s}(\mathbf{z}, \mathbf{x})}{1 - \gamma_{i,s}(\mathbf{z}, \mathbf{x})} \right] = \alpha_{is} - \sum_{j=1}^{q} \alpha_{ij} z_j + \sum_{l=1}^{k} \beta_{il(s)} x_l$$

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

The α_{is}: threshold parameters.

$$\alpha_{i1} < \alpha_{i2} \cdots < \alpha_{im_{i-1}} < \alpha_{im_i} = \infty$$

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

where $y_{i,s}$ takes the value 1 or 0.

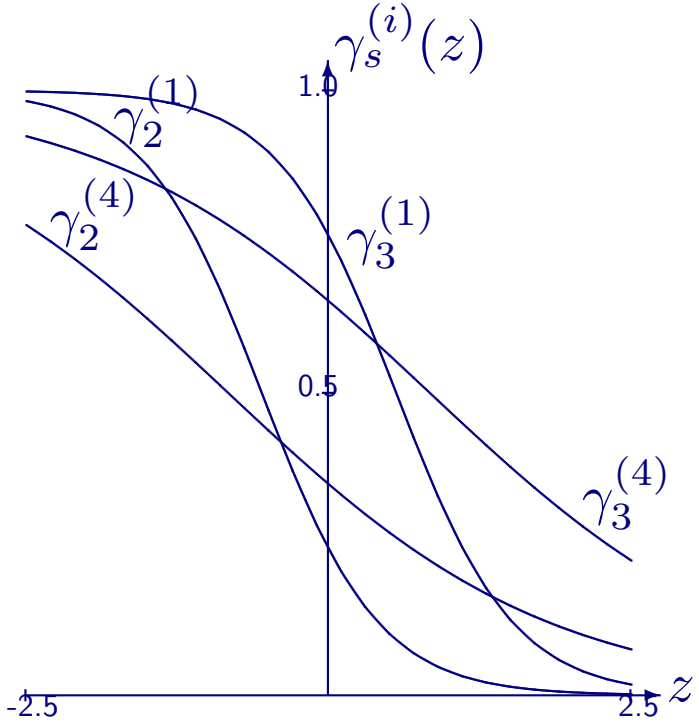


Figure 1: Logit: Four Cumulative Response Functions $\gamma_s^{(i)}$

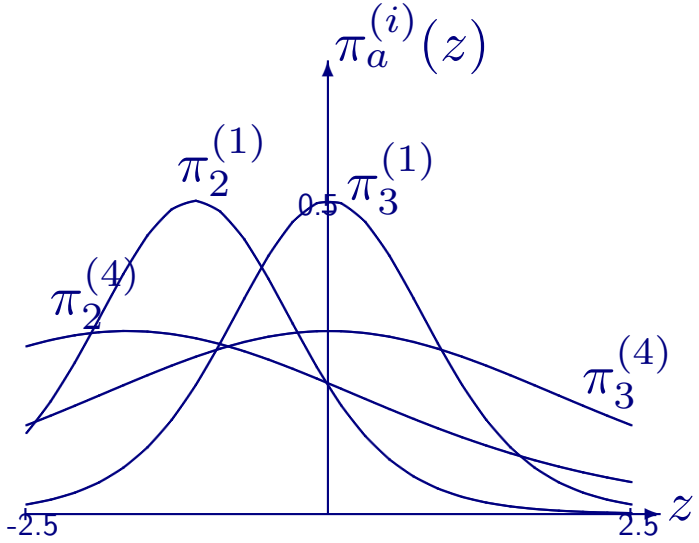


Figure 2: Logit: Four Category Response Functions $\pi_a^{(i)}$

Uncensored data

Survival times are measured in discrete time units. Let γ_t denote the probability that a randomly selected individual will have the event of interest by time t.

We model γ_t as a function of time t, the latent variables $(z_1, \ldots, z_q)$ and a set of covariates $(x_1, \ldots, x_k)$ using a logit, probit or complementary log-log link function.

