Latent variable models

Silvia Cagnone

Master in Statistical Sciences, Department of Statistical Sciences
University of Bologna, a.y. 2019-2020

20 settembre 2021

Context

Latent variable models provide an important tool for the analysis of multivariate data. They offer a conceptual framework within which many disparate methods can be unified. Two reasons why a latent variable should be introduced into a model.

1. Reduce dimensionality. Idea behind factor analysis. (Large-scale statistical enquires, such as social surveys, generate much more information than can be easily absorbed).
2. In social and economical sciences not observable variables (*latent*) for two reasons
   - measurement errors (age, salary...)
   - theoretical constructs not directly observable or not measurable (satisfaction, intelligence, racial prejudice,...).

Objective of the analysis

- Identify the underlying factors that explain the relationships among the observed variables: fitting a specific latent variable model (latent trait model, latent class model)
- Scale individuals on the latent factor space (posterior mean or component score)

CLASSIFICATION OF LATENT VARIABLE METHODS

		Manifest variables	
		Metrical	Categorical
Latent variables	Metrical	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

Theoretical framework

- $x_1, x_2, \ldots, x_p$ manifest variables (or items or indicators)
- $y_1, y_2, \ldots, y_q$ latent variables

Both the manifest and the latent variables are random variables. Hence the relationship between them must be expressed in terms of probability distributions.

R_x, R_y range spaces of $\mathbf{x}$ and $\mathbf{y}$ respectively

1. *Prior distribution* (pmf or pdf) of the latent variables y:

$$h(\mathbf{y}), \quad \mathbf{y} = (y_1, y_2, \ldots, y_q) \in R_y$$

2. *Conditional distribution* (pmf or pdf) of the observed variables $\mathbf{x}$ given the latent variables $\mathbf{y}$

$$g(\mathbf{x}|\mathbf{y}), \quad \mathbf{x} = (x_1, x_2, \ldots, x_p) \in R_x$$

3. *Joint distribution* (pmf of pdf) of the observed and latent variables $\mathbf{x}$ and $\mathbf{y}$

$$f(\mathbf{x}, \mathbf{y}) = h(\mathbf{y})g(\mathbf{x}|\mathbf{y})$$

Theoretical framework

Since only $\mathbf{x}$ can be observed any inference must be based on the *joint marginal distribution* of $\mathbf{x}$

$$f(\mathbf{x}) = \int_{R^y} g(\mathbf{x} \mid \mathbf{y}) h(\mathbf{y}) d\mathbf{y} \tag{1}$$

Aim: obtain information on $\mathbf{y}$ after $\mathbf{x}$ has been observed. From the Bayes' theorem we obtain *the posterior distribution* of the latent variables given the observed variables

$$h(\mathbf{y} \mid \mathbf{x}) = g(\mathbf{x} \mid \mathbf{y}) h(\mathbf{y}) / f(\mathbf{x}) \tag{2}$$

In order to find $h(\mathbf{y} \mid \mathbf{x})$ we need to estimate $f(\mathbf{x})$, $g(\mathbf{x} \mid \mathbf{y})$ and $h(\mathbf{y})$.
$f(\mathbf{x})$ can be estimated from the observed variables, whereas $g(\mathbf{x} \mid \mathbf{y})$ and $h(\mathbf{y})$ are not uniquely determined.

$\Rightarrow$ We must introduce further restrictions on the classes of functions considered

Assumption of conditional independence

If the dependencies among the x's are induced by a set of latent variables $\Rightarrow$ when all y's are accounted for the x's will be independent if all the y's are held fixed.
q must be chosen so that:

$$g(\mathbf{x} \mid \mathbf{y}) = \prod_{i=1}^{p} g_i(x_i \mid \mathbf{y}) \tag{3}$$

$\mathbf{y}$ is sufficient to explain the dependencies among the $\mathbf{x}$ variables. Assumption of *local or conditional independence*.

If we replace equation (3) in equation (1) we obtain:

$$f(\mathbf{x}) = \int_{R^y} \prod_{i=1}^{p} g_i(x_i \mid \mathbf{y}) h(\mathbf{y}) d\mathbf{y} \tag{4}$$

for some small value of q.

The aim of the analysis is to find the smallest q for which model (4) is adequate, that is it has a goodness of fit to the data.

Assumption of General Linear Latent Variable Model (GLLVM)

For the conditional distribution in equation (3) we assume a generalized linear model that consists of three components.

1. *Random component*: for each conditional distribution $g_i(x_i \mid \mathbf{y}), \quad i = 1, \ldots, p$, a convenient family of distributions (with many useful properties) is the *one-parameter exponential family*.

$$g_i(x_i \mid \theta_i) = F_i(x_i)G_i(\theta_i)\exp(\theta_i u_i(x_i)) \quad i = 1, \ldots, p \tag{5}$$

   The exponential family (5) includes the Bernoulli, poisson, normal and gamma distributions as special cases. If we allow x_i and θ to be vector-valued it also includes the multinomial distribution.

2. *Systematic component*: the latent variables $y_1, \ldots, y_q$ produce a linear predictor η_i given by:

$$\eta_i = \alpha_{i0} + \alpha_{i1}y_1 + \ldots + \alpha_{iq}y_q$$

Assumption of General Linear Latent Variable Model (GLLVM)

3 A *link function* between the random and the systematic component that relates θ_i to the latent variables y's.
The simplest form we can assume for θ_i is

$$\theta_i = \alpha_{i0} + \alpha_{i1} y_1 + \ldots + \alpha_{iq} y_q$$

In this case the link function is the identity function.
More in general, any monotonic differentiable function can be assumed as a link between the systematic component and the conditional mean of the random component.

This is called *general(ized) linear latent variable model*. The term *linear* refers to its linearity in the α's.

The main differences with the classical generalized linear model (GLM) are two: a set of dependent variables $x_1, \ldots, x_p$, the covariates $y_1, \ldots y_q$ are unobservable.

Properties of the GLLVM

Replacing the exponential family density (5) to $g_i(x_i \mid \mathbf{y})$ in equation (2) we get:

$$h(\mathbf{y} \mid \mathbf{x}) = \frac{h(\mathbf{y})[\prod_{i=1}^{p} F_i(x_i)G_i(\theta_i)] \exp \sum_{i=1}^{p} \theta_i u_i(x_i)}{f(\mathbf{x})}$$

Collecting all the terms that does not depend on $\mathbf{y}$ we obtain

$$\begin{aligned}
h(\mathbf{y} \mid \mathbf{x}) &\propto h(\mathbf{y}) \left[\prod_{i=1}^{p} G_i(\theta_i)\right] \exp \sum_{i=1}^{p} u_i(x_i)\theta_i && (6)\\
&\propto h(\mathbf{y}) \left[\prod_{i=1}^{p} G_i(\theta_i)\right] \exp \sum_{i=1}^{p} u_i(x_i)(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j) && (7)\\
&\propto h(\mathbf{y}) \left[\prod_{i=1}^{p} G_i(\theta_i)\right] \exp \sum_{j=1}^{q} y_j \sum_{i=1}^{p} \alpha_{ij} u_i(x_i)\\
&\propto h(\mathbf{y}) \left[\prod_{i=1}^{p} G_i(\theta_i)\right] \exp \sum_{j=1}^{q} y_j X_j
\end{aligned}$$

where $X_j = \sum_{i=1}^{p} \alpha_{ij} u_i(x_i) \quad (j = 1, \ldots, q)$.

The Sufficiency Principle

$h(\mathbf{y} \mid \mathbf{x})$ depends on $\mathbf{x}$ only through the q-dimensional vector $\mathbf{X} = (X_1, X_2, \ldots, X_q)$.

$\mathbf{X}$ is a *minimal sufficient statistic* for $\mathbf{y}$. It represents a reduction in dimensionality from the p- vector $\mathbf{x}$ to the q-vector $\mathbf{X}$.

The elements of $\mathbf{X}$ are called *components*. They play a similar role to principal components.

Equation (6) does not depend on the prior distribution of the latent variables. Whatever $h(\mathbf{y})$ is chosen, $\mathbf{X}$ represents the best reduction of information on the posterior distribution of the latent variables.

Example: Bernoulli random variable

x_i Bernoulli random variable, 1=agree, 0=disagree

$$g(x_i \mid \mathbf{y}) = \pi_i^{x_i}(1-\pi_i)^{1-x_i} = (1-\pi_i)\exp(x_i \text{logit}(\pi_i))$$

π_i is the probability of agreeing.

$$\theta_i = \text{logit}(\pi_i) = \ln[\pi_i/(1-\pi_i)] \quad G_i(\theta_i) = 1-\pi_i \quad u_i(x_i) = x_i$$

The GLLVM is

$$\text{logit}(\pi_i) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j$$

and the components are

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i$$

The logit is the *link function*. If $q = 1$ we obtain the logistic latent trait model with *response function*

$$\pi_i(y) = 1/(1+\exp(-\alpha_{i0} - \alpha_{i1} y))$$

Estimation

In this framework we can consider two approaches:

1. Maximum likelihood estimation
2. Bayesian methods

We refer to the first one that we will see in detail for the specific models.

Rotation

If $\mathbf{y}$ is continuous, no latent variable model is unique. There are different models which lead to exactly the same joint distribution of the observed variables.

$$f(\mathbf{x}) = \int_{R^q} \prod_{i=1}^{p} g(x_i \mid \mathbf{y}) h(\mathbf{y}) d\mathbf{y}$$

Any transformation of the latent variable leaves $f(\mathbf{x})$ unchanged. In general, the transformation changes both $h(\cdot)$ and $g(\cdot)$.

Example: Logit model for binary data:

$$\text{logit}(\pi_i(y)) = \alpha_{i0} + \alpha_{i1} y \quad i = 1, 2, \ldots, p$$

We transform $z = H(y)$ where H is the cumulative distribution function of the prior distribution. Then, instead a model with prior $h(y)$ we have one with a uniform prior on (0,1) and response function

$$\text{logit}(\pi_i(z)) = \alpha_{i0} + \alpha_{i1} H^{-1}(z) \quad i = 1, 2, \ldots, p$$

Both models lead to exactly the same distribution of x's.

Rotation

There are transformations that leave the form of the prior distribution and of the natural parameter unchanged $\Rightarrow$ *orthogonal rotation*.

Consider the following GLLVM

$$\boldsymbol{\theta} = \mathbf{A}\mathbf{y}$$

$$\mathbf{z} = \mathbf{M}\mathbf{y}$$

(non-singular transformation)

$$\boldsymbol{\theta} = \mathbf{A}\mathbf{M}^{-1}\mathbf{z}$$

if $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$ then $\mathbf{z} \sim N(\mathbf{0}, \mathbf{M}\mathbf{M}')$. If the transformation is orthogonal $\mathbf{M}\mathbf{M}' = \mathbf{I}$ $\Rightarrow$ $\mathbf{y}$ and $\mathbf{z}$ have the same joint distribution.

Thus we cannot distinguish between a GLLVM with loading matrix $\mathbf{A}$ and one with $\mathbf{A}\mathbf{M}'$.
Rotation has an important role in the interpretation of latent variables.

If $\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Phi})$ then any non-singular transformation (not necessarily orthogonal), would produce $\mathbf{z} \sim N(\mathbf{0}, \mathbf{M}\boldsymbol{\Phi}\mathbf{M}')$ $\Rightarrow$ *oblique rotation*.

Interpretation

Interpretation= Naming the latent variables. Two approaches

- From the tradition of the factor analysis α_{ij} are known as *factor loadings*.
  We refer to the notion of the *simple structure* for the loading matrix because it is easy to interpret.

$$\begin{bmatrix} + & . & . \\ + & . & . \\ + & . & . \\ . & + & . \\ . & + & . \\ . & + & . \\ . & + & . \\ . & . & + \\ . & . & + \\ . & . & + \end{bmatrix}$$

  $+$ large positive loadings
  . small loadings

- Components: for the GLLVM family of models the jth component is

$$X_j = \sum_{i=1}^{p} \alpha_{ij} u_i(x_i)$$

  The set of sufficient statistics $\mathbf{X}$ is not unique.

SUFFICIENCY PRINCIPLE

$$h(\underline{y}/\underline{x}) = \frac{g(\underline{x}/\underline{y})\,h(\underline{y})}{f(\underline{x})}$$

$$g(\underline{x}/\underline{y}) = \prod_{i=1}^{p} g_i(x_i/\underline{y})$$

$$g_i(x_i/\underline{y}) = F_i(x_i)\,G_i(\theta_i)\exp(\theta_i\, u_i(x_i))$$

$$g(\underline{x}/\underline{y}) = \underbrace{\left[\prod_{i=1}^{p} F_i(x_i)\right]}_{A}\left[\prod_{i=1}^{p} G_i(\theta_i)\exp(\theta_i\, u_i(x_i))\right]$$

$$h(\underline{y}/\underline{x}) = \frac{A\cdot h(\underline{y})}{f(\underline{x})}$$

$$\left(\alpha_{i0} + \sum_{j=1}^{q}\alpha_{ij}y_j\right)$$

$$= \frac{h(\underline{y})\left[\prod_{i=1}^{p} F_i(x_i)\right]\left[\prod_{i=1}^{p} G_i(\theta_i(\underline{y}))\right]\exp\left(\sum_{i=1}^{p}\theta_i\, u_i(x_i)\right)}{f(\underline{x})}$$

$$h(\underline{y}/\underline{x}) \propto h(\underline{y})\left[\prod_{i=1}^{p} G_i(\theta_i(\underline{y}))\right]\exp\left[\sum_{i=1}^{p} u_i(x_i)\left(\alpha_{i0}+\sum_j \alpha_{ij}y_j\right)\right]$$

$$\propto h(\underline{y})[\ldots]\exp\left[\sum_i u_i(x_i)\alpha_{i0} + \sum_i u_i(x_i)\sum_j \alpha_{ij}y_j\right]$$

$$\propto h(\underline{y})[\ldots]\exp\left[\sum_i u_i(x_i)\sum_j \alpha_{ij}y_j\right]$$

$$\propto h(\underline{y})[\ldots]\exp\left[\sum_i\sum_j y_j\, u_i(x_i)\alpha_{ij}\right]$$

$$\propto h(\underline{y})[\ldots]\exp\left[\sum_j y_j \sum_i u_i(x_i)\alpha_{ij}\right]$$

$$\propto h(\underline{y})[\ldots]\exp\left[\sum_j y_j \sum_i \alpha_{ij}\, u_i(x_i)\right]$$

$$\underset{p\times p}{\Sigma} = \underset{p\times q}{\Lambda}\ \underset{q\times p}{\Lambda'} + \underset{p\times p}{\Psi}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_2^2 & & \\ & & \ddots & \\ \sigma_{p1} & \cdots & & \sigma_p^2 \end{bmatrix} = \begin{bmatrix} \sum_j \lambda_{1j}^2 & \sum_j \lambda_{1j}\lambda_{2j} & \cdots & \sum_j \lambda_{1j}\lambda_{pj} \\ & \sum_j \lambda_{2j}^2 & & \\ & & \ddots & \\ \sum_j \lambda_{pj}\lambda_{1j} & \cdots & & \sum_j \lambda_{pj}^2 \end{bmatrix} + \begin{bmatrix} \psi_{11} & 0 & \cdots & 0 \\ & \psi_{22} & & \\ & & \ddots & \\ 0 & & & \psi_{pp} \end{bmatrix}$$

$$\sigma_i^2 = \mathrm{var}(x_i) = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_{ii}$$

$\sum_{j=1}^{q} \lambda_{ij}^2$ communality

ψ_{ii} specific variance

$$\Lambda\Lambda' = \begin{bmatrix} \sum_j \lambda_{1j}^2 & \sum_j \lambda_{1j}\lambda_{2j} & \cdots & \sum_j \lambda_{1j}\lambda_{pj} \\ & \sum_j \lambda_{2j}^2 & & \\ & & \ddots & \\ \sum_j \lambda_{pj}\lambda_{1j} & \cdots & & \sum_j \lambda_{pj}^2 \end{bmatrix}$$

$\mathrm{cov}(x, y) =$ $\qquad E(X) = E(E(X/y))$

$= E[(x-\mu)\, y'] =$

$= E[E\{(x-\mu)/y\}\, y'] =$

$= E(E\{(\Lambda y + \varepsilon)/y\}\, y') =$

$= E(\Lambda y\, y') = \Lambda E(yy') = \Lambda$

$\{\mathrm{diag}\,\Sigma\}^{-\frac{1}{2}}\Lambda$ $\qquad X^* = \Lambda'\Psi_x^{-1}$

$\mathrm{cov}(X^*, y) =$

$$\text{cov}(X, y) =$$

$$= E\left[(X - E(X))\, y' \right] =$$

$$= E\left[(\Lambda' \Psi^{-1} x - \Lambda' \Psi^{-1} E(x))\, y' \right] =$$

$$= \Lambda' \Psi^{-1} E\left[(x - E(x))\, y' \right] =$$

$$= \Lambda' \Psi^{-1} E\left[(x-\mu)\, y' \right] = \Lambda' \Psi^{-1} \Lambda$$

$$\nu = My \qquad y = M^{-1}\nu = M'\nu \qquad M'M = I \qquad M^{-1} = M'$$

$$\nu \sim N(0, I)$$

$$x \mid \nu \sim N(\mu + \underbrace{\Lambda M'}_{\Lambda^*}\nu, \Psi)$$

$$\Lambda^* = \Lambda M'$$

$$\Sigma = \Lambda^* \Lambda^{*\prime} + \Psi = \Lambda \overbrace{M' M}^{I} \Lambda' + \Psi =$$

$$= \Lambda\Lambda' + \Psi$$

$$\sum_{h=1}^{n} \underset{1\times p}{(x_h - \mu)'} \underset{p\times p}{\Sigma^{-1}} \underset{p\times 1}{(x_n - \mu)} =$$

$$\text{trace}(a) = a$$

$$\text{trace}(ABC) = \text{trace}(CAB) = \text{trace}(BCA)$$

$$= \text{trace}\left[\sum_{h=1}^{n} \underbrace{(x_n-\mu)'}_{A} \underbrace{\Sigma^{-1}}_{B} \underbrace{(x_n-\mu)}_{C} \right] \quad 1\times 1$$

$$= \text{trace}\left[\Sigma^{-1} \underbrace{\sum_{h=1}^{n} (x_h-\mu)(x_n-\mu)'}_{A} \right] =$$

B

$$= \mathrm{tr}\left[\Sigma^{-1} S' \right] \quad S'$$

3.36pt

The normal linear factor model

The normal linear factor model (NLFM) is the oldest and most widely used latent variable model (Spearman, 1904).

