Alma Mater Studiorum - Università di Bologna

Departement of Statistical Science
Curriculum Data Science

Latent Variables

Presented by:
Alberto Trashaj

Second edition 2022-2023

Contents

Chapter 1

Introduction

Latent variable models are a powerful and versatile class of statistical models that play a crucial role in various fields, including machine learning, statistics, psychology, and economics. These models are designed to capture hidden or unobservable factors, referred to as latent variables, which influence the observed data. Latent variable models provide a framework for understanding complex relationships within data and are widely used for tasks such as dimensionality reduction, clustering, and probabilistic modeling.
At the heart of latent variable models is the recognition that not all aspects of a system or process are directly measurable or observable. There exist underlying factors that influence the observed variables, and these hidden variables often provide a more nuanced and comprehensive understanding of the data. The incorporation of latent variables allows researchers and practitioners to model complex systems in a more realistic and interpretable manner.
Latent variable models come in various forms, including probabilistic graphical models, factor analysis, and latent class models, each tailored to address specific challenges in different domains. In probabilistic graphical models, for instance, nodes represent variables, both observed and latent, and edges encode dependencies between them. Factor analysis aims to identify a smaller number of latent factors that can explain the observed variability in the data. Latent class models, on the other hand, focus on identifying unobserved subgroups within a population based on patterns of observed variables.

These models have found applications in diverse areas such as image and speech processing, social sciences, finance, and healthcare. In the realm of machine learning, latent variable models contribute to unsupervised learning tasks, allowing for more robust representations of complex data structures.

In this exploration of latent variable models, we will delve into the fundamental concepts, methodologies, and applications of these models. Understanding latent variable models not only provides a powerful analytical toolset but also offers deeper insights into the underlying structures of complex systems, contributing to advancements in both research and practical problem-solving.

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

Table 1.1: Example Table

1.1 Objective of the analysis

Our aim will be to identify the underlying factors that explain relationship among the observed variables: fitting a specific latent variable model. In particular, the latent variables can be metrical or categorical, but also the manifest variables can be metrical or categorical.
Therefore, depending on the combination of the type of variables, we have different analysis:

1.2 Theoretical framework

Now, let's build our notation in order to be consistent and to not get confused when things are going to get more complicated: we will call

- the set of $x_1, x_2, ..., x_p$ manifest variables
- the set of $y_1, y_2, ..., y_q$ latent variables

So remember, p is the dimensionality of the manifest variables and q is the dimensionality of the latent variables.
Since both of them are random variables, we can express the relationship between them in terms of probability distributions: the domain of the random variables will be

- R_x for $\mathbf{x}$
- R_y for $\mathbf{y}$
-

Now, we can define the prior distribution, the conditional distribution and the joint distribution of our model:

- Prior distribution of the latent variables:

$$h(y),\ y = (y_1, y_2, ..., y_q) \in R_y$$

- Conditional distribution of the observed variables x given y

$$g(x|y),\ x = (x_1, x_2, ..., x_p) \in R_x$$

- Joint distribution of the observed and latent variables x and y

$$f(x, y) = h(y)g(x|y)$$

Since only x can be observed any inference must be based on the joint marginal distribution of x:

$$f(x) = \int_{R_y} g(x|y)h(y)dy$$

And from the Bayes theorem we obtain the posterior distribution of the latent variables given the observed variables,

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

To find the posterior distribution that we just wrote, we need to find all the elements in the right part of the equation: although, only $f(x)$ can be estimated from the observed variables, whereas the other temrs are not uniquely determined. So, we will make some assumptions in order to restrict the classes of functions considered: in fact, the first two equations do not specify a model, but they just represent the fact that x and y are random variables mutually dependent on one another.

1.3 Assumption of conditional independence

If the dependencies among the xs are induced by a set of latent variables y then when all ys are accounted for, the xs will be independent if all the ys are held fixed.

Chapter 2

Normal linear factor model with example

I personally started my journey on latent variables recently and I struggled a lot on understanding not the mathematichs behind but particularly the concrete real world application of what I was studying: so, I did a bit of research and I found out that there are many applications involving the study of the latent variables. I found particularly fascinating the biostatistics prevalence of this topic: so, I want to begin this dissertation, in which I will try to explain to you what we are speaking about to increase my understaning in a Feymann fashion, with an example taken from the biology field.

2.1 Plant's height

Let's consider a population of a certain plant species that exhibits variability in height. Researchers are interested in understanding the genetic factors influencing plant height. The genetic variation is assumed to be governed by Quantitative Trait Loci (QTLs), which are specific genomic regions associated with variations in the trait.
In genetics, quantitative traits like plant height are often influenced by multiple genes, each contributing in a quantitative manner. These specific genomic regions associated with quantitative traits are called Quantitative Trait Loci (QTLs). QTLs play a key role in shaping the observed variability in traits within a population.
The height of a plant is a complex trait influenced by a combination of genetic and environmental factors. In this example, we focus on the genetic aspect and aim to uncover the specific genetic factors (QTLs) that contribute to the observed variation in plant height.

2.1.1 Objectives of the Study:

1. Identify QTLs: Pinpoint specific genomic regions associated with variations in plant height.

2. Understand Genetic Effects: Investigate how latent genetic effects contribute to the overall genetic variation in plant height.

2.1.2 Hypotetical experiment

We can imagine the following scenario: we select a diverse population of the crop species with known genetic variability and we ensure that there is a representative sample that captures a range of plant heights.
Then we conduct what is known as "genotyping" to identify genetic markers distributed across the genome, so we can build a dataset of genetic marker information; now we can finally measure the heights of the individuals plants in the population and create a database with this information.

The researcher can observe and measure the height of the plants: this quantitative measurement serves as the observed variable x in the statistical model.

2.2 Data scientists are coming!

Now that our beloved biologists friends did a great job in measuring the heights of the plants, we have a database which can be analyzed with our latent variables models!

2.2.1 Why a latent variable model?

The use of a latent variable model in the context of QTL mapping is driven by the desire to explicitly model unobservable genetic effects (latent variables) that contribute to observed trait variation (height). Latent variable models offer a flexible framework to represent complex relationships and account for unobserved factors that influence the traits of interest: genetic traits, especially in quantitative genetics, are often influenced by multiple genes, each contributing to a small extent. Latent variables allow us to model the collective effects of these genes, providing a more comprehensive representation of genetic influence.

2.2.2 Which model are we going to use?

We are going to use this example to introduce the Normal linear factor model.

The Normal Linear Factor Model is a specific type of latent variable model that assumes a linear relationship between latent variables and observed variables. It is commonly used when dealing with continuous observed variables and is particularly well-suited for modeling complex structures in multivariate data.

That seems our case!

Usually, the biological data presents a complex structure and our observed variable can be assumed to be continuous. So, let's delve into some key factors of this model:

- Linear relationship: this model assumes that there is a linear relationship between latent variables and observed variables. In our case the latent genetic effects (represented by latent variables) contribute additively to the observed variation in plant height
- Normality assumption: both the latent variables and the residuals (unexplained variability) are assumed to follow a multivariate normal distribution. This normality assumption simplifies statistical inference and estimation procedures, facilitating the use of maximum likelihood estimation.

In other, more concise, words we can consider the following model:

$$X = \mu + \Lambda y + e \tag{2.1}$$

where

- $y \sim N_q(0, I)$
- $e \sim N_p(0, \Psi)$
- $y \perp e$
- Λ is a $p \times q$ matrix called "matrix of factor loadings"
- Ψ is a $p \times q$ matrix diagonal called "matrix of specific variances"

In our example:

- x is the observed plant height
- μ is the overall mean height
- Λ is the loading matrix representing the weights of the latent genetic effects on plant height.
- y is the latent genetic effects (unobservable)
- e is the environmental effects and measurement error.

This model that we just introduced can be written also in the following way that someone can find more easy to digest:

$$x|y \sim N_p(\mu + \Lambda y; \Psi) \tag{2.2}$$

and

$$y \sim N_q(0; I) \tag{2.3}$$

Why someone can find more interesting writing the model in this way? The reason why is because one can notice that the conditional distribution $(x|y)$ is normally distributed and therefore it belongs to the exponential family.

What is the exponential family?

