ISSN 0104-0499

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

DEs-UFSCar e SME-ICMC-USP

A NEW BAYESIAN APPROACH TO ONE-PARAMETER IRT MODELS TO LARGE DATA SET

Jorge Luis Bazán
Marcos Oliveira Prates

RELATÓRIO TÉCNICO

TEORIA E MÉTODO – SÉRIE A

Julho/2016
nº 263

A new Bayesian approach to one-parameter IRT models to large data set

Jorge Luis Bazán[1] and Marcos Oliveira Prates[2]

[1] email: jlbazan@icmc.usp.br. Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo. C.P. 668 - São Carlos, SP, Brazil - CEP 13560-970.

[2]email: marcosop@est.ufmg.br. Departamento de Estatística, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627 - Predio do ICEx - Belo Horizonte, MG, Brazil - CEP 31270-901

April 15, 2016

Abstract

One-parameter Item response models including the Rasch model are frequently used in large scale assessments and estimated commonly by Maximum Likelihood (ML) estimation methods since that Bayesian estimation considering MCMC methods are usually slow for large data sets. In this work, a new Bayesian estimation method considering Integrated Nested Laplace Approximations is proposed. The method is tested through simulation studies. Additionally, the method was also evaluated using a large data corresponding to a ENEM exam in Brazil. The main conclusion is that the new method provides a Bayesian alternative to analyze large-scale data sets which provide very similar results but it is faster than traditional MCMC method and perform similarly or better in terms of parameter estimation than know ML estimation methods.

Key Words: Bayesian estimation, INLA, item response theory, MCMC, Rasch model.

1. INTRODUCTION

The estimation procedure in IRT models has been dominated by the maximum likelihood (ML) approach. Using the this methodology, several approaches have been proposed such as joint likelihood, marginal likelihood and conditional likelihood (Baker and Kim, 2004). The most used approach is the marginal likelihood approach using the EM-type algorithm with Gaussian quadrature for approximating the integrals needed for implementing the E step of the algorithm involved in the estimation of the item parameters (Bock and Aitkin, 1981). This estimation procedure is implemented in software, as, for example, in the IRT-PRO software (Toit, 2003). The estimation of ability parameters is performed in a second stage with item parameters replaced by estimates computed previously. Limitations of this methodology are discussed in Patz and Junker (1999) and Sahu (2002).

To overcome this problem, Bayesian estimation was initially proposed by Swaminathan and Gifford (1982). Bayesian estimation can be distinguished in estimation with and without Markov chain Monte Carlo (MCMC). In the case of non MCMC Bayesian estimation, Bayesian marginal estimation is used with maximum and expected a posteriori estimates for latent trait considering hierarchical models or not. For example, see Baker and Kim (2004).

In Bayesian estimation the parameters are considered as random variables and then properties of the estimates can be obtained since sample from the posterior distribution if available. In fact, all the properties of a posterior can be approximated to any degree of accuracy by drawing a sample that is sufficiently large. This approach to statistical estimation is called sampling-based estimation. Sampling-based estimation allows one to study distributions that are analytically intractable, given that one can sample from them. In the last decades, much attention has been given to MCMC methods for generating a sample from a posterior.

These methods involve(s) setting up a Markov chain which in the limit generates a dependent identically distributed (*did*) sample from the posterior, and (b) the use of the Monte Carlo method for estimating properties of the posterior by properties of the *did* sample. As indicated by Maris and Maris (2002): three questions present themselves to the scientist wishing to use a MCMC-method for a particular problem: (a) how to set up a Markov chain which converges to a did sample from the posterior, (b) how to assess whether the length of the Markov chain is sufficient for it to be sufficiently close to its stationary distribution, and (c) how to assess whether the sample size (after convergence) is sufficient for the Monte Carlo estimates to be sufficiently precise. However, MCMC approach in IRT models still is standard, see for example Curtis (2010); Stone and Zhu (2015).

As indicated by Levy (2009) it is readily acknowledged that MCMC is difficult, both computationally in terms of necessary resources and conceptually in terms of constructing the chains, making relevant choices, and understanding the results. Additionally it may be extremely slow when the number of examinees and / or items increases substantially.

In Large Scale Assessments like the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) or The Progress in International Reading Literacy Study (PIRLS) a large number of students are evaluated. To obtain student competence estimates requires IRT models, a famous one is the simple logistic or Rasch model (Rasch, 1960). For this type of data or genetic data where large number of genes are available, MCMC methods are much slower then Maximum Likelihood and then they are not considered to estimation.

On the other hand, an approximate method named integrated nested Laplace (INLA) was developed by Rue et al. (2009); Martino and Rue (2010) to deterministically approximate the posterior marginal distribution of interest under the family of latent Gaussian models. A major advantage of INLA, following the authors and considering several implemented models (see by example Grilli et al. (2015)), is that computational time is short and approximations are precise;

However, to the best of our knowledge, this approach has not been used yet to one-parameter IRT models.

In this paper, we present a new Bayesian estimation for one-parameter IRT models considering the INLA method. The paper is organized as follows. In Section 2, the one-parameter IRT model is described. In Section 3, Bayesian estimation using MCMC methods and INLA are presented to the IRT model. We continue in Section 3.3 presenting different Model comparison Criteria for IRT models introducing Widely Applicable Information Criteria (WAIC) (see for example Gelman et al., 2014). Section 4 presents two simulation schemes to assess the the correctness of the INLA method when compared to MCMC and ML methods. In Section 5 the methodology is illustrated considering a large data set of the Exame Nacional do Ensino Médio (ENEM, in English: High School National Exam) with 45 items of 29442 examinees of the state of Minas Gerais from Brazil in the Mathematics exam. We conclude with a final remarks indicating future works.