$\mathbf{x}, \mathbf{y}$: continuous. $\mathbf{x}$ are correlated. Here we observe that

$$f(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

We ask whether the multivariate normal distribution admits a representation in terms of $h(\mathbf{y})$ and $g(\mathbf{x}|\mathbf{y})$.

We start considering the linear factor model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \mathbf{e} \tag{1}$$

We assume:

- $\mathbf{y} \sim N_q(\mathbf{0}, \mathbf{I})$ and $\mathbf{e} \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$
- $\mathbf{y}$ and $\mathbf{e}$ independent

$\boldsymbol{\Lambda}$ is the $p \times q$ matrix of the *factor loadings*

$\boldsymbol{\Psi}$ is the $p \times p$ diagonal matrix of *specific variances*.

The normal linear factor model

An equivalent way of writing the model is

$$\mathbf{x}|\mathbf{y} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y}, \boldsymbol{\Psi}) \tag{2}$$

and

$$\mathbf{y} \sim N_q(\mathbf{0}, \mathbf{I}) \tag{3}$$

Whether we write the model in terms of equation in random variables as in (1), or as probability distributions (2) and (3) is equivalent, it is a matter of taste.

The second representation is the general one given in terms of prior distribution $h(\mathbf{y})$ and conditional distribution $g(\mathbf{x}|\mathbf{y})$ that is normal and hence belongs to the exponential family.

Indeed the marginal conditional distribution of x_i is given by

$$g_i(x_i|\mathbf{y}) = \frac{1}{\sqrt{2\pi\psi_{ii}}} \exp\left[-\frac{1}{2\psi_{ii}}(x_i - \mu_i - \sum_j \lambda_{ij} y_j)^2\right] \tag{4}$$

and the linear predictor is

$$\eta_i = \mu_i + \sum_j \lambda_{ij} y_j \quad i = 1, \ldots, p \tag{5}$$

The normal linear factor model

Replacing (5) in (4) we obtain

$$g_i(x_i|\mathbf{y}) = \frac{1}{\sqrt{2\pi\psi_{ii}}} \exp\left[-\frac{1}{2}\left(\frac{x_i^2}{\psi_{ii}} - \frac{2x_i\eta_i}{\psi_{ii}} + \frac{\eta_i^2}{\psi_{ii}}\right)\right] \tag{6}$$

If Ψ_{ii} is known this may be written in the GLLVM form by setting:

$\theta = \eta_i/\sqrt{\psi_{ii}}$, $\quad u_i(x_i) = x_i/\sqrt{\psi_{ii}}$, $\quad G_i(\theta_i) = \exp\left(-\frac{1}{2}\eta_i^2/\psi_{ii}\right)$, $\quad F_i(x_i) = \frac{1}{\sqrt{2\pi\psi_{ii}}} \exp\left(-\frac{1}{2}x_i^2/\psi_{ii}\right)$

The sufficient statistics are

$$X_j = \sum_i \lambda_{ij} x_i/\sqrt{\psi_{ii}} \quad j = 1, \ldots, q$$

The normal linear factor model

From (2) and (3) it follows that

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.$$

The posterior distribution is given by

$$\mathbf{y}|\mathbf{x} \sim N_q(\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1})$$

The normal linear factor model: properties

From (1) it follows

- Variance of each x_i

$$var(x_i) = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i$$

  $\sum_{j=1}^{q} \lambda_{ij}^2$ are called *communality*, ψ_i the variance specific to x_i.

- Covariance between $\mathbf{x}$ and $\mathbf{y}$

$$E(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}' = E[E\{(\mathbf{x} - \boldsymbol{\mu})|\mathbf{y}\}\mathbf{y}'] = E(\boldsymbol{\Lambda}\mathbf{y}\mathbf{y}') = \boldsymbol{\Lambda}$$

  The factor loadings can be interpreted as covariances between the manifest variables and factors. The correlations are given by

$$\{diag\boldsymbol{\Sigma}\}^{-1/2}\boldsymbol{\Lambda}$$

- Covariance between the components and the factors

$$E(\mathbf{X} - E(\mathbf{X}))\mathbf{y}' = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}E[(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}'] = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$$

  If $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal there are no cross-correlations between components and factors.

Constraints on the model

Since $\mathbf{y}$ is continuous any one-to-one transformation from $\mathbf{y}$ to $\mathbf{v}$ has no effect on $f(\mathbf{x})$. In particular, both the functions h and g will be changed but $f(\mathbf{x})$ does not change.

The indeterminacy of h leave us free to adopt a metric for $\mathbf{y}$ such that h has a convenient form. The more convenient is the standard normal distribution.

If $\mathbf{x}$ is normal the indeterminacy refers to *rotation*.

Consider the orthogonal transformation $\mathbf{v} = \mathbf{M}\mathbf{y}(\mathbf{M}'\mathbf{M} = \mathbf{I})$. It follows that

$$\mathbf{v} \sim N(\mathbf{0}, \mathbf{I})$$

and

$$\mathbf{x}|\mathbf{v} \sim N(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{M}'\mathbf{v}, \boldsymbol{\Psi})$$

This model is undistinguishable from the first one.

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\mathbf{M}'\mathbf{M}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

Thus we cannot distinguish between a model with loading matrix $\boldsymbol{\Lambda}$ and one with $\boldsymbol{\Lambda}\mathbf{M}'$. Rotation has an important role in the interpretation of latent variables.

Constraints on the model

To remove indeterminacies in the model we can place constraints on the parameters of the model

- If $\mathbf{\Gamma} = \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal the y's will be independent a posteriori and the relation between the y's and the components is particularly simple. This constraint removes the freedom to arbitrarily rotate $\mathbf{\Lambda}$.
- In confirmatory factor analysis, constrain some values of $\mathbf{\Lambda}$ to zero. This usually remove the rotation freedom
- Standardized x's. It removes the arbitrariness of the scale of the manifest variables. Any change of scale in x will be reflected in the covariances and hence in the parameter estimates and their interpretation. Standardizing the x's has the effect to make all the diagonal elements of $\mathbf{\Sigma}$ equal to one and hence

$$\sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i = 1$$

Estimation methods

Different estimation methods can be applied according to the different assumptions on the observed data. All of them aims at finding parameter estimates such that the discrepancy between the observed covariance matrix $\mathbf{S}$ and the theoretical matrix impled by the model $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is as smaller as possible.

The differences among the estimation methods depends on the different discrepancy measures between the observed and the theoretical covariance matrix.

Here we see the *maximum likelihood method* and *maximum likelihood method by the EM algorithm*. Both are based on the normality assumption of the observed variables.

Other estimation methods in absence of normality of the observed variables are *least squares methods* and the *principal axis method* (see Bartholomew, Knott, Moustaki, 2011).

Maximum likelihood estimation

If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and given a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the likelihood can be written as follows

$$L(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\}^{-n} \exp\{-1/2 \sum_{h=1}^{n} (\mathbf{x}_h - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_h - \boldsymbol{\mu})\}$$

$$\begin{aligned} l = \ln L &= -n/2 \ln\{(2\pi)^p |\boldsymbol{\Sigma}|^{1/2}\} - 1/2 \sum_{h=1}^{n} (\mathbf{x}_h - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_h - \boldsymbol{\mu})\} = \\ &= \text{constant} + n/2[\ln |\boldsymbol{\Sigma}^{-1}| - \text{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]] \end{aligned}$$

where $S = \sum_{h=1}^{n} (\mathbf{x}_h - \boldsymbol{\mu})(\mathbf{x}_h - \boldsymbol{\mu})'/n$.

Maximum likelihood estimation

1. First step: compute the maximum likelihood estimate of $\boldsymbol{\mu}$: $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and replace the estimate in $\mathbf{S}$.

2. Second step: maximize $l = \ln L$ with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$.

$$\ln L(\mathbf{x}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = +\frac{n}{2} \ln |(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}| - \frac{n}{2} \text{trace}[(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}\mathbf{S}]$$

$$\frac{\partial \ln L(\mathbf{x}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = 0 \quad \frac{\partial \ln L(\mathbf{x}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Lambda}} = 0$$

We obtain the maximum likelihood estimates of $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$:

$$\mathbf{S}\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Lambda}}(\mathbf{I} + \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}) \quad \hat{\boldsymbol{\Psi}} = diag(\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}')$$

Since analytical solutions do not exist for the above equations, iterative numerical procedures have to be applied

Maximum likelihood estimation by the E-M algorithm

The Maximum likelihood method by the E-M algorithm is an iterative technique well suited to maximum likelihood estimation for models where there is missing information (in this the latent variables).

It consists of two steps:

- E step: it consists in computing the expected value of the joint log-likelihood of $(\mathbf{x}_h, \mathbf{y}_h), \quad h = 1, \ldots, n$ conditional on the $\mathbf{x}_h$
- M step: it consists in maximizing the modified log-likelihood with respect to the parameters of the model

The two steps are repeated iteratively until convergence.

In practice, it is easier to set the conditional expected value of the score function from the joint likelihood of $(\mathbf{x}_h, \mathbf{y}_h)$ given $\mathbf{x}_h$ equal to zero.

Maximum likelihood estimation by the E-M algorithm

The complete log-likelihood of the $(\mathbf{x}_h, \mathbf{y}_h), \quad h = 1, \ldots, n$ is

$$\begin{aligned} l_c &= \text{constant} - n/2\{\ln|\boldsymbol{\Psi}| + \text{trace}\boldsymbol{\Psi}^{-1}[1/n \sum_h (\mathbf{x}_h - \boldsymbol{\mu} - \boldsymbol{\Lambda} y_h)(x_h - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{y}_h)'] \\ &+ \text{trace}(1/n) \sum_h \mathbf{y}_h \mathbf{y}_h'\} \end{aligned}$$

which gives the conditioned expected score functions

- for $\boldsymbol{\mu}$

$$E(Sc_{\boldsymbol{\mu}}) = E\left[\frac{\partial l}{\partial \boldsymbol{\mu}} | \mathbf{x}_h\right] = nE[\boldsymbol{\Psi}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu} - \boldsymbol{\Lambda}\bar{\mathbf{y}}) | \mathbf{x}_h] \tag{7}$$

- for $\boldsymbol{\Lambda}$

$$E(Sc_{\boldsymbol{\Lambda}}) = E\left[\frac{\partial l}{\partial \boldsymbol{\Lambda}} | \mathbf{x}_h\right] = nE[\boldsymbol{\Psi}^{-1}(\mathbf{S}'_{xy} - \boldsymbol{\mu}\bar{\mathbf{y}} - \boldsymbol{\Lambda}\mathbf{S}'_{yy}) | \mathbf{x}_h] \tag{8}$$

- for $\boldsymbol{\Psi}$

$$\begin{aligned} E(Sc_{\boldsymbol{\Psi}}) &= E\left[\frac{\partial l}{\partial \boldsymbol{\Psi}} | \mathbf{x}_h\right] = -n/2E[\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}(\mathbf{S}'_{xx} - \boldsymbol{\mu}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\boldsymbol{\mu}' - \mathbf{S}'_{xy}\boldsymbol{\Lambda}' - \\ &- \boldsymbol{\Lambda}\mathbf{S}'_{yx} + \boldsymbol{\mu}\bar{\mathbf{y}}'\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\bar{\mathbf{y}}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Lambda}\mathbf{S}_{yy}\boldsymbol{\Lambda}')\boldsymbol{\Psi}^{-1} | \mathbf{x}_h] \end{aligned} \tag{9}$$

Maximum likelihood estimation by the E-M algorithm

The only quantities that depend on the random variables y's are the sufficient statistics:

$$\bar{\mathbf{y}} = 1/n \sum_h y_h \quad \mathbf{S}'_{xy} = 1/n \sum_h \mathbf{x}_h \mathbf{y}'_h \quad \mathbf{S}'_{yy} = 1/n \sum_h \mathbf{y}_h \mathbf{y}'_h$$

Thus, it is enough to compute the conditional expected values only for the sufficient statistics.

- $$E[\bar{\mathbf{y}}|\mathbf{x}_h] = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = \hat{\bar{\mathbf{y}}} \tag{10}$$
- $$\begin{aligned} E[\mathbf{S}'_{xy}|\mathbf{x}_h] &= 1/n \sum_h \mathbf{x}_h E[\mathbf{y}'_h|\mathbf{x}_h] = 1/n \sum_h \mathbf{x}_h (\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_h - \boldsymbol{\mu}))' = \\ &= (\mathbf{S}_{xx} - \mathbf{x}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \hat{\mathbf{S}}'_{xy} \end{aligned} \tag{11}$$
- $$\begin{aligned} E[\mathbf{S}'_{yy}|\mathbf{x}_h] &= (\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{S}'_{xx} - \boldsymbol{\mu}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}'))\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \\ &= \hat{\mathbf{S}}'_{yy} \end{aligned} \tag{12}$$

Maximum likelihood estimation by the E-M algorithm

Replacing $\hat{\bar{\mathbf{y}}}$, $\hat{\mathbf{S}}'_{xy}$ and $\hat{\mathbf{S}}'_{yy}$ in formulas (7), (8) and (9) and setting the conditioned score functions to zero $E(S_{\boldsymbol{\mu}}) = 0$, $E(S_{\boldsymbol{\Lambda}}) = 0$ and $E(S_{\boldsymbol{\Psi}})$ we obtain

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \boldsymbol{\Lambda}\hat{\bar{\mathbf{y}}} \tag{13}$$

$$\hat{\boldsymbol{\Lambda}} = (\hat{\mathbf{S}}'_{xy} - \bar{\mathbf{x}}\hat{\bar{\mathbf{y}}}')(\hat{\mathbf{S}}'_{yy} - \hat{\bar{\mathbf{y}}}\hat{\bar{\mathbf{y}}}')^{-1} \tag{14}$$

$$\begin{aligned} \hat{\boldsymbol{\Psi}} \quad &= \quad diag(\mathbf{S}'_{xx} - \hat{\boldsymbol{\mu}}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\hat{\boldsymbol{\mu}}' - \hat{\mathbf{S}}'_{xy}\hat{\boldsymbol{\Lambda}}' - \\ &- \quad \hat{\boldsymbol{\Lambda}}\hat{\mathbf{S}}'_{yx} + \hat{\boldsymbol{\mu}}\hat{\mathbf{y}}'\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' + \hat{\boldsymbol{\Lambda}}\hat{\mathbf{S}}'_{yy}\hat{\boldsymbol{\Lambda}}') \end{aligned} \tag{15}$$

The E-M algorithm consists of the following steps:

1. Choose starting values for $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$
2. Compute the conditioned expected score function (7), (8), (9)
3. Compute the maximum likelihood estimates (13), (14), (15)
4. Iterate steps 2 and 3 until convergence

Goodness of fit and choice of q

If q is specified a priori, we can use the likelihood ratio statistic:

$$H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$$

against

$$H_1 : \mathbf{\Sigma} \text{ is unconstrained}$$

The likelihood ratio statistic is

$$\begin{aligned} -2[l(H_0) - l(H_1)] &= n[\log|\hat{\mathbf{\Sigma}}| + \text{trace}\hat{\mathbf{\Sigma}}^{-1}\mathbf{S} - \log|\mathbf{S}| - p] = \\ &= n[\text{trace}\hat{\mathbf{\Sigma}}^{-1}\mathbf{S} - \log|\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}| - p] \end{aligned}$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$. If $\mathbf{\Psi} > 0$

$$-2[l(H_0) - l(H_1)] \approx \chi^2$$

with

$$df = 1/2p(p+1) - [pq + p - 1/2q(q-1)] = 1/2[(p-q)^2 - (p+q)]$$

Comparing models: model selection criteria

Akaike's information criterion for model selection.

$$AIC = -2l + 2\nu$$

Bayesian information criterion (BIC)

$$BIC = -2l + \nu \ln n$$

where ν is the number of free parameters in the model, l is the log-likelihood function.

Identifiability

A necessary condition for identifiability of the model and consistency of the parameter estimators is that there are at least as many sample statistics as the parameters.

The number of parameters in a q-factor model is $pq + p$. It can be greater than $p(p+1)/2$ that is the number of variances and covariances of the sample covariance matrix $\mathbf{S}$.

If $\mathbf{\Lambda}\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal we introduce $1/2q(q-1)$ constraints $\Rightarrow$ the number of free parameters are $pq + p - 1/2q(q-1)$.

A necessary but not sufficient condition for consistent estimation is

$$1/2p(p+1) - pq - p + 1/2q(q-1) = 1/2[(p-q)^2 - (p+q)] \geq 0$$

This condition implies that there is an upper bound to the number of factors which can be fitted that is

$$q \leq [2p + 1 - (8p+1)^{1/2}]$$

Scale-Invariant Estimation

The factor model can be fitted using the sample correlation matrix rather than the covariance matrix. Using correlations standardized variables x's.

Changing the scale of the x's has no effect on the analysis.