Unfortunately, I don't have much time to cover this topic, since probably you already found the exponential families in other context before.
In particular the conditional distribution takes the following form:

$$g_i(x_i|y) = \frac{1}{\sqrt{2\pi\Psi_{ii}}} exp\{-\frac{1}{2\Psi_{ii}}(x_i - \mu_i - \sum_{j=1}^{q} \lambda_{ij} y_j)^2\} \tag{2.4}$$

A brief jump on generalized linear models

In this model we are assume a GLM for each component of the problem: in particular, a GLM consists of three components

1. Random component: for each conditional distribution, one convenient family of distributions is the one-parameter exponential family

$$g_i(x_i|\theta_i) = F_i(x_i)G_i(\theta_i)exp(\theta_i u_i(x_i)) \tag{2.5}$$

2. Systematic component: the y's produce a linear predictor η_i given by

$$\eta_i = \alpha_{i0} + \alpha_{i1} y_1 + ... + \alpha_{iq} y_q \tag{2.6}$$

3. A link function: between the random and the systematic component the link function relates the θ_i (parameters) to the latent variables y's.
   For example:

$$\theta_i = \alpha_{i0} + \alpha_{i1} y_1 + ... + \alpha_{iq} y_q \tag{2.7}$$

2.2.3 Back to our problem

Now we are ready to formalize the three component of the GLM (which from now on will be GLLVM, generalized linear latent variable model):
the linear predictor of the conditional distribution $g_i(x_i|y)$ is

$$\eta_i = \mu_i + \sum_{j=1}^{q} \lambda_{ij} y_j \tag{2.8}$$

which can be seen as

$$\eta_i - \mu_i = \sum_{j=1}^{q} \lambda_{ij} y_j \tag{2.9}$$

and substitute into the original conditional distribution:

$$g_i(x_i|y) = \frac{1}{\sqrt{2\pi\psi_{ii}}} exp\{-\frac{1}{2}(\frac{x_i^2}{\psi_{ii}} - \frac{2x_i\eta_i}{\psi_{ii}} + \frac{\eta_i^2}{\psi_{ii}})\} \tag{2.10}$$

and if ψ_{ii} is known the equation can be seen as part of the exponential family where

- $\theta = \frac{\eta_i}{\sqrt{\psi_{ii}}}$

- $u_i(x_i) = \frac{x_i}{\sqrt{\psi_{ii}}}$
- $G_i(\theta_i) = exp(-\frac{1}{2}\frac{\eta_i^2}{\psi_{ii}})$
- $F_i(x_i) = \frac{1}{\sqrt{2\pi\psi_{ii}}exp(-\frac{1}{2}\frac{x_i^2}{\psi_{ii}})}$

and the sufficient statistics are

$$X_j = \frac{\sum_{i=1}^{p} \lambda_{ij} x_i}{\sqrt{\psi_{ii}}}, j = 1, ..., q \tag{2.11}$$

Then we can say that

$$x \sim N_p(\mu, \Lambda\Lambda' + \Psi) \tag{2.12}$$

and

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2.13}$$

Therefore, by applying the Bayes theorem we can find the posterior distribution which will have the following form

$$y|x \sim N_q(\Lambda'(\Lambda\Lambda' + \Psi)^{-1}(x - \mu), (\Lambda'\Psi^{-1}\Lambda + I)^{-1}) \tag{2.14}$$

Before applying this model to the data, let's see some properties of the model.

2.2.4 Properties of the linear factor model

1. The variance of the observed variables x_i is a combination of the variance explained by the latent factors (y) and the unique variance e not accounted for by the latent factors:

$$Var(x_i) = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i \tag{2.15}$$

   which the first term is referred as "communality" and the second as "specific variances".
   In the context of plant height, this equation suggests that the variability in observed plant height is a combination of the genetic effects represented by latent variables (y) and other unobserved or unaccounted influences (Ψ).
   $\Lambda\Lambda'$: this term represents the contribution of the latent genetic effects (represented by $y$to the variance of plant height. The higher the values in $\Lambda\Lambda'$ the more impact the corresponding latent variables have on the observed trait.
   ψ_{ii}: this term represents the specific variance of the observed variable not explained by latent factors. It includes measurement error and other unobserved environmental influences. A higher ψ_{ii} indicates a larger contribution of unaccounted factors to the observed variability.

2. Covariance between x and y

$$E(x - \mu)y' = E[E(x - \mu)|yy'] = E(\Lambda yy') = \Lambda \tag{2.16}$$

the factor loadings can be interpreted as covariances between the manifest variables and the factors.
Covariance between x(plant height) and y (latent genetic effects) captures the extent to which changes in plant height are associated with changes in the latent genetic factors. Λ_{ij} signifies the covariance between the i-th QTL (latent variable) and the observed plant height. A higher Λ_{ij} implies a stronger influence of the latent genetic factor on the corresponding plant height component. Each element Λ_{ij} in the loading matrix represents the covariance between the observed variable $x_i$and the latent factor y_j.

3. Correlations are given by

$$Corr(x_i, y_j) = \{diag\Sigma\}^{-\frac{1}{2}}\Lambda \tag{2.17}$$

The resulting matrix provides a compact representation of the correlations between plant height variables and latent genetic factors. It highlights the relationships between each observed variable and each latent variable.

4. Covariance between the components and the factors

$$Cov(x - E(X), y') = E(x - E(x))y' = \Lambda'\Psi^{-1}E[(x - \mu)y'] = \Lambda'\Psi^{-1}\Lambda \tag{2.18}$$

$\Lambda'\Psi^{-1}\Lambda$ represents the covariance between the centered plant height variables and the latent genetic factors. Each element $(\Lambda'\Psi^{-1}\Lambda)_{ij}$ signifies the covariance between the i-th plant height component and the j-th latent genetic variable (QTL).
Note that if the resulting matrix is diagonal there are no correlations among components and factors.

2.2.5 Constraint on the model

In the context of the Linear Factor Model, the continuous nature of the latent variable y implies that any one-to-one transformation from y to v does not affect the overall model $f(x)$. While such transformations may alter the specific forms of $h(y)$ and $g(v)$, the relationship between latent and observed variables remains unchanged. The indeterminacy in h allows for the adoption of a metric for y that results in a convenient form. A common and convenient choice is to constrain h so that the resulting distribution of y follows a standard normal distribution. This choice simplifies the interpretation and estimation of parameters in the model.

The mention of indeterminacy referring to rotation implies that even if the loading matrix Λ undergoes an orthogonal transformation, the resulting model remains indistinguishable. This rotation does not impact the fundamental relationships between latent and observed variables. Therefore, the flexibility in choosing a metric for y and the constraint to a standard normal distribution enhance the interpretability and estimation procedures in the Normal Linear Factor Model.

Statistically speaking, if

$$v = My \tag{2.19}$$

it follows that

$$v \sim N(0, I) \tag{2.20}$$

and

$$x|v \sim N(\mu + \Lambda M' v; \Psi) \tag{2.21}$$

So, this model is indistinguishable from the first one

$$\Sigma = \Lambda M' M \Lambda' + \Psi = \Lambda\Lambda' + \Psi \tag{2.22}$$

Therefore, it's impossibile to distinguish between a model with loading matrix Λ and one with $\Lambda M'$.

Addressing indeterminacy and placing constraints on model parameters is crucial for model identifiability and interpretability: indeterminacy in the model refers to situations where multiple sets of parameter values can generate the same observed data, making it challenging to uniquely estimate the model parameters.

Remove the indeterminacy

1. Diagonal constraint: if $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal, it implies that the elements outside the main diagonal are zero. This condition ensures that the latent variables y are independent a posteriori, removing the freedom to arbitrarily rotate the loading matrix Λ.

2. Confirmatory factor analysis: specific elements of the loading matrix Λ can be constrained to zero. This allows for a more targeted and interpretable representation of the relationships between latent and observed variables.

3. Standardizing x's: standardizing the observed variables x's is another strategy. This involves ensuring that any change of scale in x is reflected in the covariance matrix, and, consequently, in the parameter estimates and their interpretations. Specifically, the sum of the squares of the loading matrix elements for each variable is set to one.

$$\sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i = 1 \tag{2.23}$$

In our specific example, the diagonal constraint can ensures that the genetic effects y are independent, enhancing the clarity of the genetic factors contributing to plant height variability; specific elements of the loading matrix Λ could be constrained to reflect known genetic relationships or interactions. Then, standardizing plant heights ensure that any scaling changes are appropriately accounted for in the covariance matrix, aiding in the meaningful interpretation of genetic effects.