The logit link function leads to the proportional odds model:

$$\text{logit}\gamma_t(\mathbf{z}, \mathbf{x}) = \alpha^{(t)} + \sum_{j=1}^{q} \alpha_j z_j + \sum_{l=1}^{k} \beta_l x_{il}$$

where $\gamma_t = Pr(y \leq t \mid \mathbf{z}, \mathbf{x})$.

A more commonly used model for survival data is the proportional hazards model. The hazard function is defined as the instantaneous failure probability at time t conditional on survival up to time t and is denoted by λ_t. The discrete-time proportional hazards model is written (Cox 1972):

$$\lambda_t(\mathbf{z}, \mathbf{x}) = \lambda_{t0} \exp\left(\sum_{j=1}^{q} \alpha_j z_j + \sum_{l=1}^{r} \beta_l x_l \right).$$

The function λ_{t0}, known as the baseline hazard function, is an unspecified function of time and is the value of the hazard function at $\mathbf{z} = \mathbf{0}$ and $\mathbf{x} = \mathbf{0}$.

An important property of the proportional hazards model is that the ratio of two hazard functions for two individuals is independent of time.

After a complementary log-log transformation,

$$\log[-\log[1-\gamma_t(\mathbf{z},\mathbf{x})]] = \alpha^{(t)} + \sum_{j=1}^{q} \alpha_j z_j + \sum_{l=1}^{r} \beta_l x_l,$$

where $\alpha^{(t)} = \log[-\log[1-\gamma_{t0}]]$.

Censored data

For each individual, y_m is converted to a sequence of binary responses $\{\delta_{mt}\}, t = 1, 2, \ldots, y_m$, where

$$\delta_{mt} = \begin{cases} 0 & t < y_m \\ 0 & t = y_m \text{ and } y_m < t_m \\ 1 & t = y_m \text{ and } y_m = t_m \end{cases}$$

Here, we model the hazard

$$\lambda_{mt} = Pr(\delta_{mt} = 1 | \delta_{m,t-1} = 0).$$

The logistic link:

$$\text{logit}\lambda_t(\mathbf{z}, \mathbf{x}) = \alpha^{(t)} + \sum_{j=1}^{q} \alpha_j z_j + \sum_{l=1}^{k} \beta_l x_l$$

The conditional density function for a randomly selected individual m is:

$$g(\boldsymbol{\delta}_m \mid \mathbf{z}_m, \mathbf{x}_m) = \prod_{t=1}^{y_m} (1 - \lambda_{mt})^{1-\delta_{mt}} {\lambda_{mt}}^{\delta_{mt}}.$$

$\boldsymbol{\delta}_m$ is the vector of binary responses $(\delta_{m1}, \ldots, \delta_{ihy_m})$ for individual m.

Structural model

$$\mathbf{z}_m = \mathbf{\Lambda}\mathbf{w}_m + \boldsymbol{\epsilon}_m, \qquad m = 1, \ldots, n$$

where $\mathbf{z}_m$ is $q \times 1$ vector, $\mathbf{\Lambda}$ is a $q \times r$ matrix of regression coefficients, $\mathbf{w}$ is a $r \times 1$ vector of covariates and the $\boldsymbol{\epsilon}_m$ is a $q \times 1$ vector of independent standard normal variables.

It follows that the distribution of the latent variables $\mathbf{z}_m$ conditional on the covariates $\mathbf{w}_m$ is normal with mean $\mathbf{\Lambda}\mathbf{w}_m$ and variance one.

Estimation

For a random sample of size n the complete log-likelihood is written as:

$$\begin{aligned} L &= \sum_{m=1}^{n} \log f(\boldsymbol{\delta}_m, \mathbf{u}_m, \mathbf{v}_m, \mathbf{z}_m \mid \mathbf{x}_m, \mathbf{w}_m) \\ &= \sum_{m=1}^{n} [\log g(\boldsymbol{\delta}_m, \mathbf{u}_m, \mathbf{v}_m \mid \mathbf{z}_m, \mathbf{x}_m) + \log h(\mathbf{z}_m \mid \mathbf{w}_m)] \end{aligned}$$

where $\mathbf{y}'_m = (\boldsymbol{\delta}_m, \mathbf{u}_m, \mathbf{v}_m)$ are the survival, nominal and ordinal items respectively. Because $\mathbf{z}$ is unknown the log-likelihood given in equation is maximized using an EM algorithm. The expectation is with respect to the posterior distribution of $\mathbf{z}$ given the observations ($h(\mathbf{z} \mid \boldsymbol{\delta}, \mathbf{u}, \mathbf{v}, \mathbf{x})$).