2. ONE-PARAMETER ITEM RESPONSE MODEL

We consider that

$$Y_{ij}|\theta_i, b_j \sim Bernoulli\,(p_{ij}), \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, I. \tag{1}$$

where Y_{ij} are the dichotomous response corresponding to the i-th individual to the j-th item, and (θ_i, b_j) is the vector of latent variables of interest with b_j, a item parameter that correspond to *item difficulty* and θ_i, the value corresponding to *latent trait* associated to examinees i, describing its personal ability in answering the test with I items and p_{ij} is the *probability of correct answer* for examinee i in the item j. Further, let

$$p_{ij} = P(Y_{ij} = 1 \mid m_{ij}) = F(m_{ij}), \tag{2}$$

with $m_{ij} = \theta_i - b_j$, a linear function of θ_i, $i = 1, \ldots, n$ and $j = 1, \ldots, I$. The function $F(.)$ is typically known as the *item response function* or *item characteristic curve* and satisfies the property of latent monotonicity (strictly nondecreasing function of θ_i) and typically is the same for all i and j with support on the whole real line. Further, b_j and θ_i also can take any real values.

IRT models typically satisfies the *conditional independence property*, that is, for examinee i, the responses Y_{ij} corresponding to items $j = 1, \ldots, I$, are conditionally independent given the values of latent trait θ_i, $i = 1, \ldots, n$. Further, it is considered independence between responses from different examinees. Under the above assumptions, the joint distributions of $\boldsymbol{Y} = (\boldsymbol{Y}'_1, \ldots, \boldsymbol{Y}'_n)'$ with $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iI})'$ given the vector of latent variables $(\boldsymbol{\theta}, \boldsymbol{b})$ with latent trait $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ and difficulty parameters $\boldsymbol{b} = (b_1, \ldots, b_I)'$ can be written as

$$p(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\theta}, \boldsymbol{b}) = \prod_{i=i}^{n} \prod_{j=1}^{I} F(m_{ij})^{y_{ij}} (1 - F(m_{ij}))^{1-y_{ij}}. \tag{3}$$

The first one-parameter IRT model was formally introduced by Lord (1952) and considers $F(.) = \Phi(.)$, $i = 1, \ldots, n$, and $j = 1, \ldots, I$, with $\Phi(.)$ the cumulative function (cdf) of the standard normal distribution. The model is known as the normal ogive model. Rasch (1960) considered $F(\cdot) = Lo(\cdot)$ with $Lo(m_{ij}) = e^{m_{ij}}/(1 + e^{m_{ij}})$, denoted the cdf of the standard logistic distribution and thus, this model is known as the one-parameter logistic model or Rasch model (Fischer, 2007). Also, by considering $F(.) = RG(.)$ the cdf of the standard reversal Gumbel distribution with $RG(m_{ij}) = 1 - \exp[-\exp(m_{ij})]$ another one-parameter model can be defined. Note that $F(.)^{-1}$ is a link function and then $\Phi(.)^{-1}$, $\log(p_{ij}/(1 - p_{ij}))$ and $\log(-\log(1 - p_{ij}))$ are respectively the probit, logit and loglog links.

The dichotomous item response model presented in (1) - (3) involves a total of $n + I$ unknown parameters being thus overparameterized. On the other hand, for a fixed number of items, item parameters are known as structural parameters and the latent trait are known as incidental parameters, because they increase with n, the sample size, and because the analysis is generally focused on the item parameters. The model is also unidentifiable, since it is preserved under a special class of transformations of the parameters (see Albert, 1992) so that maximum likelihood estimates may not be unique. One way of contouring such difficulties is to impose restrictions on the item parameters as considered, for example, in Bock and Aitkin (1981). Another way follows by specifying a distribution for the latent trait. Lord and Novick (1968), Albert (1992) consider

$$\theta_i \overset{iid}{\sim} N(\mu, \sigma^2),\ i = 1, \ldots, n. \tag{4}$$

This assumption establishes that it is believed that the latent trait are well behaved and that are a random sample from this distribution. We can consider in this paper that $\mu = 0$ and $\sigma^2 = 1$. In more general situations, the prior structure needs to be enlarged so that hyper prior information can also be considered for σ^2 parameters.

We find it more appropriate using the notation probit-normal and logit-normal models, respectively and then the loglog-normal models complete the list of one-parameter IRT models analyzed in this paper.

3. BAYESIAN ESTIMATION

We start considering the following general class of independent prior distributions:

$$\pi(\boldsymbol{\theta}, \boldsymbol{b}) = \prod_{i=1}^{n} g_1(\theta_i) \prod_{j=1}^{I} g_2(b_j),$$

where g_1 and g_2 are specified probability *density functions* for θ_i and b_j, respectively, $i = 1, \ldots, n$ and $j = 1, \ldots, I$. Following Rupp et al. (2004) and Sahu (2002), we take the normal distribution as common prior for b_j, that is $b_j \sim N(\mu_b, \sigma_b^2)$ for g_2 and to g_1 we consider the specification in (4). The vector of hyperparameters of the one parameter IRT model is $\boldsymbol{\Omega} = (\mu, \sigma^2, \mu_b, \sigma_b^2)$.