2.2.6 Estimation methods

In classical statistical methods, particularly focusing on maximum likelihood estimation, the primary objective is to determine parameter estimates that minimize the discrepancy between the observed covariance matrix S and the theoretically implied covariance matrix Σ from the model.

The goal of these estimation methods is to find parameter values that bring the model's predicted covariance matrix as close as possible to the observed covariance matrix obtained from the data. This process involves optimizing the parameters to minimize the difference between the model's predictions and the actual data.

In the context of maximum likelihood estimation, the procedure aims to identify parameter values that maximize the likelihood of observing the given data. The likelihood represents the probability of obtaining the observed data under the assumed statistical model.

Overall, the focus is on achieving parameter estimates that result in the best possible fit between the model's predictions and the empirical covariance structure, ensuring that the model accurately captures the relationships present in the data.

In particular, in this model our aim will be to find the parameters of:

- Loading matrix Λ
- Specific variances ψ
- Mean vector μ

Maximum likelihood estimation

To find the maximum likelihood estimation we proceed as usual: find the log-likelihood function, compute the derivative with respect to the parameters and set it to 0 to find the maximum.

Likelihood function

Since we are in the normal case, I will skip some tedious computation because they might be cumbersome and you can find them everywhere:
let $x \sim N_p(\mu, \Sigma)$ and a random sample $x_1, ..., x_n$, the log-likelihood can be written as

$$l(x_1, ..., x_n, \mu, \Sigma) = -\frac{n}{2} ln\{(2\pi)^p |\Sigma|^{\frac{1}{2}}\} - \frac{1}{2} \sum_{h=1}^{n} (x_h - \mu)' \Sigma^{-1} (x_h - \mu) \tag{2.24}$$

Let

$$S(\mu) = \sum_{h=1}^{n} \frac{(x_h - \mu)(x_h - \mu)'}{n} \tag{2.25}$$

and the log-likelihood can be written as

$$l(x_1, ..., x_n, \mu, \Sigma) = c + \frac{n}{2}\{ln|\Sigma^{-1}| - trace[\Sigma^{-1}S]\} \tag{2.26}$$

where c is a constant.
The MLE of $\hat{\mu} = \bar{x}$, i.e. the sample mean: then the sample covariance matrix is

$$S(\hat{\mu}) = \sum_{h=1}^{n} \frac{(x_h - \hat{\mu})(x_h - \hat{\mu})'}{n} \tag{2.27}$$

Now it's time to estimate Λ and Ψ, we proceed by maximising the log-likelihood function and taking the derivative with respect to Λ and Ψ:

1.
$$\frac{\partial l(x, \Lambda, \Psi)}{\partial \Lambda} = 0 \tag{2.28}$$

2.
$$\frac{\partial l(x, \Lambda, \Psi)}{\partial \Psi} = 0 \tag{2.29}$$

And we obtain the following solutions:

$$S\hat{\Psi}^{-1}\hat{\Lambda} = \Lambda(I + \hat{\Lambda}'\hat{\Psi}\hat{\Lambda}) \tag{2.30}$$

$$\hat{\Psi} = diag(S - \hat{\Lambda}\hat{\Lambda}') \tag{2.31}$$

unfortunately, since the complexity of the expressions that makes it hard to find closed form solutions, iterative numerical procedures are applied. These algorithms iteratively update the parameter estimates in an attempt to optimize the likelihood function until a convergence criterion is met.

EM-algorithm

The EM algorithm is a popular (in statistics, not on Instagram) method used as a numerical optimization algorithm: it consists of two steps

1. E-step: in this step we compute the expected value of the joint log-likelihood of (x_h, y_h) conditional on the x_h.

2. M-step: in this step we maximise the modified log-likelihood with respect to the parameters of the model

Let's apply the first step:

- for μ vector we have:

$$E[\frac{\partial l}{\partial \mu}|x_h] = nE[\Psi^{-1}(\bar{x} - \mu - \Lambda\bar{y})|x_h] \tag{2.32}$$

- for Λ we have:

$$E[\frac{\partial l}{\partial \Lambda}|x_h] = nE[\Psi^{-1}(S'_{xy} - \mu\bar{y} - \Lambda S'_{yy})|x_h] \tag{2.33}$$

- for Ψ we have:

$$E[\frac{\partial l}{\partial \Psi}|x_h] = -\frac{n}{2}E[\Psi^{-1} - \Psi^{-1}()] \tag{2.34}$$

But the only quantities that depends on the random variables y's are the sufficient statistics, so it is enough to compute the conditional expected values only for them:

$$\hat{y} = \frac{1}{n}\sum_h y_h \tag{2.35}$$

$$S'_{xy} = \frac{1}{n}\sum_h x_h y'_h \tag{2.36}$$

$$S'_{yy} = \frac{1}{n}\sum_h y_h y'_h \tag{2.37}$$

Computing the conditional expected values we have

$$E[\bar{y}|x_h] = \Lambda'\Sigma^{-1}(\bar{x} - \mu) = \hat{\bar{y}} \tag{2.38}$$

$$E[S'_{xy}|x_h] = \hat{S'}_{xy} \tag{2.39}$$

$$E[S'yy|x_h] = \hat{S'}_{yy} \tag{2.40}$$

Those last results can be replaced inside the formulas 2.32, 2.33, 2.34 and setting the conditioned score functions to zero, i.e. $E(S_\mu) = 0$, $E(S_\Lambda) = 0$ and $E(S_\Psi) = 0$ we obtain the following results:

$$\hat{\mu} = \bar{x} - \Lambda\hat{\bar{y}} \tag{2.41}$$

$$\hat{\Lambda} = (\hat{S'_{xy}} - \bar{x}\,\hat{\bar{y}})(\hat{S'_{yy}} - \hat{\bar{y}}\,\hat{\bar{y}})^{-1} \tag{2.42}$$

$$\hat{\Psi} = diag(S'_{xx} - \hat{\mu}\bar{x}' - \bar{x}\hat{\mu}' - \hat{S'_{xy}}\hat{\Lambda}' - \hat{\Lambda}\hat{S'_{yx}} + \hat{\mu}\hat{y}'\hat{\Lambda}' + \hat{\mu}\hat{\mu}' + \hat{\Lambda}\hat{S'_{yy}}\hat{\Lambda}') \tag{2.43}$$

The EM
This part need to be finished

2.2.7 Evaluation of the model

Goodness of fit and choice of q

The goodness of fit in the Normal Linear Factor Model can be assessed using the likelihood ratio statistic: compare the fit of the hypothesized model H_0 against the unconstrained model H_1.

$$H_0 : \Sigma = \Lambda\Lambda' + \Psi \tag{2.44}$$

$$H_1 : \Sigma unconstrained \tag{2.45}$$

The likelihood ratio statistics is given by

$$-2[l(H_0 - l(H_1)] = n(trace(\Sigma^{-1}S) - log|\Sigma^{-1}S| - p) \tag{2.46}$$

and under the null hypothesis, the statistics follows a chi-square distribution χ^2 :

$$-2[l(H_0 - l(H_1)] \approx \chi^2 \tag{2.47}$$

with degrees of freedom given by

$$df = \frac{1}{2}[(p-q)^2 - (p+q)] \tag{2.48}$$

In our QTL mapping example for plant height, we can apply the likelihood ratio test to compare the fit of models with different latent variable dimensions. The model with the most appropriate q will yield a higher likelihood ratio and a better fit to the observed data, reflecting the underlying genetic factors influencing plant height.

Comparing models

When comparing different models in statistical modeling, information criteria provide a principled way to balance goodness of fit and model complexity. Two widely used criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

1. AIC:

$$AIC = -2L + 2\nu \tag{2.49}$$

   The AIC penalizes models with a higher number of parameters, aiming to find a balance between model fit and simplicity. Lower AIC values indicate a better trade-off between fit and complexity.

2. BIC:

$$BIC = -2l + \nu ln(n) \tag{2.50}$$

   Similar to the AIC, the BIC penalizes models for increased complexity. However, the penalty term in the BIC is stronger than in the AIC, making the BIC more stringent in favor of simpler models.

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

In summary:

- Lower AIC values suggests better-fitting models;
- Lower BIC values indicate better model fit;
- A trade-off between lies in finding a model that explains the data while avoiding unnecessary complexity.

Hypothetical example

Imagine we are testing latent variable models with different values of q to understand the genetic factors affecting plant height in a population. We fit models with $q = 1, ..., 5$ and want to determine which q provides the best balance between model fit and complexity.
Then we fit a latent variable model with all the q's where each model includes latent genetic effects that captures different aspects of genetic variation: now we compute the AIC and the BIC and we select the optimal q which is the one that minimizes both AIC and BIC.

