A general Cholesky decomposition based modeling of longitudinal IRT data: Handling skewed latent traits distributions.

José Roberto Silva dos Santos;*Caio Lucidius Naberezny Azevedo †

Abstract

In this work we develop a longitudinal IRT model considering skewed latent traits distribution, based on the work of Pourahmadi (1999), which uses the Cholesky decomposition of the matrix of variance and covariance (dependence) of interest related to the latent traits. A kind of multivariate skew-normal distribution for the latent traits is induced by an antedependence model with centered skew-normal erros. We focus on dichotomous responses considering skewed latent traits distributions and a single group of individuals followed over several evaluation conditions (time-points). In each of these evaluation conditions the subjects are submitted to a (possibly different along these time-points) measuring instruments which have some common items structure. Using an appropriate augmented data framework, a longitudinal IRT model is developed through the Pourahmadi's approach. The parameter estimation, model fit assessment and model comparison were implemented through a hybrid MCMC algorithm, such that when the full conditionals are not known, the SVE (Single Variable Exchange) algorithm is used. Simulation studies indicate that the parameters are well recovered. In addition, a longitudinal study in education, promoted by the Brazilian federal government, is analyzed to illustrate the methodology developed.

keywords: longitudinal IRT data, Bayesian inference, antedependence models, SVE algorithm, MCMC algorithms, Cholesky decomposition.

*Department of Statistics, University of Campinas, Brazil
†Department of Statistics, University of Campinas, Brazil

1 Introduction

Longitudinal data are characterized when experimental units are followed along different measurement occasions (or time-points) that can not be mutually randomized. Along these measurement occasions, characteristics of interest of those experimental units are measured. In the IRT context the main interest lies on the response of subjects to items belonging to some measurement instrument (cognitive tests, psychiatric questionnaires, educational tests, among others) along different occasions (as scholar grades). These measurement instruments, in each time point, can be partially or totaly different from each other, but they must present some structure of common items. Due to this nested structure, that is, the time-specific measurements within-subjects, it is expected to observe some within-subject dependence. Within LIRT (longitudinal IRT) data, the within subject responses to the items are assumed to be conditionally independent given the item parameters and the latent traits, whereas it is expected that some dependence structure will be observed for the latter, see Andrade and Tavares (2005), Azevedo et al. (2012b).

Some longitudinal item response theory models have been proposed in the literature. For example, Conaway (1990) proposed a Rasch model to analyze panel data using the marginal maximum likelihood approach, see Bock and Aitkin (1981). Eid (1996) defined a longitudinal model for confirmatory factorial analysis for polytomous response data. Andrade and Tavares (2005) and Tavares and Andrade (2006) developed an IRT model to estimate the paramemters of the latent distribution (also known as population parameters), considering several covariance structures, when item parameters are know and unknown, respectively. Recently, Azevedo et al. (2016) proposed a general class of IRT longitudinal models with multivariate normal distributions for the latent traits, considering a Bayesian framework. This class take into account important features of the longitudinal data, as time-heterogeneous latent trait variances and serial correlation. Furthermore, the authors proposed some Bayesian model fit assessment tools. However, the multivariate distribution considered to model the latent traits, make the development of MCMC algorithms (and even the obtaining of maximum likelihood estimates) computationally cumbersome, as the number of subjects and/or the number of measurement occasions increase. In addition, even though the authors explored some structured covariance matrices, their approach to estimate them involves a non structured full matrix. That is, despite of the restricted correlation pattern of the dependence matrix, the number of parameters to be estimated are not reduced and they increase as the number of measurement occasions increases. In this sense, it can be also difficult to handle unbalanced data, induced by dropouts and/or inclusion of subjects during the study, for example. Indeed, in this sense, the authors considered only the balanced case. In addition, they did not consider the guessing parameter in the item response function.

Another common assumption of the previous longitudinal IRT models is to consider the symmetric normal distribution (either the univariate or the multivariate) to model the latent traits structure. This assumption is often unrealistic and can lead to misleading inference, see Azevedo et al. (2011) and Azevedo et al. (2012a) and references therein. In longitudinal studies, is very common to observe asymmetry of the marginal latent traits distributions in several time-points, also due to inclusion/exclusion of the subjects along the study. Moreover, from educational point of view, is expected to observe a growing and/or decreasing in the latent traits, resulting in negative/positive asymmetry on the latent traits distribution.

Our goal is to develop a general Cholesky decomposition based modeling of longitudinal IRT data under asymmetry of the latent traits distribution. To accomplish for that, the Antedependence Models, see Pourahmadi (1999) and Nunez-Anton and Zimmerman (2000) were considered. This approach is very flexible and allows for handling multivariate distributions through univariate conditional distributions. This feature can reduce the computational cost in running MCMC algorithms, compared to the multivariate approach. It also allows to represent properly a wide range of specific correlation patterns, without considering additional random effects as in Azevedo (2008). In addition, it is quite useful to develop diagnostic tools as the residuals-based ones and those based on posterior predictive techniques. With respect to prior specifications, the antedependence modeling is quite interesting, since it allows to define more flexible priors for structured covariance matrices. Furthermore it can easily handle unbalanced data and different latent traits distribution. In this work the centered skew-normal (CSN) distribution (Azzalani, 1985) is considered for the error term, in order to characterize asymmetric behaviors of the latent traits. As pointed out in Azevedo et al. (2012b) the CSN distribution allows to identify the model straightforwardly. Also, regression structures for the latent traits, as growth curve models and for the item parameters, as differential item functioning, are more easily accommodate. All of these features can be considered as advantages of our approach, compared with the previous works.

Concerning to the item response function (IRF), most of the previous works did not the consider the effect of guessing on the probability of response for multiple choice items. This effect is important to improve the estimation of the latent traits, specially those related to subjects with low latent trait values. In this work we consider the three-parameter probit model (Baker and Kim, 2004). However, other IRF's can be also considered and properly accommodated by using the MCMC algorithms developed here through suitable augmented data structures.

This paper is outlined as follows. In Section 2 a review of the skew-normal distribution is presented. In Section 3, the IRT longitudinal antedependence model is presented along with some of its properties. In Section 4, we describe all steps of our MCMC algorithm. In the Section 5 some simulation studies are conducted to access the accuracy of our model and MCMC algorithm, concerning some features of interest. In Section 6 some model fit assessment tools are presented and a real data from the Brazilian school development program is analyzed. Finally, in Section 7 we presented some conclusions and remarks.

2 The Skew-Normal Distribution Under the Centered Parametrization

In order to make this work reasonably self-contained, we begin with an introduction to the skew-normal model. The skew-normal distribution belongs to a subclass of the elliptical distributions (Branco and Arellano-Valle, 2004). It has been used for modeling asymmetric data in many fields, including the psychometrical one, according to Bazán et al. (2004) and Azevedo et al. (2011).

A random variable θ follows the skew-normal (SN) distribution with location parameter $\xi \in \mathbb{R}$, scale parameter $\omega \in \mathbb{R}^+$ and shape parameter $\lambda \in \mathbb{R}$ (notation: $\theta \sim \mathrm{SN}(\xi, \omega, \lambda)$) if its p.d.f is given by

$$p(\theta; \xi, \omega, \lambda) = 2\omega^{-1}\phi_d\left(\frac{y-\xi}{\omega}\right)\Phi_d\left(\lambda\frac{\theta-\xi}{\omega}\right), \text{ for all } \quad \theta \in \mathbb{R}, \tag{1}$$

where ϕ_d and Φ_d denote the p.d.f and c.d.f of the standard normal distribution, respectively. The mean and variance of θ are given by, respectively

$$\mathbb{E}(\theta) = \xi + \omega\delta r \qquad \text{and} \qquad \mathrm{Var}(\theta) = \omega^2(1 - r^2\delta^2),$$

where,

$$r = \sqrt{\frac{2}{\pi}} \quad \text{and} \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$$

.

The parameter δ lies in the interval $(-1, 1)$ and it can be used in an alternative parametrization of the SN. Azzalani (1985) has been introduced the so called centered parametrization defined as follows,

$$\theta_c = \xi + \omega\theta_z = \mu + \sigma\theta_0, \tag{2}$$

where

$$\theta_z \sim \text{SN}(0, 1, \lambda), \quad \theta_0 = \sigma_z^{-1}(\theta_z - \mu_z)$$

and

$$\mu_z = \mathbb{E}(\theta_z) = r\delta, \qquad \sigma_z^2 = \text{Var}(Z) = 1 - \mu_z^2.$$

Then, θ_c follows centered skew-normal distribution (CSN) with parameters defined as:

$$\begin{aligned} \mu = \mathbb{E}(\theta_c) &= \xi + \omega\mu_z, \\ \sigma = \sqrt{\text{Var}(\theta_c)} &= \omega\sqrt{1-\mu_z^2}, \\ \gamma = \frac{\mathbb{E}[(\theta_c - \mathbb{E}(\theta_c))^3]}{\text{Var}(\theta_c)^{3/2}} &= r\delta^3 \left[\frac{4}{\pi} - 1\right] \left[1 - \mu_z^2\right]^{-3/2}. \end{aligned} \tag{3}$$

Therefore, the parameters μ and σ^2 are, respectively, mean and variance of the random variable θ_c. The parameter γ stands for the Pearson's asymmetry coefficient and lies in the interval $(-0.99527, 0.99527)$. The closer to -1 or 1 is the parameter γ the negative or positive skewed is the CSN distribution. The CSN distribution is (approximately) symmetric when $\gamma \in (-0.13, 0.13)$. It will be denoted as: $(\theta_c \sim CSN(\mu, \sigma^2, \gamma))$.

Furthermore, Fisher information matrix obtained through the CSN distribution is nonsingular for all γ and the likelihood is well behaved, unlike the usual skew-normal distribution. For more details, see Azevedo et al. (2011) and Pewsey (2000).

The p.d.f of the centered skew-normal distribution is given by

$$
\begin{aligned}
p(\theta_c) &= 2\frac{\sqrt{\sigma_z^2}}{\sigma}\phi\left[\frac{\sqrt{\sigma_z^2}}{\sigma}\left(\theta_c-\mu+\frac{\sigma}{\sqrt{\sigma_z^2}}\mu_z\right)\right]\Phi\left(\lambda\left[\frac{\sqrt{\sigma_z^2}}{\sigma}\left(\theta_c-\mu+\frac{\sigma}{\sqrt{\sigma_z^2}}\mu_z\right)\right]\right) \\
&= 2\omega^{-1}\phi\left(\omega^{-1}(\theta_c-\xi)\right)\Phi\left[\lambda(\omega^{-1}(\theta_c-\xi))\right], \qquad (4)
\end{aligned}
$$

which correspond to a usual skew-normal distribution with parameters defined as:

$$
\begin{aligned}
\xi &= \mu-\sigma\gamma^{1/3}\kappa, && (5)\\
\omega &= \sigma\sqrt{1+\gamma^{2/3}\kappa^2}, &&\\
\lambda &= \frac{\gamma^{1/3}\kappa}{\sqrt{r^2+\kappa^2\gamma^{2/3}(r^2-1)}} \text{ where} && (6)\\
\kappa &= (\frac{2}{4-\pi})^{1/3}. && (7)
\end{aligned}
$$

Another important result concerns to the Henze's stochastic representation (Henze, 1986). It means that a random variable $\theta \sim CSN(\mu,\sigma^2,\gamma)$ can be represented by

$$\theta=\xi+\omega\left(\delta X_1+\sqrt{1-\delta^2}X_2\right), \qquad (8)$$

where $X_1 \sim HN(0,1)$ and $X_2 \sim N(0,1)$ are independent random variables with half-normal and normal distribution, respectively. Conditional to X_1 we have

$$
\begin{aligned}
\theta|X_1 &\sim N(\xi+\omega\delta X_1,\omega^2(1-\delta^2)),\\
X_1 &\sim HN(0,1). \qquad (9)
\end{aligned}
$$

or

$$
\begin{aligned}
\theta|X_1 &\sim N(\alpha+\tau X_1,\varsigma^2), \qquad (10)\\
X_1 &\sim HN(0,1).
\end{aligned}
$$

where $\tau=\omega\delta$ and $\varsigma=\omega\sqrt{1-\delta^2}$. The last parametrization is more attractive by computational point of view, see Rodriguez (2005).

3 Modeling

The IRT data structure consists in T time-points which one with n_t subjects ($t = 1, 2, \ldots, T$), answering tests with I_t items. Common items are defined across the tests, and it can be recognized as an incomplete block design. Then, the total number of items is $I \leq \sum_{t=1}^{T} I_t$ and the total number of latent traits is $n = \sum_{t=1}^{T} n_t$. Dropouts and inclusions of subjects during the study are allowed. Let us define the following notation: θ_{jt} is the latent trait of the subject j ($j = 1, 2, \ldots, n_t$) at the time-point t, $\boldsymbol{\theta}_{j.}$ is the latent traits vector of the subject j, and $\boldsymbol{\theta}_{..}$ is the vector of all latent traits. Let Y_{ijt} denoting the response of the subject j to the item i ($i = 1, 2, \ldots, I$) at the time-point t, $\mathbf{Y}_{.jt} = (Y_{1jt}, \ldots, Y_{I_t jt})'$ is the response vector of subject j at the time-point t, $\mathbf{Y}_{...} = (\mathbf{Y}'_{..1}, \ldots, \mathbf{Y}'_{..T})'$ is the entire response matrix and $(y_{ijt}, \mathbf{y}'_{.jt}, \mathbf{y}'_{...})'$ are the respective observed values. Let $\boldsymbol{\zeta}_i$ represents the vector of item parameters of the item i, $\boldsymbol{\zeta}$ the vector of all item parameters and $\boldsymbol{\eta}_{\boldsymbol{\theta}}$ the vector of population parameters, related to the latent trait distribution.

Our IRT longitudinal model is defined in two levels: the level of responses and the level of the latent traits. In the first level is considered a probit three-parameter IRT model, which is suitable for dichotomous or dichotomized responses. In the second level we are assuming some appropriate multivariate distribution, that is

$$
\begin{aligned}
Y_{ijt}|\theta_{jt}, \boldsymbol{\zeta}_i &\sim \text{Bernoulli}(P_{ijt}), \\
P_{ijt} = \mathbb{P}(Y_{ijt}|\theta_{jt}, \boldsymbol{\zeta}_i) &= c_i + (1 - c_i)\Phi\left(a_i\theta_{jt} - b_i\right), \qquad (11) \\
\boldsymbol{\theta}_{j.}|\boldsymbol{\eta}_{\boldsymbol{\theta}} &\sim D_T(\boldsymbol{\eta}_{\boldsymbol{\theta}}), \qquad (12)
\end{aligned}
$$

where $D_T(.)$ stands for som T-dimensional skew-normal distribution indexed by the parameters $\boldsymbol{\eta}_{\boldsymbol{\theta}}$. In equation (11), a_i denote the discrimination parameter, $b_i = a_i b_i^*$, where b_i^* is the original difficulty parameter and c_i is the so called guessing parameter, see Baker and Kim (2004).