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \mathbf{e}$$

We can transform $\mathbf{x}^* = \mathbf{C}\mathbf{x}$ where $\mathbf{C}$ is diagonal with positive elements. We obtain

$$\mathbf{x}^* = \mathbf{C}\boldsymbol{\mu} + \mathbf{C}\boldsymbol{\Lambda}\mathbf{y} + \mathbf{C}\mathbf{e}$$

and

$$var(\mathbf{x}^*) = \mathbf{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{C} + \mathbf{C}\boldsymbol{\Psi}\mathbf{C}$$

If $\mathbf{C} = \text{diag}(\boldsymbol{\Sigma})^{-1/2}$ we obtain

$$\mathbf{x}^* = \boldsymbol{\mu}^* + \boldsymbol{\Lambda}^*\mathbf{y} + \mathbf{e}^*$$

where $var(\mathbf{x}^*) = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}'^* + \boldsymbol{\Psi}^*$ and $\boldsymbol{\Lambda}^* = \mathbf{C}\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}^* = \mathbf{C}\boldsymbol{\Psi}\mathbf{C}$

Heywood case

The parameter space is restricted by the condition $\mathbf{\Psi} \geq 0$. However in some cases we can have a negative or zero ψ. The minimum we seek will lie on a boundary of the admissible region (one or more of the ψ's is zero) $\Rightarrow$ Heywood case.

Possible causes:

- Result of sampling error. A key factor is the sample size. It can occur with small samples. For a given sample size th erisk deceases as p increases.
- High correlations between variables.
- Attempt to extract more factors than are present.

Possible remedies:

- Choose a large sample with a good number of variables.
- Avoid to introduce new variables which little to those already there. This will create high correlations without contributing significantly to the information about latent variables.
- Avoid over-factoring using adequate criteria
- Consider a Bayesian approach by using a prior distribution for the ψ's which assigns zero probability to negative values.
- Stop the iteration at some arbitrary small value of ψ_i such as 0.05 or 0.01.

Rotation and related matters

As illustrated above, one of the property of the factor model is that it is invariant to orthogonal rotation of the factor loadings. It means that we can get infinite solutions all equivalent.

In practice, factor rotation consists in finding a matrix $\mathbf{M}$ such that the factor loadings $\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{M}'$ can be more easily interpreted than the original factor loadings $\mathbf{\Lambda}$.

Is it a good idea to look for such rotations?

- Cons: One can keep rotating the factors until one finds an interpretation that one likes.
- Pros: Factor rotation does not change the overall structure of a solution. It only changes how the solution is described, and finds the simplest description.

Rotation and related matters

What do we look for?

Factor loadings can often be easily interpreted if:

- Each variable is highly loaded on at most one factor
- All factor loadings are either large and positive, or close to zero.

We look for a *simple structure*.

Example: unrotated and rotated factor loadings.

	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}^*_{i1}$	$\hat{\lambda}^*_{i2}$
x_1	0.2	0.3	0.0	0.4
x_2	0.4	0.5	0.0	0.6
x_3	0.6	0.7	0.0	0.9
x_4	0.7	0.7	0.0	1.0
x_5	0.5	-0.5	0.7	0.0
x_6	0.7	-0.6	0.9	0.0
x_7	0.3	-0.2	0.4	0.0

Rotation and related matters

Two types of rotations

1. *Orthogonal rotation*: the factors are restricted to be uncorrelated.
2. *Oblique rotation*: the factors may be correlated. When the factors are uncorrelated the standardized factor loadings could be interpreted as the correlation coefficients of the manifest variables and the factors. Under oblique rotation this is no longer true. Indeed:

$$E(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}' = \boldsymbol{\Lambda}\boldsymbol{\Phi}$$

where $\boldsymbol{\Phi}$ is the correlation matrix of the y's. The matrix of correlation is called *structured loading matrix* to distinguish it from the *pattern loading matrix* $\boldsymbol{\Lambda}$.

- Advantage of orthogonal rotation: For orthogonal rotation (based on standardized variables), the factor loadings represent correlations between factors and observed variables. This is not the case for oblique rotations.
- Advantage of oblique rotation: May be unrealistic to assume that factors are uncorrelated. One may obtain a better fit by dropping this assumption.

Types of rotation

- Orthogonal:
  - Varimax: aims at factors with a few large loadings, and many near-zero loadings.
  - Quartimax: aims to evidence the correspondence between observed and latent factors
- Oblique:
  - Promax: aims at simple structure with low correlation between factors.
  - Oblimin: tends to produce varimax-looking factors, but which are oblique

Factor scores

Once we found the estimation of the factor loadings and specific variances we can estimate the factor scores for each individual. Different estimation methods exist.

- *Bartlett's method*: it is based on the maximum likelihood estimation hence requires the normality distribution of the observed variables. For the hth individual they are given by:

$$\hat{\mathbf{f}}_h = (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}})\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-1}\mathbf{x}_h$$

- *Thompson's method*: the estimates of the factor scores are obtained in a Bayesian framework

$$\hat{\mathbf{f}}_h = (\mathbf{I} + \hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}})^{-1}\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-1}\mathbf{x}_h$$

Both methods have advantages and disadvantages, no clear favorite.

Example

In the Table the correlation matrix between marks obtained in 6 subjects by a sample of 220 students is reported (Lawley e Maxwell, 1971).

	Gaelic	English	History	Arithmetic	Algebra	Geometry
Gaelic	1.00					
English	0.44	1.00				
History	0.41	0.35	1.00			
Arithmetic	0.29	0.35	0.16	1.00		
Algebra	0.33	0.32	0.19	0.59	1.00	
Geometry	0.25	0.33	0.18	0.47	0.46	1.00

We can observe the correlations are all positive: students are good or not in all the subjects. The scientific subjects present high correlation and the humanistic subjects as well. It seems there are two groups.

Example

We decided to estimate a 2 factor model with the MLE method. In the Table the estimates of the factor loadings are reported.

Materia	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$
Gaelic	0.56	0.43
English	0.57	0.29
History	0.39	0.45
Arithmetic	0.74	-0.28
Algebra	0.72	-0.21
Geometry	0.60	-0.13

Example

Interpretation:

The factor loadings can be interpreted as correlation between marks and latent variables. For example, the correlation between Gaelic and the first factor is 0.56.

If we want to interpret the latent factors we have to look at the magnitude of the loadings.

The loadings related to the first factor are all positive and quite high. Thus the first factor represents a general performance of the students.

The second factor is a contrast between the humanistic and the scientific subjects.

Communalities: the communality of a manifest variable represents the percentage of the variability explained by the common factor.

Example

Interpretation

Subjects	Communality
Gaelic	0.49
English	0.41
History	0.36
Arithmetic	0.62
Algebra	0.56
Geometry	0.37

For example, the 49% of the variance of the mark in Gaelic is explained by the 2 factors. This communality is given by $0.49 = 0.56^2 + 0.43^2$ where 0.56 and 0.43 are the factor loadings associated to the 2 factors. If a manifest variable presents a high communality, it will be a pure indicator of the common factor. The sum of the communality represents the variance explained by the model. In this case this value is 2.18 that is the 47% of 6 (total variance)

Example

Goodness of fit and choice of the number of factors q.

Reproduced correlation matrix One way to test the goodness of the estimated model is to compare the estimated (reproduced)and the observed correlation matrix of $\mathbf{x}$. The elements of the reproduced matrix (for example between Gaelic and English) are computed as $corr(x_2, x_1) = \hat{\lambda}_{21}\hat{\lambda}_{11} + \hat{\lambda}_{22}\hat{\lambda}_{12} = (0.57 \times 0.56) + (0.29 \times 0.43) = 0.44$

Correlation	Gaelic	English	History	Arithmetic	Algebra	Geometry
Gaelic	0.49	0.44	0.41	0.29	0.31	0.28
English	0.44	0.41	0.35	0.34	0.35	0.30
History	0.41	0.35	0.36	0.16	0.19	0.17
Arithmetic	0.29	0.34	0.16	0.62	0.59	0.48
Algebra	0.31	0.35	0.19	0.59	0.56	0.46
Geometry	0.28	0.30	0.17	0.48	0.46	0.37

Discrepancy	Gaelic	English	History	Arithmetic	Algebra	Geometry
Gaelic		0.00	0.00	0.00	0.02	- 0.03
English	0.00		0.00	0.01	- 0.03	0.03
History	0.00	0.00		0.00	0.00	0.00
Arithmetic	0.00	0.01	0.00		0.00	0.00
Algebra	0.00	-0.03	0.00	0.00		0.00
Geometry	-0.03	0.03	0.00	0.00	0.00	0.00

The 2 factor model seems to be a good model

Example

Test of goodness of fit

Under the normality assumption of the data we test the hypothesis

$$H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$$

that is the covariance/correlation matrix has the specified form.
We use the test W that is distributed as a chi square with $[(p-q)^2 - (p+q)]/2$ degrees of freedom.

In the example W is 2.18 with 4 degrees of freedom (p-value=0.70), that suggests that the fit of the model to the data is good

Factor analysis: example

Solutions under different rotations

Sublects								
	None		Varimax		Quartimax		Oblimin	
Gaelic	0.56	0.43	0.23	**0.66**	0.32	**0.62**	0.04	**0.68**
English	0.57	0.29	0.32	**0.55**	0.39	**0.50**	0.17	**0.53**
History	0.39	0.45	0.08	**0.59**	0.16	**0.57**	-0.11	**0.65**
Arithmetic	0.74	-0.28	**0.77**	0.17	**0.79**	0.06	**0.81**	-0.04
Algebra	0.72	-0.21	**0.72**	0.22	**0.74**	0.12	**0.73**	0.03
Geometry	0.60	-0.13	**0.57**	0.22	**0.56**	0.13	**0.57**	0.07

With Oblimin the correlation between the factors is 0.52.

EM algorithm

- The EM algorithm (Dempster et al., 1977) is one of the most used algorithm in statistics for maximum likelihood estimation in complex problems. In particular, it is used in the cases in which the model depends on unobserved latent variables.

- Iterative algorithm composed by two steps: E-step and M-step.

  - E step: Compute the conditional expected value to the observed data given the current values of the parameter estimates.
  - M step: the conditioned expected value is maximized with respect to the parameters of interest.

Motivating example

Have two coins: Coin 1 and Coin 2

Each has it's own probability of seeing 'H' on any one flip. Let

$p_1 = P(H \ on \ Coin \ 1)$
$p_2 = P(H \ on \ Coin \ 2)$

Select a coin at random and flip that one coin m times.

Now have data

$$\begin{array}{ll} X_{11}X_{12}\ldots X_{1m} & Y_1 \\ X_{21}X_{22}\ldots X_{2m} & Y_2 \\ X_{31}X_{32}\ldots X_{3m} & Y_3 \\ \vdots & \\ X_{n1}X_{n2}\ldots X_{nm} & Y_n \end{array}$$

Here, the X_{ij} are Bernoulli random variables taking values in $\{0, 1\}$ where

$$X_{ij} = \begin{cases} 1, & \text{if the } j\text{th flip for the } i\text{th coin chosen is H;} \\ 0, & \text{if the } j\text{th flip for the } i\text{th coin chosen is T.} \end{cases}$$

and the Y_i live in $\{1, 2\}$ and indicate which coin was used on the nth trial.

Motivating example

Note that all the X's are independent and in particular

$$X_{i1}X_{i2}\dots X_{im}|Y_i = j \sim Ber(p_j)$$

We can write out the joint pdf of all $nm + n$ random variables and formally come up with MLE's for p_1 and p_2. They will turn out as expected

$$\hat{p}_1 = \frac{total\ \sharp\ of\ times\ Coin\ 1\ came\ up\ H}{total\ \sharp\ of\ times\ Coin\ 1\ was\ flipped}$$

$$\hat{p}_2 = \frac{total\ \sharp\ of\ times\ Coin\ 2\ came\ up\ H}{total\ \sharp\ of\ times\ Coin\ 2\ was\ flipped}$$

Motivating example

Now suppose that the Y_i are not observed but we still want MLE's for $p_1$and p_2. The data set now consists of only the X's and is 'incomplete'.

The goal of the EM Algorithm is to find MLEs for p_1 and p_2 in this case.

Let X be observed data, generated by some distribution depending on some parameters. These data may or may not be iid. X will be called an 'incomplete data set'.

Let Y be some 'hidden', 'latent' or 'unobserved data' depending on some parameters.

Let $Z = (X, Y)$ represent the 'complete' data set. We say that it is a 'completion' of the data given by X.

Assume that the distribution of Z (likely a big fat joint distribution) depends on some(likely high-dimensional) parameter θ and that we can write the pdf for Z as

$$f(z, \theta) = f(x, y, \theta) = f(y|x, \theta) f(x, \theta)$$

We usually use $L(\theta)$ to denote a likelihood function and it always depends on some random variables which are not shown by this notation. Because there are many groups of random variables here, we will be more explicit and write $L(z, \theta)$ or $L(x, \theta)$ to denote the **complete likelihood** and **incomplete likelihood functions**, respectively.

The complete likelihood function is

$$L_c(\theta) = L(z, \theta) = L(x, y, \theta) = f(x, y, \theta)$$

The incomplete likelihood function is

$$L(\theta) = L(x, \theta) = f(x, \theta)$$

EM algorithm (Dempster, Laird, Rubin, 1977)

The EM algorithm allows to simplify the problem by using the complete distributions of z without requiring the knowledge of the latent variables.

It starts with the aim of maximizing the incomplete log-likelihood $\max_\theta \ln L(\theta) = \max_\theta \ln L(x;\theta) = \max_\theta \ln f(x;\theta)$ that can be rephrased as follows:

$$\begin{aligned}
\max_\theta \ln f(x;\theta) &= \max_\theta \ln \sum_y f(x,y;\theta) \\
&= \max_\theta \ln \sum_y f(y|x;\theta') \frac{f(x,y;\theta)}{f(y|x;\theta')} \\
&= \max_\theta \ln E_{y|x;\theta'} \left[\frac{f(x,y;\theta)}{f(y|x;\theta')} \right]
\end{aligned}$$

where $f(x|y;\theta')$ depends on a set of parameters θ' that are fixed and known and hence valuable.

EM algorithm (Dempster, Laird, Rubin, 1977)

Apply the *Jensen inequality*:

if g is a concave function (i.e. its second derivative is negative) then we have: $g(E[x]) \geq E[g(x)]$.

In our case $g(x) = \ln x$ and $g(x)'' = -\frac{1}{x^2} < 0$, thus we can apply the inequality and obtaining:

$$\begin{aligned} \ln \sum_y f(y|x;\theta') \frac{f(x,y;\theta)}{f(y|x;\theta')} &\geq \sum_y f(y|x;\theta') \ln \frac{f(x,y;\theta)}{f(y|x;\theta')} \\ \ln L(\theta) &\geq h(\theta|\theta') \end{aligned}$$

$h(\theta|\theta')$ is a lower bound function of the incomplete log-likelihood that we want to maximize.
In order to maximize $\ln L(\theta)$ we can maximize $h(\theta|\theta')$ with respect to θ.

EM algorithm (Dempster, Laird, Rubin, 1977)

We have that

$$h(\theta|\theta') = \sum_y f(y|x;\theta') \ln f(x,y;\theta) - \sum_y f(y|x;\theta') \ln f(y|x;\theta')$$

We can notice that the second term does not depend on θ but only on θ' that we suppose to be known. This term is the *entropy* of the posterior $f(y|x;\theta')$.

Maximizing $h(\theta|\theta')$ with respect to θ is equivalent to maximize only $\sum_y f(y|x;\theta') \ln f(x,y;\theta)$.

EM algorithm (Dempster, Laird, Rubin, 1977)

We want to maximize:

$$\begin{aligned}
\arg\max_{\theta} h(\theta|\theta') &= \arg\max_{\theta} \sum_{y} f(y|x;\theta') \ln f(x,y;\theta) \\
&= \arg\max_{\theta} \sum_{y} f(y|x;\theta') \ln L_c(\theta) \\
&= \arg\max_{\theta} E_{y|x;\theta'}[\ln L_c(\theta)]
\end{aligned}$$

the estimation problem is equivalent to maximize (M-STEP) the conditional expected value (E-STEP) of the complete density, instead of the incomplete one.

Formalization of the EM algorithm

The algorithm can be formalized in the following way:

- Chose $\theta^{(0)} = \theta^{(h)}$ with $h = 0$ (initialization)
- Repeat the following steps until convergence:
  - E-STEP

    Compute $Q(\theta; \theta^{(h)}) = E_{y|x;\theta^{(h)}} [\ln L_c(\theta)]$
  - M-STEP

    Find $\theta^{(h+1)}$ that maximize $Q(\theta; \theta^{(h)})$, that is the value such that

    $Q(\theta^{(h+1)}; \theta^{(h)}) \geq Q(\theta; \theta^{(h)}) \quad \forall \theta \in \Omega$
  - $h = h + 1$

Monotonic property

The EM algorithm satisfies the monotonic condition of the log-likelihood, that is the algorithm guarantees that, in each iterative step, the log-likelihood does not decrease:

$$\ln L\left(\theta^{(h+1)}\right) \geq \ln L\left(\theta^{(h)}\right) \qquad \forall\ h$$

Monotonic property: proof

We can observe that the joint density $f(x, y; \theta)$ can be arbitrarily decomposed as $f(x, y; \theta) = f(x|y; \theta) f(x; \theta)$ or as $f(x, y; \theta) = f(y|x; \theta) f(x; \theta)$.

Considering the last expression we can write the conditional expected value as:

$$\begin{aligned} Q(\theta; \theta^{(h)}) &= E_{y|x;\theta^{(h)}} [\ln L_c(\theta)] \\ &= \sum_y f(y|x; \theta^{(h)}) \ln f(y|x; \theta) + \sum_y f(y|x; \theta^{(h)}) \ln f(x; \theta) \\ &= \sum_y f(y|x; \theta^{(h)}) \ln f(y|x; \theta) + \ln f(x; \theta) \sum_y f(y|x; \theta^{(h)}) \\ &= H(\theta; \theta^{(h)}) + \ln f(x; \theta) = H(\theta; \theta^{(h)}) + \ln L(\theta) \end{aligned}$$

where the first term is given in compact form as $H(\theta; \theta^{(h)})$ that is a kind of entropy.