Albert and Ghosh (2000) mention that the choice of a proper prior distribution on the latent trait resolves particular identification problems, and, further, informative prior distributions placed for b_j can be used to reflect the prior belief that the values of the item parameters are not extreme (in the frontier of the parametric space). In general is assumed $\mu_b = 0$. In the common situation where little prior information is available about the difficulty parameters, one can chose σ_b^2 to be a large value. This choice will have a modest effect on the posterior distribution for non extreme data, and will result in a proper posterior distribution when extreme data (where students are observed to get correct or incorrect answers to every item) is observed (Albert and Ghosh, 2000), also, Sahu (2002) states that larger values of the variance led to unstable estimates. This priors are denominated as vague priors in the literature. Another situation, is to consider a hyper prior for σ_b^2, thus, Swaminathan and Gifford (1982) use IG(m, n), the inverted gamma distribution with (known) hyperparameters m and n.

Let $D_{obs} = \boldsymbol{Y}$, the observed data. Hence, the likelihood function for the One-parameter Dichotomous model is given by

$$L(\boldsymbol{\theta}, \boldsymbol{b}|D_{obs}) = \prod_{i=1}^{n}\prod_{j=1}^{I} F(m_{ij})^{y_{ij}}(1 - F(m_{ij}))^{1-y_{ij}}, \tag{5}$$

and consequently the likelihood function and prior specification, a joint posterior distribution is given by:

$$f(\boldsymbol{\theta}, \boldsymbol{b}|D_{obs}) \propto L(\boldsymbol{\theta}, \boldsymbol{b}|D_{obs}) \times \pi(\boldsymbol{\theta}, \boldsymbol{b}). \tag{6}$$

3.1 MCMC Estimation

As the joint posterior distributions above are complex to be dealt with, note that all full conditional distributions are non-standard. Hence straightforward implementation of the Gibbs sampler using standard sampling distributions is not possible. However, all the full conditional distributions for the probit-normal model are log-concave (log of the density is concave) according to (Sahu, 2002). Exact sampling from one dimensional log-concave distributions can be performed using rejection sampling, even when the normalizing constants are unknown (Gilks and Wild, 1992). These authors also develop an adaptive rejection sampling (ARS) scheme. ARS dynamically constructs two envelopes (one lower and one upper) for the distribution to be sampled from using successive evaluations of the density at the rejected points. The algorithm stops when one proposed point has been accepted.

An alternative MCMC method to estimate the One-parameter Dichotomous model was initially proposed by Albert (1992) to the probit-nomal model and extended to the logit-normal model by Maris and Maris (2002). Introducing a auxiliary latent variable yields a model equivalent to the One-parameter Dichotomous model. Then, a "augmented" likelihood function is obtained and Data Augmentation Gibbs Sampling (DAGS) or Data Augmented Transformation Gibbs Sampling (DATGS) MCMC methods can be proposed by probit-normal and logit-normal models respectively. Routines in R and Matlab to both algorithm are available in the Web and also can be easily formulated in WinBUGS or OpeBUGS (see for example Curtis, 2010) and SAS (see for example Stone and Zhu, 2015) and then MCMC methods are standard procedures to fit IRT model under a Bayesian approach. However it is known that MCMC methods are slow when the number of individuals and/or the number of items are increased dramatically making unfeasible to fit IRT models to large data set (see for example Levy, 2009). In the next section we present an alternative for this.

3.2 INLA Estimation

Rue et al. (2009) introduced the INLA approach to perform Bayesian analysis for a broad class of models where the response y_{ij} assumes independence conditional on some latent field $\boldsymbol{\theta}, \boldsymbol{b}$ and a vector of hyperparameters $\boldsymbol{\Omega}$, these models are called latent Gaussian models. As indicated by Rue et al. (2009), latent Gaussian models are a structured additive regression models where the observation (or response) variable is assumed to belong to an exponential family distribution, and the conditional mean μ_{ij} is linked to a structured additive predictor η_i through a link function $g(.)$, so that $g(\mu_{ij}) = \eta_{ij}$. As known, see by example Fischer (2007), Rasch model, conditionally belong to the exponential family, in consequence are latent Gaussian models. This is also the case of the one-parameter IRT models when we use the probit and loglog as link function.

In order to propose the INLA estimation method for the one-parameter IRT model we formulate this model as a Bayesian hierarchical model with a latent Gaussian random field $\boldsymbol{\theta}, \boldsymbol{b}$. Therefore, this characteristics are clearly suitable for the use of the INLA framework. For the dichotomous IRT model, the first stage is the observational model $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b})$, where $\boldsymbol{y}$ denotes the binary response and then the vector $\boldsymbol{\theta}, \boldsymbol{b}$ corresponds to the latent Gaussian Markov Random Field (GMRF) which is responsible for all latent components of the model, $\pi(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{\Omega})$ with $\boldsymbol{\Omega}$ the hyper parameters defined in Section 3. The GMRF is typically controlled by a few hyperparameters $\boldsymbol{\Omega}$, that can be held fixed or not as discussed above.

To estimate the item difficulty b_j and the latent trait i (θ_i) we are interested in the posterior marginal of the elements of (θ_i, b_j) give $\boldsymbol{\Omega}$ (i.e the latent effects and hyperparameters) give by

$$\pi(\theta_i, b_j|\boldsymbol{y}) = \int \pi(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})\pi(\boldsymbol{\Omega}|\boldsymbol{y})d\boldsymbol{\Omega}, \tag{7}$$

$$\pi(\Omega_k|\boldsymbol{y}) = \int \pi(\Omega|\boldsymbol{y})d\boldsymbol{\Omega}_{-k}, \tag{8}$$

where Ω_k is the kth entry of vector $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{-k}$ is the vector $\boldsymbol{\Omega}$ with the kth entry removed.