Let's say for example that the model with $q = 3$ minimizes both the criterion: $q = 3$ is the best compromise between explaining the genetic factors influencing plant height and avoiding unnecessary complexity.

2.2.8 Identifiability

Identifiability is a crucial concept in latent variable modeling to ensure that the parameters of the model can be uniquely estimated from the observed data.
Identifiability requires that there are at least as many sample statistics as there are parameters in the model: for a latent variable model with q factors the number of parameters is $pq + p$.
The total number of parameters can be greated than the number of variances and covariances in the sample covariance matrix S, denoted by $\frac{p(p+1)}{2}$, which are the number of elements in a $p \times p$ matrix.

Remember the constraint on the model? If $\Lambda\Psi^{-1}\Lambda$ is diagonal, $\frac{1}{2q(q-1)}$ are imposed, which leads to a number of free parameters of

$$pq + p - \frac{1}{2q(q-1)} \tag{2.51}$$

A necessary but not sufficient condition for consistent estimation is expressed as

$$\frac{1}{2p(p+1)} - pq - p + \frac{1}{2q(q-1)} \geq 0 \tag{2.52}$$

this condition is interesting because it implies an upper bound on the number of factors q that can be reliably estimated.
This upper bound ensures that the latent variable model remains identifiable and the estimation process is consistent.

2.2.9 Scale-invariant estimation

Factor models can be fitted using the sample correlation matrix instead of the covariance matrix, allowing for scale-invariant estimation. This approach involves using correlations of standardized variables, which proves advantageous for various reasons.
The key benefit is that changing scale has no effects on the analysis:
let's consider the already described factor model equation

$$x = \mu + \Lambda y + e \tag{2.53}$$

and let's take the following transformation of the observed variables

$$x^* = Cx \tag{2.54}$$

where C is a diagonal matrix with positive elements, then the equation becomes

$$x^* = C\mu + C\Lambda y + Ce \tag{2.55}$$

Now, let's compute the variance of this transformed model

$$var(x^*) = C\Lambda\Lambda' C + C\Psi C \tag{2.56}$$

and if $C = diag(\Sigma)^{-\frac{1}{2}}$ where Σ is the sample covariance matrix we obtain

$$x^* = \mu^* + \Lambda^* y + e^* \tag{2.57}$$

where the variance is

$$var(x^*) = \Lambda^* \Lambda^{*'} + \Psi^* \tag{2.58}$$

with

- $\Lambda^* = C\Lambda$
- $\Psi^* = C\Psi C$

For example, in our case of QTL mapping, the choice of scale-invariant estimation proves valuable for robust and versatile analyses: In a diverse population of a plant species, the observed heights x might inherently have varying scales due to factors like genetic diversity or environmental influences. Instead of relying on the covariance matrix, which is sensitive to the scale of variables, the biologyst can opt for the sample correlation matrix.
In summary, applying scale-invariant estimation techniques to the plant height QTL mapping example enhances the robustness and adaptability of the factor model. Researchers can confidently explore genetic influences on plant height while mitigating challenges posed by variations in observation scales within diverse plant populations.

2.2.10 Heywood case

In factor analysis the Heywood case refers to a situation where the parameter space is restricted by the condition $\Psi \geq 0$, but the estimation procedure yields a negative ψ. This scenario can arise due to various reasons and needs careful consideration.

- Possible causes:
  - With small sample size, sampling error can lead to improper estimation, including negative ψ's
  - Elevated correlations among variables can contribute to the Heywood case
  - Attempting to extract more factors than genuinely present in the data can lead to over-factoring
- Possible remedies:
  - Opt for a larger sample size can reduce the impact of sampling error, particularly as the number of variables (p) increases
  - Avoiding the introduction of new variables that contribute with little information but increase correlations

- Use appropriate criteria to determine the optimal number of factors
- Consider a Bayesian approach where in the prior distribution of ψ's there is zero probability of negative values (Bayesian approach always the best ones).
- Implement a stopping criteria during the estimation process, so when ψ reach a small predetermined value like $\psi = 0.05$

2.2.11 Rotation

The invariance of orthogonal rotation of factor loadings is a notable property: this property implies that there are infinite solutions to the factor model, all of which are equivalent.
Since there are infinite solutions, among them we can find some solutions that enhance interpretability: this is involved by the search for a matrix M such that the transformed factor loadings $\Lambda^* = \Lambda M'$ offer a more straightforward interpretation than the original Λ.

The downside is the subjectivity implied: researches might continue rotating factors until they find an interpretation that aligns with their preferences.
The pursuit of interpretability in factor loadings involves ensuring distinctiveness, considering magnitude and sign, and ultimately aiming for a simple structure. Factor rotation emerges as a valuable tool in refining loadings to enhance the clarity of relationships between variables and factors.
We are going to see more in details what I mean by magnitude and sign with an example, but first let me explain how we usually implement the rotations.

Types of rotations

There are two main types of rotations:

1. Orthogonal rotation: in this case the factors are restricted to be uncorrelated. The standardized factor loadings can be interpreted as correlation coefficients between manifest variables and factors. This rotation leads to clear interpretability since the loadings directly represent correlations between factors and observed variables. Although the assumption here is quite strong: we assume independence between factor, which may be a realistic or unrealistic assumption depending on the context.

2. Oblique rotation: in this case the factors might be correlated. The standardized factor loadings no longer directly represent correlations: the relationship is captured by a correlation matrix Φ of the latent variables y's called "structured loading matrix". This matrix is given by

$$E(x - \mu)y' = \Lambda\Phi \tag{2.59}$$

The orthogonal transformations are divided by:

- Varimax: which aims to achieve factors with a few large loadings and many near-zero loadings. This rotation seeks to enhance the interpretability of factors by emphasizing clear, distinct loadings

- Quartimax: which aims to highlight the correspondence between observed and latent factors. Focuses on maximizing the variance of each observed variable explained by the factors

The oblique transformations are divided by:

- Promax: seeks to establish a simple structure with low correlation between factors. Aims to achieve a balance between simplicity and capturing correlations among factors.

- Oblimin: tends to produce factors with a varimax-like appearance but with oblique (correlated) relationships. Aims to strike a balance between achieving simplicity in structure and acknowledging potential correlations between factors.

Chapter 3

Latent trait model

Latent trait models, also known as latent variable models or item response models, are a class of statistical models used to analyze the relationships between observed variables and unobservable (latent) traits. These models are commonly employed in various fields, including psychometrics, education, and social sciences, to understand and quantify unobservable characteristics or abilities possessed by individuals.
Binary responses are extremely common, especially in the social sciences.
Individuals can be classified according to whether or not they belong to a trade union or take holidays abroad. They can be recorded as agreeing or disagreeing with some proposition or as getting some item in an educational test right or wrong. Such binary variables are often supposed to be indicators of more fundamental attitudes or abilities, and it is in these circumstances that latent variable modelling is relevant. Even when the observed responses fall into more than two categories, it is often useful to conflate them into two categories.

3.1 Model setup

Let's denote p as the number of items/questions in a test or survey: each item has a binary response typically coded as 0 (wrong/disagree) or 1 (right/agree).
A respondent's set of responses to the p items can be represented as a binary response pattern: if coded with 0 and 1, a response pattern is a string of p binary digits. With p items, each having two possible outcomes, there are 2^p different response patterns.

3.1.1 Two approaches

The two approaches to latent trait models for binary data are

1. The response function approach
2. The underlying variable approach (SEM)

Response function approach

Let's begin with the response function approach: the key idea here is that this approach analyzes the observed binary data directly without assuming an underlying

continuous variable. It models the probability of observing specific response patterns given the latent trait.
In this context, the joint probability distribution $f(x)$ is the distribution that assigns probabilities to the various response patterns. However, for certain analytical purposes, it's convenient to express the joint distribution in terms of a set of marginal probabilities.
The marginal probabilities represent the probabilities of individual items being answered affirmatively: $P(x_i = 1)$ as well as the joint probabilities of $P(x_i = 1, x_j = 1)$ and so on...
The total number of probabilities involved in specifying the joint distribution in this way is given by the sum of the combinations $\binom{p}{k}$ for $k = 1, 2, ..., p$:

$$\binom{p}{1} + \binom{p}{2} + ... + \binom{p}{p} = 2^p - 1 \tag{3.1}$$

This formula represents the total number of probabilities needed to fully characterize the joint distribution of binary responses for p items.
The dependence among the binary responses x_i is attributed to a vector y of latent variables. The relationship between them is expressed through the item response function or item characteristic curve, denoted as $\pi_i(y)$:

$$P(x_i = 1|y) = \pi_i(y) \tag{3.2}$$

Here $\pi_i(y)$ is a monotonically increasing function with $0 \leq \pi_i(y) \leq 1$.
This function is often referred to as the item characteristic curve in the context of the Item Response Theory (IRT).
The latent variable y is assumed to follow a normal distribution

$$y \sim N(0, I) \tag{3.3}$$

The latent variables are standardized and independent normal variables. This assumption simplify the modelling process and the estimation of the parameters.