From the previous equation we have:

$$\ln L(\theta) = Q(\theta; \theta^{(h)}) - H(\theta; \theta^{(h)}) \tag{1}$$

Monotonic property: proof

The monotonic property is verified if

$$\ln L(\theta^{(h+1)}) - \ln L(\theta^{(h)}) \geq 0$$

Let's start from the difference that we can rephrase using equation(1):

$$\begin{aligned} \ln L(\theta^{(h+1)}) - \ln L(\theta^{(h)}) &= \left\{ Q(\theta^{(h+1)}; \theta^{(h)}) - Q(\theta^{(h)}; \theta^{(h)}) \right\} - \\ &- \left\{ H(\theta^{(h+1)}; \theta^{(h)}) - H(\theta^{(h)}; \theta^{(h)}) \right\} \end{aligned} \tag{2}$$

In (2) the first difference is greater or equal to 0 because of the M-step.

Monotonic property: proof

The monotonic property is verified if we can prove

$$H(\theta^{(h+1)};\theta^{(h)}) - H(\theta^{(h)};\theta^{(h)}) \leq 0$$

for every θ we have

$$\begin{aligned} H(\theta^{(h+1)};\theta^{(h)}) - H(\theta^{(h)};\theta^{(h)}) &= E_{y|x;\theta^{(h)}}\left[\ln \frac{f(y|x;\theta^{(h+1)})}{f(y|x;\theta^{(h)})}\right] \\ &\leq \ln\left(E_{y|x;\theta^{(h)}}\left[\frac{f(y|x;\theta^{(h+1)})}{f(y|x;\theta^{(h)})}\right]\right) \\ &= \ln \sum_{y} f(y|x;\theta^{(h+1)}) = 0 \end{aligned}$$

where in the previous development the Jensen inequality has been applied.

Immagine

martedì 20 ottobre 2020 11:22

Addendum EM

mercoledì 23 settembre 2020 14:11

Monotonic Property:

$$H(\theta, \theta^{(h)}) = \sum_y f(y|x,\theta^{(h)}) \ln f(y|x;\theta)$$

$$H(\theta^{(h+1)}, \theta^{(h)}) - H(\theta^{(h)}, \theta^{(h)}) =$$

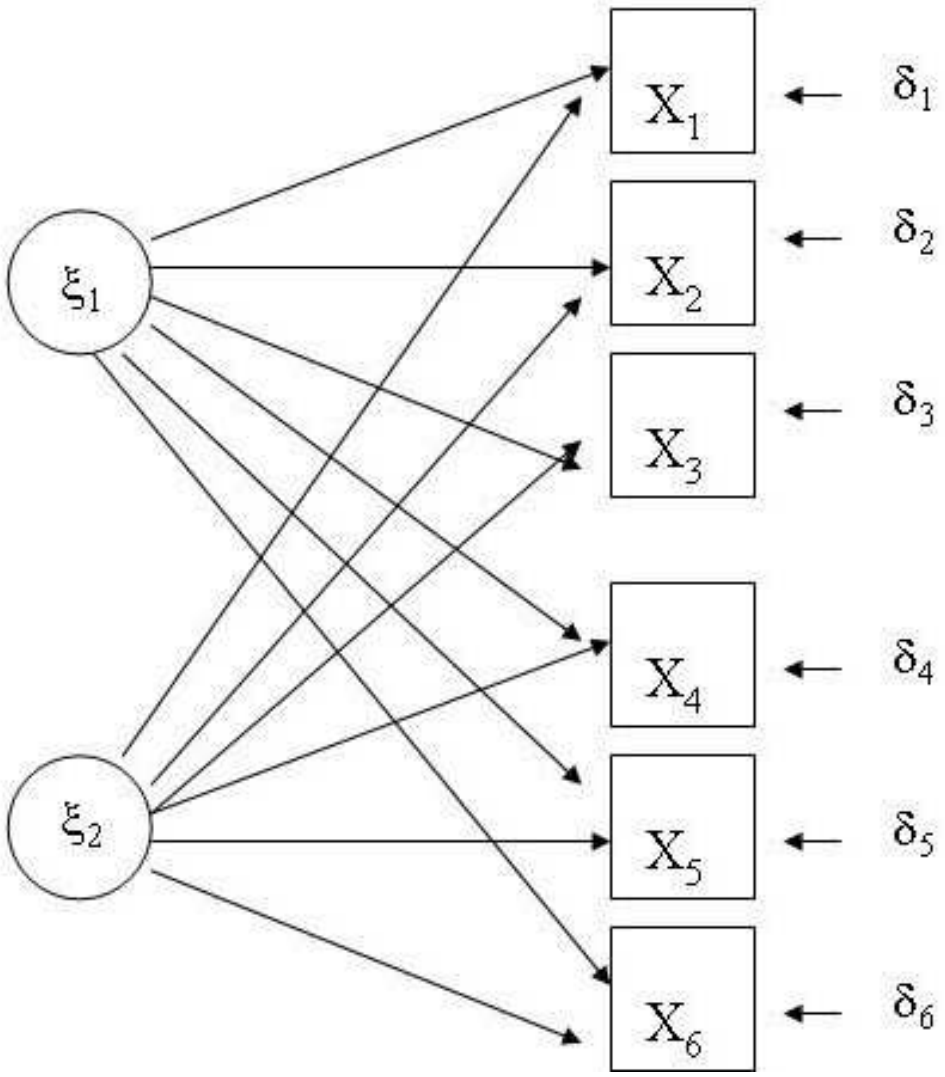
$$= \sum_y f(y|x,\theta^{(h)}) \left[\ln \frac{f(y|x,\theta^{(h+1)})}{f(y|x,\theta^{(h)})} \right] \le$$

$$\le \ln E_{y|x,\theta^{(h)}} \left[\frac{f(y|x,\theta^{(h+1)})}{f(y|x,\theta^{(h)})} \right] =$$

$$= \ln \sum_y f(y|x,\theta^{(h)}) \frac{f(y|x,\theta^{(h+1)})}{f(y|x,\theta^{(h)})} =$$

$$= \ln \underbrace{\sum_y f(y|x,\theta^{(h+1)})}_{1} = 0$$

Exploratory factor analysis



Exploratory factor analysis

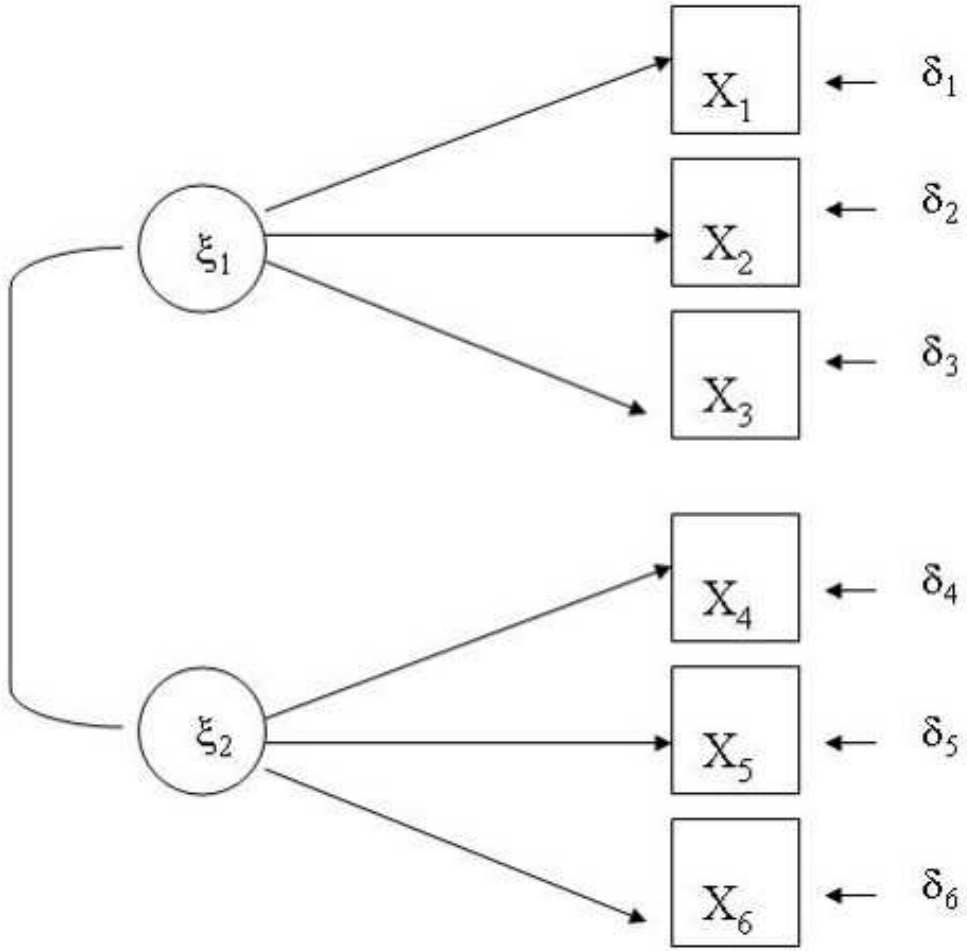
Characteristics:

1. No a priori hypothesis non the models, all the observed variables are influenced by all the factors

2. all the parameters have to be estimated, they are free parameters.

3. we get equivalent solutions at each rotation of the axis (invariance to orthogonal rotations)

Confirmatory factor analysis



Confirmatory factor analysis

Characteristics:

1. Specific hypothesis on the number of factors, on the relations between factors, on the relations between factors and variables coming from previous analysis, existent theories, or from the characteristics of the experimental design or from an inspection of the correlation matrix

2. Some parameters are free and some are fixed, we choose the parameters to be estimated according to the hypothesis of the theoretical model

3. The solution is unique

Binary data: latent trait models

Binary responses are very common in social and educational sciences. Such binary variables are often supposed to be indicators of more fundamental attitudes or abilities.

Suppose there are p items or questions to which the respondent is required to give a binary response: right/wrong, agree/disagree, yes/no.

If we code the possible responses with 0 and 1, the *response pattern* consists of string of zeros and ones.

With p variables, each having two outcomes, 2^p different responses consists of string of zeros and ones.

The number of observed response patterns depends on n.

Binary data: latent trait models

Example: Attitude towards abortion

- $p = 4$ items (0=agree, 1=disagree)
- 2^4 possible response patterns
- $n = 379$ individuals after listwise deletion

The four items are

1. The woman decides on her own that she does not [WomanDecide]
2. The couple agree that they do not wish to have the child [CoupleDecide]
3. The woman is not married and does not wish to marry the man [NotMarried]
4. The couple cannot afford any more children [CannotAfford]

Objectives of the analysis

1. Measure a single latent variable expressing the attitude toward abortion
2. Ranking the individuals according to the latent variable

Binary data: latent trait models

Response pattern	Frequency
1111	141
0000	103
0111	44
0011	21
0001	13
1110	12
0010	10
0110	7
1011	6
0101	6
1101	3
1100	3
1000	1
1010	0
1001	0
Total	379

Percentage of individual agreeing that abortion should be legal under circumstances described by item 1 to 4 are: 43.8, 59.4, 63.6, 61.7.

Cross tabulation of items 1 and 2

	Yes	No
Yes	159	7
No	66	147

Binary data: latent trait models

Two approaches.

1. *Response function approach*: it analyzes the data as they are by imposing distributions on the observed variables.

2. *Underlying variable approach* (SEM): it assumes that the data have been produced by dichotomizing underlying continuous variables.

Response function approach

The joint probability distribution $f(\mathbf{x})$ assigns probabilities to the 2^p response patterns. For some purposes it is useful to specify the joint distribution in terms of set of *marginal probabilities*.

$$P(x_i = 1) \quad (i = 1, 2, \ldots, p)$$

$$P(x_i = 1, x_j = 1) \quad (i, j = 1, 2, \ldots, p) \quad (i \neq j)$$

$$P(x_i = 1, x_j = 1, x_k = 1) \quad (i, j, k = 1, 2, \ldots, p) \quad (i \neq j \neq k)$$

$$\vdots$$

$$P(x_1 = 1, x_2 = 1, \ldots, x_p = 1).$$

Total number of probabilities are

$$\binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{p} = 2^p - 1$$

Response function approach: logit/normal model

If the dependence among the x's is wholly explained by a vector $\mathbf{y}$ of latent variables they may be regarded as mutually independent random variables with

$$P(x_i = 1|\mathbf{y}) = \pi_i(\mathbf{y}) \quad (x_i = 0, 1; i = 1, \ldots, p)$$

$\pi_i(\mathbf{y})$: *item response function or item characteristic curve* with $0 \leq \pi_i(\mathbf{y}) \leq 1$ (item characteristic curve in IRT) and it is a not decreasing monotonic function.

The prior distribution of the latent variables $h(\mathbf{y})$ is assumed to be normal

$$h(\mathbf{y}) \sim N(\mathbf{0}, \mathbf{I})$$

that is the latent variables are standardized independent normal variables.

Response function approach: logit/normal model

The conditional distribution $g(\mathbf{x}|\mathbf{y})$ is assumed to a be a GLLVM as follows:

- Random component

$$g(x_i|\mathbf{y}) = \pi_i(\mathbf{y})^{x_i}(1-\pi_i(\mathbf{y}))^{1-x_i}$$

  that is $g_i(x_i|\mathbf{y}) \sim Ber(\pi_i(\mathbf{y}))$ with $\pi_i(\mathbf{y})$ probability of success.

- Systematic component

$$\eta_i = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j \quad i = 1, \ldots, p$$

- Link function: logit $\Rightarrow \theta_i = \text{logit}(\pi_i(\mathbf{y})) = \eta_i = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j$

  Equivalently

$$\pi_i(\mathbf{y}) = \frac{\exp(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j)}{1 + \exp(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j)}$$

This model is called *logit/normal model*.

Parameter interpretation

α_{i0} determines the probability of the median individual

$$\pi_i(0) = \frac{1}{1 + \exp(-\alpha_{i0})}$$

$\pi_i(0)$ probability of a positive response from an individual at $y = 0$ (median).

α_{ij} determines the effect of the latent variable y_j on the probability of success with respect to x_i. A given change of y_j will produce a larger change in the probability of success of the variable x_i when this parameter is large than when it is small.

- IRT (Item Response Theory: educational context)

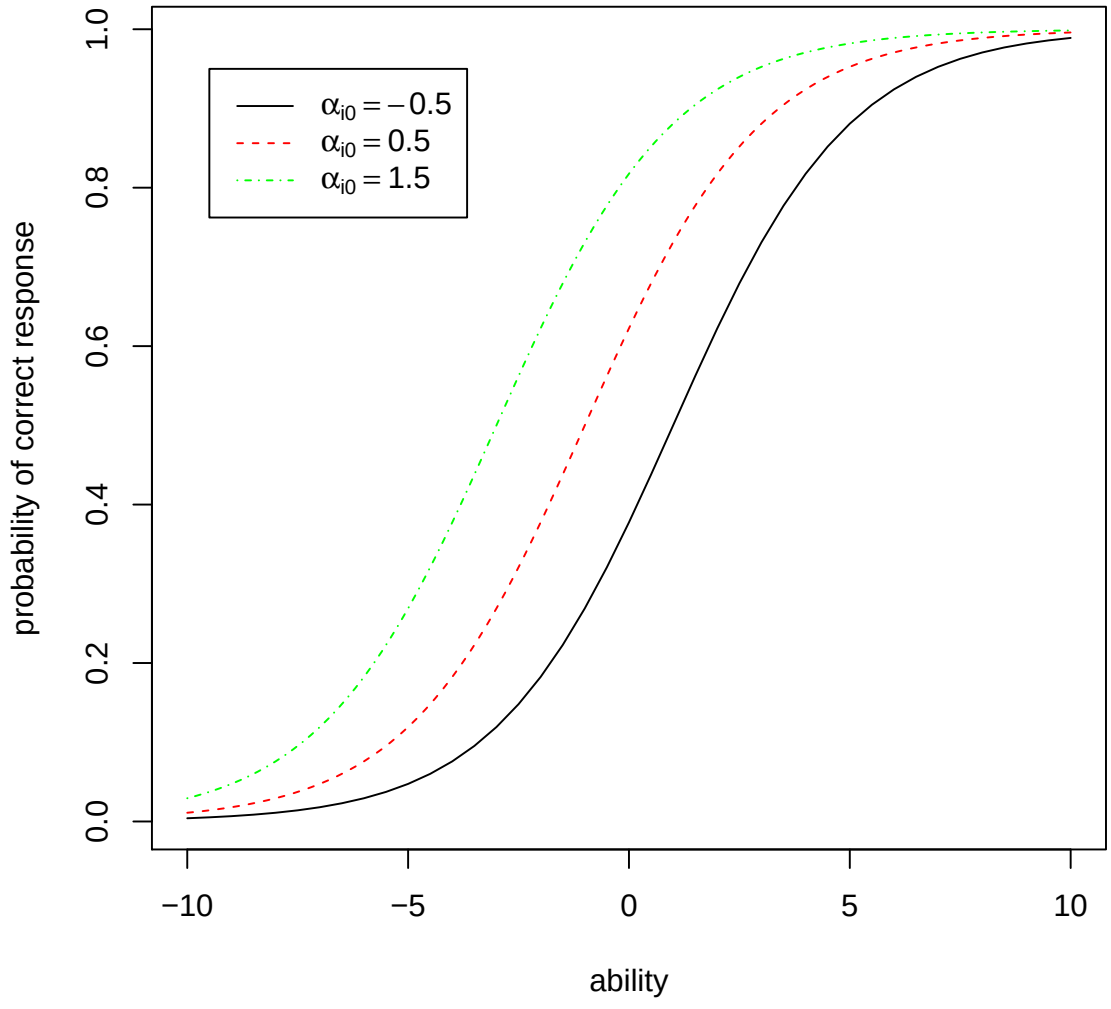
  α_{i0} *difficulty* parameter: an increasing of the value of this parameter increases the probability of positive response for all values of y_j

  α_{ij} *discrimination* parameter: the bigger is the value of this parameter the easier it will be to discriminate between a pair of individuals a given distance apart on the latent scale.