In order to estimate the parameters of the model, we first approximates $\pi(\boldsymbol{\Omega}|\boldsymbol{y})$, using the a Gaussian approximation to the full conditional distribution of $\boldsymbol{\theta}, \boldsymbol{b}$ by a multivariate Gaussian density $\tilde{\pi}_G(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\Omega})$ (for details see Rue and Held, 2005) evaluated at its mode $(\boldsymbol{\theta}, \boldsymbol{b})^\star(\boldsymbol{\Omega})$. Then the posterior density of $\boldsymbol{\Omega}$ is approximated by using the Laplace approximation (Tierney and Kadane, 1986)

$$\tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}, \boldsymbol{\Omega})}{\tilde{\pi}_G(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\Omega})} \right|_{\boldsymbol{\theta}, \boldsymbol{b} = (\boldsymbol{\theta}, \boldsymbol{b})^\star(\boldsymbol{\Omega})}. \tag{9}$$

In a second step is to compute the Laplace approximation of $\pi(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})$ for selected values of $\boldsymbol{\Omega}$, which will be used to perform a numerical integration to obtain the posterior marginals of θ_i, b_j presented in (7). The density $\pi(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})$ is approximated by

$$\tilde{\pi}_{LA}(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega}) \propto \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}, \boldsymbol{\Omega})}{\tilde{\pi}_G((\boldsymbol{\theta}, \boldsymbol{b})_{-ij}|\theta_i, b_j, \boldsymbol{y}, \boldsymbol{\Omega})} \right|_{(\boldsymbol{\theta}, \boldsymbol{b})_{-ij} = (\boldsymbol{\theta}, \boldsymbol{b})^\star_{-ij}(\theta_i, b_j, \boldsymbol{\Omega})}, \tag{10}$$

where $(\boldsymbol{\theta}, \boldsymbol{b})_{-ij}$ denotes the vector $\boldsymbol{\theta}, \boldsymbol{b}$ without the $(i,j)^{th}$ component, $\tilde{\pi}_G((\boldsymbol{\theta}, \boldsymbol{b})_{-ij}|\theta_i, b_j, \boldsymbol{y}, \boldsymbol{\Omega})$

is the Gaussian approximation of $\pi((\boldsymbol{\theta}, \boldsymbol{b})_{-ij}|\theta_i, b_j, \boldsymbol{y}, \boldsymbol{\theta})$, treating θ_i, b_j as observed and $(\boldsymbol{\theta}, \boldsymbol{b})^{\star}_{-ij}(\theta_i, b_j, \boldsymbol{\Omega})$ is the mode of $\pi((\boldsymbol{\theta}, \boldsymbol{b})_{-ij}|\theta_i, b_j, \boldsymbol{y}, \boldsymbol{\Omega})$.

The approximation $\tilde{\pi}_{LA}(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})$ in (10) can be quite expensive, since is necessary to recompute $\tilde{\pi}_G((\boldsymbol{\theta}, \boldsymbol{b})_{-ij}|\theta_i, b_j, \boldsymbol{y}, \boldsymbol{\Omega})$ for all θ_i, b_j and $\boldsymbol{\Omega}$. Rue et al. (2009) proposes two alternatives to obtain these full conditionals in a cheaper way. We focus our analysis in the simplified Laplace approximation defined as the series expansion of $\tilde{\pi}_{LA}(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})$ (for details see Rue et al., 2009).

Finally, the full posteriors approximations obtained previously are combined and the marginal posterior densities of m_{ij} and θ_k are obtained by numerically integrating out the irrelevant terms. Therefore, the marginal approximation of the latent variables using (9) and (10) can be obtained by

$$\pi(\theta_i, b_j|\boldsymbol{y}) = \int \pi(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega})\pi(\boldsymbol{\Omega}|\boldsymbol{y})d\boldsymbol{\Omega} \approx \sum_l \tilde{\pi}_{LA}(\theta_i, b_j|\boldsymbol{y}, \boldsymbol{\Omega}_l)\tilde{\pi}(\boldsymbol{\Omega}_l|\boldsymbol{y},)\Delta_l,$$

which is evaluated using a finite sum on a set $\boldsymbol{\Omega}_l$ of grid points, with area weights l for $l = 1, 2, \ldots, L$. Rue et al. (2009) argue that because the points $\boldsymbol{\Omega}_l$ are selected in a regular grid, it is feasible to take all the area weights l to be equal. In a similar way, the posterior marginal of $\pi(\Omega_k|\boldsymbol{y})$ is obtained.

Hierarchical extensions of the Rasch model as the proposed in Maier (2001) are immediate in INLA approximation considering adequate specification of the vector of hyperparameters $\boldsymbol{\Omega}$ in the model.