An important issue in longitudinal data analysis, concerns to the appropriate modeling of the covariance structure. A suitable specification of the correlation pattern is very important to explain the growth in latent traits, as pointed out by Azevedo et al. (2016). In this work, we will adapt for IRT context, the general procedure of covariance matrix estimation proposed by Pourahmadi (1999). Such approach is based on the Cholesky decomposition of the inverse of the covariance matrix (precision matrix) and allows to represent a wide range of the variance-covariance structures.

3.1 Antedependence models

To handling the multivariate structure of latent traits, we consider the so-called antedependence models, see Zimmerman and Núñez-Antón (2009). This approach offers a flexible way to deal with multivariate distribution and to represent covariance structures. In a general way, consider the latent traits $\boldsymbol{\theta}_{j.}$ with $\mathbb{E}(\boldsymbol{\theta}_{j.}) = \boldsymbol{\mu}_\theta$ and $\mathrm{Cov}(\boldsymbol{\theta}_{j.}) = \boldsymbol{\Sigma}_\theta$.

Then, we can write the latent trait of the subject j $(j = 1, \ldots, n_t)$ at the time-point t as:

$$\theta_{jt} = \mu_{\theta_t} + \sum_{k=1}^{t-1} \phi_{tk}(\theta_{jk} - \mu_{\theta_k}) + \varepsilon_{jt}, \quad t = 1, 2, \ldots, T. \tag{13}$$

where ϕ_{tk} are the so-called generalized autoregressive parameters, see Pourahmadi (1999). The parameters ϕ_{tk} should not be confused with notation ϕ_d related to the standard normal distribution. In addition, consider $\sum_{k=1}^{0} k = 0$.

In matrix form, we have:

$$\boldsymbol{\varepsilon}_{j.} = \mathbf{L}(\boldsymbol{\theta}_{j.} - \boldsymbol{\mu}_\theta), \tag{14}$$

This model was named by Zimmerman and Núñez-Antón (2009) the unstructured antedependence model. The random variables $\boldsymbol{\varepsilon}_{j.} = (\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jT})'$ are uncorrelated with $\mathrm{Cov}(\boldsymbol{\varepsilon}_j) = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix $\mathrm{diag}(d_1, d_2, \ldots, d_T)$ and $\mathbf{L}$ is a $(T \times T)$ lower-triangular matrix having the following form,

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{21} & 1 & 0 & \cdots & 0 \\ -\phi_{31} & -\phi_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\phi_{T1} & -\phi_{T2} & \cdots & -\phi_{T(T-1)} & 1 \end{pmatrix} \tag{15}$$

Then, from (14) and using the definition of $\mathbf{D}$ we have that,

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_{j.}) = \mathbf{L}\mathrm{Cov}(\boldsymbol{\theta}_{j.} - \boldsymbol{\mu}_\theta)\mathbf{L}' = \mathbf{L}\boldsymbol{\Sigma}_\theta\mathbf{L}' = \mathbf{D}. \tag{16}$$

Therefore the matrix $\mathbf{L}$ diagonalizes the covariance matrix $\boldsymbol{\Sigma}_\theta$. This result is related with a variant of the classical Cholesky decomposition (Newton, 1988) of the $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_\theta^{-1}$.

More parsimonious models, can be obtained by considering some specific correlation patterns. When the restricted covariance model is supported by the data, we can reduce, considerably, the number of parameters to be estimated and it can improve the model fit compared to the unstructured model. Furthermore, the unstructured pattern might not be appropriate in more complex situations as unbalanced data design, small sample sizes with respect to the number of subjects and items and many measurement occasions (or time-points), see Azevedo et al. (2016) and Jennrich and Schluchter (1986) for more details.

For example, consider $T = 3$ time-points and the following structured matrix:

$$\boldsymbol{\Sigma}_\theta = \begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1}\sigma_{\theta_2}\rho_\theta & \sigma_{\theta_1}\sigma_{\theta_3}\rho^2_\theta \\ \sigma_{\theta_1}\sigma_{\theta_2}\rho_\theta & \sigma^2_{\theta_2} & \sigma_{\theta_2}\sigma_{\theta_3}\rho_\theta \\ \sigma_{\theta_1}\sigma_{\theta_3}\rho^2_\theta & \sigma_{\theta_2}\sigma_{\theta_3}\rho_\theta & \sigma^2_{\theta_3} \end{pmatrix}$$

This is an extension of the *first-order autoregressive* matrix that allow heteroscedasticity. The $\mathbf{L}$ and $\mathbf{D}$ matrices are given by:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{\sigma_{\theta_2}}{\sigma_{\theta_1}}\rho_\theta & 1 & 0 \\ 0 & -\frac{\sigma_{\theta_3}}{\sigma_{\theta_2}}\rho_\theta & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} \sigma^2_{\theta_1} & 0 & 0 \\ 0 & \sigma^2_{\theta_2}(1-\rho^2_\theta) & 0 \\ 0 & 0 & \sigma^2_{\theta_3}(1-\rho^2_\theta) \end{pmatrix}$$

By induction and by using equation (14) we can obtain the following linear model:

$$\begin{aligned}\theta_{j1}-\mu_{\theta_1} &= \varepsilon_{j1},\\ \theta_{jt}-\mu_{\theta_t} &= \frac{\sigma_{\theta_t}}{\sigma_{\theta_{t-1}}}\rho_\theta(\theta_{j(t-1)}-\mu_{\theta_{t-1}})+\varepsilon_{jt}, \quad t=2,\ldots,T.\end{aligned} \tag{17}$$

Note that, the parameters $(d_1, d_2, \ldots, d_T, \phi_{21}, \phi_{31}, \phi_{32}, \ldots, \phi_{T(T-1)})'$ are an one-to-one mapping of the parameters of interest $(\sigma^2_{\theta_1}, \sigma^2_{\theta_2}, \ldots, \sigma^2_{\theta_T}, \rho_\theta)'$. This results is convenient to specify flexible prior distributions for the covariance parameters and also to implement MCMC algorithms.

Mostly, the matrices $\mathbf{L}$ and $\mathbf{D}$ do not have recognizable form. Then, for a more complex structured matrix it is very difficult or not possible to obtain a general expression for the antedependence model (as the expression (17)). However, the Cholesky decomposition can be obtained numerically. Table 1 presents some examples of structured matrices.

Table 1: Structured covariance matrices used in this work. The σ-parameters are related to variances, while ρ-parameters are related to correlations.

Structure	**Matrix form**
First-order Heteroscedastic Autoregressive: ARH(1)	$\begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1}\sigma_{\theta_2}\rho_\theta & \cdots & \sigma_{\theta_1}\sigma_{\theta_T}\rho_\theta^{T-1} \\ \sigma_{\theta_1}\sigma_{\theta_2}\rho_\theta & \sigma^2_{\theta_2} & \cdots & \sigma_{\theta_2}\sigma_{\theta_T}\rho_\theta^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\theta_1}\sigma_{\theta_T}\rho_\theta^{T-1} & \sigma_{\theta_2}\sigma_{\theta_T}\rho_\theta^{T-2} & \cdots & \sigma^2_{\theta_T} \end{pmatrix}$
First-order Heteroscedastic Autoregressive Moving-Average: ARMAH(1,1)	$\begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \cdots & \sigma_{\theta_1}\sigma_{\theta_T}\rho_{\theta_1}\rho_{\theta_2}^{T-2} \\ \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \sigma^2_{\theta_2} & \cdots & \sigma_{\theta_2}\sigma_{\theta_T}\rho_{\theta_1}\rho_{\theta_2}^{T-3} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\theta_1}\sigma_{\theta_T}\rho_{\theta_1}\rho_{\theta_2}^{T-2} & \sigma_{\theta_2}\sigma_{\theta_T}\rho_{\theta_1}\rho_{\theta_2}^{T-3} & \cdots & \sigma^2_{\theta_T} \end{pmatrix}$
Heteroscedastic Toeplitz: HT	$\begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \cdots & \sigma_{\theta_1}\sigma_{\theta_T}\rho_{\theta_{(T-1)}} \\ \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \sigma^2_{\theta_2} & \cdots & \sigma_{\theta_2}\sigma_{\theta_T}\rho_{\theta_{(T-2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\theta_1}\sigma_{\theta_T}\rho_{\theta_{(T-1)}} & \sigma_{\theta_2}\sigma_{\theta_T}\rho_{\theta_{(T-2)}} & \cdots & \sigma^2_{\theta_T} \end{pmatrix}$
Antedependence Matrix: AD	$\begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \cdots & \sigma_{\theta_1}\sigma_{\theta_T}\prod_{t=1}^{T-1}\rho_{\theta_t} \\ \sigma_{\theta_1}\sigma_{\theta_2}\rho_{\theta_1} & \sigma^2_{\theta_2} & \cdots & \sigma_{\theta_2}\sigma_{\theta_T}\prod_{t=2}^{T-1}\rho_{\theta_t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\theta_1}\sigma_{\theta_T}\prod_{t=1}^{T-1}\rho_{\theta_t} & \sigma_{\theta_2}\sigma_{\theta_T}\prod_{t=2}^{T-1}\rho_{\theta_t} & \cdots & \sigma^2_{\theta_T} \end{pmatrix}$

In order to represent the multivariate skew-normal structure of the latent traits, we will consider the following antedependence model:

$$\theta_{jt} = \mu_{\theta_t} + \sum_{k=1}^{t-1} \phi_{jk}(\theta_{jk} - \mu_{\theta_k}) + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim CSN(0, d_t, \gamma_{\varepsilon_t}), \quad t = 1, 2, \ldots, T. \tag{18}$$

Equivalently, we have

$$\boldsymbol{\varepsilon}_{j.} = \mathbf{L}(\boldsymbol{\theta}_{j.} - \boldsymbol{\mu}_\theta), \tag{19}$$

where $\mathbf{D}$ is a diagonal matrix as described before and $\boldsymbol{\varepsilon}_{j.}$ is a vector of independent centered skew-normal distributions as defined in equation (18). Then we are considering centered skew-normal distribution for the error term, with mean zero variances d_t and skewness coefficient γ_{ε_t}, that will be called conditional skewness coefficient. The multivariate distribution of the latent trait $\boldsymbol{\theta}_{j.}$ of the subject j can be characterized by the density bellow. According to the result (4) and (18) we have:

$$\begin{aligned} p(\boldsymbol{\theta}_{j.}) &= p(\theta_{j1}) \prod_{t=2}^{T} p(\theta_{jt} | \theta_{j(1:t-1)}) \\ &= 2^T \prod_{t=1}^{T} \omega_t^{-1} \phi_d \left(\omega^{-t}(\theta_{jt} - \beta_{jt}) \right) \Phi_d \left[\lambda_t(\omega_t^{-1}(\theta_{jt} - \beta_{jt})) \right], \quad j = 1, \ldots, n_t, \end{aligned} \tag{20}$$

where subscript $(1 : t-1)$ stands for the preceding latent traits and β_{jt} is defined as:

$$\begin{aligned} \beta_{j1} &= \xi_1 \\ \beta_{jt} &= \xi_t + \sum_{k=1}^{t-1} \phi_{jk}(\theta_{jk} - \mu_{\theta_k}), \quad j = 2, \ldots, n_t. \end{aligned} \tag{21}$$

To characterize marginal asymmetry, that is, the asymmetry of latent traits at the time-point t, we will also use the Pearson's skewness coefficient. This coefficient is given in terms of the centered moment of the random variable θ_{jt}. The following proposition provides a general expression for the third centered moment.

Proposition 3.1. *Consider the antedependence model defined in equation (18), and let l_{tk} denoting the entries of matrix $\mathbf{L}^{-1}$. Then the third centered moment of the random variable θ_{jt} is given by:*

$$\mathbb{E}[(\theta_{jt}-\mu_{\theta_t})^3] = d_t^{3/2}\gamma_{\varepsilon_t} + \sum_{k=1}^{t-1} l_{tk}^3 d_k^{3/2}\gamma_{\varepsilon_k}, \quad j=1,\dots,n_t \textit{ and } t=1,\dots,T. \tag{22}$$

Proof. By equation (14) we can write,

$$(\boldsymbol{\theta}_{j.} - \boldsymbol{\mu}_\theta) = \mathbf{L}^{-1}\boldsymbol{\varepsilon}_{j.}.$$

By definition, matrix $\mathbf{L}$ is a lower triangular matrix with ones in the main diagonal (see matrix (15)). Therefore, $\mathbf{L}^{-1}$ has the same form. Therefore we can write

$$(\theta_{jt}-\mu_{\theta_t}) = (\varepsilon_{jt} + \sum_{k=1}^{t-1} l_{tk}\varepsilon_{jk}) \Leftrightarrow (\theta_{jt}-\mu_{\theta_t})^3 = (\varepsilon_{jt} + \sum_{k=1}^{t-1} l_{tk}\varepsilon_{jk})^3, \quad t=1,\dots,T.$$

Taking expectations we have,

$$\mathbb{E}[(\theta_{jt}-\mu_{\theta_t})^3] = \mathbb{E}[(\varepsilon_{jt} + \sum_{k=1}^{t-1} l_{tk}\varepsilon_{jk})^3], \quad t=1,\dots,T.$$

By multinomial theorem, the term $(\varepsilon_{jt} + \sum_{k=1}^{t-1} l_{tk}\varepsilon_{jk})^3$ can be rewrite as:

$$(\varepsilon_{jt} + \sum_{k=1}^{t-1} l_{tk}\varepsilon_{jk})^3 = \sum_{k_1+k_2+\cdots+k_t=3} \binom{3}{k_1,k_2,\dots,k_t} \varepsilon_{jt}^{k_t} \prod_{1\le m\le t-1} l_{tm}^{k_m}\varepsilon_{jm}^{k_m}, \quad t=1,\dots,T,$$

where

$$\binom{3}{k_1,k_2,\dots,k_t} = \frac{3!}{k_1!\ k_2!\cdots k_t!}$$

is a multinomial coefficient. Indices k_1 through k_t are nonnegative integer, such that the sum of all k_m is 3. That is, for each term in the expansion, the exponents of the ε_{jm} must add up to 3. We can also see that,

$$\varepsilon_{jt}^3 + \sum_{k=1}^{t-1} l_{tk}^3\varepsilon_{jk}^3 + \sum_{\substack{k_1+k_2+\cdots+k_t=3,\\ k_m\in\{0,1,2\}}} \binom{3}{k_1,k_2,\dots,k_t} \varepsilon_{jt}^{k_t} \prod_{1\le m\le t-1} l_{tm}^{k_m}\varepsilon_{jm}^{k_m}, \quad t=1,\dots,T.$$