- Factor analysis
  α_{i0} intercept, α_{ij} factor loading

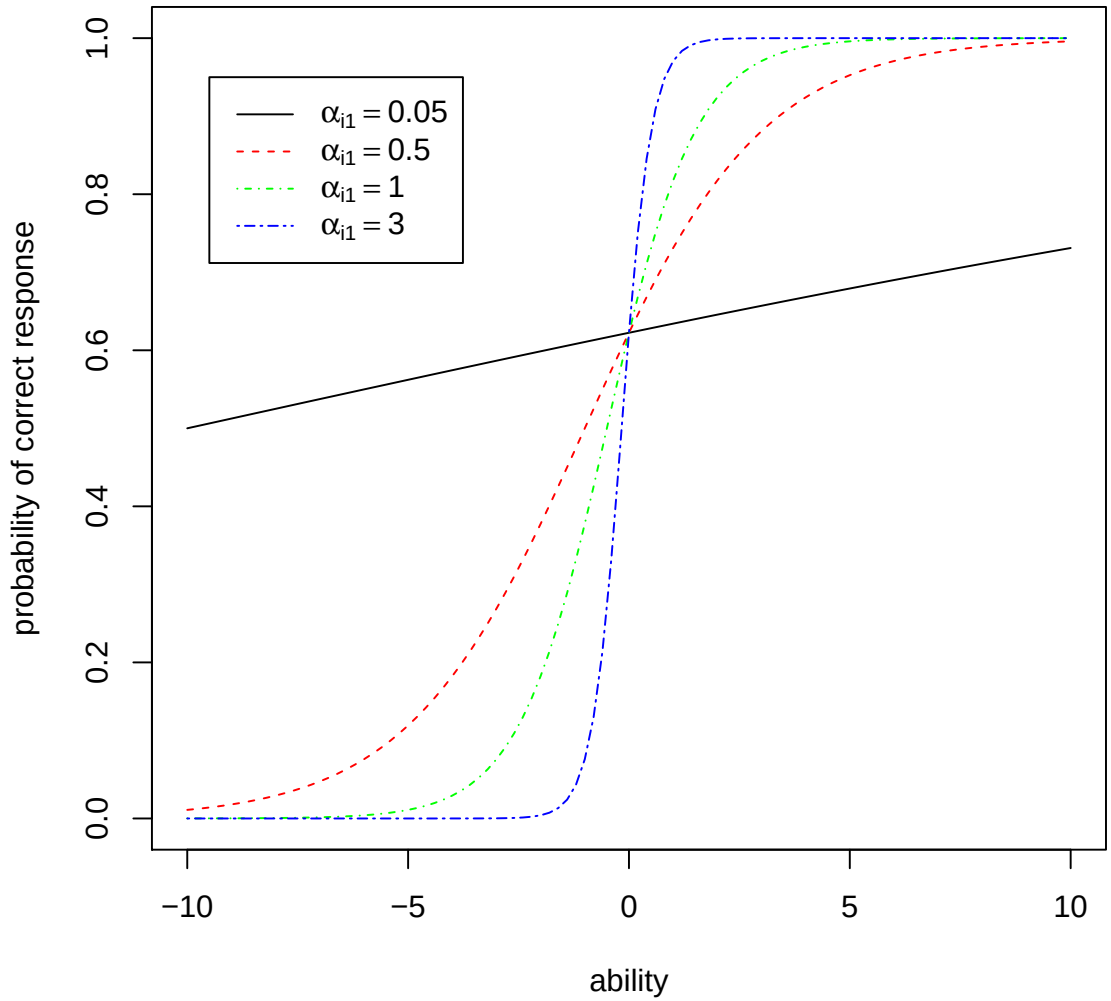
Parameter interpretation

Item response function for different values of the difficulty parameter and $\alpha_{i1} = 0.5$



Parameter interpretation

Item response function for different values of the discrimination parameter and $\alpha_{i0} = 0.5$



Component scores

Recall the expression of the component scores

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i \quad (j = 1, \ldots, q)$$

Since in latent trait models x_i are either 0 or 1, X_j is simply the sum of those α_{ij} for which $x_i = 1$.

Example:

Given the individuals h, h' we have the two response patterns:

$x_h = (0, 1, 1, 1, 0, 1, 0)$
$x_{h'} = (0, 0, 1, 1, 0, 1, 0)$

The component scores for the jth latent variable are :

$X_{hj} = \alpha_{2j} + \alpha_{3j} + \alpha_{4j} + \alpha_{6j}$
$X_{h'j} = \alpha_{3j} + \alpha_{4j} + \alpha_{6j}$

If α_{2j} is large the scores of the two individuals will be very different.

The Rasch model

A special case of the unidimensional model is obtained when all the discrimination parameters are equal to 1. This model has been introduced by Rasch in 1960 and it is usually written as:

$$P(x_i = 1, \beta_{i0}, y_h) = \frac{\exp(y_h - \beta_{i0})}{1 + \exp(y_h - \beta_{i0})}$$

This model is used in educational testing to study the abilities of a particular set of individuals. The y_h's measure each individual's ability, and the parameter β_{i0} remains the difficulty parameter.

The Rasch model

The Rasch model is still quite popular in educational testing because of its simplicity and its attractive theoretical properties. In particular:

- The total score $\sum_{i=1}^{p} x_{ih}$ is sufficient for y_h- that is, it contains all the information in the data about the β's if the model is true.
  $\sum_{i=1}^{p} x_{ih}$ is total number of correct responses given by individual h.

- The total number of positive/correct responses for item i,$\sum_{h=1}^{n} x_{ih}$ is sufficient for β_{i0}.
  $\sum_{h=1}^{n} x_{ih}$ is the total number of correct answer given to item i.

2-Parametric Logistic model

An extension of the Rasch model in the case in which the discrimination parameters are all equal is the 2-parameter logistic model (2PL) introduced by Birnbaum in 1969. The model is written as

$$P(x_i = 1, \beta_{i0}, \beta_{i1}, y_h) = \frac{\exp(\beta_{i1}(y_h - \beta_{i0}))}{1 + \exp(\beta_{i1}(y_h - \beta_{i0}))}$$

The parameters β_{i0}, β_{i1} remain the difficulty and the discrimination parameters of the item. This model aims at shifting and scaling the ability according to the 2 parameters β_{i0},and β_{i1}.
It represents a different parametrization of the latent trait model. Indeed

$\beta_{i0} = -\alpha_{i0}/\alpha_{i1}$
$\beta_{i1} = \alpha_{i1}$

In the IRT framework there are models with more parameters than two parameters (3PL model) and extensions to more than one ability.

Model estimation

Maximum likelihood estimation via the E-M algorithm. One factor model.
E-M algorithm is used when there are missing observations. Here the missing observations are the values of y.
The E-M algorithm starts from the log-likelihood l_c for the complete data $(x_{ih}, y_h)(i = 1, \ldots, p, h = 1, \ldots, n)$

$$l_c = \sum_{h=1}^{n} \ln f(\mathbf{x}_h, y_h) = \sum_{h=1}^{n} \left[\sum_{i=1}^{p} \ln g(x_{ih}|y_h) + \ln h(y_h) \right]$$

Two steps:

1. E-step: evaluate $E[l_c|\mathbf{x}_1, \ldots, \mathbf{x}_n]$
2. M-step: maximise $E[l_c|\mathbf{x}_1, \ldots, \mathbf{x}_n]$ over the parameters.

Since $h(y)$ does not depend on the parameters we can replace l_c with

$$l = \sum_{h=1}^{n} \ln g(\mathbf{x}_h|y_h) = \sum_{h=1}^{n} \sum_{i=1}^{p} \ln g(x_{ih}|y_h)$$

Model estimation

We can write

$$g(x_{ih}|y_h) = (1 - \pi_i(y_h)) \exp(x_{ih}(\alpha_{i0} + \alpha_{i1} y_h))$$

$$\begin{aligned} l &= \sum_{h=1}^{n} \left[\sum_{i=1}^{p} (\ln(1 - \pi_i(y_h)) + x_{ih}(\alpha_{i0} + \alpha_{i1} y_h)) \right] = \\ &= \sum_{i=1}^{p} \left[\sum_{h=1}^{n} \ln(1 - \pi_i(y_h)) + \alpha_{i0} \sum_{h=1}^{n} x_{ih} + \alpha_{i1} \sum_{h=1}^{n} x_{ih} y_h \right] \end{aligned}$$

E-step

$$\begin{aligned} E[l|\mathbf{x}] &= \sum_{i=1}^{p} [\sum_{h=1}^{n} E(\ln(1 - \pi_i(y_h))|\mathbf{x}_1, \ldots, \mathbf{x}_n) + \alpha_{i0} \sum_{h=1}^{n} x_{ih} + \\ &+ \alpha_{i1} \sum_{h=1}^{n} x_{ih} E(y_h|\mathbf{x}_1, \ldots, \mathbf{x}_n)] \end{aligned}$$

Model estimation

M-step
For each item i we maximize

$$
\begin{aligned}
E[l_i|\mathbf{x}] &= \int_R \left[\ln(1-\pi_i(y)) \sum_{h=1}^{n} h(y|\mathbf{x}_h) + \alpha_{i0} \sum_{h=1}^{n} x_{ih} + \alpha_{i1} \sum_{h=1}^{n} x_{ih} h(y|\mathbf{x}_h) \right] dy \\
&= \int_R \left[\ln(1-\pi_i(y)) N(y) + \alpha_{i0} \sum_{h=1}^{n} x_{ih} + \alpha_{i1} R_i(y) \right] dy \qquad (1)
\end{aligned}
$$

where

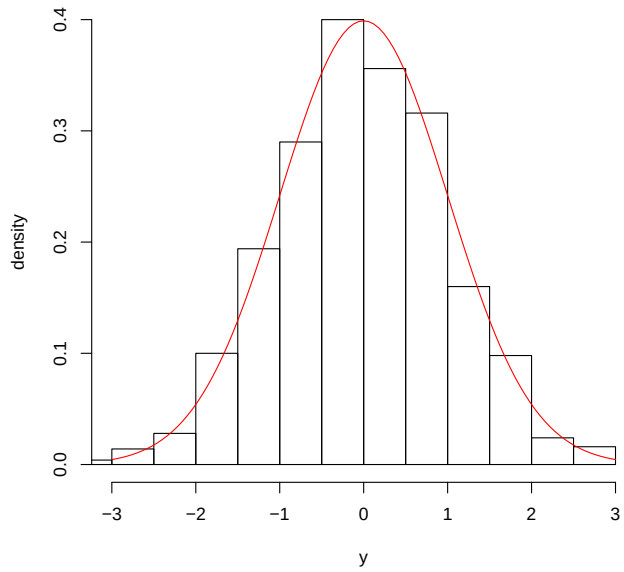
$$R_i(y) = \sum_{h=1}^{n} x_{ih} h(y|\mathbf{x}_h)$$

$$N(y) = \sum_{h=1}^{n} h(y|\mathbf{x}_h)$$

The integral in equation (1) is not solvable analytically $\Rightarrow$ numerical approximation

Model estimation: approximation of the integral

It is possible to approximate any density $f(\cdot)$ with finite moments with a discrete distribution with given level of accuracy.



The latent variable y is treated as discrete with values (nodes) $y_1, \ldots, y_k$ and corresponding weights $h(y_1), \ldots, h(y_k)$.

When the density is standard normal the appropriate approximation method is called Gauss-Hermite quadrature.

Model estimation

Thus the approximated likelihood for each item i becomes

$$\sum_{t=1}^{k}\left[\ln(1-\pi_i(y_t))n_t+\alpha_{i0}\sum_{h=1}^{n}x_{ih}+\alpha_{i1}r_{it}\right]$$

with

$$r_{it}=\sum_{h=1}^{n}x_{ih}h(y_t|\mathbf{x}_h)\quad n_t=\sum_{h=1}^{n}h(y_t|\mathbf{x}_h)$$

r_{it} expected number of those predicted to be at y_t who will respond positively.
$n(t)$ expected number of individuals in latent position y_t.

Differentiating the approximated $E[l_i|\mathbf{x}]$ with respect to α_{i0} and α_{ij} leads to non linear equations in the parameters. $\Rightarrow$ Newton-Raphson iterative procedure.

Model estimation

E-M algorithm:

1. Step 1: choose starting values for α
2. Step 2: compute r_{it} and $n(t)$
3. Step 3: obtain improved estimates of the α's treating r_{it} and $n(t)$ as given numbers
4. Step 4: return to Step 2 until convergence

Remarks:

- It can happen that the EM algorithm does not converge. Possible causes are small sample size, small number of items, the discrimination parameters are very different.
- The number of nodes k can affect the final estimation. It is recommended to repeat the estimation procedure with increasing number of quadrature points.

Underlying variable approach: Probit/normal model

Alternative model for binary data is the probit model

$$\Phi^{-1}(\pi_i(\mathbf{y})) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j \quad (i = 1, \ldots, p)$$

It follows that

$$\pi_i(\mathbf{y}) = P(x_i = 1|\mathbf{y}) = \Phi(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j) \quad (i = 1, \ldots, p)$$

It is very close to the logit model since

$$\text{logit}(u) = (\pi/\sqrt{3})\Phi^{-1}(u)$$

The two models are virtually equivalent but the probit/normal model lacks the sufficiency property of the components X.

Underlying variable approach

Probit model related to the standard normal linear model $\Rightarrow$ *Underlying variable approach (UVA)*

Given an the observed binary variable x_i, we suppose that underlying to this indicator, there is a continuous variable, ξ.

$$x_i = \begin{cases} 1, & \xi_i \leq \tau_i; \\ 0, & \text{otherwise} \end{cases}$$

where τ_i is a threshold parameter and ξ_i is standard normal.

The variable ξ_i is assumed to be generated by a standard normal factor model. In matrix form

$$\boldsymbol{\xi} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{R}_{\boldsymbol{\xi}})$. $\boldsymbol{\Lambda}$ is the matrix of the factor loadings.
$\boldsymbol{\Psi}$ is a diagonal matrix as in the classical normal linear factor model.

$\mathbf{R}_{\boldsymbol{\xi}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$

$\mathbf{R}_{\boldsymbol{\xi}}$ is a unknown correlation matrix and has to be estimated
In the binary case the correlations are called *tetrachoric correlations*.

Underlying variable approach: estimation procedure steps

The parameters to be estimated are the thresholds, the loadings and the specific variances.

The estimation procedure consists of three steps:

1. Maximum likelihood estimate the thresholds $\tau_1, \tau_2, \ldots, \tau_p$ using each marginal distribution of the latent variables.

$$\pi_i = P(x_i = 1) = \int_{-\infty}^{\tau_{(i)}} \phi(\xi_i) \mathrm{d}\xi_i = \Phi(\tau_{(i)})$$

   from which, given the sample proportion P_i, that is the maximum likelihood estimation of π_i, we obtain

$$\hat{\tau}_{(i)} = \Phi^{-1}(P_i)$$

2. Estimate of $\mathbf{R}_{\boldsymbol{\xi}}$, the matrix of tetrachoric correlations using the bivariate distribution of the latent variables.

$$\pi_{ij} = P(x_i = 1, x_j = 1) = \int_{-\infty}^{\tau_{(i)}} \int_{-\infty}^{\tau_{(j)}} \phi(\xi_i, \xi_j, \rho_{ij}) \mathrm{d}\xi_i$$

3. Estimate the loadings and the specific variances using the classical methods of the classical linear factor analysis.

Equivalence between the response function and the underlying variable approach

Let S be any non-empty subset of $\{1, 2, \ldots, p\}$ then

$$P(x_i = 1, x_j = 1, \ldots) = P\left\{\bigcap_{i \in S} (x_i = 1)\right\} = \int \ldots \int P\left\{\bigcap_{i \in S} (x_i = 1|\mathbf{y}) h(\mathbf{y})\right\} d\mathbf{y} =$$
$$= \int \ldots \int \prod_{i \in S} P\left\{(x_i = 1|\mathbf{y}) h(\mathbf{y})\right\} d\mathbf{y}$$

$$P\left\{\bigcap_{i \in S} (x_i = 1)\right\} = P\left\{\bigcap_{i \in S} (\xi_i \leq \tau_i)\right\} =$$
$$= \int \ldots \int \prod_{i \in S} P\left\{(\xi_i \leq \tau_i)|\mathbf{y}) h(\mathbf{y})\right\} d\mathbf{y}$$

Equivalence between the response function and the underlying variable approach

$$
\begin{aligned}
P((\xi_i \leq \tau_i)|\mathbf{y}) &= P\left\{\mu_i + e_i + \sum_{j=1}^{q} \lambda_{ij} y_j \leq \tau_i |\mathbf{y}\right\} = \\
&= P\left\{\frac{e_i}{\psi_i^{1/2}} \leq \frac{\tau_i - \mu_i - \sum_{j=1}^{q} \lambda_{ij} y_j}{\psi_i^{1/2}} |\mathbf{y}\right\} \\
&= R\left(\frac{\tau_i - \mu_i - \sum_{j=1}^{q} \lambda_{ij} y_j}{\psi_i^{1/2}}\right)
\end{aligned}
$$

$R(\cdot)$ is the distribution function of $e_i/\psi_i^{1/2}$.
Remind that for the response function model

$$
P(x_i = 1|\mathbf{y}) = G\left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j\right)
$$

G: inverse logit or probit.