3.3 Models comparison Criteria

In order to compare alternative one-parameter IRT models, we make use of some model comparison criteria discussed in Gelman et al. (2013). Specifically, we consider the Deviance Information Criterion (DIC) which is defined by DIC $= \overline{\mathrm{D}}(\boldsymbol{\theta}) + \rho_{\mathrm{DIC}} = 2\overline{\mathrm{D}}(\boldsymbol{\theta}) -$

$\mathrm{D}(\tilde{\boldsymbol{\theta}})$, where the term ρ_{DIC} is a measure of the effective number of parameters in the model and it is defined as $\rho_{\mathrm{DIC}} = \overline{\mathrm{D}}(\boldsymbol{\theta}) - \mathrm{D}(\tilde{\boldsymbol{\theta}})$, with $\overline{\mathrm{D}}(\boldsymbol{\theta})$ is the *posterior expectation of the deviance* estimated using the MCMC sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ from the posterior distribution as $\overline{\mathrm{D}}(\boldsymbol{\theta}) = -2\frac{1}{M}\sum_{m=1}^{M} \log f(\mathbf{y}|\boldsymbol{\theta}_m)$ and $\mathrm{D}(\tilde{\boldsymbol{\theta}})$ is the *deviance evaluated at the posterior mean* of the parameters $\tilde{\boldsymbol{\theta}} = \mathrm{E}[\boldsymbol{\theta}|\mathbf{y}]$. Additionally we consider the Expected Akaike Information Criterion (EAIC) and the Expected Bayesian Information Criterion (EBIC) which are calculated penalizing the $\overline{\mathrm{D}}(\boldsymbol{\theta})$ by using $2p$ and $p\log\, n$ as penalties function, respectively, where p is the number of parameters in the model and n is the sample size. For all criteria the smaller values indicate better fit.

In addition we propose to use the Widely Applicable Information Criterion WAIC which is based in the log pointwise posterior predictive density ($lppd$) given by $lppd = \sum_{i=1}^{n}\sum_{j=1}^{I} \log\left(\frac{1}{M}\sum_{m=1}^{M} L(\boldsymbol{\theta}_m, \boldsymbol{b}_m|y_{ij})\right)$, and then, to adjust for overfitting, add a term to correct for effective number of parameters

$\rho_{\mathrm{WAIC}} = \sum_{i=1}^{n}\sum_{j=1}^{I} V_{m=1}^{M}(\log L(\boldsymbol{\theta}_m, \boldsymbol{b}_m)|y_{ij})$, where $V_{m=1}^{M}(a) = \frac{1}{M-1}\sum_{m=1}^{M}(a_m - \bar{a})^2$. Finally, as proposed by Gelman et al. (2014), the WAIC is calculated by $\mathrm{WAIC} = -2(lppd - \rho_{\mathrm{WAIC}})$.

On the other hand, the conditional predictive ordinate CPO is another common approach to compare models. The CPO_{ij} follow the idea of the leave one out cross validation, where each value is an indicator of the likelihood value given all the other observations. Thus, low values of CPO_{ij} must correspond to poorly fitted observations. For the ij-th observation, the CPO_{ij} can be written as

$$CPO_{ij} = \pi\left(y_{ij}|\boldsymbol{y}_{(-ij)}\right) = \int\int L\left(\boldsymbol{\theta}, \boldsymbol{b}|y_{ij}\right) f\left(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{y}_{(-ij)}\right) d\boldsymbol{\theta} d\boldsymbol{b} = \left\{\int\int \frac{f\left(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{y}\right)}{L\left(\boldsymbol{\theta}, \boldsymbol{b}|y_{ij}\right)} d\boldsymbol{\theta} d\boldsymbol{b}\right\}^{-1},$$

where $\boldsymbol{y}_{(-ij)}$ is the $\boldsymbol{y}$ without the ij-th observation. Dey et al. (1997) showed that an harmonic mean approach can be used to do a Monte Carlo approximation of the

CPO_{ij} by using a posterior sample as $\widehat{CPO}_{ij} = \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{1}{L(\boldsymbol{\theta}_m, \boldsymbol{b}_m | y_{ij})} \right\}^{-1}$. Since the CPO_{ij} is defined for each observation, the log-marginal pseudo likelihood (LPML) given as LMPL $= \sum_{i=1}^{n} \sum_{j=1}^{I} \log\left(\widehat{CPO}_{ij}\right)$, is used to summarize the CPO_{ij} information and the larger the value of LMPL is, the better the fit of the model under consideration. For a revision of these criteria, one may refer to Gelman et al. (2013)

The codes for fitting the one-parameter IRT model using INLA method and to calculate model comparison criteria are available under requirement.

4. SIMULATION STUDY

4.1 Comparison between the INLA and ARS MCMC method

Four different scenarios were created with different sample and item sizes: 1) $n = 500$, $I = 15$; 2) $n = 500$, $I = 30$; 3) $n = 2000$, $I = 15$; 4) $n = 2000$, $I = 30$; and for each scenario 100 data sets of the Rasch model were generated. In all cases $\theta_i \overset{iid}{\sim} N(0,1)$ and $b_j \overset{iid}{\sim} N(0,2)$ is the baseline of our simulation.

To comparison with the INLA method we estimate difficulties and latent trait considering ARS MCMC method. The `R-INLA` and the OpenBUGS software in conjunction with the R library `rbugs` were considered respectively. To guarantee convergence in the simulation, each MCMC chain have 25,000 iterations with a burning period of 5,000 iterations and thinning of 10 resulting in 2,000 posterior samples. Since we only have one chain the convergence test was verified using Geweke's criterion.

To carry out the comparison for Bayesian inference using INLA and MCMC estimation method we used the root mean square error (RMSE), the mean absolute error (MAE) and the average correlation (Corr) between the estimated and true parameters. Table 1 presents the summarized results obtained by fitting both methods for 100 simulated data set under the specified scenarios. For parameters b the results showed similar performance in terms of the RMSE, MAE and Corr. The same pattern is observed for parameters θ, the results using INLA methods are very similar to those obtained for MCMC methods It could be argued that if we allow the MCMC chain to run longer enough we could improve the fit to become as good as desired. However, it is well known that for MCMC methods there is a trade-off between precision and waiting time.