Now taking expectations,

$$\mathbb{E}[(\theta_{jt}-\mu_{\theta_t})^3] = \mathbb{E}(\varepsilon_{jt}^3) + \sum_{k=1}^{t-1} l_{tk}^3 \mathbb{E}(\varepsilon_{jk}^3) + \sum_{\substack{k_1+k_2+\cdots+k_t=3,\\ k_m\in\{0,1,2\}}} \binom{3}{k_1,k_2,\ldots,k_t} \mathbb{E}(\varepsilon_{jt}^{k_t}) \prod_{1\leq m\leq t-1} l_{tm}^{k_m}\mathbb{E}(\varepsilon_{jm}^{k_m}).$$

Since ε_{jt} are independent random variables with zero mean, the last term of the sum is equal to zero. Then,

$$\mathbb{E}[(\theta_{jt}-\mu_{\theta_t})^3] = \mathbb{E}(\varepsilon_{jt}^3) + \sum_{k=1}^{t-1} l_{tk}^3 \mathbb{E}(\varepsilon_{jk}^3).$$

Expectations $\mathbb{E}(\varepsilon_{jt}^3)$ are obtained in terms of the conditional skewness parameters. As we defined in equation (3) the skewness parameter is defined as:

$$\gamma_{\varepsilon_t} = \frac{\mathbb{E}(\varepsilon_{jt}^3)}{d_t^{3/2}} \quad \Leftrightarrow \quad \mathbb{E}(\varepsilon_{jt}^3) = d_t^{3/2}\gamma_{\varepsilon_t}.$$

Thus,

$$\mathbb{E}[(\theta_{jt}-\mu_{\theta_t})^3] = d_t^{3/2}\gamma_{\varepsilon_t} + \sum_{k=1}^{t-1} l_{tk}^3 d_k^{3/2}\gamma_{\varepsilon_k}. \tag{23}$$

□

The marginal skewness coefficient, denoted by γ_{θ_t}, is defined as: $\gamma_{\theta_t} = \mathbb{E}\left[\left(\frac{\theta_{jt}-\mu_{\theta_t}}{\sigma_{\theta_t}}\right)^3\right]$. According to Proposition 3.1 it is given by:

$$\begin{aligned}
\gamma_{\theta_t} &= \gamma_{\varepsilon_1} \\
\gamma_{\theta_t} &= \frac{1}{\sigma_{\theta_t}^3}\left[d_t^{3/2}\gamma_{\varepsilon_t} + \sum_{k=1}^{t-1} l_{tk}^3 d_k^{3/2}\gamma_{\varepsilon_k}\right], \quad t=2,\ldots,T,
\end{aligned} \tag{24}$$

where $\sigma_{\theta_t}^2$ are the marginal variances defined as in chapter 2. That is,

$$\begin{aligned}
\sigma_{\theta_1}^2 &= d_1 \\
\sigma_{\theta_t}^2 &= d_t + \sum_{k=1}^{t-1} l_{tk}^2 d_t, \quad t=2,\ldots,T.
\end{aligned} \tag{25}$$

It is importante to notice that the multivariate skew-normal (20) and its respective marginal distributions, are different of those know in the literature.

3.2 Model identification

To ensure identification we assume that the expectation and standard deviation of a reference time-point (in this case, time-point 1) are fixed, for example, at zero and one, respectively. In other words, we need to ensure that,

$$\theta_{j1} \sim CSN(0, 1, \gamma_{\theta_1}), \tag{26}$$

with a suitable common items structure along the administered tests. Therefore, the metric (scale) defined in model (11) is identified due to the fact that such model is no longer invariant to location-scale transformations, since that the expected value and the standard deviation of the latent distribution of the reference group (in this case, group 1) are fixed and also due to the linking design. This ensures that the metric for the latent traits is well defined and the results related to all tests (item parameters) and groups (latent traits and population parameters) lie on the same scale. In addition, the likelihood of our model is much improved compared with the one based on the ordinary skew-normal distribution.

4 Bayesian Estimation and MCMC Algorithms

In order to facilitates the implementation of the MCMC algorithms, particularly, aiming to obtain full conditional distribution with know form and to develop properly model-fit assessment tools; we will use the augmented data approach to represent our IRT model, see Tanner and Wong (1987). For the three-parameter probit model we can use the augmented data scheme proposed by Béguin and Glas (2001). This methodology consist on define a vector of binary variables W_{ijt} and continuous variables Z_{ijt} such that

$$W_{ijt} = \begin{cases} 1, & \text{if the subject } j\text{, at time-point } t \text{ knows the right response to the item } i \\ 0, & \text{if the subject } j\text{, at time-point } t \text{ does not know the right response to the item } i. \end{cases}$$

Consequently, the conditional distribution of W_{ijt} given $Y_{ijt} = y_{ijt}$ corresponds to

$$\begin{aligned} \mathbb{P}(W_{ijt} = 1 | Y_{ijt} = 1, \theta_{jt}, \boldsymbol{\zeta}_i) &\propto \Phi(a_i\theta_{jt} - b_i) \\ \mathbb{P}(W_{ijt} = 0 | Y_{ijt} = 1, \theta_{jt}, \boldsymbol{\zeta}_i) &\propto c_i(1 - \Phi(a_i\theta_{jt} - b_i)) \\ \mathbb{P}(W_{ijt} = 1 | Y_{ijt} = 0, \theta_{jt}, \boldsymbol{\zeta}_i) &= 0 \\ \mathbb{P}(W_{ijt} = 0 | Y_{ijt} = 0, \theta_{jt}, \boldsymbol{\zeta}_i) &= 1. \end{aligned} \tag{27}$$

Therefore the augmented variables $\mathbf{Z} = (Z_{111}, \ldots, Z_{1n_11}, \ldots, Z_{Tn_TI})'$, are given by

$$Z_{ijt}|(\theta_{jt}, \zeta_i, w_{ijt}) = \begin{cases} N(a_i\theta_{jt} - b_i, 1)\mathbb{I}_{(Z_{ijt}\geq 0)}, & \text{if } W_{ijt} = 1, \\ N(a_i\theta_{jt} - b_i, 1)\mathbb{I}_{(Z_{ijt}<0)}, & \text{if } W_{ijt} = 0. \end{cases} \tag{28}$$

The original data can be represented by

$$Y_{ijt} = \mathbb{I}(Z_{ijt} > 0)\mathbb{I}(W_{ijt} = 1) + \mathbb{I}(Z_{ijt} \leq 0)\mathbb{I}(W_{ijt} = 0), \tag{29}$$

where, $\mathbb{I}$ denotes the indicator function. To handle incomplete block design an indicator variable $\mathbf{I}$ is defined as:

$$\mathbf{I}_{ijt} = \begin{cases} 1, & \text{if item } i\text{, was administrated to the respondent } j \text{ at time-point } t, \\ 0, & \text{if item } i\text{, was not administrated to the respondent } j \text{ at time-point } t. \end{cases}$$

To describe possible omissions on the data, caused by uncontrolled events, such that, non-response or errors in recoding data, we defined another variable as follows,

$$V_{ijt} = \begin{cases} 1, & \text{if observed response of respondent } j \text{ at time-point } t \text{ on item } i, \\ 0, & \text{otherwise.} \end{cases}$$

We assumed that the missing data are missing at random (MAR), such that the missing data patterns distribution does not depend on the unobserved data. Therefore, the augmented likelihood is given by

$$L(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta, |\mathbf{z}_{...}, \mathbf{y}_{...}, \mathbf{I}) \propto \prod_{t=1}^{T} \prod_{j=1}^{n_t} \prod_{i \in I_{jt}} \exp\left\{-0.5(z_{ijt} - a_i\theta_{jt} + b_i)^2\right\} \mathbb{I}_{(z_{ijt}, y_{ijt})}, \tag{30}$$

where $\mathbb{I}_{(Z_{ijt}, y_{ijt})}$ stands for the indicator function $\mathbb{I}_{(z_{ijt}<0, y_{ijt}=0)} + 1\!\!\mathrm{l}_{(z_{ijt}\geq 0, y_{ijt}=1)}$ and I_{jt} is the set of items answered by the subject j at time t.

4.1 Prior specification and posterior distributions

The joint prior distribution of the unknown parameters is assumed to be

$$p(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_{\theta}|\boldsymbol{\eta}_{\zeta}, \boldsymbol{\eta}_{\eta}) = \left\{ p(\theta_{j1}|\boldsymbol{\eta}_{\theta_1}) \prod_{t=2}^{T} \prod_{j=1}^{n_t} p(\theta_{jt}|\theta_{j(1:t-1)}, \boldsymbol{\eta}_{\theta_t}) \right\} \left\{ \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i|\boldsymbol{\eta}_{\zeta}) \right\} \left\{ \prod_{t=1}^{T} p(\boldsymbol{\eta}_{\theta_t}|\boldsymbol{\eta}_{\eta}) \right\}. \tag{31}$$

where the subscript $(1 : t-1)$ denotes the preceding latent traits. The prior distributions of the latent traits are defined in equation 18. For the item parameters we have:

$$p(\boldsymbol{\zeta}_{i(-c_i)}) \propto \exp\left[-0,5(\boldsymbol{\zeta}_{i(-c_i)} - \boldsymbol{\mu}_{\zeta})' \boldsymbol{\Psi}_{\zeta}^{-1} (\boldsymbol{\zeta}_{i(-c_i)} - \boldsymbol{\mu}_{\zeta})\right] \mathbb{I}_{(a_i>0)}. \tag{32}$$

and,

$$c_i \sim \text{Beta}(a_c, b_c). \tag{33}$$

where $\boldsymbol{\zeta}_{i(-c_i)} = (a_i, b_i)$. In order to obtain full conditional distributions for the population parameters, we consider the Henze's stochastic representation (34). Therefore, the model (18) can be rewrite as:

$$\begin{aligned} \theta_{j1}|h_{j1} &\sim N(\xi_1 + \tau_1 h_{j1}, \varsigma_1^2), \\ \theta_{jt}|h_{jt} &\sim N\left(\xi_t + \sum_{k=1}^{t-1} \phi_{tk}(\theta_{jk} - \mu_k) + \tau_t h_{jt}, \varsigma_t^2\right), \\ H_{jt} &\sim HN(0,1), \quad t = 1, \ldots, T \text{ and } j = 1, \ldots, n_t. \end{aligned} \tag{34}$$

where $\omega_t = \sqrt{\tau_t^2 + \varsigma_t^2}$ and $\lambda_t = \frac{\tau_t}{\varsigma_t}$. The original parameters can be recovered by using equations (3), that is:

$$\mu_t = \xi_t + r\delta_t\omega_t; \quad d_t = \omega_t^2(1 - r^2\delta^2)^2; \quad \gamma_t = r\delta_t^3 \left[\frac{4}{\pi} - 1\right] \left[1 - r^2\delta^2\right]^{-3/2}. \tag{35}$$

where $\delta_t = \frac{\lambda_t}{\sqrt{1+\lambda^2}}$. For the Henze's parameters we are considering the following conjugate priors:

$$
\begin{aligned}
\xi_t &\sim N(\mu_\xi, \sigma_\xi^2), \\
\tau_t &\sim N(\mu_\tau, \sigma_\tau^2), \\
\varsigma_t^2 &\sim IG(a_\varsigma, b_\varsigma), \quad t = 1, \ldots, T. \qquad (36) \\
& \qquad (37)
\end{aligned}
$$

This parameters will be considered just in the MCMC sampling process and the original parameters (35) will be recovered later. For the generalized autoregressive parameters we define:

$$
\phi_{tk} \sim N(\mu_\phi, \sigma_\phi^2) \quad t = 2, \ldots, T \text{ and } k = 1, \ldots, t-1. \tag{38}
$$

In the case of structured matrix, the prior distributions for correlation parameters are directly specified as:

$$
\rho_{\theta_t} \sim N(\mu_\rho, \sigma_\rho^2)\mathbb{I}_{[0,1]}, \quad t = 1, 2, \ldots, T-1. \tag{39}
$$

That is, a truncated normal distribution on the interval $[0, 1]$. This interval was consider since negative correlations are rarely observed in longitudinal studies.