Measures of Goodness-of-Fit

• Compare the observed (O) and expected (E) frequencies of the 2^p response patterns by means of a X^2 Pearson Goodness-of-fit or a likelihood ratio test G^2.

$$X^2 = \sum_{i=1}^{2^p} \frac{(O_i - E_i)^2}{E_i}$$

$$G^2 = 2 \sum_{i=1}^{2^p} O_i \log \frac{O_i}{E_i}$$

When n is large and p small the above statistics follow a chi-square distribution. As the number of items increases the chi-square approximation to the distribution of either goodness-of-fit statistic ceases to be valid. Parameter estimates are still valid but it is difficult to assess the model.

- Examination of residuals.

Compare the observed and expected frequencies for pair and triplets of responses. If these differences are small it means that the associations between all pairs of responses are well predicted by the model. Check whether pairs or triples of responses occur more or less, often than the model predicts. The above given discrepancy measures can be used to measure discrepancies in the margins. The residuals are not independent and so not a formal test can be applied. However, if we consider the distribution of each residual as a chi-square with 1 degree-of-freedom then a residual with a X^2 or G^2 value greater than 4 will indicate a poor fit.

Diagnostics procedures based on residuals:

- Give reasons for poor fit.

- Suggest ways in which the scales may be improved.

• Tests based on the bivariate margins Christoffersson (1975, Psychometrika), Muthen (1978, Psychometrika), Reiser and VandenBerg (1994, BJMSP), Bartholomew and Leung (2002, BJMSP), Maydeu-Olivares (2003, working paper).

$$H_0 : \boldsymbol{\epsilon} = 0$$

where

$$\epsilon_{ij} = \sqrt{N}(p_{ij} - \pi_{ij}),$$

p_{ij} is the observed frequency and π_{ij} is the expected from the model. All tests are based on

$$W = \boldsymbol{\epsilon}' S^{-1} \boldsymbol{\epsilon}$$

where S is the covariance matrix of $\boldsymbol{\epsilon}$. W is a Wald statistic.

The differences among the tests is related to the way the covariance matrix is estimated.

Application

The data are from the Bangladesh Demographic and Health Survey of 1996. This is a nationally representative survey of ever married women aged 10-49. The sample size was 9127; our analysis sample is a random sample of 800.

The variables used are:

1. Ever use of modern method of contraception (1=No, 0=Yes)[EverUse]

2. Sex preference in ideal number of children (1=Male, 0=Female or no preference=0) [SonPref]

3. Ideal number of children (1,2,3,more than 4) [FamilySize]

4. Time of the 2nd birth within 5 years of the 1st birth [SecondBirth].

Item 3 is treated as ordinal and item 4 is the survival variable that models the hazard of a 2nd birth within 5 years of the 1st birth (duration between the 1st birth and the conception of the 2nd). A woman who had a 2nd birth after 5 years or had a miscarriage is treated as censored. The duration variable is in years, starting from 1.

The covariates included in the model are:

- Age in years
- Woman's educational level (0='None', 1='Primary', 2='Secondary')
- Urban (0='Rural', 1='Urban')

Table 1: ML estimates for the one factor model without covariates

		$\hat{\alpha}_{i0}$	$\hat{\alpha}_{i1}$	$\mathrm{st.}\hat{\alpha}_{\mathrm{i1}}$
Binary items				
EverUse		0.88	-0.67	-0.56
SonPref		-1.78	1.08	0.73
Ordinal item				
FamilySize	No. children			
	3	1.06	-2.94	-0.95
	4+	3.15		
	'Up to God'	5.47		
Survival item				
SecondBirth	Year			
	1	-2.03	0.03	
	2	-1.06		
	3	-1.04		
	4	-1.21		
	5	1.20		