[Table 1 about here.]

To confirm the results presented in Table 1 we focus in the simpler scenario, N = 500 and I = 15, where the INLA performs closer to the MCMC analysis. Thus, in Figure 1 we present a boxplot to compare the bias of the estimates of each parameter b_j for the INLA and MCMC for the 100 simulations. We can see that INLA and MCMC bias are very similar for parameters b_j, $j = 1, \ldots, I$, for the 100 simulations. Reassuring that INLA performs estimation of the parameters as well as MCMC.

[Figure 1 about here.]

On the other hand, the mean computational time to run each of the 100 data sets of each scenario is also presented in Table 1. From Table 1 it is clear that INLA performs the analysis much faster than the MCMC method. This gain is essential when dealing with larger data sets. In our biggest scenario (N = 2000 and I = 30) the MCMC chain took in average about 11 hours to run while INLA provided results in 13 minutes. The MCMC analysis becomes impracticable when dealing with larger data sets.

4.2 Comparison between the INLA and ML methods

Six different scenarios were created with different sample and item sizes: 1) N = 100, I = 10; 2) N = 100, I = 20; 3) N = 500, I = 10; 4) N = 500, I = 20; 5) N = 1000, I = 10; 3) N = 1000, I = 20; and for each scenario 50 data sets of the Rasch model were generated. In all cases $\theta_i \overset{iid}{\sim} N(0,1)$ and $b_j \overset{iid}{\sim} U(-2,2)$ is the baseline of our simulation.

To comparison with the INLA method we estimate difficulties and latent trait considering traditional Maximum Likelihood estimation approachs. Specifically we consider: (a) `MLM estimation + MAP` (we adopt marginal maximum likelihood (MML) method to the estimation of difficulties and maximum a posteriori (MAP) to estimation of abilities. To details see (Tong and Coombes, 2012) and Rizopoulos (2006)) (b) `CML` (we adopt conditional maximum likelihood (CML) methods to the estimation of difficulties and abilities, see Mair and Hatzinger (2007)) (c) `RMLM + MAP` (we adopt Restricted Maximum Likelihood (RML) method to the estimation of difficulties and MAP to the estimation of abilities. For details see De Boeck et al. (2011).

To carry out the comparison between the different methods considered we used the root mean square error (RMSE), the average correlation (Corr) between the estimated parameters and the true one and computational time. Table 2 presents the summarized results obtained by fitting all methods for 50 simulated data set under the specified scenarios.

[Table 2 about here.]

For both parameters b and θ we found that `MMLM + MAP` and INLA methods presents the best perfomance and very similar results in terms of the RMSE and Corr among then. In addition it is possible to see that INLA method is slightly slower than `MMLM+MAP` and `CLM` methods but slightly faster than the `RMLM + MAP` method. Although it could be argued that `R-INLA` is slower it performs full Bayesian inference and provide marginal posterior distributions for all parameters.

5. ANALYSIS OF THE ENEM MATHEMATICS EXAM

In order to evaluate the performance of the proposed approach with a large data set we analyze data set of the ENEM. The ENEM is the most important exam non-mandatory, standardized Brazilian national exam, which evaluates high school students in Brazil as a standard university entrance qualification test. For illustration of the approach development in this paper we consider only a version of a subtest. Items 136 (1 here) to 180 (45 here) of the sub test Mathematics and its Technologies correspondent to the Mathematics Area of the blue version are considered. The total time to the subtest considering items 91 to 180 was 5 hours with 30 minutes. The items and the answer correct are available in `http://portal.inep.gov.br/web/enem/edicoes-anteriores/provas-e-gabaritos`.

Official organization responsible for the ENEM, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) indicate that in the first step of the analysis is performed a review of the behavior of all the items, the known parameters of the items are reevaluated and estimates of the parameters in which items were not known yet are performed. This step is named of calibration. Only after the calibration phase is consolidated, starts then, the phase of estimation of the proficiency of the participants. Both procedures, item analysis and calculation of abilities in ENEM are based on IRT. Specifically the three-parameter logistic model is considered and the method used for calculation of abilities is called Expected A Posteriori (EAP). In this method, the expected a posteriori estimate of ability for response pattern is obtained using Hermite-Gauss quadrature approximation. For details see Baker and Kim (2004).

In order to illustrate the methodology we consider only candidates in Minas Gerais state. A total of 29442 individuals and 45 items corresponding to the Mathematics Area is the data set considered.

To evaluate alternative one-parameter IRT models, we consider logit-normal, probit-normal and loglog-normal models. Table 3 shows the fit comparison using several model comparison criteria discussed in section 4.2. We found, clearly, that the logit-normal or

Rasch model has a best fit for the ENEM data considering all criteria.

[Table 3 about here.]

This results is expected since that ENEM Test is based in the three-parameter model considering also the logit link, and then rasch model is a particular model.

Considering the Rasch model we show a Boxplot of the estimates of parameter b for the items of ENEM Mathematics Exam in Figure 2 using the ENEM scale $(500 + 100\hat{b}_j)$. With the ENEM scale is possible to classify the items in four groups. Then, we found that 25 of the 45 items of intermediate difficulty (ENEM scores between 400 and 600). We also identified 18 difficult items (ENEM scores between 600 and 700) and one with high difficulty (item 28 with ENEM scores between 600 and 700). Also we found one item (item 1) being the easiest in the Test.

[Figure 2 about here.]