Equivalence between the response function and the underlying variable approach

The equivalence exists if

- $G \equiv R$
- $\alpha_{i0} = (\tau_i - \mu_i)/\psi_i^{1/2}$
- $\alpha_{ij} = -\lambda_{ij}/\psi_i^{1/2}$

The parameters τ_i, μ_i and λ_{ij} are not individually estimable because no information about the standard deviation of ξ_i. We assume that

$$var(\xi_i) = 1 = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i$$

$$\alpha_{ij} = -\frac{\lambda_{ij}}{\sqrt{1 - \sum_{j=1}^{q} \lambda_{ij}^2}}$$

$$st\alpha_{ij} = \frac{\alpha_{ij}}{\sqrt{1 + \sum_{j=1}^{q} \alpha_{ij}^2}}$$

Standard errors

The determination of exact standard errors for the parameter estimates is not possible $\Rightarrow$ asymptotic variance-covariance matrix using the information matrix.

If we recall the set of parameters $\boldsymbol{\beta}$ we have

$$var(\hat{\boldsymbol{\beta}})^{-1} = E\left[-\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right]$$

It can be approximated by using the observed matrix

$$var^*(\hat{\boldsymbol{\beta}})^{-1} = \left[\sum_{h=1}^{n} \frac{1}{f^2(\mathbf{x}_h)} \frac{\partial f(\mathbf{x}_h)}{\partial\beta_i} \frac{\partial f(\mathbf{x}_h)}{\partial\beta_j}\right]$$

Example: parameter estimates in Attitude towards abortion

Parameter estimates and standard errors in brackets and standardized loadings for the one-factor model

Item	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	st$\hat{\alpha}_{i1}$	$\hat{\pi}_i(0)$
WomanDecide	-0.72	(0.33)	4.15	(0.85)	0.97	0.33
CoupleDecide	1.11	(0.35)	4.50	(0.81)	0.98	0.75
NotMarried	2.18	(0.61)	6.21	(1.54)	0.99	0.90
CannotAfford	1.15	(0.28)	3.49	(0.50)	0.96	0.76

Goodness of fit

The theoretical hypothesis is $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\alpha})$ where $\boldsymbol{\pi}$ is the true probability and $\boldsymbol{\pi}(\boldsymbol{\alpha})$ is computed from the model. In practice, the sample proportion $\mathbf{P}$ are compared with $\boldsymbol{\pi}(\hat{\boldsymbol{\alpha}})$ where $\hat{\boldsymbol{\alpha}}$ is the estimator with some method. The classical test statistics are the Pearson χ^2 statistic and the likelihood ratio test G^2:

$$\chi^2 = \sum_{r=1}^{2^p} \frac{(nP_r - n\pi_r(\hat{\boldsymbol{\alpha}}))^2}{n\pi_r(\hat{\boldsymbol{\alpha}})}$$

$$G^2 = 2\sum_{r=1}^{2^p} nP_r \ln\left(\frac{nP_r}{n\pi_r(\hat{\boldsymbol{\alpha}})}\right)$$

Under regular condition, when n is large and p is small, the two statistics follow a chi square distribution with degrees of freedom $2^p - p(q+1) - 1$. When p increases $\Rightarrow$ sparse data, chi square distributing does not hold anymore.

Example: $p = 10$ $2^p = 1024$ $n = 1000$. We expect many responses with $n\pi_i(\hat{\boldsymbol{\alpha}}) \leq 1$

Remedies

- Group the response patterns with expected frequencies less than 5. Risk to have degrees of freedom equal to 0.
- Empirical sampling distribution of the test statistics using the parametric bootstrap method.
- Consider residuals calculated from marginal frequencies:

  $P(x_i = 1)$ first-order margins, contain no information about the dependencies among the x's.
  $P(x_i = 1, x_j = 1)$ second-order margins, information about pairwise association
  $P(x_i = 1, x_j = 1, x_k = 1)$ third-order margins.
  Second and third-order margins usually suffice.

O: observed frequencies for any marginal probability
E: corresponding expected frequency

$$R = \frac{(O - E)^2}{E}$$

Large values of R for the second-order margins identify pairs of x's responsible of the bad fit.
Rule of thumb: $R > 4$ indicates bad fit.

Example: goodness of fit in Attitude towards abortion

$G^2 = 17.85$ and $\chi^2 = 15.09$, 3 degrees of freedom (p-values= 0.0005, 0.001).

Residuals for second order margins

Response	Item i	Item j	O	E	O-E	$(O-E)^2/E$
(0,0)	2	1	147	143.74	3.26	0.07
	3	1	131	133.17	-2.17	0.04
	3	2	117	119.69	-2.69	0.06
	4	1	129	133.68	-4.68	0.16
	4	2	114	116.09	-2.09	0.04
	4	3	116	111.79	4.21	0.16
(0,1)	2	1	7	11.30	-4.30	1.64
	3	1	7	5.94	1.06	0.19
	3	2	21	19.42	1.58	0.13
	4	1	16	11.99	4.01	1.34
	4	2	31	29.58	1.42	0.07
	4	3	29	33.88	-4.88	0.70
(1,0)	2	1	66	69.89	-3.89	0.22
	3	1	82	80.46	1.54	0.03
	3	2	37	35.35	1.65	0.08
	4	1	84	79.95	4.05	0.21
	4	2	40	38.95	1.05	0.03
	4	3	22	27.32	-5.32	1.04
(1,1)	2	1	159	154.07	4.93	0.16
	3	1	159	159.43	-0.43	0.00
	3	2	204	204.54	-0.54	0.00
	4	1	150	153.58	-3.38	0.07
	4	2	194	194.38	-0.38	0.00
	4	3	212	206.01	5.99	0.17

Posterior analysis

As in the general GLLVM, all the information about $\mathbf{y}$ is contained in the posterior distribution $h(\mathbf{y}|\mathbf{x})$.

In the case of the logit/normal model the distribution depends on $\mathbf{x}$ only through the q-variate sufficient statistic $\mathbf{X} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} = \{\alpha_{ij}\}$.

Component scores are often used for scaling, expecially in measurement abilities.

The posterior mean $E(y_j|\mathbf{x}_h)$ and the components give the same ranking to response patterns/individuals.

Example: factor scores in Attitude towards abortion

Response Pattern	Observed frequency	Expected frequency	$E(y\|\mathbf{x})$	$\sigma(y\|\mathbf{x})$	Component score	Total score
0000	103	100	-1.19	0.55	0.00	0
0001	13	16.6	-0.61	0.32	3.49	1
1000	1	1.7	-0.55	0.30	4.15	1
0100	9	9.1	-0.52	0.29	4.50	1
0010	10	12.3	-0.38	0.26	6.21	1
1001	0	1.3	-0.29	0.24	7.64	2
0101	6	7,4	-0.27	0.24	7.99	2
1100	3	1.0	-0.24	0.24	8.65	2
0011	21	14.8	-0.18	0.24	9.70	2
1010	0	2.0	-0.14	0.25	10.37	2
0110	7	2.0	-0.12	0.26	10.71	2
1101	3	1.9	-0.01	0.28	12.14	3
1011	6	6.2	0.14	0.32	13.86	3
0111	44	41.1	0.17	0.32	14.20	3
1110	12	7.2	0.24	0.34	14.87	3
1111	141	143.9	0.95	0.61	18.35	4
Total	379					

Latent class models for binary data

Suppose there are p binary variables denoted with $x_1, x_2, \ldots, x_p$ where $x_i = 0, 1$.

Assume that the mutual association of these variables is accounted for a single latent binary variable y.

We assume to divide the population into two parts so that the x's are mutually independent in each group.

Labels 0 and 1 for each group. The prior distribution of y is

$$h(1) = P(y = 1) = \eta \qquad h(0) = 1 - h(1) = 1 - \eta$$

The conditional distribution of x_i given y is a Bernoulli variable

$$g(x_i|y) = \pi_{iy}^{x_i}(1 - \pi_{iy})^{1-x_i} \quad (x_i, y = 0, 1)$$

It follows that

$$f(\mathbf{x}) = \eta \prod_{i=1}^{p} \pi_{i1}^{x_i}(1 - \pi_{i1})^{1-x_i} + (1 - \eta) \prod_{i=1}^{p} \pi_{i0}^{x_i}(1 - \pi_{i0})^{1-x_i}$$

Latent class models for binary data

In order to evaluate if the model has a good fit to the data we can use the goodness of fit tests.

if the model is not adequate we can

- extend the number of classes to a number of classes $K \geq 3$
- consider models that do not assume the conditional independence
- consider models with continuous latent variables

If the model is adequate we define a rule to allocate the individuals in the two latent classes based on the observed values $\Rightarrow$ posterior distribution of the latent variables given the observed variables

Latent class models for binary data

The posterior probability of belonging to the two classes given the observed variables are the following

$$
\begin{aligned}
h(1|\mathbf{x}) &= P(y=1|x_1\ldots,x_p) = \frac{P(y=1,x_1\ldots,x_p)}{f(x_1\ldots,x_p)} = \\
&= \frac{h(y=1)g(x_1\ldots,x_p|y=1)}{f(\mathbf{x})} = \\
&= \frac{\eta\prod_{i=1}^{p}\pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i}}{\eta\prod_{i=1}^{p}\pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i}+(1-\eta)\prod_{i=1}^{p}\pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}}
\end{aligned}
$$

$$
h(0|\mathbf{x}) = \frac{(1-\eta)\prod_{i=1}^{p}\pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}}{\eta\prod_{i=1}^{p}\pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}+(1-\eta)\prod_{i=1}^{p}\pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}}
$$

The allocation rule is the following:

$h(1|\mathbf{x}) > h(0|\mathbf{x})$ the unit is allocated in latent class 1.

$h(1|\mathbf{x}) \leq h(0|\mathbf{x})$ the unit is allocated in latent class 0.

Latent class models for binary data

The extension to a K-class model is almost immediate. In an exploratory analysis K is a parameter to be determined, otherwise it is given.

π_{ij} probability of a positive response, variable i, category j
$(i = 1, \ldots, p; j = 0, 1, \ldots, K-1)$.

η_j prior probability that a randomly chosen individual is in class $j (\sum_{j=0}^{K-1} \eta_j = 1)$.

Assumption of the model: conditional independence

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

It follows that

$$f(\mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{(1-x_i)}$$

Latent class models for binary data

The posterior probability that an individual with response vector $\mathbf{x}$ belongs to category j is

$$h(j|\mathbf{x}) = \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} / f(\mathbf{x}) \quad (j = 0, 1, \ldots, K-1)$$

The posterior distribution can be used to construct an allocation rule $\Rightarrow$ an individual is allocated in the class with the highest posterior probability.

The vector of the unknown parameters to be estimated is
$\theta = (\eta_0, \ldots, \eta_{K-1}, \pi_{10}, \ldots, \pi_{p,K-1})$

Maximum likelihood estimation

As before, we can use the E-M algorithm. For a random sample of size n the log-likelihood can be written as

$$l = \sum_{h=1}^{n} \ln \left\{ \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}} (1 - \pi_j)^{1 - x_{ih}} \right\}$$

It is maximized subject to $\sum \eta_j = 1$, so we find the maximum of

$$\phi = l + \gamma \sum_{j=0}^{K-1} \eta_j$$

γ is an undetermined multiplier (it is equivalent to the Lagrange multiplier).

Maximum likelihood estimation: EM algorithm

Finding partial derivatives we have

$$\frac{\partial \phi}{\partial \eta_j} = \sum_{h=1}^{n} g(\mathbf{x}_h|j)/f(\mathbf{x}_h) + \gamma = 0$$

Given that

$$h(j|\mathbf{x}_h) = \frac{\eta_j g(\mathbf{x}_h|j)}{f(\mathbf{x}_h)} \Rightarrow \frac{h(j|\mathbf{x}_h)}{\eta_j} = \frac{g(\mathbf{x}_h|j)}{f(\mathbf{x}_h)}$$

It follows that

$$\frac{\partial \phi}{\partial \eta_j} = \sum_{h=1}^{n} \frac{h(j|\mathbf{x}_h)}{\eta_j} + \gamma = 0 \Rightarrow \sum_{h=1}^{n} h(j|\mathbf{x}_h) = -\gamma\eta_j$$

Summing up over the $K-1$ categories

$$\sum_{h=1}^{n}\sum_{j=0}^{K-1} h(j|\mathbf{x}_h) = -\gamma \sum_{j=0}^{K-1} \eta_j \Rightarrow n = -\gamma$$

and thus we obtan

$$\hat{\eta}_j = \sum_{h=1}^{n} h(j|\mathbf{x}_h)/n$$

Maximum likelihood estimation: EM algorithm

In the same way

$$\frac{\partial \phi}{\partial \pi_{ij}} = \sum_{h=1}^{n} \eta_j \frac{\partial \phi}{\partial \pi_{ij}} \frac{g(\mathbf{x}_h|j)}{f(\mathbf{x}_h)} = \frac{\eta_j}{\pi_{ij}(1-\pi_{ij})} \sum_{h=1}^{n} (x_{ih} - \pi_{ij}) \frac{g(\mathbf{x}_h|j)}{f(\mathbf{x}_h)} = 0$$

$$\frac{\eta_j}{\pi_{ij}(1-\pi_{ij})} \sum_{h=1}^{n} (x_{ih} - \pi_{ij}) \frac{h(j|\mathbf{x}_h)}{\eta_j} = 0$$

$$\sum_{h=1}^{n} x_{ih} h(j|\mathbf{x}_h) = \pi_{ij} \sum_{h=1}^{n} h(j|\mathbf{x}_h)$$

and thus we obtain

$$\hat{\pi}_{ij} = \frac{\sum_{h=1}^{n} x_{ih} h(j|\mathbf{x}_h)}{\sum_{h=1}^{n} h(j|\mathbf{x}_h)}$$

where

$$h(j|\mathbf{x}_h) = \frac{\eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}} (1-\pi_{ij})^{1-x_{ih}}}{\sum_{k=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ik}^{x_{ih}} (1-\pi_{ik})^{1-x_{ih}}}$$

Maximum likelihood estimation: EM algorithm

The steps of the E-M algorithm are the following:

1. Choose an initial set of posterior probabilities $h(j|\mathbf{x}_h)$
2. Compute a first approximation of $\hat{\eta}_j$ and $\hat{\pi}_{ij}$
3. Obtain improved estimates of $h(j|\mathbf{x}_h)$
4. Return to Step 2 and continue until convergence

In the latent class model there is the problem of likelihood convergence into local maxima. The risk increases as K increases and decreases with increasing sample size. Many different starting values need to be used. The problem of convergence to local maxima occurs especially when the number of variables is large and the sample size is small.

Goodness of fit

As for the latent trait model we consider also in this case the Pearson χ^2 statistic and the likelihood ratio test G

$$\chi^2 = \sum_{r=1}^{2^p} \frac{(nP_r - n\hat{\pi}_r)^2}{n\hat{\pi}_r)}$$

$$G^2 = 2\sum_{r=1}^{2^p} nP_r \ln\left(\frac{nP_r}{n\hat{\pi}_r}\right)$$

where in this case

$$\hat{\pi}_r = f(\mathbf{x}_r) = \sum_{j=0}^{K-1} \hat{\eta}_j \prod_{i=1}^{p} \hat{\pi}_{ij}^{x_{ir}} (1 - \hat{\pi}_{ij})^{(1-x_{ir})}$$

Under regular conditions, χ^2 and G^2 converge to a chi square with $df = 2^p - (K-1) - Kp - 1$.

In presence of sparse data we use the low-order residuals as before.

Latent class model vs latent trait model

Latent class model is special case of latent trait model in which prior distribution consists of discrete probability masses.
$\Rightarrow$ all the general results not depending on the prior can be applied to the latent class model
$\Rightarrow$ The form of the sufficient statistic will be linear combination of the x's.

Consider the case of $K = 2$.
Logit/normal model:

$$\text{logit}(\pi_i(y)) = \alpha_{i0} + \alpha_{i1} y$$

where y has 0 mean and unit standard deviation.
Suppose that y takes two values $\sqrt{(1-\eta)/\eta}$ and $-\sqrt{\eta/(1-\eta)}$ with probabilities η and $1-\eta$ respectively.
$\Rightarrow$ It follows that $E(y) = 0$ and $Var(y) = 1$.

$$\text{logit}\pi_{i0} = \alpha_{i0} - \alpha_{i1}\sqrt{\frac{\eta}{1-\eta}} \quad \text{logit}\pi_{i1} = \alpha_{i0} + \alpha_{i1}\sqrt{\frac{1-\eta}{\eta}}$$

Latent class models for binary data: example1

We apply the latent class model to the attitude to abortion data.
Estimated conditional probabilities and prior probabilities for the two-class model

Item	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$
WomanDecide	0.01 (0.01)	0.71 (0.03)
CoupleDecide	0.09 (0.03)	0.91(0.02)
NotMarried	0.12 (0.04)	0.96 (0.02)
CannotAfford	0.15 (0.04)	0.91 (0.02)
$\hat{\eta}_j$	0.39 (0.03)	0.61 (0.03)

Latent class models for binary data: example1

Chi-squared residuals greater than 3 for the second order margins, for the two class model

	Items	O	E	$O-E$	$(O-E)^2/E$
Response (0,0)	2,1	147	137.79	9.21	0.62
	3,1	131	130.16	0.84	0.05
	3,2	117	117.58	-0.58	0.00
	4,1	129	129.12	-0.12	0.00
	4,2	114	114.61	-0.61	0.00
	4,3	116	109.97	6.03	0.33
Response (0,1)	1,2	66	75.21	-9.21	1.13
	1,3	82	82.84	-0.84	0.01
	1,4	84	83.88	-0.12	0.00
	2,1	7	16.21	-9.21	5.24
	2,3	37	36.42	0.58	0.01
	2,4	40	39.89	0.61	0.01
	3,1	7	7.84	-0.84	0.09
	3,2	21	20.42	0.58	0.02
	3,4	22	28.03	-6.03	1.30
	4,1	16	15.88	0.12	0.00
	4,2	31	30.39	0.61	0.01
	4,3	29	35.03	-6.03	1.04
Response (1,0)	2,1	159	149.79	9.21	0.57
	3,1	159	158.16	0.84	0.01
	3,2	204	204.58	-0.58	0.00
	4,1	150	150.12	-0.12	0.00
	4,2	194	194.61	-0.61	0.00
	4,3	212	205.97	6.03	0.18

$\chi^2 = 44.81, df = 6$ $G^2 = 37.02, df = 4$

Latent class models for binary data: example1

Estimated posterior probabilities of class membership

Response Pattern	$\hat{h}(j=1\|\mathbf{x})$	$\hat{h}(j=2\|\mathbf{x})$	Class allocation
0000	1.00	0.00	1
0001	0.99	0.01	1
0010	0.96	0.04	1
0100	0.98	0.02	1
1000	0.95	0.05	1
0011	0.31	0.69	2
0101	0.45	0.55	2
0110	0.21	0.79	2
1100	0.16	0.84	2
0111	0.00	1.00	2
1011	0.00	1.00	2
1101	0.00	1.00	2
1110	0.00	1.00	2
1111	0.00	1.00	2

Latent class models for binary data: example2

Attitude to science and technology data

The 7 items were originally measured on a four point scale. For the purpose of the present analysis, response categories are dichotomised as follows: categories 'strongly disagree' and 'disagree to some extent' are coded as 0 and categories 'agree to some extent' and 'strongly agree' are coded as 1.