The logit/normal model

The conditional distribution $g(x|y)$ is specified as a GLLVM:

- The random component is:

$$g(x_i|y) = \pi_i(y)^{x_i}(1 - \pi_i(y))^{1-x_i} \tag{3.4}$$

  which expresses a Bernoulli distribution where $\pi_i(y)$ is the probability of success, representing the probability that $x_i = 1$ given the latent variable y. In other words, the observed binary response x_i follows a Bernoulli distribution with parameter $\pi_i(y)$

- Systematic component:

$$\eta_i = \alpha_{i0} + \sum_{j=1}^{p} \alpha_{ij} y_i \tag{3.5}$$

  this part involves a linear combination of the latent variables y_j with corresponding coefficients α_{ij}. The linear predictor η_i is a systematic component capturing the influence of latent variables on the response x_i.

- Link function In this case the link function used here is the logit function. It transform the probability $\pi_i(y)$ to the log-odds scale, ensuring that the linear predictor η_i is linearly related to the log-odds of success.

$$\pi_i(y) = \frac{exp(\eta_i)}{1 + exp(\eta_i)} \tag{3.6}$$

3.1.2 Parameters interpretation

- α_{i0}(intercept) determines the probability of a positive response for the median individual, where

$$\pi_i(0) = \frac{1}{1 + exp(-\alpha_{i0})} \tag{3.7}$$

  This is the probability of a positive response when the latent variable y is at its median value.

- α_{ij}(latent variable effects) measures the effect of the latent variable y_i on the probability of success for x_i: a larger value of α_{ij} indicates that a change in y_j will have a more pronounced effect on the probability of success for x_i. In other words, it represents the sensitivity of x_i to changes in y_j.
  In the contest of specific latent trait models in the IRT (Item Response Theory):

  - α_{i0} is known as "difficulty parameter": an increse in α_{i0} makes it easier to obtain a positive response for all values of y_j. It reflects the overall difficulty level of the item.
  - α_{ij} is known as "discrimination parameter": it makes it easier to discriminate between individuals who are at different points on the latent scale. It indicates how well the item discriminates between individuals with different abilities.

  In the contest of the factor analysis

  - The intercept α_{i0} acts as an intercept of the model
  - The factor loadings α_{ij} represents the degree to which the latent variable y_j influences the observed variable x_i in factor analysis.

  These interpretations provide insights into the role of parameters in the latent trait model and their implications for understanding the relationship between latent variables and observed binary responses.

3.1.3 Component scores

Component scores in latent trait models play a crucial role in summarizing the information about individuals' latent traits based on their observed responses.
Denoted with X_j, the component scores are calculated for each latent variable y_j in the model:

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i \tag{3.8}$$

represents the component scores, where X is a vector of component scores with α_{ij} is the parameter corresponding to the effect of y_j on x_i, and x_i is the observed response for the item.
Since x_i is binary, the calculation simplifies and X_j is the sum of all α_{ij} for which $x_i = 1$. For example if $x_h = (0, 1, 1, 0)$, $X_h = \alpha_{2j} + \alpha_{3j}$.
The component scores provide a numerical summary of the influence of the latent variable y_j on an individual's response pattern. Larger values of component scores indicate a stronger influence of the corresponding latent trait on the individual's responses.
In summary, component scores condense the information about the latent variables into a numerical representation, facilitating the interpretation and comparison of individuals based on their response patterns.

3.2 The Rasch model

The Rasch model is a widely used model in educational testing due to its simplicity and appealing theoretical properties.
In this model, the total score for an individual h, denoted as $\sum_{i=1}^{p} x_{ih}$ is considered sufficient for capturing all the information about the latent trait y_h: it contains all the information about β's if the model is true. This implies that once you know an individual's total score, you have all the information about their ability or trait.
Similarly, the total number of positive or correct responses for a specific item i is sufficient for β_{i0}.
The Rasch model's focus on total scores and sufficiency simplifies the modeling process and makes it an attractive choice, particularly in educational contexts where simplicity and interpretability are crucial.

3.2.1 Two parametric logistic model

The 2-parameter logistic model (2PL) is an extension of the Rasch model that introduces discrimination parameters to account for variations in how well items differentiate between individuals. The probability of a correct response is given by the logistic function defined as

$$P(x_i = 1, \beta_{i0}, \beta_{i1}, y_h) = \frac{exp(\beta_{i1}(y_h - \beta_{i0})}{1 + exp(\beta_{i1}(y_h - \beta_{i0}))} \tag{3.9}$$

β_{i0} is the analogous to the Rasch model's difficulty parameter α_{i0} and β_{i1} serves as the discrimination parameter, indicating how well item i discriminates between individuals with different trait levels.
The 2PL model aims at shifting and scaling the latent trait according to the difficulty and discrimination parameters. It allows for a more flexible representation of item characteristics compared to the Rasch model.
Those parameters in the 2PL model can be related to the Rasch model as follows:

$$\beta_{i0} = -\frac{\alpha_{io}}{\alpha_{i1}} \tag{3.10}$$

and

$$\beta_{i1} = \alpha_{i1} \tag{3.11}$$

In the broader context of Item Response Theory, models with more than two parameters exist, allowing for additional complexities in modeling item characteristics.

3.2.2 Model estimation

The model estimation process involves maximum likelihood estimation using the Expectation-Maximization (E-M) algorithm, particularly for a one-factor model where there are missing observations (in this case, the latent trait values y).

The log-likelihood function l for the complete data (both observed responses x and latent traits y) is given by:

$$l_c = \sum_{h=1}^{n} \sum_{i=1}^{p} lng(x_{ih}|y_h) + lnh(y_h) \tag{3.12}$$

The EM algorithm involves two steps:

1. Evaluate the expected log-likelihood: $E[l_c|x_1, ..., x_n]$
2. Maximise $E[l_c|x_1, ..., x_n]$ over the parameters

But since $h(y)$ does not depend on the parameters we can replace l_c with the following equation

$$l = \sum_{h=1}^{n} ln(g(x_h|y_h)) = \sum_{h=1}^{n} \sum_{i=1}^{p} ln(g(x_{ih}|y_h)) \tag{3.13}$$

we can write $g(x_{ih}|y_h)$ as

$$g(x_{ih}|y_h) = (1 - \pi_i(y_h))exp(x_{ih}(\alpha_{i0} + \alpha_{i1}y_h)) \tag{3.14}$$

And the log-likelihood looks like this

$$l = \sum_{h=1}^{n} (\sum_{i=1}^{p} (ln(1 - \pi_i(y_h)) + x_{ih}(\alpha_{i0} + \alpha_{i1}y_h))) \tag{3.15}$$

Now we need to compute the conditional expected log-likelihood and maximise it: I will skip for now those passage and go directly to the results.
We arrive at the point where the integral needed to solve the maximisation is not analytically solvable: therefore, we apply numerical approximation.
The latent variable y is treated as discrete, with values $y_1, ..., y_k$ and corresponding weights $h(y_1), ..., h(y_k)$. The specific approximation used is Gausse-Hermite quadrature, particularly useful when the density is standard normal.

So, the approximated likelihood for each item i becomes

$$\sum_{t=1}^{k} [ln(1 - \pi_i(y_t)\eta_t + \alpha_{i0} \sum_{h=1}^{n} x_{ih} + \alpha_{i1}r_{it})] \tag{3.16}$$

with

$$\eta_t = \sum_{h=1}^{n} h(y_t|x_h) \tag{3.17}$$

and

$$r_{it} = \sum_{h=1}^{n} x_{ih} h(y_t|x_h) \tag{3.18}$$

The expected number of positive responses is indicated by r_{it} and it is particularly important as it represents the expected number of individuals predicted to be at y_t who will respond positively.
Instead, $\eta(t)$ is the expected number of individuals in latent position y_t.
Differentiating the approximated expected log-likelihood $E[l_i|x]$ with respect to α_{i0} and α_{ij} leads to non linear equations in the parameters. To solve these equations and estimate the parameters, an iterative optimization method like Newton-Raphson iterative procedure is employed.