In Table 4 we show statistics of the posterior distribution of some items picked from each group: very difficult (item 28), difficult (item 23), intermediate (item 45), easy (item 1). We show the posterior mean and HPD interval to the correspondent item. Also, we show the probability of correct response to a student with trait=500. Note that when the item is easy, then the probability of correct response is 0.80 but when the item is very difficult then the probability of correct response is 0.08.

[Table 4 about here.]

In addition we show in the Figure 3 the correspondent Item Characteristic Curves of the items choose considering the usual scale. Less difficult items are located on the left of the scale, most difficult items on the right.

[Figure 3 about here.]

Finally, since that marginal posterior distributions for all parameters are available, other summary measures can be obtained to the difficulties of the itens and trait latent variables of the students.

6. FINAL REMARKS

In this paper was presented how estimate the parameters of the IRT models INLA method (Rue et al., 2009), as an alternative method to commonly MCMC and ML methods in the literature. One-parameter IRT models with logit (Rasch model), probit, and loglog links were easily implemented.

In a first simulation study we compare the obtained results from both methodologies INLA and MCMC showing that both methods provide very similar estimates of the posterior distribution of the parameters of the models studied. However, the computational time of INLA method is much smaller than the traditional MCMC using for example OpenBUGS. For this reason, we advocate that INLA can be used to Bayesian estimation in IRT modeling for large data sets in a reasonable time as presented in Section 5 with a ENEM application. In addition, in a second simulation study we show that the INLA method performed similarly or more favorably than some of the ML methods in the literature briging additional information since we have the posterior distribution of the parameters.

Further extensions for two and three parameters IRT models using INLA methodology are under investigation and are not so trivial. This theme is investigated in a different manuscript and it is out of the scope of the current that aims to present how to perform in a simple manner large data set fitting of one parameter IRT models.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, 17:251–269.

Albert, J. H. and Ghosh, M. (2000). Item response modeling. In D.K. Dey, S. G. and Mallick, B., editors, *Generalized Linear Models: A Bayesian Perspective*, pages 173–193. Marcel-Dekker, New York.

Baker, F. and Kim, S.-H. (2004). *Item response theory: parameter estimation techniques*. Statistics: A Series of TextBooks and Monographs. Marcel Dekker, 2nd edition.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46:443–459.

Curtis, S. M. (2010). Bugs code for item response theory. *Journal of Statistical Software*, 36(1):1–34.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., and Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, 39(1):1–28.

Dey, D. K., Chen, M. H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 53:1239–1252.

Fischer, G. H. (2007). Rasch models. In Rao, C. R. and Sinharay, S., editors, *Psychometrics*, pages 515–586. Elsevier.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC Texts in Statistical Science. Taylor and Francis.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41:337–348.

Grilli, L., Metelli, S., and Rampichini, C. (2015). Bayesian estimation with integrated nested laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 85:2718–2726.

Levy, R. (2009). The rise of markov chain monte carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, 2009:1–18.

Lord, F. and Novick, M. (1968). *Statistical Theories of Mental Test Scores Reading*. M.A. Addison-Wesley.

Lord, F. M. (1952). A theory of test scores. Technical Report 7, Psychometric Corporation, Richmond, VA.

Maier, K. S. (2001). A rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, 26(3):307–330.

Mair, P. and Hatzinger, R. (2007). Extended rasch modeling: The erm package for the application of irt models in r. *Journal of Statistical Software*, 20(1):1–20.

Maris, G. and Maris, E. (2002). A mcmc-method for models with continuous item responses. *Psychometrika*, 67:335–350.

Martino, S. and Rue, H. (2010). Implementing approximate bayesian inference using integrated nested laplace approximation: a manual for the inla program. *Department of Mathematical Sciences, Norwegian University of Science and Technology.*

Patz, R. J. and Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response. *Journal of Educactional and Behavioral Statistics*, 24:146–178.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danmarks Paedogogiske Institut, Copenhangen.

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

Rupp, A., Dey, D., and Zumbo, B. (2004). To bayes or not to bayes, from whether to when: Applications of bayesian methodology to modeling. *Structural Equations Modeling*, 11:424–451.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, 72:217–232.

Stone, C. A. and Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models Using SAS.* SAS Institute Inc., Cary, NC, USA.

Swaminathan, H. and Gifford, J. (1982). Bayesian estimation in the rasch model. *Journal of Educational Statistics*, 7(3):175–191.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal American Statistical Association.*, 81:82–86.

Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Scientific Software International.

Tong, P. and Coombes, K. R. (2012). integirty: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory. *Bioinformatics*, 28(22):2861–2869.

Table 1: Comparison between INLA and MCMC methods to the performance on the recovery of the b's and θ's parameters and computational time in the Rasch model to different sample (n) and item sizes (I). Mean absolute error(MAE), root mean square error (RMSE) and correlation (Corr) based on 100 runs of the simulation.

n	I	b difficulties						time (min)	
		RMSE		MAE		Corr			
		INLA	MCMC	INLA	MCMC	INLA	MCMC	INLA	MCMC
500	15	0.418	0.421	1.291	1.298	0.997	0.997	0.39	48.32
	30	0.628	0.630	2.822	2.852	0.997	0.997	2.00	85.31
2000	15	0.250	0.250	0.777	0.780	0.999	0.999	3.10	277.14
	30	0.353	0.354	1.570	1.570	0.999	0.999	13.30	667.15
n	I	θ trait persons							
		RMSE		MAE		Corr			
		INLA	MCMC	INLA	MCMC	INLA	MCMC		
500	15	11.821	11.810	209.829	209.630	0.836	0.836		
	30	9.130	9.109	162.383	162.053	0.907	0.907		
2000	15	24.110	24.046	856.541	855.254	0.844	0.844		
	30	18.580	18.543	660.129	659.695	0.911	0.911		