Given the augmented likelihood in equation (30) and the prior distribution in equations (34), (32), (33), (36) and (38), the joint posterior distribution is given by:

$$
\begin{aligned}
p(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}, \boldsymbol{\eta}_\theta, |\mathbf{z}_{...}, \mathbf{y}_{...}, \mathbf{I}) \;\propto\; & \left\{ \prod_{t=1}^{T} \prod_{j=1}^{n} \prod_{i \in I_{jt}} \exp\left\{-0.5(z_{ijt} - a_i\theta_{jt} + b_i)^2\right\} \mathbb{I}_{(Z_{ijt}, y_{ijt})} \right\} \\
\times & \left\{ \prod_{j=1}^{n_t} p(\theta_{j1}, h_{jt}, \boldsymbol{\eta}_{\theta_t}) p(h_{j1}) \prod_{t=2}^{T} p(\theta_{jt}|\theta_{j(1:t-1)}, h_{jt}, \boldsymbol{\eta}_{\theta_t}) p(h_{jt}) \right\} \\
\times & \left\{ \prod_{i=1}^{I} p(\boldsymbol{\zeta}_i | \boldsymbol{\eta}_\zeta) \right\} \left\{ \prod_{t=1}^{T} p(\boldsymbol{\eta}_{\theta_t} | \boldsymbol{\eta}_\eta) \right\}
\end{aligned} \tag{40}
$$

where $\boldsymbol{\eta}_\zeta$ and $\boldsymbol{\eta}_\eta$ are hyperparameters associated with $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}_\theta$, respectively, and subscript $(1 : t-1)$ denotes the preceding latent traits. In addition, we are assuming independence between items and population parameters. Since the posterior distribution has an intractable analytical form, we will use MCMC algorithms in order to obtain empirical approximation for the marginal posterior distribution of interest. In this sense, we can see that the full conditional distribution of the model parameters $\boldsymbol{\zeta}, \boldsymbol{\eta}_\theta$ are easy to sample from, at least for the unstructured covariance matrix. For the items parameters we have that:

$$\boldsymbol{\zeta}_{i(-c_i)}|(.) \sim N(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{\zeta}_i}\widehat{\boldsymbol{\zeta}}_i, \widehat{\boldsymbol{\Psi}}_{\boldsymbol{\zeta}_i}) \text{ where,} \tag{41}$$

$$\begin{aligned}
\widehat{\boldsymbol{\zeta}}_i &= (\boldsymbol{\Theta}_{i.})' \mathbf{z}_{i.} + \boldsymbol{\Psi}_{\boldsymbol{\zeta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\zeta}}, \\
\widehat{\boldsymbol{\Psi}}_{\boldsymbol{\zeta}_i} &= \left[(\boldsymbol{\Theta}_{i.})' (\boldsymbol{\Theta}_{i.}) + \boldsymbol{\Psi}_{\boldsymbol{\zeta}}^{-1}\right]^{-1}, \\
\boldsymbol{\Theta}_{i.} &= [\boldsymbol{\theta} - \mathbf{1}_n] \bullet \mathbb{1}_i,
\end{aligned}$$

(.) denotes the set of all others parameters, $\mathbb{1}_i$ is a $(n \times 2)$ matrix with lines, equals to 1 or 0, according to the response/missing response of the subject j to the item i at time-point t and $\bullet$ denotes the *Hadamard* product and for *guessing parameters*,

$$c_i|(.) \sim \text{Beta}(s_i + a_c - 1, t_i - s_i + b_c - 1) \text{ where,} \tag{42}$$

$$s_i = \sum_{j|w_{ijt}=0}^{n} \mathbf{y}_{ij.}; \quad \sum_{j=1}^{n} \mathbb{I}(w_{ijt} = 0).$$

Considering β_{jt} defined as before, for the random effect H_{jt} we have:

$$H_{jt}|(.) \sim N\left[\frac{\tau_t(\theta_{jt} - \beta_{jt})}{\tau_t^2 + \varsigma_t^2}, \frac{\varsigma_t^2}{\tau_t^2 + \varsigma_t^2}\right] \mathbb{I}(h_{jt} > 0) \tag{43}$$

For the Henze's stochastic representation parameters we have:

$$\xi_t|(.) \sim N(aA; A), \tag{44}$$

where

$$a = \frac{1}{\varsigma_t^2}\sum_{j=1}^{n_t}\left[\theta_{jt} - \sum_{k=1}^{t-1}\phi_{tk}\left(\theta_{jk} - \mu_{\theta_k}\right) - \tau_t h_{jt}\right] + \frac{\mu_\xi}{\sigma_\xi^2} \quad \text{and} \quad A = \left(\frac{n_t}{\varsigma_t^2} + \frac{1}{\sigma_\xi^2}\right)^{-1}$$

$$\tau_t|(.) \sim N(bB; B), \tag{45}$$

where

$$b = \frac{1}{\varsigma_t^2} \sum_{j=1}^{n_t} (\theta_{jt} - \beta_{jt})\, h_{jt} + \frac{\mu_\tau}{\sigma_\tau^2} \quad \text{and} \quad B = \left(\frac{\sum_{j=1}^{n_t} h_{jt}^2}{\varsigma_t^2} + \frac{1}{\sigma_\tau^2} \right)^{-1}$$

and

$$\varsigma_t^2 |(.) \sim IG \left\{ \frac{n_t}{2} + a_\varsigma; \frac{1}{2} \sum_{j=1}^{n_t} (\theta_{jt} - \beta_{jt} - \tau_t h_{jt})^2 + b_\varsigma \right\} \tag{46}$$

where IG stands for the inverse gamma distribution. The generalized autoregressive parameters are simulated by

$$\phi_{tk}|(.) \sim N(Q_{tk} q_{tk}, Q_{tk}) \text{ where,}$$

$$\begin{aligned} Q_{tk} &= \left(\frac{\sum_{j=1}^{n_t} (\theta_{j(t-1)} - \mu_{\theta_{t-1}})^2}{\varsigma_t^2} + \frac{1}{\sigma_\phi^2} \right)^{-1} \\ q_{tk} &= \frac{1}{\varsigma_t^2} \sum_{j=1}^{n_t} (\theta_{jk} - \mu_{\theta_k})(\theta_{jt} - \xi_t - \sum_{k \neq t} \phi_{tk}(\theta_{jk} - \mu_{\theta_k}) - \tau_t h_{jt}). \end{aligned} \tag{47}$$

for all $t = 2, \ldots, T$ and $k = 1, \ldots, (t-1)$. In the next sections we will discuss the sampling of the latent traits and correlation parameters of the structured matrices.

4.2 The latent traits sampling

A common way to sample from the $\boldsymbol{\theta}_{j.}$ posterior distribution it is to consider univariate full conditional distributions, in order to sample a time-specific latent trait given all other ones, using Gibbs sampling. It means that: $\theta_{jt}|\boldsymbol{\theta}_{j(-t)}$ for all $t = 1, \ldots, T$, where $\boldsymbol{\theta}_{j(-t)}$ is the latent traits vector without the t-th component. However, this procedure can generate chains with high autocorrelations, specially in the presence of many time-points (Gamerman and Lopes, 2006). Carter and Kohn (1994) and Frühwirth-Schnatter (1994), have proposed a sampling scheme for *dynamic models*, which allows to sample the so-called state parameters jointly, based on the Kalman filter. It is called **F***orward* **F***iltering* **B***ackward* **S***ampling*.

4.2.1 Dynamic Models

Dynamic models are defined by the pair of equations called *observation* and *system or evolution* equations. Using the notation of the Gamerman and Lopes (2006) we have:

$$
\begin{aligned}
u_t &= F_t^{\top}\theta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2), && (48)\\
\theta_t &= G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, \Omega_t). && (49)
\end{aligned}
$$

where u_t is a sequence of observations at the time, conditionally independent given θ_t and σ_t^2. The model is completed with the prior $\theta_1 \sim N(r, R)$. For our IRT model, considering the augmented data structure, we have the following representation in terms of dynamic model:

$$
\begin{aligned}
Z_{jit} &= a_i\theta_{jt} - b_i + \xi_{jit}, \quad \xi_{jit} \sim N(0, 1), && (50)\\
\theta_{jt} &= \beta_{jt} + \tau_t h_{jt} + \varepsilon_{jt}, \quad \varepsilon_{jt}|h_{jt} \sim N(0, \varsigma_t^2). && (51)\\
& && (52)
\end{aligned}
$$

where β_{jt} is as defined before and $H_{jt} \sim HN(0, 1)$, for all $i \in I_{jt}$, $j = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$.

4.2.2 FFBS Algorithm

The FFBS algorithm basically consist on two steps: *forward* and *backward*. The forward step is performed by the Kalman filter procedure described below. Following Gamerman and Lopes (2006) consider the conditional distribution $\theta_{j(t-1)}|\mathbf{Z}_{j.}^{t-1} \sim N(m_{j(t-1)}, C_{j(t-1)})$, where $\mathbf{Z}_{j.}^{t-1}$ refer to the information until $t-1$. The system equations (52) can be written as $\theta_{jt}|\theta_{j(t-1)} \sim N(\mu_t + \sum_{k=1}^{t-1}\phi_{tk}(\theta_{jk} - \mu_k), d_t)$. By proprieties of the normal distribution, these specifications can be combined leading to the marginal distribution:

$$
\theta_{jt}|\mathbf{Z}_{j.}^{t-1} \sim N(a_{jt}, R_{jt}) \text{ where,} \tag{53}
$$

$$
a_{jt} = \xi_t + \sum_{k=1}^{t-1}\phi_{tk}(m_{jk} - \mu_k) + \tau_t h_{jt} \quad \text{and} \quad R_{jt} = \varsigma_t^2 + \sum_{k=1}^{t-1}\phi_{tk}^2 C_{jk}.
$$

Thus,

$$
\theta_{jt}|\mathbf{Z}_{j.}^{t} \sim N(m_{jt}, C_{jt}) \text{ where,} \tag{54}
$$

$$C_{jt} = \left(\sum_{i\in I_{jt}} a_i^2 + \frac{1}{R_{jt}}\right)^{-1} \quad \text{and} \quad m_{jt} = \left(\sum_{i\in I_{jt}} a_i(z_{ijt} + b_i) + \frac{a_{jt}}{R_{jt}}\right) C_{jt}.$$

Equation 54 is referred in the literature as *Kalman Filter*. Therefore, the backward distributions are given by:

$$\theta_{jt}|\theta_{j(t+1)}, \mathbf{Z}_{j.}^t \sim N(m_{\theta_t}, C_{\theta_t}) \text{ where,} \tag{55}$$

$$C_{\theta_{jt}} = \left(\frac{\phi_{t+1,t}^2}{\varsigma_{t+1}^2} + \frac{1}{C_{jt}}\right)^{-1} \quad \text{and} \quad m_{\theta_{jt}} = \left(\frac{\phi_{t+1,t}\left(\theta_{j(t+1)} - \alpha_{j(t+1)}\right)}{\varsigma_{t+1}^2} + \frac{m_{jt}}{C_{jt}}\right) C_{\theta_{jt}}.$$

where $\alpha_{j(t+1)} = \xi_{t+1} - \phi_{t+1,t}\mu_t + \tau_t h_{j(t+1)} + \sum_{k=1}^{t-1} \phi_{t+1,t}(\theta_{jk} - \mu_k)$, for all $j = 1, \ldots, n_t$ and $t = 1, \ldots, T$. Then, a scheme to sample from the full conditional distribution of $\boldsymbol{\theta}_{j.}$ is given by

Algorithm 4.1 FFBS algorithm

1: Sample θ_{jT} from $\theta_{jT}|\mathbf{Z}_{j.}^T$ and set $t = T - 1$.
2: Sample θ_{jt} from $\theta_{jt}|\theta_{j(t+1)}, \boldsymbol{Z}_{j.}^t$.
3: Decrease t to $t - 1$ and return to step 2 until $t = 1$.

Step 1 is obtained by running the Kalman filter from $t = 1$ to $t = T$.

4.3 An alternative correlation parameters sampler

When a specific structure is imposed, the full conditional distributions of the correlation parameters are difficult to obtain, unlike to the generalized autoregressive parameters ϕ_{tk}. In this section we proposed an auxiliary sampler for the correlation parameters for structured matrices.

An usual way for sampling from an intractable posterior distribution it is to use the *Metropolis-Hastings* algorithm, see Gamerman and Lopes (2006). It requires a proposal distribution (a common choice is a random walk centred on the previous result) to generate proposed values. This value is accepted or rejected according to a transition probability. Then, the acceptation rate of the proposed values will depend on the variance of the proposal distribution, such that, when it is too large most values are rejected, whereas when it is too small only small steps are taken and the chain does not mix properly, see Marsman (2014). An alternative to overcome this problem was prosed by Murray et al. (2012), see algorithm (4.2). The so called *Single-Variable Exchange* algorithm (SVE) differs in the way to propose *candidates* values.

Specifically, consider the following posterior distribution:

$$p(\vartheta|\mathbf{x}) \propto p(\mathbf{x}|\vartheta)p(\vartheta), \tag{56}$$

where ϑ and $\mathbf{x}$ represent a parameter and a set of observations, respectively. Therefore, $p(\mathbf{x}|\vartheta)$ and $p(\vartheta)$ denote the likelihood and prior distributions, respectively. Then an scheme to sample from posterior distribution (56) is given by algorithm 4.2.

Algorithm 4.2 The Single-Variable Exchange algorithm

1: Draw $\vartheta^{(m)} \sim p(\vartheta)$
2: Draw $\mathbf{x}^{(m)} \sim p(\mathbf{x}|\vartheta^{(m)})$
3: Draw $u \sim U(0,1)$
4: **if** $(u < \pi(\vartheta^{(m-1)} \to \vartheta^{(m)}))$ **then**
5: $\quad \vartheta^{(m-1)} = \vartheta^{(m)}$
6: **end if**

where,

$$\pi(\vartheta^{(m-1)} \to \vartheta^{(m)}) = \min\left\{1, \frac{p(\mathbf{x}|\vartheta^{(m)})p(\mathbf{x}^{(m)}|\vartheta^{(m-1)})}{p(\mathbf{x}|\vartheta^{(m-1)})p(\mathbf{x}^{(m)}|\vartheta^{(m)})}\right\}.$$

and $\vartheta^{(m)}$, $\vartheta^{(m-1)}$ denotes a proposed and current values, respectively, for all simulation $m = 1, \ldots, M$.

Although the prior distribution closely resembles the target one, the SVE algorithm tends to frequently generate transition kernels for which the acceptance probability is low. To improve the efficiency of the SVE algorithm, in the sense to concentrate more probability mass on transition kernels with high acceptance probability, Marsman (2014) has been proposed an oversampling approach. It consists to simulate a number of i.i.d proposal values, instead of simulate a single one. Let $t(\mathbf{x})$ a sufficient statistic for the parameter ϑ and $t(\mathbf{x}^{(m)})$ the corresponding sufficient statistic of the proposed value $\vartheta^{(m)}$. Let s being the number of proposed values each one with its own sufficient statistic, we have to choose the one whose statistic $t(\mathbf{x}^{(m)})$ is closest to the observed $t(\mathbf{x})$, see algorithm (4.3).