3.3 Probit-Normal model

In the Probit model, the probability of a binary response x_i taking the value 1 given the underlying latent variable y is modeled using the cumulative distribution function of the standard normal distribution denoted here with Φ:

$$\Phi^{-1}(\pi_i(y)) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j \tag{3.19}$$

where

- $\pi_i(y)$ is the probability of x_i being 1 given y
- Φ^{-1} is the inverse of the standard normal CDF
- y is the vector of latent variables
- α_{i0} is the intercept for item i
- α_{ij} are the coefficients associated with the latent variables

This formulation is quite analogous to the logit model but the uses the normal CDF instead of the logistic function:

$$\pi_i(y) = \Phi(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j) \tag{3.20}$$

Comparing this with the logit model we can highlight the relationship

$$logit(u) = \frac{\pi}{\sqrt{3}} \Phi^{-1}(u) \tag{3.21}$$

While the Probit and Logit models are virtually equivalent in many aspects, the Probit/normal model lacks the sufficiency property of the components X. In the logit model X contains all the information about the latent traits.

3.3.1 Underlying variable approach

This model assumes that the observed binary variable x_i is related to an underlying continuous variable ξ_i with the following expression:

$$x_i = \begin{cases} 1, \xi_i \leq \tau_i \\ 0, \text{otherwise} \end{cases} \tag{3.22}$$

Here, τ_i is a threshold parameter, and ξ_i follows a standard normal distribution: this formulation implies that the binary outcome x_i is determined by whether ξ_i falls below a certain threshold τ_i.
We can express this in matrix form:

$$\xi = \mu + \Lambda y + \epsilon \tag{3.23}$$

where

- ξ is a vector of the continuous latent variables and $\xi \sim N(0, R_\xi)$
- $R_\xi = \Lambda\Lambda' + \Psi$ and this matrix is a unknown correlation matrix and has to be estimated.
- Λ is the matrix of factor loadings
- $\epsilon \sim N(0, \Psi)$ with Ψ a diagonal matrix as in the normal linear factor model

3.3.2 Estimation procedure steps

In this model we need to estimate

- The thresholds τ_i
- The loadings
- The specific variances

There estimation procedure consists in three steps:

1. Estimate the thresholds $\tau_1, ..., \tau_p$ with maximum likelihood estimation by using each marginal distribution of the latent variables:

$$\pi_i = P(x_i = 1) = \int_{-\infty}^{\tau_i} \phi(\xi_i) d\xi_i = \Phi(\tau_i) \tag{3.24}$$

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

$$\hat{\tau}_i = \Phi^{-1}(P_i) \tag{3.25}$$

2. Estimate of R_ξ using the bivariate distribution of the latent variables

$$\pi_{ij} = P(x_i = 1, x_j = 1) = \int_{-\infty}^{\tau_i} \int_{-\infty}^{\tau_i} \phi(\xi_i, \xi_i) d\xi_i \tag{3.26}$$

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

3.3.3 Equivalence between response function approach and underlying variable approach

Let S be any non-empty subsets of $\{1, 2, ..., p\}$ then we can write the joint distribution as

$$P(x_i = 1, x_j = 1) = P\{\bigcap_{i \in S}(x_i = 1)\} \tag{3.27}$$

we can express the intersection as the product of the marginals

$$= \int ... \int \prod_{i \in S} P\{(x_i = 1|y)h(y)\}dy \tag{3.28}$$

Now, we previously defined the x_i with the indicator function, therefore, when $x_i = 1$ we are saying that $\xi_i \leq \tau_i$

$$= P\{\bigcap_{i \in S}(\xi_i \leq \tau_i)\} \tag{3.29}$$

$$= ... = \int ... \int \prod_{i \in S} P\{(\xi_i \leq \tau_i|y)h(y)\}dy \tag{3.30}$$

We know that x_i is defined as a linear relationship with μ, e, Λ, so

$$P((\xi_i \leq \tau_i)|y) = P\{\mu_i + e_i + \sum_{j=1}^{q} \lambda_{ij} y_j \leq \tau_i|y\} \tag{3.31}$$

we can re-arrange this equation to get

$$= P\{\frac{e_i}{\psi^{\frac{1}{2}}} \leq \frac{\tau_i - \mu_i - \sum_{j=1}^{q} \lambda_{ij} y_j}{\psi_i^{\frac{1}{2}}}|y\} \tag{3.32}$$

this probability can be seen as the distribution function of $\frac{e_i}{\psi^{\frac{1}{2}}}$

$$R(\frac{\tau_i - \mu_i - \sum_{j=1}^{q} \lambda_{ij} y_j}{\psi^{\frac{1}{2}}}) \tag{3.33}$$

In the response function model we had that

$$P(x_i = 1|y) = G(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j) \tag{3.34}$$

where G is the inverse logit function.
Therefore we can say that there is an equivalence if:

- $G = R$
- $\alpha_{i0} = \frac{(\tau_i - \mu_i)}{\psi^{\frac{1}{2}}}$
- $\alpha_{ij} = -\frac{\lambda_{ij}}{\psi^{\frac{1}{2}}}$

Since we don't have any information about the standard deviation of ξ_i, the parameters τ_i, μ_i and λ_{ij} are not individually estimable, so we assume that:

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

3.3.4 Standard errors

Unfortunately, is not possible to exactly determine the standard errors for the parameter estimates: therefore, we can use asymptotic variance-covariance matrix using the information matrix. Recall that

$$var(\hat{\beta})^{-1} = E[-\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}] \tag{3.35}$$

which can be approximated by using the observed matrix

$$var^*(\hat{\beta})^{-1} = [\sum_{h=1}^{n} \frac{1}{f^2(x_h)} \frac{\partial f(x_h)}{\partial\beta_i} \frac{\partial f(x_h)}{\partial\beta_j}] \tag{3.36}$$

3.3.5 Goodness of fit

Hypoteshis testing is used to provide a measure of how well the model fits the observed data: the theoretical hypotesis is

$$H_0 : \pi = \pi(\alpha) \tag{3.37}$$

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

$$\chi^2 = \sum_{r=1}^{2^p} \frac{(nP_r - n\pi_r(\hat{\alpha}))^2}{n\pi_r(\hat{\alpha})} \tag{3.38}$$

$$G^2 = 2\sum_{r=1}^{2^p} nP_r ln(\frac{nP_r}{n\pi_r(\hat{\alpha})}) \tag{3.39}$$

Under regular condition, when n is large and p is small, the two statistics follow a χ^2 distribution with degrees of freedom $2^p - p(q+1) - 1$.
As p increases, leading to sparse data, the χ^2 distribution assumption may not hold: for example if $p = 10$, $2^p = 1024$, $n = 1000$, there may be many responses with $n\pi_i(\hat{\alpha}) \leq 1$.
The comparison of observed proportions with model-predicted probabilities helps assess the adequacy of the latent trait model.

3.3.6 Remedies of goodness of fit issues

1. Issue: when expected frequencies are less than 5, there's a risk of having degrees of freedom equal to 0 in statistical tests.
   Remedy: Group response patterns with expected frequencies less than 5 to address the issue

2. Issue: Bootstrap methods are employed to create an empirical sampling distribution of the test statistics.
   Remedy: Use the parametric bootstrap method to obtain a more robust estimation of the distribution of test statistics

Let's consider residuals calculated from marginal frequencies, which contain information about associations among variables:

1. First order margins: $P(x_i = 1)$ contain no information about the dependencies among the x's
2. Second order margins: $P(x_i = 1, x_j = 1)$ provide information about pairwise association
3. Third order margins: $P(x_i = 1, x_j = 1, x_k = 1)$ capture information about three-way associations.

Now, to compute the values of the residual we have that:

- O are the observed frequencies for any marginal probability
- E are the corresponding expected frequency

and R is given by

$$R = \frac{(O - E)^2}{E} \tag{3.40}$$

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

3.3.7 Posterior analysis

As in the general GLLVM, all the information about y is contained in the posterior distribution $h(y|x)$: in the case of the logit/normal model, the distribution depends on x only through the q-variate sufficient statistic defined as

$$X = Ax \tag{3.41}$$

where $A = \{\alpha_{ij}\}$.
The component scores are often used for scaling, expecially in measurement abilities: the posterior mean $E(y_j|x_h)$ and the components give the same ranking to response patterns/indviduals.