The items Environment, Technology and Industry are negatively expressed. They are recoded so that a high score corresponds to positive attitude towards science and technology.

The number of respondents is 392. There are $2^7 = 128$ possible response patterns. Many patterns do not occur.

First we fit a two-class model.

Latent class models for binary data: example2

Estimated conditional probabilities and prior probabilities for the two-class model

Item	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$
Comfort	0.75 (0.06)	0.95 (0.02)
Environment	0.76 (0.06)	0.68 (0.03)
Work	0.36 (0.09)	0.75 (0.04)
Future	0.20 (0.18)	0.94 (0.04)
Technology	0.74 (0.06)	0.72 (0.03)
Industry	0.84 (0.05)	0.86 (0.02)
Benefit	0.41 (0.09)	0.77 (0.03)
$\hat{\eta}_j$	0.21 (0.07)	0.79 (0.07)

Latent class models for binary data: example2

Chi-squared residuals greater than 3 for the second order margins, for the two class model

	Items	O	E	$O-E$	$(O-E)^2/E$
Response (0,0)	5,1	17	10.05	6.95	4.81
	5,2	52	33.18	18.82	10.67
	6,1	10	5.56	4.44	3.54
	6,2	31	17.23	13.77	11.00
	6,5	26	15.83	10.17	6.53
Response (0,1)	2,5	67	85.82	-18.82	4.13
	5,2	57	75.82	-18.82	4.67
	6,2	26	39.77	-13.77	4.77
	6,5	31	41.17	-10.17	2.51
Response (1,0)	2,5	57	75.82	-18.82	4.67
	2,6	26	39.77	-13.77	4.77
	5,2	67	85.82	-18.82	4.13
	5,6	31	41.17	-10.17	2.51

$\chi^2 = 118.48, df = 4$ $G^2 = 109.03, df = 4$

Latent class models for binary data: example2

Estimated conditional probabilities and prior probabilities for the three-class model

Item	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
Comfort	0.63	0.80	0.95
Environment	0.74	0.29	0.76
Work	0.28	0.92	0.67
Future	0.16	0.95	0.82
Technology	0.74	0.00	0.83
Industry	0.74	0.55	0.92
Benefit	0.00	0.72	0.77
$\hat{\eta}_j$	0.09	0.12	0.79

This differs from the two-class model defined above only in the way we designate the classes and in the parameterisation. For class 1, corresponding to $y = \sqrt{(1-\eta)/\eta}$,

$$\pi_{i1} = \frac{1}{1 + \exp\left\{-\alpha_{i0} - \alpha_{i1}\sqrt{\frac{\eta}{1-\eta}}\right\}},$$

and for class 0,

$$\pi_{i0} = \frac{1}{1 + \exp\left\{-\alpha_{i0} + \alpha_{i1}\sqrt{\frac{1-\eta}{\eta}}\right\}}.$$

These equations relate the class-specific response probabilities of the latent class model to the intercept and slope parameters of the latent trait model. If, for example, we had estimated π_{i0}, π_{i1} and η using the theory for the latent class model, we could express them as latent trait parameters. They will not, of course, be equal to the estimates of α_{i0} and α_{i1} obtained from IRTPRO (Cai *et al.* 2011) because that program assumes a normal prior distribution. How close they turn out to be will give some indication of how sensitive the estimates are to the choice of prior distribution.

We have used this method for the Law School Admission Test data given in Table 4.2. The results are set out in Table 6.1.

The two sets of estimates are remarkably close, especially in the case of the α_{i0}. The first set of estimates uses the normal prior and the second the two-point distribution appropriate for the latent class model. We can interpret this result either by saying that the prior distribution has little effect or that it is difficult to distinguish empirically between the latent trait and latent class models. The reader can confirm the latter interpretation by using the Latent GOLD program (Vermunt and Magidson 2000) to fit a two-class model to the Law School Admission Test data. The predicted

Table 6.1 Parameter estimates for the two-class model fitted to the Law School Admission Test VI data together with a comparison of the estimates of $\{\alpha_{i0}\}$ and $\{\alpha_{i1}\}$ obtained from (6.4) and (6.6) and from the logit/normal model (from Table 4.1).

			Logit/normal estimates		Latent class estimates	
i	$\hat{\pi}_{i0}$	$\hat{\pi}_{i1}$	$\hat{\alpha}_{i0}$	$\hat{\alpha}_{i1}$	$\hat{\alpha}_{i0}$	$\hat{\alpha}_{i1}$
1	0.852	0.965	2.77	0.83	2.75	0.74
2	0.528	0.812	0.99	0.72	0.97	0.64
3	0.301	0.693	0.25	0.89	0.22	0.79
4	0.611	0.849	1.28	0.69	1.27	0.61
5	0.776	0.924	2.05	0.66	2.03	0.59

frequencies are as close to the observed, as in the case of the latent trait model. The $\chi^2 = 19.42$ and the log-likelihood ratio $G^2 = 22.76$ on 20 degrees of freedom, with $P = 0.49$ and 0.30 respectively, indicate a good fit.

There is a second way in which a two-class model can be expressed as a latent trait model. In this case we retain the standard normal prior but choose the item response function to be

$$\pi_i(y) = \begin{cases} \pi_{i0} & \text{if } y \le y_0, \\ \pi_{i1} & \text{if } y > y_0, \end{cases}$$

where $\pi_{i1} > \pi_{i0}$ $(i = 1, 2, \ldots, p)$. All individuals with $y \le y_0$ will thus have the same response probability and the proportion of the population in this class will be $\Phi(y_0) = \eta$, say. The difference between the latent class model formulated in this way and the logit/normal model is in the form of the *response function*. This example illustrates very clearly the intimate relationship between the prior and the response function. We have two substantively different hypotheses which give rise to the same model. In one the prior is normal with a step-function as its item response curve. In the other the prior is a two-point distribution with an arbitrary response function constrained only by the values it takes at two values of y.

The step-function approach can easily be extended to models with more than two latent classes by increasing the number of segments into which the latent scale is divided.

6.4 K latent classes within the GLLVM

We now return to the representation in (6.4) and (6.5) of the two-class model as a latent trait model and extend it to the general case of K classes. It has to be shown that the model as defined in Section 6.2 can be written in the form (4.3) for some q. This was easy to do when $K = 2$ because we could immediately see how to construct a single binary variable with zero mean and unit standard deviation so that the response probabilities coincided. In the general case we proceed as follows. Let $\mathbf{y}$ be an indicator vector showing into which class an individual falls. Thus, for $j = 0, \ldots, K-1$, $y_j = 1$ if the individual is in class j and zero otherwise. It follows that $\sum y_j = 1$. So far as the formal treatment is concerned, we can treat $\mathbf{y}$ just like any vector of latent variables because the general theory places no restrictions whatsoever on the form of their distribution. This would lead to a GLLVM of the form

$$\text{logit}\,\pi_i(\mathbf{y}) = \sum_{j=0}^{K-1} \pi_{ij} y_j \quad (i = 1, 2, \ldots p). \tag{6.6}$$

The prior distribution $h(\mathbf{y})$ is highly degenerate. For $j = 0, \ldots, K-1$ it puts probability η_j at the point where $y_j = 1$ and all other ys are 0.

Polytomous data: latent trait model

The models for polytomous data are a generalization of those for binary data.

c_i number of categories of the variable i with $(i = 1, \ldots, p)$ and $s = 0, \ldots, c_i - 1$. We recode x_i as follows:

$$x_i(s) = \begin{cases} 1, & \text{if the response falls in category } s; \\ 0, & \text{otherwise.} \end{cases}$$

The c_i vector is denoted by $\mathbf{x}_i$ and $\sum_s x_i(s) = 1$. Full response pattern $\mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_p)$.
If $c_i = 2 \Rightarrow$ binary case $(\mathbf{x}_i = x_i(1))$.

As in the binary case we define the response function as

$$P(x_i(s) = 1|\mathbf{y}) = \pi_{is}(\mathbf{y}) \quad s = 0, \ldots, c_i - 1 \quad i = 1, \ldots, p.$$

with $\sum_s \pi_{is}(\mathbf{y}) = 1$.

Polytomous data: latent trait model

In the classical GLLVM x_i and θ_i are supposed to be scalars.

In this case they are vectors. Thus GLLVM generalise to provide a vector-valued GLLVM. We obtain

- Random part
  The conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}$ is taken from the multinomial distribution

$$
\begin{aligned}
g_i(\mathbf{x}_i|\mathbf{y}) &= \prod_{s=0}^{c_i-1} \pi_{is}(\mathbf{y})^{x_{i(s)}} = \\
&= \prod_{s=0}^{c_i-1} \pi_{i0}(\mathbf{y})^{x_{i(s)}} \left(\pi_{is}(\mathbf{y})/\pi_{i0}(\mathbf{y})\right)^{x_{i(s)}} = \\
&= \pi_{i0}(\mathbf{y}) \exp \sum_{s=0}^{c_i-1} x_{i(s)} \ln(\pi_{is}(\mathbf{y})/\pi_{i0}(\mathbf{y})) = \\
&= \pi_{i0}(\mathbf{y}) \exp \boldsymbol{\theta}_i' \mathbf{x}_i
\end{aligned}
$$

Polytomous data: latent trait model

where

$$\boldsymbol{\theta}_i' = \left(0, \ln \frac{\pi_{i1}(\mathbf{y})}{\pi_{i0}(\mathbf{y})}, \ln \frac{\pi_{i2}(\mathbf{y})}{\pi_{i0}(\mathbf{y})}, \ldots, \ln \frac{\pi_{i(c_i-1)}(\mathbf{y})}{\pi_{i0}(\mathbf{y})}\right)$$

The first category of the polytomous variable is arbitrarily selected to be the reference category.

- Systematic part

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha}_{i0} + \boldsymbol{\alpha}_{i1} y_1 + \ldots + \boldsymbol{\alpha}_{iq} y_q$$

- Link function: logit

$$\boldsymbol{\theta}_i = \boldsymbol{\alpha}_{i0} + \boldsymbol{\alpha}_{i1} y_1 + \ldots + \boldsymbol{\alpha}_{iq} y_q$$

  where $\boldsymbol{\alpha}_{ij} = (\alpha_{ij}(0) = 0, \alpha_{ij}(1), \ldots, \alpha_{ij}(c_i - 1))$

Moreover

$$\pi_{is}(\mathbf{y}) = \frac{\exp(\alpha_{i0}(s) + \sum_{j=1}^{q} \alpha_{ij}(s) y_j)}{\sum_{r=0}^{c_i-1} \exp(\alpha_{i0}(r) + \sum_{j=1}^{q} \alpha_{ij}(r) y_j)}$$

Polytomous data: latent trait model

Remarks:

- It is easily to show that the components are given by

$$X_j = \sum_{i=1}^{p} \sum_{s=0}^{c_i - 1} \alpha_{ij}(s) x_i(s)$$

  The X_j is the total score for an individual
- The interpretation of π's is clear from their definition as median response probabilities.
  The α's may be interpreted as in the binary case but, separately, as functions of s and i.
  Each α's can be also interpreted as category scores.
- In the polytomous case the equivalence between UVA and RF breaks down.

Model estimation

The E-M algorithm for binary data is easily generalised to the polytomous case.
As before we start from

$$l = \sum_{h=1}^{n} \ln g(\mathbf{x}_h | y_h)$$

we define the E-step in terms of $E(L|\mathbf{x})$ and we end up to an expression that depends on

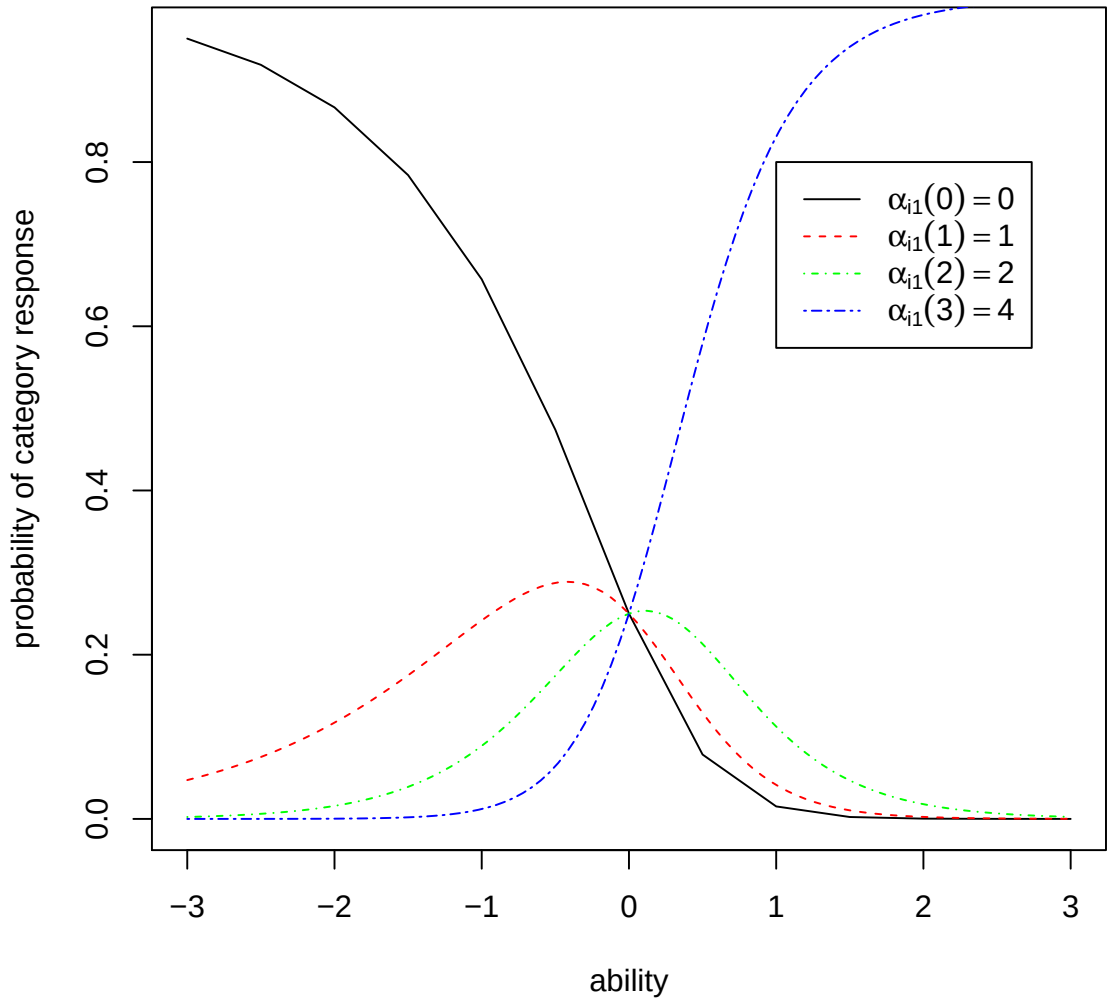
$$r_{it}(s) = \sum_{h=1}^{n} x_{ih}(s) h(y_t | \mathbf{x}_h)$$

$$n_t = \sum_{h=1}^{n} h(y_t | \mathbf{x}_h)$$

In the M-step we maximize the $E(l|\mathbf{x})$ with respect to the parameters. The procedure is the same as in the binary case.

Probabilities of response categories

Probabilities of response categories for $\alpha_{i0}(s) = 0, \quad s = 0, \ldots, 3$



Latent trait model for polytomous data: example

Attitudes to the environment.

- $n = 291$ individuals
- $p = 6$ items (possible responses: 'very concerned','slightly concerned', 'not very concerned', 'not at all concerned')

The items are the following

- Lead from petrol [LeadPetrol]
- River and sea pollution [RiverSea]
- Transport and storage of radioactive waste [RadioWaste]
- Air pollution[AirPollution]
- Transport and disposal of poisonous chemicals [Chemical]
- Risks from nuclear power station [Nuclear]

Since the proportion of individuals falling into the 'not very concerned' is less than 10%, categories 'not very concerned' and 'not al all concerned' were collapsed. We have 3 categories for each items. The first category is assumed as reference category.

Latent trait model for polytomous data: example

The items are assumed to be indicators of individuals' attitudes towards environmental issues.