To illustrate this algorithm consider, for example, an ARH(1) matrix, see Table 1. Let $\boldsymbol{\theta}_{..}$ being the set of all latent traits, as defined in Section 3 and $p(\boldsymbol{\theta}_{..}|\boldsymbol{\mu}_\theta, \boldsymbol{\phi}, \mathbf{d}, \boldsymbol{\gamma}_\theta)$ denoting the likelihood generated by the antedependence model defined in equation (18), that is

Algorithm 4.3 The Single-Variable Exchange algorithm with Oversampling

```
for s = 1 to S do
    Draw ϑ_s^(m) ~ p(ϑ)
    Draw x_s^(m) ~ p(x|ϑ_s^(m))
    Compute t(x_s^(m))
end for
Choose the ϑ_s^(m) whose t(x_s^(m)) is closest to t(x)
Draw u ~ U(0,1)
if (u < π(ϑ^(m-1) → ϑ^(m))) then
    ϑ^(m-1) = ϑ^(m)
end if
```

$$
\begin{aligned}
p(\boldsymbol{\theta}_{..}|\boldsymbol{\mu}_\theta, \boldsymbol{\phi}, \mathbf{d}, \boldsymbol{\gamma}_\theta) &= \prod_{j=1}^{n_t} p(\theta_{j1}|\mu_{\theta_1}, d_1, \gamma_{\theta_1}) \prod_{t=2}^{T} p(\theta_{jt}|\mu_{\theta_1}, \boldsymbol{\phi}_t, d_t, \gamma_{\theta_t}) \\
&\propto 2^{Tn_t} \prod_{j=1}^{n_t} \prod_{t=1}^{T} \omega_t^{-1} \phi_d \left(\omega_t^{-1}(\theta_{jt} - \beta_{jt})\right) \Phi_d \left[\lambda_t(\omega_t^{-1}(\theta_{jt} - \beta_{jt}))\right]. \qquad (57)
\end{aligned}
$$

where $\boldsymbol{\phi} = (\phi_{21}, \phi_{31}, \phi_{32}, \ldots, \phi_{T(T-1)})'$, $\mathbf{d} = (d_1, \ldots, d_T)'$, $\boldsymbol{\gamma}_\theta = (\gamma_{\theta_1}, \ldots, \gamma_{\theta_T})'$, $\boldsymbol{\mu}_\theta = (\mu_{\theta_1}, \ldots, \mu_{\theta_t})$ and $\boldsymbol{\phi}_t$ denotes the elements correspondents to the time-point t. Let us remember that β_{jt} are defined as:

$$
\begin{aligned}
\beta_{j1} &= \xi_1 \\
\beta_{jt} &= \xi_t + \sum_{k=1}^{t-1} \phi_{jk}(\theta_{jk} - \mu_{\theta_k}), \quad j = 2, \ldots, n_t, \qquad (58)
\end{aligned}
$$

and

$$
\begin{aligned}
\xi_t &= \mu_{\theta_t} - \sqrt{d_t}\gamma_{\theta_t}^{1/3}\kappa, \qquad (59) \\
\omega_t &= \sqrt{d_t(1 + \gamma_{\theta_t}^{2/3}\kappa^2)}, \qquad (60) \\
\lambda_t &= \frac{\gamma_{\theta_t}^{1/3}\kappa}{\sqrt{r^2 + \kappa^2\gamma_{\theta_t}^{2/3}(r^2 - 1)}}, \qquad (61)
\end{aligned}
$$

where r and κ are the constants defined in Section 2. Algorithm 4.4 present the correlation parameters sampler for the ARH(1) model for all simulation $m = 1, \ldots, M$.

Algorithm 4.4 The SVE algorithm with oversampling to sample a correlation parameter considering ARH(1) matrix

Require: A function `chol()` to perform the Cholesky decomposition
Require: A function `AR1.matrix()` to build the ARH(1) matrix
1: **for** $s = 1$ to S **do**
2: Draw $\rho_{\theta_s}^{(m)} \sim p(\rho_\theta)$
3: Draw $\boldsymbol{\theta}_{..s}^{(m)}$ from the model (18)
4: Compute $r_1(\boldsymbol{\theta}_{..s}^{(m)})$ the first-order sample correlation
5: **end for**
6: Choose the $\rho_{\theta_s}^{(m)}$ and $\boldsymbol{\theta}_{..s}^{(m)}$ whose $r_1(\boldsymbol{\theta}_{..s}^{(m)})$ is closest to $r_1(\boldsymbol{\theta}_{..}^{(m-1)})$
7: Set $\rho_\theta^{(m)} = \rho_{\theta_s}^{(m)}$ and $\boldsymbol{\theta}_{..}^{(m)} = \boldsymbol{\theta}_{..s}^{(m)}$ the candidate values
8: Build the ARH(1) proposed matrix $\Sigma_{\rho_\theta}^{(m)}$ using `ARH1.matrix()`
9: Perform the Cholesky decomposition of $\Sigma_{\rho_\theta}^{(m)}$ to obtain the matrices $\mathbf{L}^{(m)}$ and $\mathbf{D}^{(m)}$
10: Draw $u \sim U(0,1)$
11: **if**

$$\left(u < \min\left\{1, \frac{p(\boldsymbol{\theta}_{..}^{(m-1)}|\boldsymbol{\mu}_\theta^{(m-1)}, \boldsymbol{\phi}^{(m)}, \mathbf{d}^{(m)}, \boldsymbol{\gamma}_\theta^{(m-1)})p(\boldsymbol{\theta}_{..}^{(m)}|\boldsymbol{\mu}_\theta^{(m-1)}, \boldsymbol{\phi}^{(m-1)}, \mathbf{d}^{(m-1)}, \boldsymbol{\gamma}_\theta^{(m-1)})}{p(\boldsymbol{\theta}_{..}^{(m-1)}|\boldsymbol{\mu}_\theta^{(m-1)}, \boldsymbol{\phi}^{(m-1)}, \mathbf{d}^{(m-1)}, \boldsymbol{\gamma}_\theta^{(m-1)})p(\boldsymbol{\theta}_{..}^{(m)}|\boldsymbol{\mu}_\theta^{(m-1)}, \boldsymbol{\phi}^{(m)}, \mathbf{d}^{(m)}, \boldsymbol{\gamma}_\theta^{(m-1)})}\right\}\right)$$

then

12: $\rho_\theta^{(m-1)} = \rho_\theta^{(m)}$
13: **end if**

For more than one correlation parameter, the algorithm 4.4 can be applied independently to each one by choosing suitable sufficient statistics. It is also possible to sample blocks of correlation parameters. This can be done by modifying line 11 of the algorithm 4.4 to allow accept/reject proposed values jointly.

In summary, a general algorithm to estimate the parameter's model it is combination of Gibbs sampling with the FFBS and SVE algorithms as we can see in algorithms 4.5 and 4.6.

Algorithm 4.5 Gibbs sampling with FFBS for unstructured matrix

1: Start the algorithm by choosing suitable initial values. Repeat steps 2-9.
2: Simulate W_{ijt} from $W_{ijt}|(.)$ for all $i = 1, \dots, I$, $j = 1, \dots, n$ and $t = 1, \dots, T$.
3: Simulate Z_{ijt} from $W_{ijt}|(.)$ for all $i = 1, \dots, I$, $j = 1, \dots, n$ and $t = 1, \dots, T$.
4: Simulate H_{jt} from $H_{jt}|(.)$ for all $j = 1, \dots, n$ and $t = 1, \dots, T$.
5: Simulate θ_{jt} using the algorithm 4.1 for all $j = 1, \dots, n$ and $t = 1, \dots, T$.
6: Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i|(.)$ for all $i = 1, \dots, I$.
7: Simulate c_i from $c_i|(.)$ for all $i = 1, \dots, I$.
8: Simulate ξ_t from $\xi_t|(.)$ for all $t = 1, \dots, T$.
9: Simulate τ_t from $\tau_t|(.)$ for all $t = 1, \dots, T$.
10: Simulate ς_t^2 from $\varsigma_t^2|(.)$ for all $t = 1, \dots, T$.
11: Simulate ϕ_{tk} from $\phi_{tk}|(.)$ for all $t = 2, \dots, T$ and $k = 1, \dots, t-1$.

Algorithm 4.6 Gibbs sampling with FFBS for structured matrices

1: Start the algorithm by choosing suitable initial values. Repeat steps 2-9.
2: Simulate W_{ijt} from $W_{ijt}|(.)$ for all $i = 1, \dots, I$, $j = 1, \dots, n$ and $t = 1, \dots, T$.
3: Simulate Z_{ijt} from $W_{ijt}|(.)$ for all $i = 1, \dots, I$, $j = 1, \dots, n$ and $t = 1, \dots, T$.
4: Simulate H_{jt} from $H_{jt}|(.)$ for all $j = 1, \dots, n$ and $t = 1, \dots, T$.
5: Simulate θ_{jt} using the algorithm 4.1 for all $j = 1, \dots, n$ and $t = 1, \dots, T$.
6: Simulate $\boldsymbol{\zeta}_i$ from $\boldsymbol{\zeta}_i|(.)$ for all $i = 1, \dots, I$.
7: Simulate c_i from $c_i|(.)$ for all $i = 1, \dots, I$.
8: Simulate ξ_t from $\xi_t|(.)$ for all $t = 1, \dots, T$.
9: Simulate τ_t from $\tau_t|(.)$ for all $t = 1, \dots, T$.
10: Simulate ς_t^2 from $\varsigma_t^2|(.)$ for all $t = 1, \dots, T$.
11: Simulate ρ_{θ_t} using a SVE procedure for all necessary correlation parameters.

5 Parameter recovery study

In this section we study the efficiency of our model and the proposed estimation algorithm concerning parameter recovery. Our algorithm allows to consider different structured covariance matrices. Some examples can be seen in Table 1. For simplicity and without lose generality, the AD matrix was chosen to procedure the parameter recovery, since it is the most general matrix considered in this work.

Responses of $n_t = 1500$ subjects, for all t along of $T = 6$ time points were simulated according to the longitudinal model described in Section 3, considering the AD matrix. The items parameters were fixed in the following intervals: $a_i \in [0.7, 2.62]$, $b_i^* \in [-1.95, 4]$ and the guessing parameter c_i assume the values $(0.20, 0.21, 0.22, 0.23, 0.24, 0.25)$. The values of the difficulty parameters were fixed in order to consider low, middle and high difficulty in the items, with respect to the mean of the latent traits along the time-points. Similarly, we fixed high, middle and low discrimination. The tests structure is described as follows:

- Test 1: 20 items;
- Test 2: Test 1 + 20 other items;
- Test 3: the last 20 items of test 2 + 20 other items;
- Test 4: the last 20 items of test 3 + 20 other items;
- Test 5: the last 20 items of test 4 + 20 other items;
- Test 6: the last 20 items of test 5 + 20 other items;

Therefore, we have a total of $I = 120$ items. The latent traits were simulated from model 18 considering: $\boldsymbol{\mu}_\theta = (0.0, 1.0, 1.4, 2.0, 2.3, 2.5)'$, $\boldsymbol{\sigma}^2_\theta = (1.00, 1.27, 0.90, 0.88, 0.70, 0.65)'$ and $\boldsymbol{\gamma}_\theta = (0.80, 0.55, 0.18, 0.27, 0.22, -0.04)'$ being the vector of marginal skewness coefficients. The correlation parameters was fixed as $\boldsymbol{\rho}_\theta = (0.81, 0.89, 0.93, 0.73, 0.89)'$. We fixed increasing values for the population means on the $(0, 1)$ scale (which correspond, respectively, to mean and variance of the latent traits in the first time-point), meaning that, the average latent traits of the respondents increased during the study. This is an expected behavior in educational longitudinal studies, for example, see Santos et al. (2013) and Azevedo et al. (2012b). The values for the population variances were fixed in order to have a increasing and then a decreasing behavior. Concerning the correlation parameters, we fixed high values in order to obtain a pattern similar to that observed in the real data.

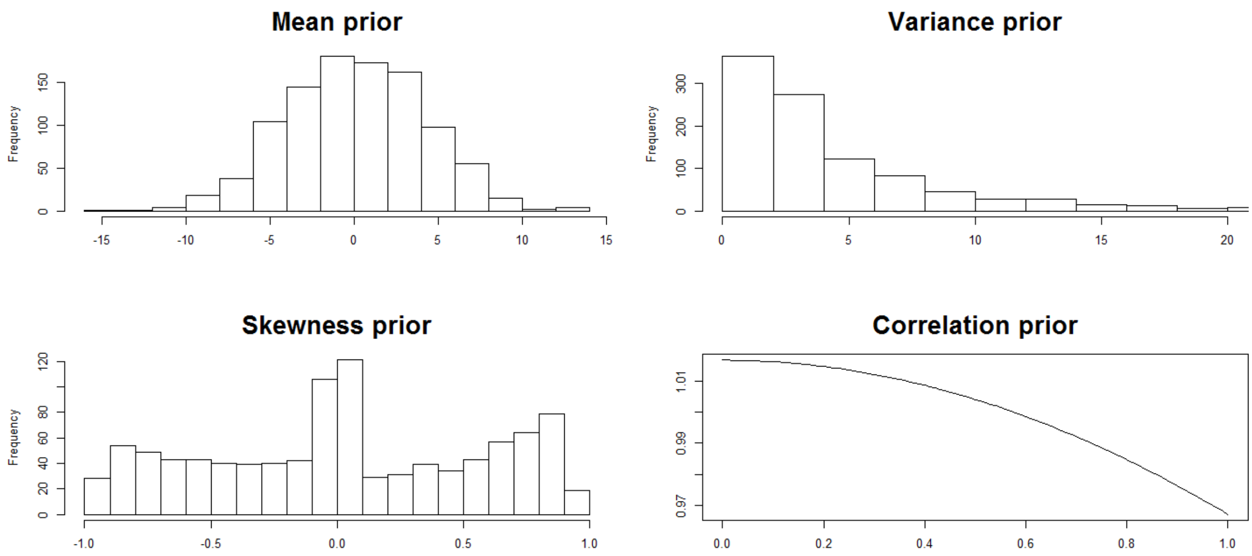

Figure 1: Prior distributions of the population parameters

Table 2 presents the hyperparameters for the prior distributions. Figure 1 presents the behavior of the prior distribution of the population parameters. The priors of the population mean, variance and skewness coefficients are presented in form of histogram of their simulated values. The correlation parameter prior correspond to a truncated normal distribution, according to the equation (39).

The prior distribution for the population mean and variance are concentrated around zero and one, respectively. For the skewness parameter, we are assuming more probability for values near zero but allowing reasonable probabilities for the others. The discrimination parameters are assumed to vary reasonably around a satisfactory discrimination power and for the difficulty parameter we assume a value above the mean of the reference time-point.

Table 2: Hyperparameters for the prior distributions

Hyperparameters					
$\boldsymbol{\mu}_\zeta$	$\boldsymbol{\Psi}_\zeta$	$(\mu_\xi;\sigma^2_\xi)$	$(\mu_\tau;\sigma^2_\tau)$	$(a_\varsigma, b_\varsigma)$	$(\mu_\rho, \sigma^2_\rho)$
(1, 0)	(0.5, 16)	(0,10)	(0,10)	(2.1, 1.1)	(0, 10)

The usual tools to investigate the MCMC algorithms convergence, that is, trace plots, Gelman-Rubin's and Geweke's statistics were monitored. We generate three chains based on three different sets of starting values. The Gelman-Runin's statistic were close to one for all parameters, indicating convergence. The trace plots and Geweke's monitoring indicate that a Burn-in of 10000 iterations was enough to reach the convergence. Further, the correlograms indicate that the samples composed by storing every 40th iteration have negligible autocorrelation. Therefore, we will work with valid samples with size 1000.