3.4 Summary

The explanation covers various aspects related to latent trait models, including the Rasch model, the 2-parameter logistic model, the Probit-Normal model, and the underlying variable approach. Here's a breakdown of the content:

1. **Latent Trait Models:** - Described latent trait models as statistical models used to analyze relationships between observed variables and unobservable traits. -

Mentioned their common applications in psychometrics, education, and social sciences, especially for analyzing binary responses.

2. **Model Setup:** - Defined (p) as the number of items/questions in a test or survey.
- Introduced binary response patterns and two approaches: response function approach and underlying variable approach.

3. **Response Function Approach:**
- Explained the response function approach, focusing on the item response function or item characteristic curve.
- Described the logit/normal model, specifying the random and systematic components, and introducing the link function.

4. **Parameters Interpretation:**
- Provided interpretations for model parameters, including intercepts and latent variable effects.
- Discussed interpretations specific to Item Response Theory (IRT) and factor analysis.

5. **Component Scores:**
- Explained the concept of component scores in latent trait models.
- Described how component scores summarize information about latent traits based on observed responses.

6. **Rasch Model:**
- Introduced the Rasch model and its focus on total scores for capturing information about latent traits.
- Mentioned the simplicity and theoretical properties that make the Rasch model attractive in educational testing.

7. **2-Parameter Logistic Model:**
- Described the 2-parameter logistic model as an extension of the Rasch model with discrimination parameters.
- Presented the logistic function for the probability of a correct response.

8. **Model Estimation:**
- Outlined the model estimation process, emphasizing maximum likelihood estimation using the Expectation-Maximization (E-M) algorithm.
- Mentioned the challenges of estimating latent variables.

9. **Probit-Normal Model:**
- Introduced the Probit-Normal model, where the probability of a binary response is modeled using the cumulative distribution function of the standard normal distribution.

10. **Underlying Variable Approach:**
- Described the underlying variable approach, relating observed binary variables to

an underlying continuous variable.
- Outlined the estimation procedure steps for thresholds, loadings, and specific variances.

11. **Equivalence Between Approaches:**
- Explored the equivalence between the response function approach and the underlying variable approach.
- Provided conditions and relationships for equivalence.

12. **Standard Errors:**
- Acknowledged the challenge of determining exact standard errors for parameter estimates.
- Mentioned the use of asymptotic variance-covariance matrices for standard error approximation.

13. **Goodness of Fit:**
- Discussed hypothesis testing for measuring how well the model fits observed data.
- Introduced Pearson's χ^2 statistic and the likelihood ratio test G^2 for model fit assessment.

14. **Remedies for Goodness of Fit Issues:**
- Identified issues related to small expected frequencies.
- Proposed remedies, such as grouping response patterns and employing bootstrap methods.

15. **Posterior Analysis:**
- Explained that all information about y is contained in the posterior distribution $h(y|x)$.
- Introduced the q-variate sufficient statistic $X = Ax$ and its role in scaling.

Chapter 4

Latent class model

4.1 Introduction

Let's begin this new topic with an example: in the context of social science research, understanding attitudes toward science and technology is a crucial aspect that provides insights into public perceptions, preferences, and beliefs. In this analysis, we focus on a dataset consisting of responses from 392 individuals regarding their attitudes toward science and technology. The dataset includes 7 items, each initially measured on a four-point scale. To simplify the analysis, the response categories have been dichotomized. The coding scheme involves assigning a value of 0 to responses falling under 'strongly disagree' and 'disagree to some extent,' while responses categorized as 'agree to some extent' and 'strongly agree' are coded as 1. The responses are organized into 128 possible response patterns, considering the binary coding for each of the 7 items. However, it is noted that many of these potential patterns do not occur in the dataset, highlighting the variability and uniqueness of individual attitudes.

4.2 Statistical representation of the problem

Suppose there are p binary variables denoted with $x_1, ..., x_p$ where x_i takes values $(0, 1)$.
Assume that the mutual association of these variables is accounted for a single latent binary variable y.
Now, we assume to divide the population into two parts so that the x's are mutually independent in each group: we label the groups as 0 and 1, and the prior distribution of y is given by

$$h(1) = P(y = 1) = \eta \tag{4.1}$$

and

$$h(0) = 1 - h(1) = 1 - \eta \tag{4.2}$$

The conditional distribution of x_i given y is a Bernoulli variable

$$g(x_i|y) = \pi_{iy}^{x_i}(1 - \pi_{iy}^{1-x_i}) \tag{4.3}$$

with x_i and y that takes values 0 or 1.
It follows that

$$f(x) = \eta \prod_{i=1}^{p} x_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} + (1-\eta)\prod_{i=1}^{p} \pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i} \tag{4.4}$$

which is the sum of conditional distribution when y takes value 1 and when y takes value 0.
When we fit a model, we always ask if the model has a good fit to the data: if the model is not adeguate we can

- extend the number of classes to a number of classes $K \geq 3$: imagine that in the previous example there is a class between the ones described as "agree" and "not agree", like for example "neither agree/disagree". In that case assuming that could be a third class could improve the goodness of fit of the model;
- Then we can consider models that do not assume the conditional independence assumption;
- Moreover, we can consider models with continuous latent variables.

But, if the model is adequate we can define a rule to allocate the individuals/observations in the two latent classes: we are going to use the posterior distribution of the latent variabels given the observed variables.

4.2.1 Posterior probability

As we understood, there are two classes: therefore, we can write the posterior distribution of belonging to the class 1 or to the class 0:

$$h(1|x) = P(y=1|x_1,...,x_p) = \frac{h(y=1)g(x_1,...,x_p|y=1)}{f(x)} = \tag{4.5}$$

$$= \frac{\eta \prod_{i=1}^{p} x_{i1}^{x_i}(1-\pi_{i1})^{1-x_i}}{\eta \prod_{i=1}^{p} x_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} + (1-\eta)\prod_{i=1}^{p} \pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}} \tag{4.6}$$

and for the other class

$$h(0|x) = \frac{(1-\eta) \prod_{i=1}^{p} x_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}}{\eta \prod_{i=1}^{p} x_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} + (1-\eta)\prod_{i=1}^{p} \pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}} \tag{4.7}$$

And the allocation rule is the following:
if

$$h(1|x) > h(0|x) \tag{4.8}$$

the unit is allocated in latent class 1
if

$$h(1|x) \leq h(0|x) \tag{4.9}$$

the unit is allocated in latent class 0.

4.2.2 The extension to a K-class model

In the two-class model, we assumed that the population is divided into two groups labeled as 0 and 1. Now, let's extend this to a K-class model where $K \geq 3$. This extension allows us to capture more nuanced patterns in the data.
Let π_{ij} be the probability of a positive response where i is the variable and j is the individual's category: $(i = 1, ..., p; j = 0, 1, ..., K-1)$ Now, η_j is the prior probability that a randomly chose individual is in class j: since we are speaking about partion of the sample space, $\sum_{j=0}^{K-1} \eta_j = 1$.
Let's assume to have conditional independence: we can write

$$g(x|j) = \prod_{i=1}^{p} g(x_i|j) \tag{4.10}$$

and it follows that

$$f(x) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} \tag{4.11}$$

The posterior probability that an individual with response vector x belongs to category j is given by the Bayes formula:

$$h(j|x) = \frac{\eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}}{f(x)} \tag{4.12}$$

with $(j = 0, 1, ..., K-1)$. As before, the posterior distribution can be used to construct the following allocation rule: an individual is allocated in the class with the highest posterior probability.
The vector of the unknown parameters to be estimated is:

$$\theta = (\eta_0, ..., \eta_{K-1}, \pi_{1,0}, ..., \pi_{p,K-1}) \tag{4.13}$$

4.3 Maximum likelihood estimation

We can use even in this case the $E - M$ algorithm:
let's take a random sample of size n, the log-likelihood can be written as

$$l = \sum_{h=1}^{n} ln\{\sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i h} (1 - \pi_j)^{1-x_{ih}}\} \tag{4.14}$$

and it is maximised subject to the constrain $\sum \eta_j = 1$, so we find that the maximum is

$$\phi = l + \gamma \sum_{j=0}^{K-1} \eta_j \tag{4.15}$$

where γ is an undetermined multiplier (which is equivalent to the Lagrange multiplier).
Now, we need to take the partial derivatives of the log-likelihood with respect to the parameters η and π and set them equal to 0.

