Mock exam *Answers*

Exercise 1

1. (a) The possible response patterns are $2^4 = 16$.

   (b) The item with the highest proportion of no is the item FRAUD (93% of no).

   (c) The observed frequency of the response pattern (0,0,0,0) is 207 and the observed frequency of the response pattern (1,1,1,1) is 2.

   (d) There are two pairs of items that are not significantly associated. The pair of items composed by LIEPAPER and COPYEXAM and the pair composed by LIEEXAM and COPYEXAM.

2. **The 2-PL mode**
   The 2-PL model was introduced by Birnbaum in 1969. The model is written as :

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

   where β_{i0} and β_{i1} are the difficulty and the discrimination parameter.
   The latent trait model
   Two approaches for the latent trait models:

   - Response function approach: it analyzes the data as they are by imposing distributions on the observed variables.
   - Underlying variable approach: it assumes that the data have been produced by dichotomizing underlying continuous variables.

   In the response function approach the dependence among the xs can be explained by a vector of $\mathbf{y}$ latent variables that can be regarded as mutually independent random variables with $P(x_1 = 1|\mathbf{y}) = \pi_i(\mathbf{y})$ $(x_i = 0, 1; i = 1, ..., p)$. In the logit/normal model the conditional distribution $g(\mathbf{x}|y)$ is assumed to be a GLLVM as follows:

   - Random component

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

     where $g(x_i|\mathbf{y}) \sim Ber(\pi(\mathbf{y}))$, where $\pi(\mathbf{y})$ is the probability of success.

   - Systematic component

$$\eta_i = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} y_j \qquad j = 1, ...p$$

- Link function: $logit \rightarrow \theta_i = logit(\pi(\mathbf{y})) = \eta_i$

In the latent trait model with the response function approach α_{i0} e α_{i1} are the difficulty and discrimination parameters, respectively. In the underlying variable approach the model for the binary data is the probit/normal model

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

The probit/normal model is related to the standard normal linear model and imply the underlying variable approach.

Relation between the 2-PL model and the latent trait model

The 2-PL model represents a different parametrization of the univariate logit/normal latent trait model. Indeed the parameter of the 2-PL model can be expressed as:

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

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

The 2-PL and latent trait model parameters can be estimated using the E-M algorithm. Both representations (2-PL and univariate logit/normal) give the same results in terms of the values of the scores, computed with different methods (total score, component score and expected mean).

3. (a) The most difficult item is LIEEXAM because it has the smallest value for the difficulty parameter, equal to -4.47. The easiest item is COPYEXAM because it has the biggest value for the difficulty parameter, equal to -1.37.

   (b) The rank of the items into decreasing order, according to the discrimination parameter values, is the following:

   i. LIEEXAM($\alpha_{11} = 3.11$)
   ii. LIEPAPER ($\alpha_{21} = 2.58$)
   iii. FRAUD ($\alpha_{31} = 1.02$)
   iv. COPYEXAM ($\alpha_{41} = 0.51$).

   LIEEXAM is the item that discriminates more between individuals that have low and high values of the latent variable.

4. The value of the standardized discrimination parameter are the following: LIEEXAM ($sd\alpha_{11} = 0.95$) LIEPAPER ($sd\alpha_{21} = 0.93$) FRAUD ($sd\alpha_{31} = 0.71$) COPYEXAM

($sd\alpha_{41} = 0.45$). These coefficients are the discrimination parameters (they give the same rank for the items obtained at point 3(b)) rescaled to be between 0 and 1. The probability that a median individual (y=0) responds positively to COPIEXAM is the highest (0.20), while the probability that a median individual(y=0) responds positively to LIEXAM is the smallest (0.01).

5. The value of the Chi-square test is 9.26 with 7 degrees of freedom and p-value 0.23. The value of the LR test is 8.16 with 7 degrees of freedom and p-value 0.31. According to both test the univariate latent trait model has a good fit to the data because we do not reject the null hypothesis ($H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\alpha)$). We cannot rely on these results because some of the expected frequencies estimated under the model of certain response patterns are <5.

6. We should consider residuals calculated from marginal frequencies. We consider the two way margins: none of the residuals is > 4. We can conclude that the univariate latent trait model has a good fit to the data.

7. The latent variable represents the cheating behaviour. There are different method for scaling individual: total, component and expected mean scores. The component and expected mean scores are the estimates of the values of the latent variable. The component and expected mean scores give the same ranking of the response pattern/individuals. For example the response pattern (0,0,0,0) has the lowest mean score (-0.354) and the lowest component score (0.00), that means that the people that show this response pattern do not have a cheating behaviour. People that have response pattern (1,1,1,1) have the highest mean score (2.14) and the highest component score (7.24), that means they have a very strong cheating behaviour. The total score rank the response patterns/individuals based on the values of the observed variables and can give a different ranking than the component and mean score methods. In this analysis the total score give a different ranking of the response patterns compared to the other two scoring methods.

Exercise 2

1. Given **x** and **y** continuous random variables, **x** correlated, the normal linear factor model is defined as

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

The assumptions are:

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

(b) **y** and **e** independent.

$\mathbf{\Lambda}$ is the pxp matrix of the factor loadings and $\mathbf{\Psi}$ is the pxp diagonal matrix of specific variances. An equivalent way of writing this model is

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

and

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

$h(\mathbf{y})$ is the prior distribution while $g(\mathbf{x}|\mathbf{y})$ is the conditional distribution.
$g_i(x_i|\mathbf{y})$ can be written as

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

where $\eta_i = \mu_i + \sum_{j=1}^{q} \lambda_{ij} y_j$, $i = 1, ..., p$.
If Ψ_{ii} is known, equation (1) can be written in the GLLVM form by setting: $\theta = \frac{\eta_i}{\sqrt{\psi_{ii}}}$, $u_i(x_i) = \frac{x_i}{\sqrt{\psi_{ii}}}$, $G_i(\theta_i) = \exp\left(\frac{\frac{-1}{2}\eta_i^2}{\psi_{ii}}\right)$ and $F_i(x_i) = \frac{1}{\sqrt{2\pi\psi_{ii}}} \exp\left(\frac{\frac{-1}{2}\mathbf{x}_i^2}{\psi_{ii}}\right)$. The sufficient statistics, written in matrix form, are:

$$\mathbf{X} = \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{x}$$

2. The parameter space is restricted by the condition $\mathbf{\Psi} > 0$. However in some cases we can have a negative or zero ψ. If one of more of the ψs are 0, we have the Heywood cases. The possible causes of the Heywood cases are:

   - Result of sampling error. A key factor is the sample size. It can occur with small samples. For a given sample size the risk 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.