In order to assess the parameter recovery we consider the following statistics: correlations (Corr), mean of the bias (MBias), mean of the absolute bias (MABias) and mean of the absolute relative bias (MAVRB). Also, mean of the variances (MVAR) and mean of the root squared mean error (MRMSE), whose definitions can be seen bellow. Let ϑ and $\hat{\vartheta}$ a parameter and its estimate (posterior mean), respectively. The comparison statistics are defined as: Mcorr: $\mathrm{cor}(\vartheta_l, \hat{\vartheta}_l)$, MBias: $\frac{1}{n_p}\sum_{l=1}^{n_p}(\vartheta_l - \hat{\vartheta}_l)$, MABias: $\frac{1}{n_p}\sum_{l=1}^{n_p}|\vartheta_l - \hat{\vartheta}_l|$, MAVRB: $\frac{1}{n_p}\sum_{l=1}^{n_p}\frac{|\vartheta_l - \hat{\vartheta}_l|}{|\vartheta_l|}$, MVAR: $\frac{1}{n_p}\sum_{l=1}^{n_p}(\hat{\vartheta}_l - \bar{\hat{\vartheta}}_l)^2$ and MRMSE: $\sqrt{\frac{1}{n_p}\sum_{l=1}^{n_p}(\mathrm{MVAR} + (\vartheta_l - \vartheta)^2)}$, with $l = 1, \ldots, n_p$ where n_p denotes the number of parameters.

Tables 3 and 4 present the results of the parameter recovery study. The mean and variance of the first time-point were fixed in 0 and 1, respectively, in order to define the latent trait's scale. This restriction along with the common items design ensure the comparability of the latent traits and the model identification. We can see in Table 3 that the estimates of the population parameters are very close to the true values and most of 95% credibility intervals are covering the parameters. Some skewness and correlation parameters present a slight deviation of the true value. Probably this is due to random fluctuations. Table 4 presents the results for the latent traits and item parameters. The results indicate that estimates were very accurate. Note that, in the case guessing parameters, the correlation is small. This is expected since the true values have low variability. However, the other results indicate that estimates were close to the true values. Figure 2 presents the estimates of the latent traits and item parameters with 95% credibility intervals for the item parameters. Considering the item parameters we can see some deviations of the true value only for the discrimination parameter. In a general way, we conclude that the parameters were properly recovered by the estimation algorithm.

Table 3: Results for the population parameters.

	True value	Mean	SD	CI(95%)
μ_{θ_1}	0.000	–	–	–
μ_{θ_2}	1.000	1.011	0.039	[0.933, 1.088]
μ_{θ_3}	1.400	1.398	0.045	[1.315, 1.486]
μ_{θ_4}	2.000	2.057	0.069	[1.934, 2.201]
μ_{θ_5}	2.300	2.347	0.086	[2.200, 2.543]
μ_{θ_6}	2.500	2.572	0.100	[2.402, 2.806]
$\sigma^2_{\theta_1}$	1.000	–	–	–
$\sigma^2_{\theta_2}$	1.270	1.212	0.111	[1.022, 1.461]
$\sigma^2_{\theta_3}$	0.900	0.947	0.140	[0.750, 1.304]
$\sigma^2_{\theta_4}$	0.880	1.050	0.188	[0.794, 1.536]
$\sigma^2_{\theta_5}$	0.700	0.984	0.193	[0.716, 1.470]
$\sigma^2_{\theta_6}$	0.650	0.956	0.189	[0.685, 1.445]
γ_{θ_1}	0.800	0.739	0.063	[0.604, 0.848]
γ_{θ_2}	0.546	0.470	0.045	[0.373, 0.547]
γ_{θ_3}	0.181	0.224	0.032	[0.156, 0.288]
γ_{θ_4}	0.275	0.244	0.027	[0.188, 0.295]
γ_{θ_5}	0.221	0.130	0.017	[0.098, 0.167]
γ_{θ_6}	-0.039	0.047	0.021	[-0.001, 0.080]
ρ_{θ_1}	0.810	0.814	0.008	[0.795, 0.828]
ρ_{θ_2}	0.890	0.874	0.008	[0.859, 0.886]
ρ_{θ_3}	0.930	0.892	0.013	[0.864, 0.916]
ρ_{θ_4}	0.730	0.801	0.021	[0.761, 0.840]
ρ_{θ_5}	0.890	0.838	0.027	[0.776, 0.881]

Table 4: Results for the estimated latent traits and item parameters.

	Parameter	Statistic				
		Corr	MBias	MABias	MVAR	MRMSE
AD	Latent trait	0.970	-0.028	0.248	0.110	0.677
	Discrimination	0.877	0.095	0.152	0.017	0.510
	Difficulty	0.993	-0.039	0.113	0.024	0.463
	Guessing	0.308	0.002	0.026	0.002	0.246

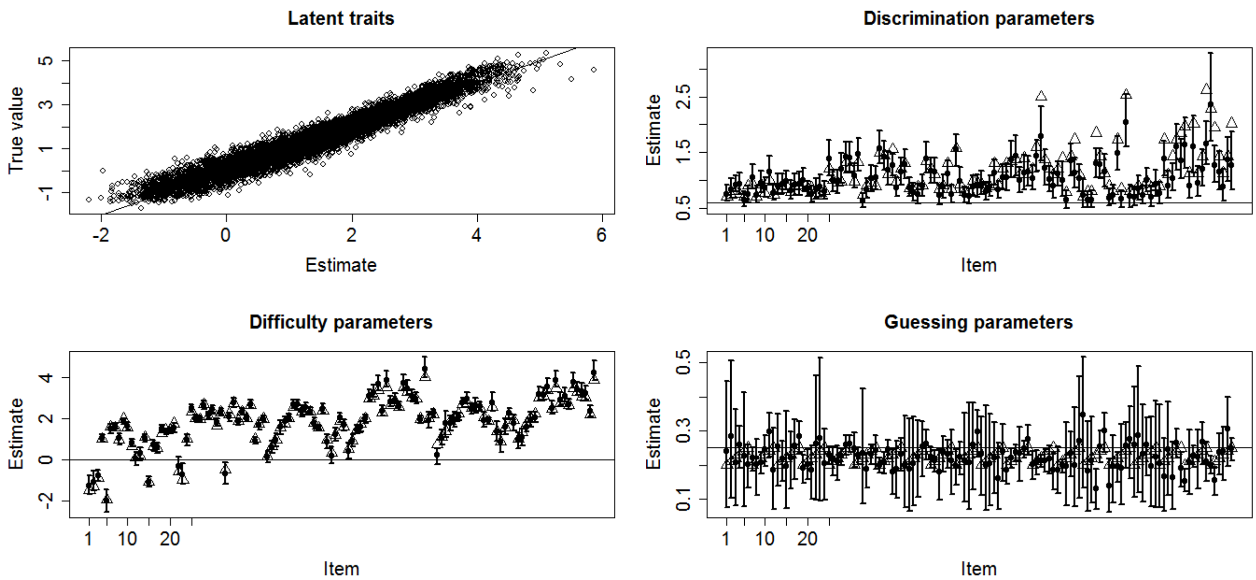

Figure 2: Estimates of latent traits and item parameters. Legend: circles denotes the estimates, triangles denotes the true values and the vertical bars denote 95% credibility intervals

6 Real Data Analysis and Model fit Assessment

6.1 Model fit assessment tools

For model fit assessment we will consider the so-called *Posterior Predictive Model Checking*, see Sinharay (2006) and Sinharay et al. (2006) for more details. The main ideia it is to compare the observed and simulated, where the former is generated by using the posterior predictive distribution. Specifically, let $\mathbf{y}^{obs}$ being the response matrix, and $\mathbf{y}^{rep}$ denoting the replicated response matrix. Then, the *posterior predictive distribution* of replicated data at the time-point t is given by

$$p(\mathbf{y}_t^{rep}|\mathbf{y}_t^{obs}) = \int p(\mathbf{y}_t^{rep}|\boldsymbol{\vartheta}_t)p(\boldsymbol{\vartheta}_t|\mathbf{y}_t^{obs})d_{\boldsymbol{\vartheta}_t}, \tag{62}$$

where $\boldsymbol{\vartheta}_t$ denotes the parameters at the time-point t. An usual method to compare the replicated and observed data, is to calculate the *Bayesian p-value* defined as

$$\mathbb{P}(D(\mathbf{y}_t^{rep}|\boldsymbol{\vartheta}_t) \geq D(\mathbf{y}_t^{obs}|\boldsymbol{\vartheta}_t)|\mathbf{y}_t^{obs}) = \int_{D(\mathbf{y}_t^{rep})\geq D(\mathbf{y}_t^{obs})} p(\mathbf{y}_t^{rep}|\mathbf{y}_t^{obs})d_{\mathbf{y}_t^{rep}}, \tag{63}$$

where D denotes a suitable statistic defined to address some aspect of interest. In practice, if we have M draws from the posterior distribution $p(\boldsymbol{\vartheta}_t|\mathbf{y}_t^{obs})$ of $\boldsymbol{\vartheta}_t$ and M draws from the likelihood distribution $p(\mathbf{y}_t^{rep}|\boldsymbol{\vartheta}_t)$, the proportion of the M replications for which $D(\mathbf{y}_t^{rep})$ exceeds $D(\mathbf{y}_t^{obs})$ provides an estimate of the Bayesian p-value. Values close to 1, or 0, indicate model misfit.

For IRT models, Béguin and Glas (2001) have proposed a posterior predictive check to compare the observed score distribution with the posterior predictive score distribution. For the longitudinal IRT model, the observed score distribution can be evaluated per time-point. Specifically, to evaluate items fit we defined the following statistic:

$$D_i = \sum_l \frac{|P_{li}^O - P_{li}^E|}{P_{li}^E}, \tag{64}$$

where P_{li}^O and P_{li}^E denote, respectively, the observed and expected proportion of respondents with scores l, that scored correctly the item i, for all $l = 1, 2, \ldots, L$ and $i = 1, 2, \ldots, I$, where L denotes the maximum score.

6.2 Model Comparison

For model comparison, where the main interest lies on the choice of the most appropriated covariance matrix, we used the approach of Spiegelhaltere et al. (2002). The related statistics are *Deviance information criteria* (DIC), and the expected values of the *Akaike's information* *criteria* (EAIC) and *Bayesian information criteria* (EBIC). These statistics are based on the ρ_D statistics defined as follows:

$$\rho_D = \overline{D(\vartheta)} - D(\bar{\vartheta}). \tag{65}$$

Again, ϑ denotes a set of interest parameters and $D(\vartheta) = -2(\text{Log likelihood})$. In our case we have,

$$D(\vartheta) = -2\text{Log}(L(\boldsymbol{\theta}_{..}, \boldsymbol{\zeta}_{.}, \boldsymbol{\eta}_{\theta})P(\theta|\boldsymbol{\eta}_{\theta})). \tag{66}$$

In practice, having M MCMC draws from the posterior distributions, the quantity $\overline{D(\vartheta)}$ can be estimated as:

$$\overline{D(\vartheta)} = \frac{1}{M}\sum_{m=1}^{M} D(\vartheta^{(m)}), \tag{67}$$

and $D(\bar{\vartheta})$ is evaluated on the estimates. Then, the estimates of the comparison statistics are give by

$$\begin{aligned}
\widehat{\text{DIC}} &= D(\bar{\vartheta}) + 2\rho_D, && (68)\\
\widehat{\text{EAIC}} &= \overline{D(\vartheta)} + 2\rho_D, && (69)\\
\widehat{\text{EBIC}} &- \overline{D(\vartheta)} + 2\text{log}(n \times I), && (70)
\end{aligned}$$

where n and I are, respectively, the number of subjects and the number of items.

6.3 The Brazilian school development study

The analyzed data concern to a major study promoted by the Brazilian Federal Government know as the School Development Program. It aims to monitor the teaching quality in Brazilian public schools. A more detailed description of this data can be found in Azevedo et al. (2016). In a general way, it is a longitudinal study, performed to evaluate children's ability in Math and Portuguese language. Only the results concerning to Math part were considered in our analysis. A total of 1987 public school's students selected from different regions of the country, were followed from fourth to eighth grade of the primary school, answering a different test in each one of these six different occasions, which are: 1999/April, 1999/November, 2000/November, 2001/November, 2002/November and 2003/November. A total of 167 items were considered in this analysis. Table 5 presents the structure of the tests, that is, the number of items per test and the number of common items across them.

Table 5: Structure of tests: real data analyze

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Test 1	34	10	5	1	0	0
Test 2	10	38	10	4	0	0
Test 3	5	10	36	7	3	1
Test 4	1	4	7	34	10	2
Test 5	0	0	3	10	40	10
Test 6	0	0	1	2	10	34

The skew antedependence IRT longitudinal model considering the unstructured covariance matrix, was applied to the data. The estimated covariance matrix can be seen in equation 71. It presents the estimated variances on the main diagonal, estimated correlations on the upper triangular and estimated covariances on the lower triangular. The estimates indicate a time-heteroscedastic structure. Moreover, the correlations are high and decay slowly. Therefore, due their quick decay, the autoregressive matrices ARH(1) and ARMAH(1,1) are not suitable, as we saw in chapter 2. Unlike the AD and HT structures can better describe the correlation pattern displayed by the data, as we can see in Figure 3.

$$\begin{pmatrix} 1.00 & .87 & .77 & .72 & .68 & .57 \\ .71 & .68 & .89 & .83 & .78 & .66 \\ .54 & .51 & .48 & .93 & .88 & .74 \\ .65 & .61 & .58 & .81 & .94 & .79 \\ .75 & .71 & .67 & .93 & 1.22 & .85 \\ 1.17 & 1.11 & 1.05 & 1.46 & 1.91 & 4.18 \end{pmatrix} \tag{71}$$

We compared skew IRT model under the AD and HT dependence structures with the symmetric IRT longitudinal model with AD structure (the selected model in Chapter 2). For short we will refer the skewed models as skewAD and skewHT and the symmetric AD as AD. Table 6 presents the associated statistics for model comparison, where all selected the skew AD model as the best model.