Table 2: Goodness-of-fit statistics for alternative models

	Covariate effects on ...			
Model	Manifest indicators	Latent variable	AIC	BIC
1	Age+Educ+Urban	None	5758.54	5785.53
2	Age+Educ	Urban	5756.26	5781.45
3	Age+Urban	Educ	5778.10	5800.60
4	Educ+Urban	Age	5786.03	5811.22
5	Age	Educ+Urban	5774.85	5794.65
6	Educ	Age+Urban	5780.96	5803.45
7	Urban	Age+Educ	5803.46	5823.25

Table 3: Likelihood ratio tests comparing Model 2 with simpler models

	Covariate effects on ...				
Model	Manifest indicators	Latent variable	-logL	$-2\Delta\log L$	Δdf
2	Age+Educ	Urban	2850.1	-	-
2-Urban	Age+Educ	None	2853.4	6.6	1
2-Educ	Age	Urban	2874.9	49.6	8
2-Age	Educ	Urban	2868.1	36.0	4

Table 4: ML estimates of covariate effects for the selected model

		Education		
	Age	Primary	Secondary	Urban
Manifest indicators				
EverUse	-0.29	0.25	1.35	–
SonPref	0.22	0.22	-0.36	–
FamilySize	-1.02	1.21	2.24	–
SecondBirth	0.21	0.04	-0.30	–
Latent variable	–	–	–	-0.42

Latent class model for categorical responses

Latent class models: Lazarsfeld (1950), Goodman (1974), Haberman (1979), Hagenaars (1990), Vermunt (1997) and Vermunt and Magisdon (2000).

Latent class models in their more classical form have been discussed for clustering binary, nominal or metric variables in Bartholomew and Knott (1999) and for clustering mixed mode data in Everitt (1988), Everitt and Merette (1990), Moustaki (1996) and Moustaki and Papageorgiou (2004).

In latent class models we assume that the factor space consists of K classes. For each class j there is an associated probability, η_j.

$$f(\mathbf{y}_h) = \sum_{j=1}^{K} \eta_j g(\mathbf{y}_h \mid j).$$

- For a binary variable:

$$g(y_i \mid j) = \pi_{ij}{}^{y_i}(1 - \pi_{ij})^{1-y_i}$$

where $\pi_{ij} = P(y_i = 1 \mid j)$.

- For Nominal variables:

The single response probability (π_{ij}) of the binary case is now replaced by a set of functions $\pi_{ij(s)}$ $(s = 1, \ldots, c_i)$ where $\sum_{s=1}^{c_i} \pi_{ij(s)} = 1$.

The distribution assumed is the multinomial:

$$g_i(\mathbf{y}_i \mid j) = \prod_{s=1}^{c_i} (\pi_{ij(s)})^{y_{i(s)}}$$

where $\pi_{ij(s)}$ is the probability that an object who is in class j will belong to category s for variable i.

- For Ordinal variables:

$$\gamma_{ij(s)} = \pi_{ij(1)} + \pi_{ij(2)} + \cdots + \pi_{ij(s)},$$

where $j = 1, \ldots, K$ and $s = 1, \ldots, m_i$.

For a manifest variable y_i the conditional distribution of $y_i \mid j$ is the multinomial:

$$g(y_i \mid j) = \prod_{s=1}^{m_i} \pi_{ij(s)}^{y_{i,(s)}} = \prod_{s=1}^{m_i} (\gamma_{ij(s)} - \gamma_{ij(s-1)})^{y_{i,(s)}}$$

where $y_{i,(s)} = 1$ if a randomly selected object belongs into category s of the ith variable and $y_{i,(s)} = 0$ otherwise.