- $$\frac{\partial \phi}{\partial \eta_j} = \sum_{h=1}^{n} \frac{g(x_h|j)}{f(x_h)} + \gamma = 0 \tag{4.16}$$

- $$\frac{\partial \phi}{\partial \pi_{ij}} = \sum_{h=1}^{n} \eta_j \frac{\partial \phi}{\partial \pi_{ij}} \frac{g(x_h|j)}{f(x_h)} = 0 \tag{4.17}$$

After some algebra we find from the two equations that

- $$\hat{\eta}_j = \sum_{h=1}^{n} \frac{h(j|x_h)}{n} \tag{4.18}$$

- $$\hat{\pi}_{ij} = \frac{\sum_{h=1}^{n} x_{ih} h(j|x_h)}{\sum_{h=1}^{n} h(j|x_h)} \tag{4.19}$$

In summary, the procedure is the following

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

4.4 Goodness of fit

As for the latent trait model, we consider the Pearson χ^2 statistic and the likelihood ratio test G:

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

where in this case

$$\hat{\pi}_r = f(x_r) = \sum_{j=0}^{K-1} \hat{\eta}_j \prod_{i=1}^{p} \pi_{ij}^{x_{ir}} (1 - \hat{\pi}_{ij})^{(1-x_{ir})} \tag{4.20}$$

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

4.5 Latent class model vs latent trait model

The latent class model is a specialized case within the broader framework of latent trait models. In this specific instance, the prior distribution is characterized by discrete probability masses. This distinction implies that all general results, not contingent on the specifics of the prior, can be seamlessly applied to the latent class model.
In the context of latent class models, the form of the sufficient statistic becomes a linear combination of the observed variables, denoted as x

Chapter 5

Latent trait model for polytomous data

In the binary case we identified two approaches to the problem. The first, via the sufficiency principle, led to the logit model in which the response function turned out to be a logistic curve. The second started by supposing that the binary observations were formed by dichotomising an underlying continuous variable and then assuming that these underlying variables were generated by a normal linear factor model. The two models arrived at from these very different starting points proved to be equivalent. Which one we chose to adopt was therefore largely a non-empirical matter. In the polytomous case this equivalence breaks down.

5.1 Introduction

In the polytomous case, we need to extend a bit the notation:

- let c_i denote the number of categories of variable i which are labelled $0, 1, ..., c_i - 1$ with $(i = 1, 2, ..., p)$ and indexed by s.
- The indicator variable x_i is now replaced by a vector-valued indicator function with its sth element defined by

$$x_i(s) = \begin{cases} 1 \text{ if the responses falls in category } s; \\ 0 \text{ otherwise} \end{cases} \tag{5.1}$$

- The c_i vector with these elements is denoted by x_i and $\sum_s x_i = 1$
- The full response pattern for an individual is denoted by $x' = (x_1', x_2', ..., x_p')$ of dimension $\sum_i c_i$.

5.1.1 A special case

If we set $c_i = 2$ for all i, this leads to the binary case.
The response function $\pi_i(y)$ is now replaced by a set of functions defined by

$$P\{x_i(s) = 1|y\} = \pi_{is}(y) \tag{5.2}$$

where $s = 0, ..., c_i - 1$ and $i = 1, ..., p$ with $\sum_s \pi_{is}(y) = 1$

5.1.2 Example

To consolid the understanding of the notation, let's consider the following example: Suppose we have a dataset that includes responses from students on three different exam questions, each with three possible levels of understanding: Low, Medium, and High.

Variables:

X_1:Level of understanding for Question 1 (Low, Medium, High)
X_2:Level of understanding for Question 2 (Low, Medium, High)
X_3:Level of understanding for Question 3 (Low, Medium, High).
Now, let's introduce the latent trait y, which represents the overall academic ability of a student. We assume that this latent trait follows a standard normal distribution.
For each question, the probability of a student falling into a specific level of understanding depends on the latent trait y. The response functions can be defined as follows:

For X_1:

$$\begin{cases} P(X_1 = Low|y) = \pi_{10}(y) \\ P(X_1 = Medium|y) = \pi_{11}(y) \\ P(X_1 = High|y) = \pi_{12}(y) \end{cases} \tag{5.3}$$

For X_2:

$$\begin{cases} P(X_2 = Low|y) = \pi_{20}(y) \\ P(X_2 = Medium|y) = \pi_{21}(y) \\ P(X_2 = High|y) = \pi_{22}(y) \end{cases} \tag{5.4}$$

For X_3:

$$\begin{cases} P(X_3 = Low|y) = \pi_{30}(y) \\ P(X_3 = Medium|y) = \pi_{31}(y) \\ P(X_3 = High|y) = \pi_{32}(y) \end{cases} \tag{5.5}$$

These probabilities are influenced by the latent trait y, representing the student's overall academic ability. In this way, the latent trait model for polytomous data helps us understand how students with different latent abilities are likely to respond to questions with multiple categories of understanding.

5.1.3 The model

We may now suppose the conditional probability function of x_i given y to be multinomial so that

$$g_i(x_i|y) = \prod_{s=0}^{c_i-1} \pi_{is}(y)^{x_i(s)} \tag{5.6}$$

And the posterior density, with the Bayes theorem is given by:

$$h(y|x) = \frac{h(y)\{\prod_{i=1}^{p}\prod_{s=0}^{c_i-1}\pi_{is}(y)^{x_i(s)}\}}{f(x)} \tag{5.7}$$

But, as we introduced in the Normal linear factor model, the conditional distribution of a manifest variable x_i had a distribution belonging to the exponential family with canonical parameter θ_i. In that case, both values supposed to be scalar, in this case they are vectors: in fact, sufficiency property still holds even when x_i, θ_i and the αs are vectors.
In the binary case we implicitly reparameterised the problem: equivalently here

$$g_i(x_i|y) = \prod_{s=0}^{c_i-1}\pi_{is}(y)^{x_i(s)} \tag{5.8}$$

and multiplying and dividing by $\pi_{i0}(y)^{x_i(s)}$

$$= \prod_{s=0}^{c_i-1}\pi_{i0}(y)^{x_i(s)}\frac{\pi_{is}(y)^{x_i(s)}}{\pi_{i0}(y)^{x_i(s)}} \tag{5.9}$$

taking the logarithm transformation

$$= \pi_{i0}(y)exp\sum_{s=0}^{c_i-1}x_i(s)ln(\frac{\pi_{is}(y)}{\pi_{i0}(y)}) \tag{5.10}$$

which can be defined as

$$= \pi_{i0}(y)exp(\theta_i^{'}x_i) \tag{5.11}$$

where the parameter $\theta_i^{'}$ is

$$\theta_i^{'} = (0, ln(\frac{\pi_{i1}(y)}{\pi_{i0}(y)}), ln(\frac{\pi_{i2}(y)}{\pi_{i0}(y)}), ..., ln(\frac{\pi_{i(c_i-1)}(y)}{\pi_{i0}(y)})) \tag{5.12}$$

This is what we generally call as "random part" of the GLLVM.
Systematic part is defined as

$$\eta_i = \alpha_{i0} + \alpha_{i1}y_1 + ... + \alpha_{iq}y_q \tag{5.13}$$

and the link function is the logit in which

$$\theta_i = \alpha_{i0} + \alpha_{i1}y_1 + ... + \alpha_{iq}y_q \tag{5.14}$$

where

$$\alpha_{ij} = (\alpha_{ij}(0) = 0, \alpha_{ij}(1), ..., \alpha_{ij}(c_j - 1)) \tag{5.15}$$

with $j = 0, 1, ..., q$.
It is easily shown that the components are given by

$$X_j = \sum_{i=1}^{p}\sum_{s=0}^{c_j-1}\alpha_{ij}(s)x_i(s) \tag{5.16}$$

The last result of the link function combined with the formulation of the conditional distribution $g_i(x_i|y)$ leads to say that

$$\pi_{is}(y) = \frac{exp(\alpha_{i0}(s) + \sum_{j=1}^{q}\alpha_{ij}(s)y_j)}{\sum_{r=0}^{c_i-1} + \sum_{j=1}^{q}\alpha_{ij}(r)y_j} \tag{5.17}$$