Table 6: Statistics for model comparison

	DIC	EAIC	EBIC
AD	496717.2	499933.9	540155.5
skewAD	**484481.7**	**484537.7**	**485237.4**
skewHT	492392.9	494778.6	524609.0

Figure 4 presents the observed and predicted scores with 95% credibility intervals for the six time-points. We can see that all the observed scores distribution are well within the intervals, indicating that the model is well fitted. Figures 5 and 6 present smoothed histograms of the latent trait estimates according to the symmetric and skew AD models, with their respective theoretical curves. The plot of the theoretical curves in the skew case was based on latent traits simulated via model 18. We can see that, the skew model represents better the latent trait distributions, specially for the time-points 1 and 2. Table 7 presents a comparison between the symmetric and skew population parameters estimations. The population variances tend to be smaller according to the skew model, indicating that the symmetric model tents to overestimate the population variance in the presence of asymmetry of the latent trais distribution. In general the standard error for the population parameters were smaller under the skew model. The latent trait distributions at the time-points 1, 2, 3 and 4 presented a moderate asymmetry.

Figure 7 presents Bayesian p-values based on the statistic given in Equation (64) for each item. Items with p-value below .05 or above .90 were assumed not well fitted. We can see that, only few items (around eight items) are not fitted properly. The misfit could be caused by DIF (*Differential item functioning*) or misspecification of the item response function.

Figures 8 to 10 present the estimates of the items parameters considering both skew and symmetric model. In this Figures circles denote estimates of the skew model, triangles denote estimates of the symmetric model and vertical bars represent 95% credibility intervals. In general we can see that all test presented good discrimination power (estimates greater than .6), except the last one, and difficulty parameters higher than the mean value of the latent traits. Also, we can notice that the discrimination parameters values tend to be higher according to the skew model. Moreover, the guessing parameters estimates indicate that the actual values are different from zero, which supports the use of the three parameters model. Some items presented guessing parameters estimates higher than .3, that is not expected indicating some inconsistency on these items formulation.

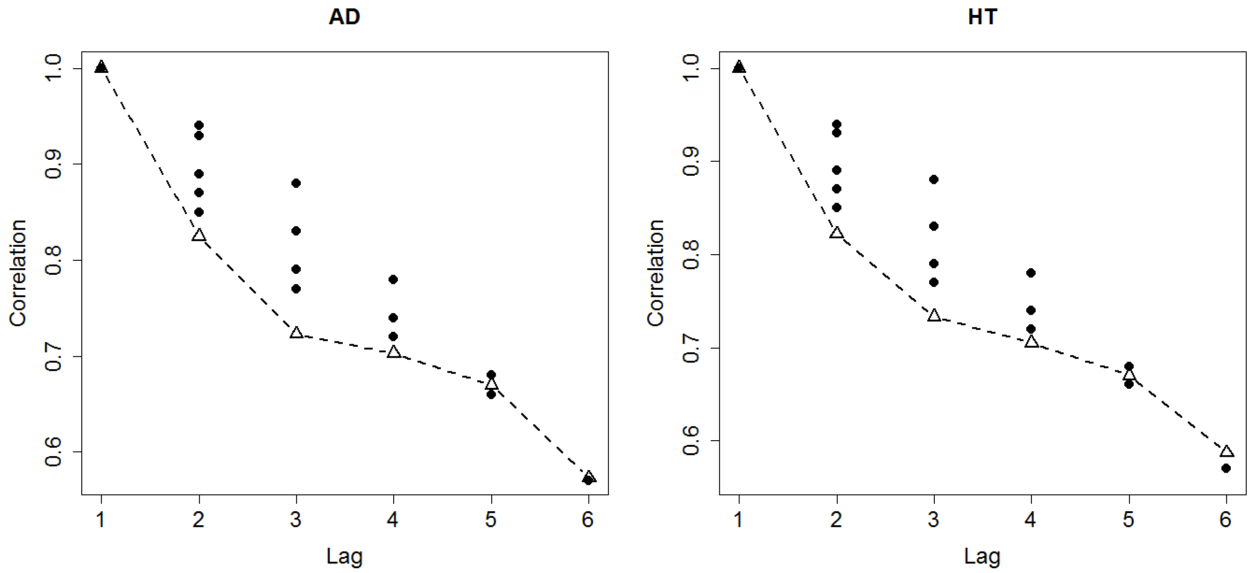

Figure 3: Correlation profiles. Legend: unstructured matrix (●), structured matrices (− △ −).

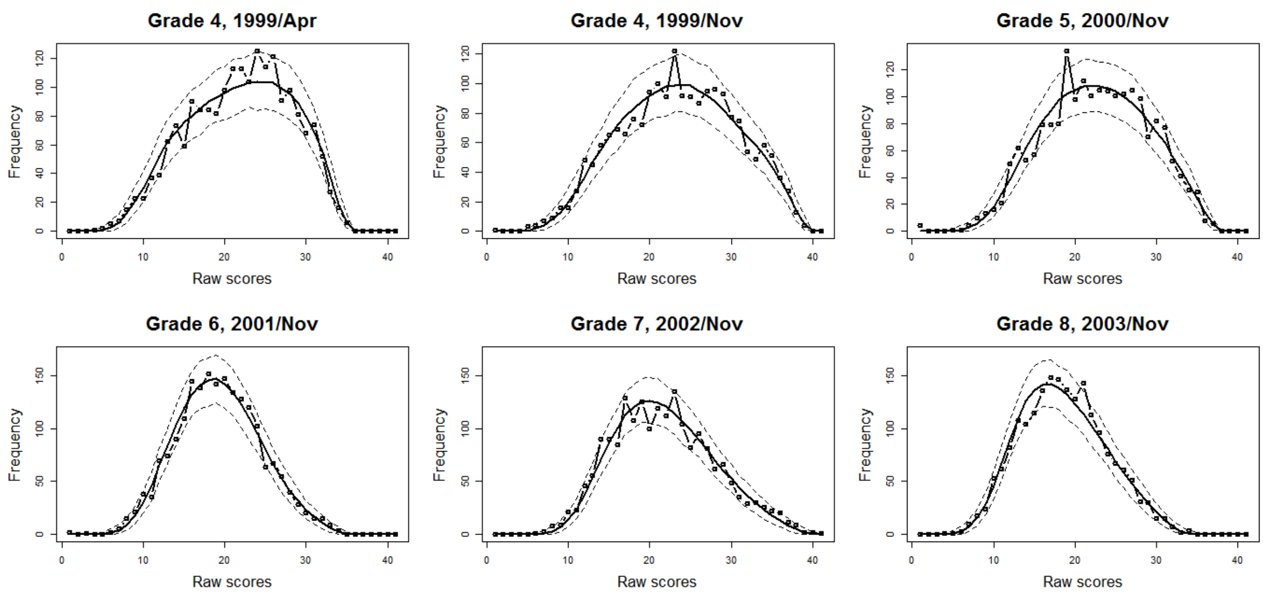

Figure 4: Observed and predicted scores distributions with 95% credibility intervals

Table 7: Estimates of the population parameters according to the AD model considering both symmetric and skew models

	Symmetric			Skew		
	Mean	SD	CI (95%)	Mean	SD	CI (95%)
μ_{θ_1}	.000	–	–	.000	–	–
μ_{θ_2}	.268	.026	[.222, .320]	.288	.023	[.243, .334]
μ_{θ_3}	.665	.030	[.604, .721]	.644	.023	[.598, .691]
μ_{θ_4}	1.115	.052	[.997, 1.210]	1.029	.036	[.972, 1.103]
μ_{θ_5}	1.290	.065	[1.133, 1.403]	1.160	.046	[1.095, 1.259]
μ_{θ_6}	.935	.074	[.784, 1.073]	1.118	.053	[1.024, 1.218]
$\sigma^2_{\theta_1}$	1.000	–	–	1.000	–	–
$\sigma^2_{\theta_2}$	.597	.057	[.483, .715]	.513	.051	[.426, .624]
$\sigma^2_{\theta_3}$	.480	.074	[.321, .667]	.391	.061	[.300, .533]
$\sigma^2_{\theta_4}$	.354	.057	[.242, .472]	.156	.028	[.111, .221]
$\sigma^2_{\theta_5}$	.219	.032	[.158, .287]	.047	.010	[.030, .070]
$\sigma^2_{\theta_6}$	2.255	.610	[1.008, 3.588]	1.429	.213	[1.092, 1.837]
ρ_{θ_1}	.784	.037	[.717, .810]	.824	.005	[.811, .833]
ρ_{θ_2}	.702	.041	[.628, .733]	.740	.008	[.728, .754]
ρ_{θ_3}	.658	.035	[.595, .697]	.687	.012	[.665, .709]
ρ_{θ_4}	.618	.022	[.580, .654]	.627	.011	[.609, .649]
ρ_{θ_5}	.546	.034	[.485, .584]	.567	.013	[.536, .593]
γ_{θ_1}	–	–	–	-.604	.063	[-.724, -.478]
γ_{θ_2}	–	–	–	-.412	.041	[-.489, -.330]
γ_{θ_3}	–	–	–	-.269	.030	[-.330, -.212]
γ_{θ_4}	–	–	–	-.205	.025	[-.252, -.155]
γ_{θ_5}	–	–	–	-.156	.020	[-.195, -.114]
γ_{θ_6}	–	–	–	-.110	.036	[-.185, -.057]

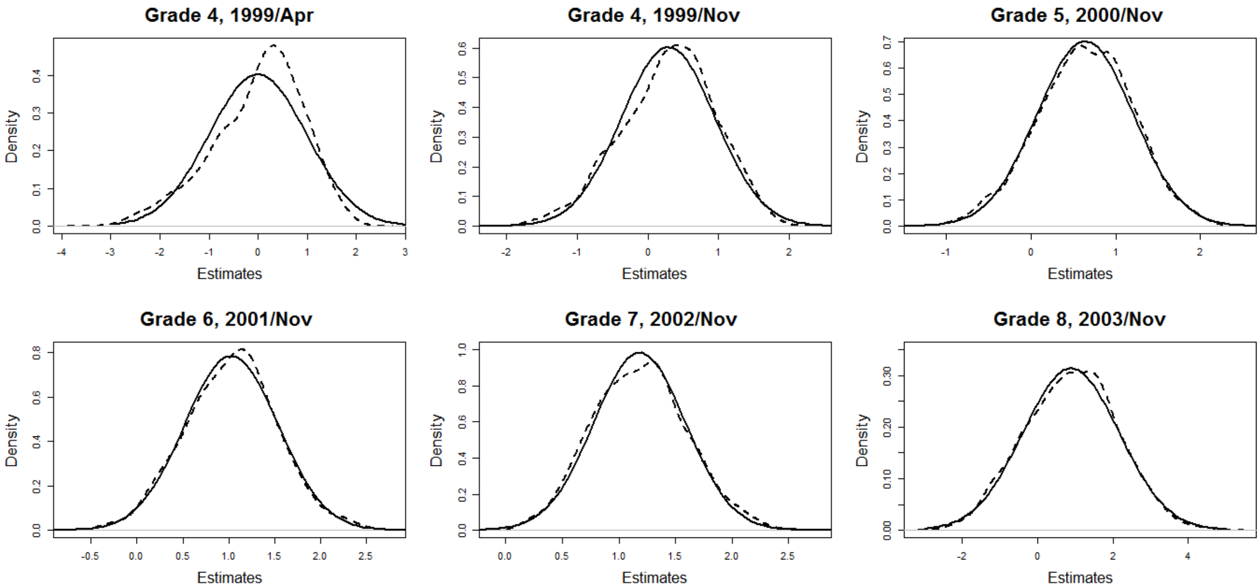

Figure 5: Estimated latent traits distributions according to the symmetric model. Legend: Smoothed histograms (dashed line), Theoretical curve (Solid line)

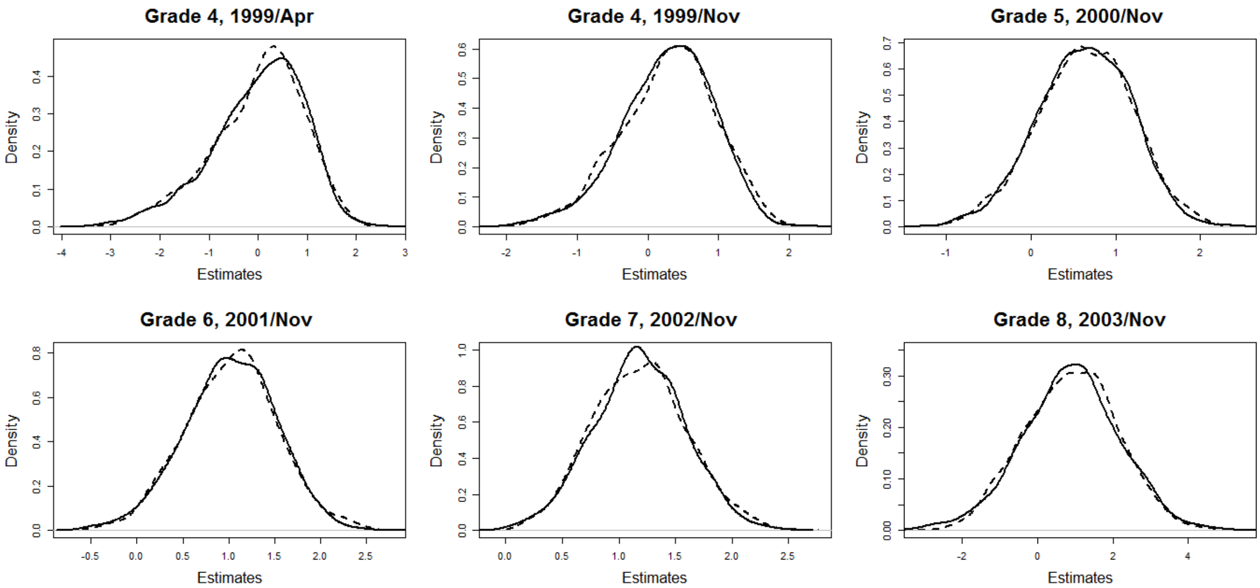

Figure 6: Estimated latent traits distributions according to the skew model. Legend: Smoothed histograms (dashed line), Theoretical curve (Solid line)