Items	Category	$\alpha_{i0}(s)$	$\alpha_{i1}(s)$
LeadPetrol	2	-0.75 (0.18)	1.34 (0.27)
	3	-3.06 (0.41)	2.06 (0.45)
RiverSea	2	-2.33 (0.37)	2.11 (0.44)
	3	-8.87 (2.60)	5.06 (1.74)
RadioWaste	2	-2.57 (0.50)	3.31 (0.73)
	3	-6.00 (1.16)	5.14 (1.12)
AirPollution	2	-1.39 (0.35)	3.10 (0.69)
	3	-8.53 (1.62)	6.52 (1.24)
Chemicals	2	-2.83 (0.62)	3.61 (0.95)
	3	-5.43 (1.08)	4.75 (1.20)
Nuclear	2	-0.33 (0.18)	1.54 (0.33)
	3	-1.95 (0.38)	2.82 (0.51)

A response function model for ordinal variables

We consider the case in which each item i has c_i ordered categories. Examples of ordered variables are often attitudinal statements with response alternative such as 'strongly disagree', 'disagree', 'agree' and 'strongly agree'.

As in the polytomous case we refer to the response category probabilities $\pi_{is}(\mathbf{y}), \quad (s = 1, \ldots, c_i)$, that is the probability that, given $\mathbf{y}$, a response falls in category s for the variable i.

As before $\sum_s \pi_{is}(\mathbf{y}) = 1$.

In order to generalize the results of the binary case, we can divide the categories into two groups:

1. Group 1: categories $1, \ldots, s$
2. Group 2: categories $s + 1, \ldots, c_i$

We want to apply the binary logit model. At this aim we define the cumulative probabilities:

$\gamma_{is}(\mathbf{y}) = P(x_i \leq s|\mathbf{y}) = \pi_{i1}(\mathbf{y}) + \pi_{i2}(\mathbf{y}) + \ldots + \pi_{is}(\mathbf{y})$ (corresponding to label 1)

$1 - \gamma_{is}(\mathbf{y}) = P(x_i > s|\mathbf{y}) = \pi_{is+1}(\mathbf{y}) + \pi_{i2}(\mathbf{y}) + \ldots + \pi_{ic_i}(\mathbf{y})$ (corresponding to label 0)

A response function model for ordinal variables

The GLLVM is defined as follows:

- Random part: the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}$ is taken from the multinomial distribution

$$g(\mathbf{x}_i|\mathbf{y}) = \prod_{s=1}^{c_i} \pi_{is}(\mathbf{y})^{x_{is}} = \prod_{s=1}^{c_i} (\gamma_{is}(\mathbf{y}) - \gamma_{i,s-1}(\mathbf{y}))^{x_{is}}$$

  where

$$\gamma_{is}(\mathbf{y}) = \pi_{i1}(\mathbf{y}) + \pi_{i2}(\mathbf{y}) + \ldots + \pi_{is}(\mathbf{y})$$

  and hence $\pi_{i1}(\mathbf{y}) = \gamma_{i1}(\mathbf{y})$ and $\pi_{is}(\mathbf{y}) = \gamma_{is}(\mathbf{y}) - \gamma_{i,s-1}(\mathbf{y})$.
  $x_{is} = 1$ if a randomly selected individual responds to category s of the item ith and $x_{is} = 0$ otherwise.

- Systematic component

$$\eta_{is} = \alpha_{i0(s)} - \sum_{j=1}^{q} \alpha_{ij} y_j$$

A response function model for ordinal variables

- Link function: logit

$$\ln\left(\frac{\gamma_{is}(\mathbf{y})}{1-\gamma_{is}(\mathbf{y})}\right) = \text{logit}(\gamma_{is}(\mathbf{y})) = \eta_{is} = \alpha_{i0(s)} - \sum_{j=1}^{q} \alpha_{ij} y_j \qquad (1)$$

The α_{ij} are interpreted as factor loadings. The negative sign in front of the factor loadings indicates that, as y_j increases, the response on the observed item x_i is more likely to fall at the high end of the scale.
The factor loadings remain the same across categories of the same variable, that is the discriminating power of the item does not depend on the category.

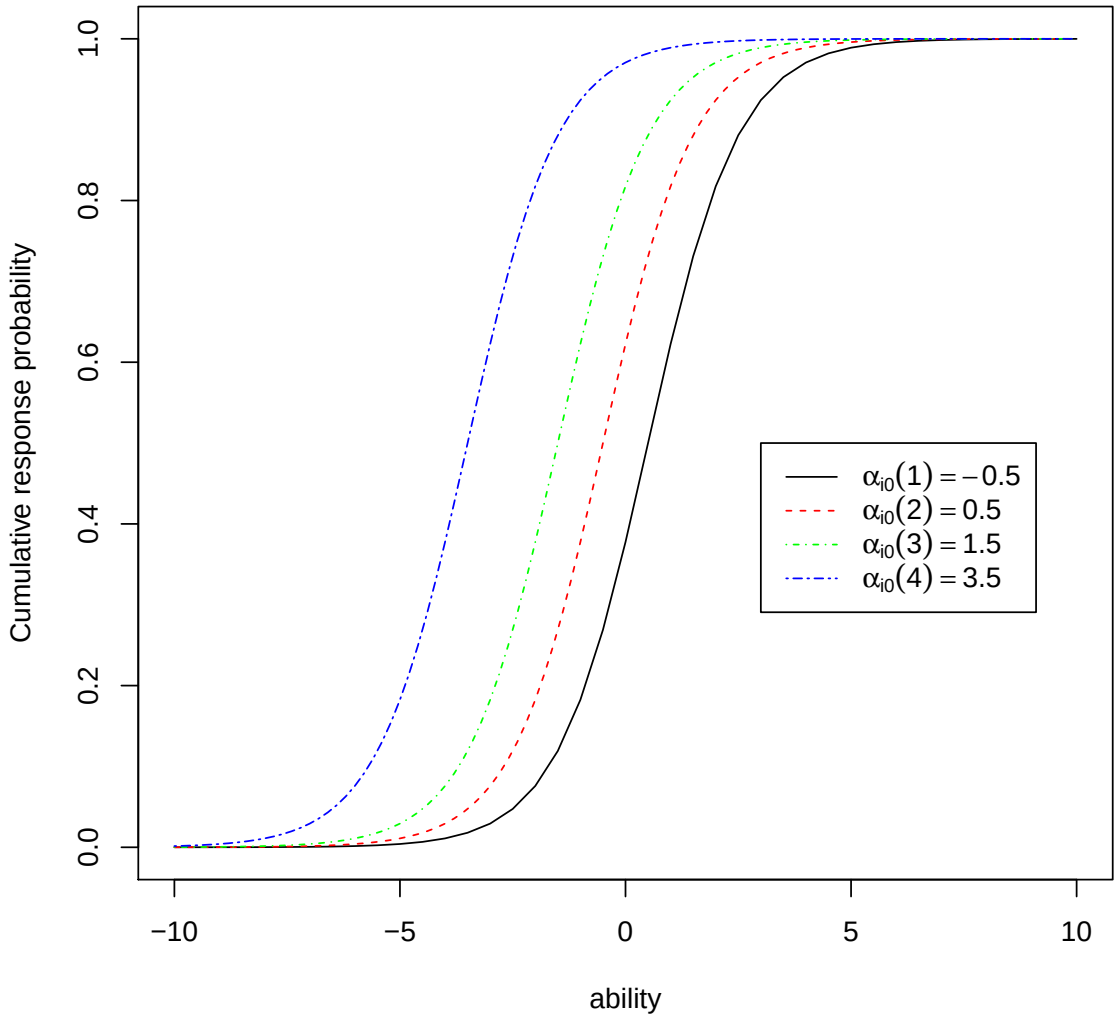
This model is called the *proportional odds model*.

This name comes from the fact that, in the one-factor case, the difference between two cumulative logits, that is, the left side of (1), for two persons with factor scores y_1 and y_2 is proportional to $y_2 - y_1$.
The intercept parameters are category dependent. This reflect the fact that as the threshold increases for the response, the difficulty will be also. Hence $\alpha_{i0(1)} \leq \alpha_{i0(2)} \leq \ldots \leq \alpha_{i0(c_i)} = +\infty$.

Model estimation can be performed via the E-M algorithm. The steps are the same as before.

Cumulative response probabilities

Probabilities of response categories for $\alpha_{i1} = 1.0$



Goodness of fit

As in the binary case we can consider the Pearson χ^2 and the likelihood ratio G^2 in this case defined as

$$\chi^2 = \sum_{r=1}^{C} \frac{(nP_r - n\pi_r(\hat{\boldsymbol{\alpha}}))^2}{n\pi_r(\hat{\boldsymbol{\alpha}})}$$

$$G^2 = 2\sum_{r=1}^{C} nP_r \ln\left(\frac{nP_r}{n\pi_r(\hat{\boldsymbol{\alpha}})}\right)$$

where $C = \prod_{j=1}^{p} c_j$. In this case C is the number of possible response patterns.

Under regular conditions, the two statistics are asymptotically distributed as a chi-square with $C - \sum_j c_j - p - pq - 1$ degrees of freedom.
However, also in this case the sparse problem is very frequent.

Bivariate residuals defined as

$$GF_{fit}^{(ij)} = n\sum_{a=1}^{c_i}\sum_{b=1}^{c_j} \frac{(P_{ab} - \pi_{ab})^2}{\pi_{ab}}$$

where P_{ab} and π_{ab} are the observed and expected bivariate residuals.

Rule of thumb: $GF_{fit}^{(ij)} > 4c_ic_j$ indicates bad fit between item i and item j.

An underlying variable model

Consider the i-th observed variable with c_i ordered categories. We define an underlying standardized normal variable ξ_i such that

$$x_i = s \quad \text{if} \quad \tau_{is-1} \leq \xi_i \leq \tau_{i,s} \quad (s = 1, \ldots, c_i)$$

where $\tau_{i0} = -\infty, \tau_{ic} = \infty$.

As in the binary case, we further assume a standard linear factor model

$$\boldsymbol{\xi} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \boldsymbol{\varepsilon}$$

We assume that the underlying bivariate distributions are normal.

We can fit the factor model provided we can estimate the correlation coefficients.

Estimates of this coefficients are called *polychoric correlations*. They are estimated using the maximum likelihood method.

An underlying variable model

The equivalence found in the binary case holds also in the ordinal case. In particular

$$\alpha_{i0(s)} = \frac{\tau_{i,s-1} - \mu_i}{\psi^{1/2}}, \quad \alpha_{ij} = \frac{\lambda_{ij}}{\psi^{1/2}}$$

The equivalence holds for the probit model and for the logit model in so far as the logit is a good approximation of the probit. As before we obtain

$$st\alpha_{ij} = \frac{\alpha_{ij}}{\sqrt{1 + \sum_j \alpha_{ij}^2}}$$

Parameter interpretation

The proportional odds model does not arise from the sufficiency principle.

The parameter π_{is} is still the probability that the median individual falls in category s on the variable i.

But the α's cannot be interpreted as category scores or as weights in a component.

The posterior density of $\mathbf{y}$ is no longer dependent on a linear function of $\mathbf{x}$.

Proportional odds model: example

Attitude to science and technology

- $n = 392$ respondents
- 7 items
  - Science and technology are making our lives healthier, easier and more comfortable. [Comfort]
  - Scientific and technological research cannot play an important role in protecting the environment and repairing it. [Environment]
  - The application of science and new technology will make work more interesting [Work]
  - Thanks to science and technology, there will be more opportunities for the future generations. [Future]
  - New technology does not depend on basic scientific research. [Technology]
  - Scientific and technological research do not play sn important role in industrial development. [Industry]
  - The benefits of science are greater than any harmful effects it may have. [Benefit]

Possible response categories are 'strongly disagree', 'disagree to some extent', 'agree to some extent', 'strongly agree'.

First analysis: items positive worded that are Comfort, Work, Future, Benefit. They are considered indicators of attitude towards science and technology.

Proportional odds model: example

Attitude to science and technology

Estimated factor loadings with standard errors in brackets and standardized loadings

	α_{i1}	$st\alpha_{i1}$
Comfort	1.04 (0.16)	0.72
Work	1.23 (0.19)	0.77
Future	2.28 (0.32)	0.92
Benefit	1.10 (0.16)	0.74

Sums of chi-square residuals for pairs of items from two-way margins for the one-factor model

	Work	Future	Benefit
Comfort	25.54	11.98	27.23
Work		9.21	23.27
Future			17.41

Chi-squared residuals for the two-way marginals of items Comfort and Work

Categories	1	2	3	4
1	0.87	1.79	0.44	11.50
2	2.43	0.01	0.14	2.09
3	0.02	0.04	0.45	2.39
4	0.66	0.26	1.51	0.94

Proportional odds model: example

Second analysis: we consider all the seven items. First we fit a one-factor model.

Sums of chi-squared residuals for pair of items from the two-way margins for the one-factor model

Items	2	3	4	5	6	7
1	13.33	26.20	12.41	18.93	9.21	24.51
2		27.54	24.09	98.36	90.76	23.52
3			10.28	21.08	35.78	23.11
4				23.23	26.90	17.17
5					103.24	17.09
6						20.28

Proportional odds model: example

Second analysis: we consider all the seven items. First we fit a second-factor model.

Items	Category	$\alpha_{i0}(s)$	$\alpha_{i1}(s)$	$\alpha_{i2}(s)$
Comfort	1	-5.00 (2.09)	0.27 (0.20)	1.16 (0.18)
	2	-2.74 (1.77)		
	3	1.53 (0.33)		
Environment	1	-3.45 (1.20)	1.61 (0.35)	0.09 (0.22)
	2	-1.26 (0.87)		
	3	0.99 (0.71)		
Work	1	-2.95 (0.86)	-0.39 (0.35)	1.20 (0.23)
	2	-0.90 (0.83)		
	3	2.30 (0.45)		
Future	1	-5.05 (1.92)	-0.30 (0.28)	2.16 (0.34)
	2	-2.13 (1.43)		
	3	1.90 (0.62)		
Technology	1	-4.17 (1.60)	1.71 (0.36)	0.08 (0.24)
	2	-1.49 (1.04)		
	3	1.07 (0.70)		
Industry	1	-4.71 (1.60)	1.55 (0.31)	0.55 (0.25)
	2	-2.53 (1.30)		
	3	0.45 (0.59)		
Benefit	1	-3.38 (1.46)	-0.08 (0.00)	1.12 (0.20)
	2	-1.00 (0.72)		
	3	1.71 (0.39)		

Proportional odds model: example

Sums of chi-squared residuals for pair of items from the two-way margins for the two-factor model

Items	2	3	4	5	6	7
1	12.36	25.14	11.99	18.80	7.37	24.58
2		21.64	23.29	20.94	31.86	22.85
3			9.44	15.31	30.54	23.64
4				21.93	26.71	17.23
5					34.87	16.67
6						20.78

Latent class models with polytomous manifest variables

- Unordered polytomous manifest variables
  As before, when $\mathbf{x}_i$ has c_i categories we define

$$x_i(s) = \begin{cases} 1, & \text{if the response falls in category } s; \\ 0, & \text{otherwise.} \end{cases}$$

  The c_i vector is denoted by $\mathbf{x}_i$ and $\sum_s x_i(s) = 1$. Full response pattern $\mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_p)$.
  $\pi_{ij}(s)$ probability of an individual in class j is in category s on variable i

$$f(\mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)}$$

$$h(j|\mathbf{x}) = \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)} / f(\mathbf{x})$$

Latent class models with polytomous manifest variables

- Ordered polytomous manifest variables

  To preserve the ordinality property of the variable x_i, as before we define

$$\gamma_{ijs} = \pi_{ij}(s) + \pi_{ij}(s+1) + \ldots + \pi_{ij}(c_i - 1)$$

$$f(\mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)} = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} (\gamma_{ij}(s) - \gamma_{ij}(s+1))^{x_i(s)}$$

Maximum likelihood estimation

The log-likelihood is written as

$$L = \sum_{h=1}^{n} \ln f(\mathbf{x}_h)$$

Now two sets of constraints:

$$\sum \eta_j = 1 \quad \sum_{s=0}^{c_i - 1} \pi_{ijs} = 1$$

The function to be maximised is

$$\phi = L + \gamma \sum_{j=0}^{K-1} \eta_j + \sum_{j=0}^{K-1} \sum_{i=1}^{p} \beta_{ij} \sum_{s=0}^{c_i - 1} \pi_{ijs}$$

where $\hat{\eta}_j = 1/n \sum_{h=1}^{n} h(j|\mathbf{x}_h)$ and $\hat{\pi}_{ijs} = \frac{\sum_{h=1}^{n} h(j|\mathbf{x}_h) x_{ihs}}{n\hat{\eta}_j}$

Identifiability

The latent class model may be not identifiable $\Rightarrow$ more than one point in the parameter space can yields the same likelihood.

If we consider the contingency table with $c_1 \times c_2 \times \ldots \times c_p$ cells, the independent probabilities are $\prod_{i=1}^{p} c_i - 1$.

These cell probabilities are function of the model parameters. If the model parameters are greater than the independent cell probabilities $\Rightarrow$ identification problem.
The number of the model parameters are

$$K \sum_{i=1}^{p} (c_i - 1) + K - 1$$

The model is unidentifiable if

$$\prod_{i=1}^{p} c_i - 1 < K \sum_{i=1}^{p} (c_i - 1) + K - 1$$

This is a necessary but not sufficient condition (probabilities are subject to other constraints) $\Rightarrow$ local identifiability.

Software

R: https://CRAN.R-project.org/view=Psychometrics

- factor analysis: factanal
- Latent trait analysis: ltm
- Underlying variable approach: lavaan
- Latent class analysis: randomLCA, poLCA

Mplus (http://www.statmodel.com/)

GLLAM (http://www.gllamm.org/)

LATENT GOLD

IRTPRO