Finally, the log-likelihood for a random sample of size n is written:

$$L = \sum_{h=1}^{n} \log f(\mathbf{y}_h) = \sum_{h=1}^{n} \log \sum_{j=1}^{K} \eta_j g(\mathbf{y}_h \mid j)$$

In addition, the observed variables have been taken as conditionally independent given each class j.

$$g(\mathbf{y} \mid j) = \prod_{i=1}^{p} g(y_i \mid j), \quad j = 1, \ldots, K.$$

The above log-likelihood can be maximized using an EM algorithm under the constraints $\sum_{j=1}^{k} \eta_j = 1$, $\sum_{s=1}^{c_i} \pi_{ij(s)} = 1$ and $\sum_{s=1}^{m_i} \pi_{ij(s)} = 1$.

An object is more likely to be in class j than class k if

$$h(j \mid \mathbf{y}) > h(k \mid \mathbf{y})$$

where

$$h(j \mid \mathbf{y}_h) = \eta_j g(\mathbf{y}_h \mid j)/f(\mathbf{x}_h).$$

Example: Can Sora Data set

The Can Sora data set comes from a ceramic assemblage found in a cistern at the Punic and Roman site of Ses Paises de Cala d'Hort in Eivissa.

Variables: 15 binary variables, 3 ordinal and 25 metric. The natural logarithms of the metric variables were taken first and they were standardized afterwards.

The metric variables measure the chemical composition of the ceramic. The categorical variables aim to derive information regarding the provenance of the objects (petrological analysis).

The AIC and BIC suggested a 6-class solution.

Table 5: Residuals for the second order margins, 5-class model, Can Sora

Response	Variable i	Variable j	Observed frequency (O)	Expected frequency (E)	$O-E$	$(O-E)^2/E$
(0,0)	8	6	4	1.62	2.37	3.46
(0,1)	15.1	5	0	2.37	-2.37	2.37
	15.2	5	4	1.62	2.37	3.46
(1,0)	5	15.1	0	2.37	-2.37	2.37
	5	15.2	4	1.62	2.37	3.46
(1,1)	15.1	5	4	1.62	2.37	3.46
	15.2	5	0	2.37	-2.37	2.37
	15.2	15.1	0	2.37	-2.37	2.37

Table 6: Residuals for the third order margins, 5-class model, response (1,1,1) to variables (i, j, k), Can Sora

Variable i	Variable j	Variable k	$O - E$	$(O - E)^2/E$
2	5	15.1	2.37	3.47
2	5	15.2	-2.37	2.37
5	6	18.1	0.75	2.25
5	7	9	2.57	4.67
5	7	10	2.57	4.67
5	7	15.1	3.07	10.27
5	9	15.1	3.07	10.27
5	10	15.1	3.07	10.27
5	13	15.1	2.37	3.47
6	8	15.2	2.79	3.55

Table 7: Classification of the 22 objects, six-class model, Can Sora

Group	Objects
Plutonic	CS2, CS3, CS4, CS5, CS6, CS14, CS23
Volcanic	CS10, CS11, CS15, CS16, CS17
Muscovite	CS18, CS19, CS20
Phyllite	CS21, CS22
Pantellerian	CS26, CS27
Outliers	CS7, CS24, CS25

Extensions

• IRT models have come closer to SEM. Still computational problems to be resolved.

• Extensions to allow for covariate effects (Sammel, Ryan, and Legler, 1997 JRSSB, Moustaki 2004, BJMSP)

• Use of IRT models to obtain information about missingness (Knott, Albanese and Galbraith 1990, The Statistician, O'Muircheartaigh and Moustaki 1999, JRSSA, Moustaki and Knott 2000, JRSSA)

• Multilevel IRT models (Fox and Glas 2001, Psychometrika, Skrondal and Esketh, Psychometrika)

• Bayesian estimation methods has allowed more complex generalizations such as to dynamic factor analysis models for longitudinal data (Dunson, 2003 JASA).

• Robust estimation for the GLLVM (Moustaki and Victoria-Feser, Working Paper).

• Search for outliers (unusual response patterns) implementing techniques from regression analysis such as forward search (Atkinson and Riani 2000, Springer).

• Goodness-of-fit measures for models with mixed type data (apply Bayesian model selection methods such as predictive distributions and reversible jump).