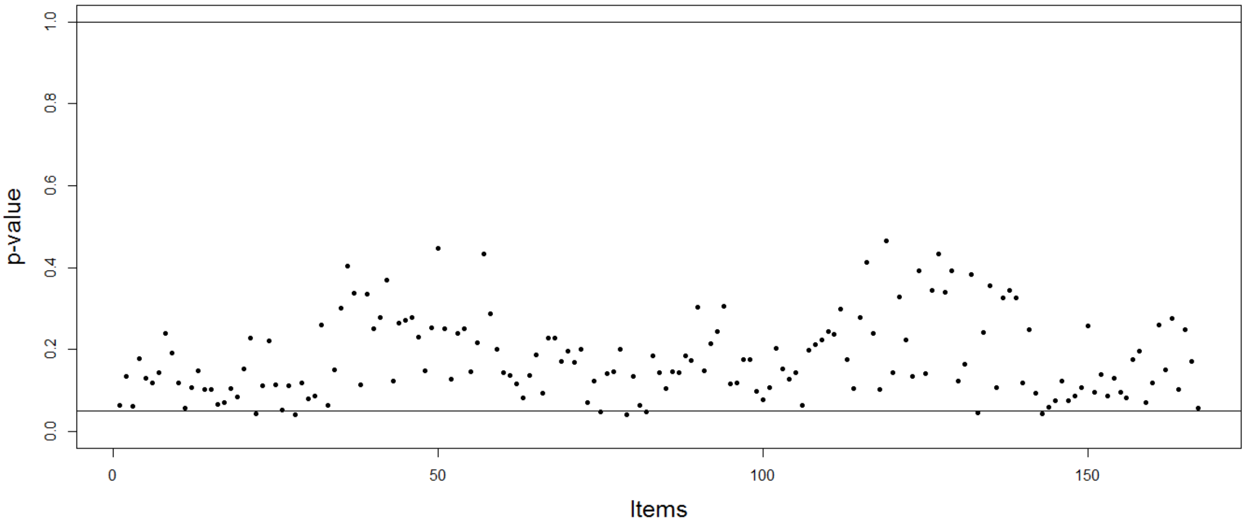

Figure 7: Bayesian p-values for the items

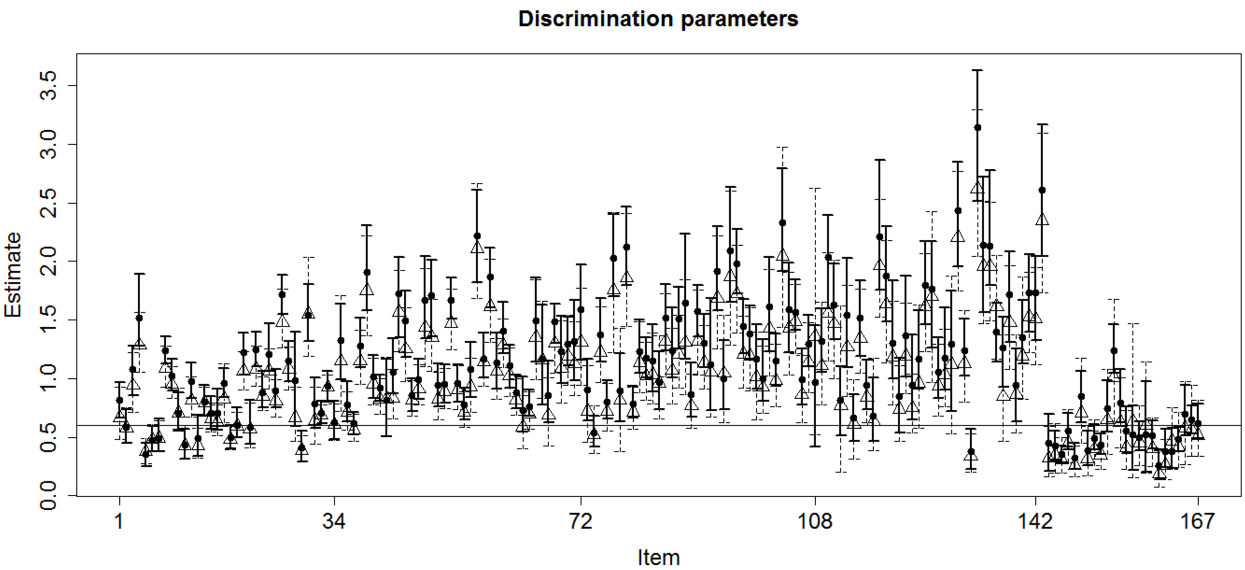

Figure 8: Posterior means and 95% central credibility intervals for discrimination parameters.

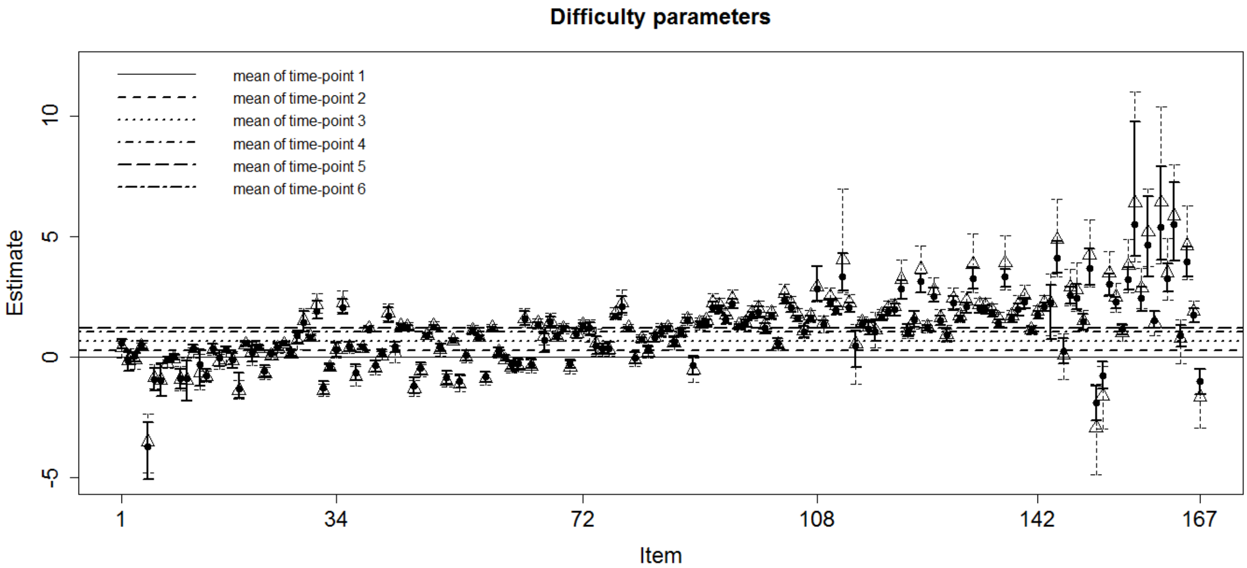

Figure 9: Posterior means and 95% central credibility intervals for difficulty parameters.

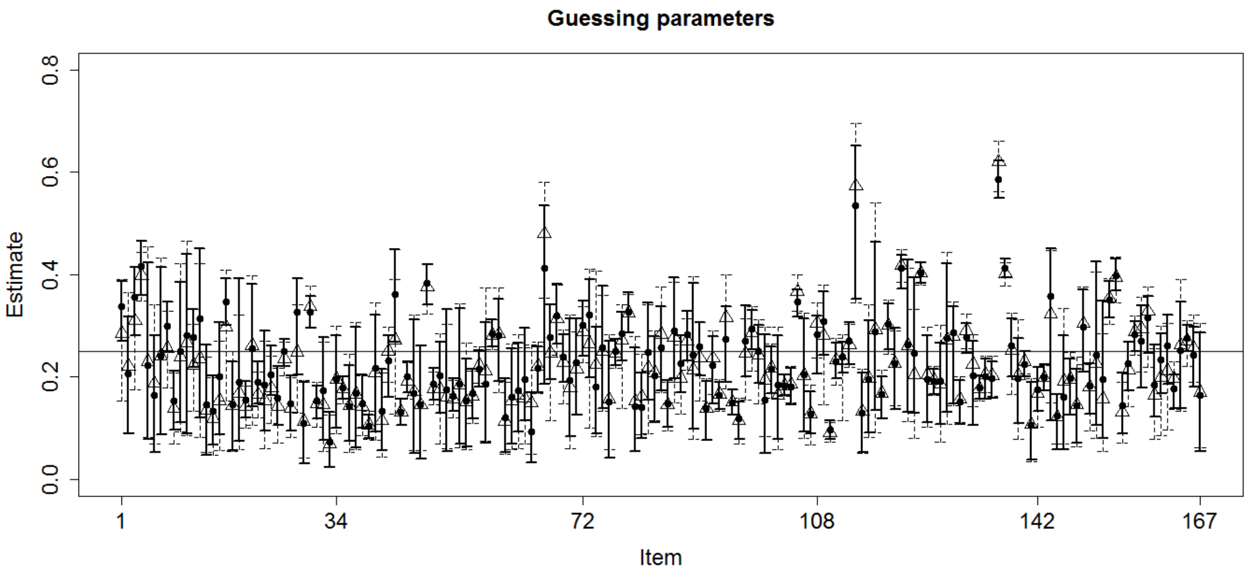

Figure 10: Posterior means and 95% central credibility intervals for guessing parameters.

7 Concluding Remarks

We presented a longitudinal three parameters IRT model based on a general Cholesky decomposition procedure with skewed latent trait distributions. Such methodology accommodate a wide range of dependence structures and allows asymmetry of the latent distributions. The univariate conditional distributions of the latent traits are assumed to be skew-normally distributed with centered parametrization. We noticed some difficulty to derive the marginal distributions of the latent traits, however, the marginal Pearson's skewness coefficient was relatively easy to obtain. An MCMC algorithm based on the FFBS and SVE procedures was developed for estimating the model parameters. It showed to be efficient in terms of parameter recovery, according to the simulation study. Furthermore, a real data concerning to a Brazilian school development study was analyzed. Some model fit assessment tools were considered, indicating that the model was well fitted. The model identified high between-time correlations of the latent traits. Also, four marginal latent trait distributions presented asymmetric behavior. Further, the skew model fitted better to the data compared to the symmetric model. In conclusion, our approach is a promising alternative to the usual ones in analyzing longitudinal IRT data. In future research we intend to explore some extensions of our model, considering growth curves and regression structures for the population mean of the latent traits distribution. The multiple group structure for longitudinal IRT data could be also considered for the next works as in Azevedo et al. (2015).

References

Andrade, D. and Tavares, H. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, 95(1):1–22. 2

Azevedo, C. (2008). *Longitudinal multiple-group multilevel IRT models: estimation methods and structural selection under a Bayesian framework (PhD thesis in portugues)*. IME-USP, São Paulo, SP. 3

Azevedo, C., Bolfarine, H., and Andrade, D. (2011). Bayesian inference for a skew-normal irt model under the centred parameterization. *Computational Statistics & Data Analysis*, 55:353–365. 3, 4, 5

Azevedo, C., Bolfarine, H., and Andrade, D. (2012a). Parameter recovery for a skew-normal irt model under a bayesian approach: hierarchical framework, prior and kernel sensitivity and sample size. *Journal of Statistical Computation & Simulation*, 82(11):1679–1699. 3

Azevedo, C. L., Andrade, D. F., and Fox, J.-P. (2012b). A bayesian generalized multiple group irt model with model-fit assessment tools. *Computational Statistics & Data Analysis*, 56(12):4399 – 4412. 2, 3, 26

Azevedo, C. L., Fox, J.-P., and Andrade, D. F. (2015). Longitudinal multiple-group irt modelling: covariance pattern selection using mcmc and rjmcmc. *International Journal of Quantitative Research in Education*, 2(3-4):213–243. 40

Azevedo, C. L. N., Fox, J.-P., and Andrade, D. F. (2016). Bayesian longitudinal item response modeling with restricted covariance pattern structures. *Statistics and Computing*, 26(1):443–460. 2, 8, 9, 32

Azzalani, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 12:171–178. 3, 5

Baker, F. B. and Kim, S.-H. (2004). *Item response theory : Parameter Estimation Techniques*. Marcel Dekker, Inc, New York, NY, second edition. 4, 7

Bazán, J. L., Bolfarine, H., and Leandro, R. A. (2004). Bayesian estimation via mcmc for probit-normal model in item response theory. *Technical report, University of São Paulo*, (15):26. 4

Béguin, A. A. and Glas, C. A. W. (2001). Mcmc estimation and some model-fit analysis of multidimensional irt models. *Psychometrika*, 66(4):541–561. 14, 31

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an em algorithm. *Psychometrika*, 46(4):433–459. 2

Branco, M. D. and Arellano-Valle, R. (2004). *Asymmetric Elipitical Distributions*. 16^0 SINAPE, ABE, first edition. 4

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81:541–553. 19

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics*, 46(2):317–328. 2

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses : Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online*, 1(4):65–85. 2

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of time series analysis*, 15:183–202. 19

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo : Stochastic simulation for bayesian inference*. Chapman & Hall/CRC, London, second edition. 19, 20, 21

Henze, N. (1986). A probalistic representation of the skew-normal distribution. *Scandinavian journal of statistics*, 13:271–275. 6

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820. 9

Marsman, M. (2014). *Plausible Values in Statistical Inference*. PhD Thesis University of Twente, Enschede., 1 edition. 21, 22

Murray, I., Ghahramani, Z., and MacKay, D. (2012). Mcmc for doubly-intractable distributions. ArXix e-prints. 22

Newton, H. (1988). *TIMESLAB: A time series analysis laboratory*. Pacific Grove, CA: Wadsworth & Brooks/Cole., 1 edition. 9

Nunez-Anton, V. and Zimmerman, D. L. (2000). Modelinng nonstationary longitudinal data. *Biometrics*, 56(3):699–705. 3

Pewsey, A. (2000). Problems of inference for azzalini's skewnormal distribution. *Journal of Applied Statistics*, 27(7):859–870. 5

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690. 1, 3, 8

Rodriguez, C. L. B. (2005). *Bayesian Inference on the Skew-normal Model*. Master dissertation (In portugues), São Paulo, SP, ime-usp edition. 7

Santos, J. R. S., Azevedo, C. L. N., and Bolfarine, H. (2013). A multiple group item response theory model with centred skew normal latent trait distributions under a bayesian framework. *Journal of Applied Statistics*, 40(10):2129–2149. 26

Sinharay, S. (2006). A bayesian item fit analysis for unidimensional item response theory models. *British journal of mathematical and statistical psychology*, 59:429–449. 31

Sinharay, S., Johnson, M. S., and Stern, H. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4):298–321. 31

Spiegelhaltere, D. J., Best, N. G., Carlin, B. P., and van der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal Royal Statistical Society*, 64(3):583–639. 31

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, 82(398):5258–550. 14

Tavares, H. R. and Andrade, D. F. (2006). Item response theory for longitudinal data: Item and population ability parameters estimation. *TEST*, 15(1):97–123. 2

Zimmerman, D. and Núñez-Antón, V. (2009). *Antedependence Models for Longitudinal Data*. Chapman and Hall/CRC, 1 edition. 8