Table 2: Comparison between INLA and Maximum Likelihood methods to the performance on the recovery of the b's and θ's parameters in the Rasch model to different sample (n) and item sizes (I). Root mean square error (RMSE), correlation (Corr) and computational time (time) based on 50 runs of the simulation.

n	I	MML+MAP				
		RMSE b	Corr b	RMSE θ	Corr θ	time
100	10	0.257	0.982	0.615	0.791	0.177
	20	0.264	0.979	0.476	0.880	0.317
500	10	0.102	0.996	0.606	0.793	0.520
	20	0.113	0.996	0.482	0.879	0.987
1000	10	0.083	0.998	0.615	0.788	0.678
	20	0.082	0.998	0.479	0.879	2.142
n	I	CML				
		RMSE b	Corr b	RMSE θ	Corr θ	time
100	10	0.378	0.982	0.917	0.787	0.192
	20	0.310	0.979	0.621	0.878	0.627
500	10	0.305	0.996	0.916	0.791	0.229
	20	0.231	0.996	0.630	0.876	0.547
1000	10	0.340	0.998	0.928	0.784	0.264
	20	0.168	0.998	0.613	0.875	0.721
n	I	RML + MAP				
		RMSE b	Corr b	RMSE θ	Corr θ	time
100	10	0.263	0.982	0.744	0.791	5.173
	20	0.264	0.979	0.580	0.879	23.352
500	10	0.102	0.996	0.758	0.793	34.081
	20	0.112	0.996	0.598	0.879	264.233
1000	10	0.084	0.998	0.778	0.787	75.669
	20	0.082	0.998	0.563	0.878	363.511
n	I	INLA				
		RMSE b	Corr b	RMSE θ	Corr θ	time
100	10	0.241	0.983	0.614	0.791	4.713
	20	0.249	0.979	0.475	0.880	3.105
500	10	0.101	0.996	0.606	0.793	7.172
	20	0.112	0.996	0.482	0.879	13.133
1000	10	0.082	0.998	0.615	0.788	15.524
	20	0.082	0.998	0.479	0.879	38.940

Table 3: Model comparison Criteria to alternative one-parameter IRT models using different links to ENEM Mathematics Exam.

Criteria	one-parameter IRT models		
	Rasch	Probit-Normal	Loglog-Normal
Dbar	1474076.7	1476924.1	1487259.5
Dhat	1448724.7	1448984.9	1459079.1
DIC	1499428.6	1504863.3	1515440.0
EAIC	1533050.7	1535898.1	1546233.5
EBIC	1889750.2	1892597.6	1902933.1
WAIC	1500754.9	1506682.3	1517661.7
-2LPML	1500740.0	1506709.2	1517694.1

Table 4: Summary of posterior distribution to some items of the ENEM Mathematics Exam.

Item Group	Item	mean	sd	Interval HPD	Prob of correct response to ENEM score = 500
very difficult	28	740.7	2.1	736.6 – 744.6	0.08
difficult	23	645.2	1.6	642.1 – 648.3	0.19
intermediate	45	503.7	1.4	501.0 – 506.3	0.49
easy	1	361.2	1.6	358.2 – 364.3	0.80

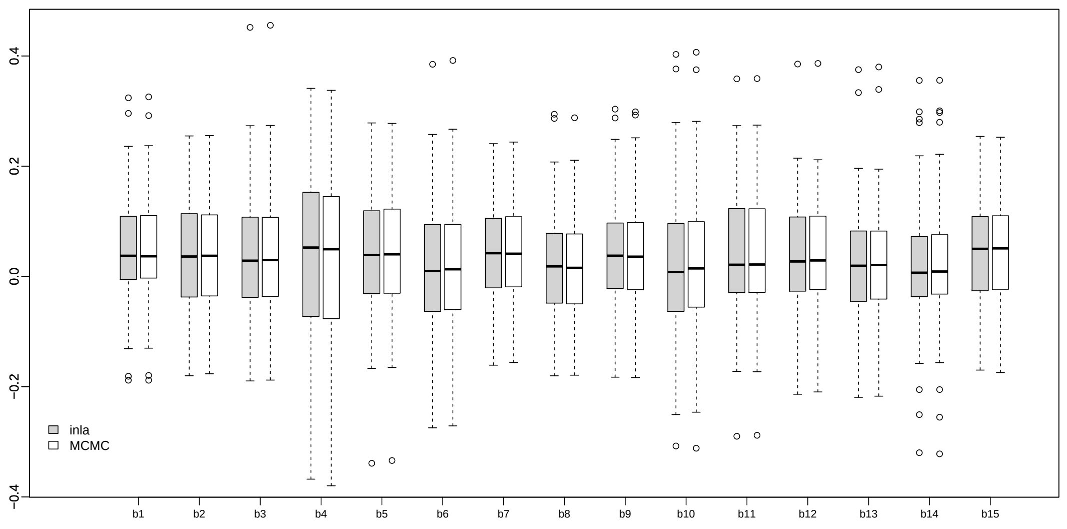

Figure 1: Boxplot of the bias of the estimates of b's parameters in the Rasch model for both INLA (light grey) and MCMC (white) methods based on 100 runs of the simulation with $n = 500$ examinees and $I = 15$ items.

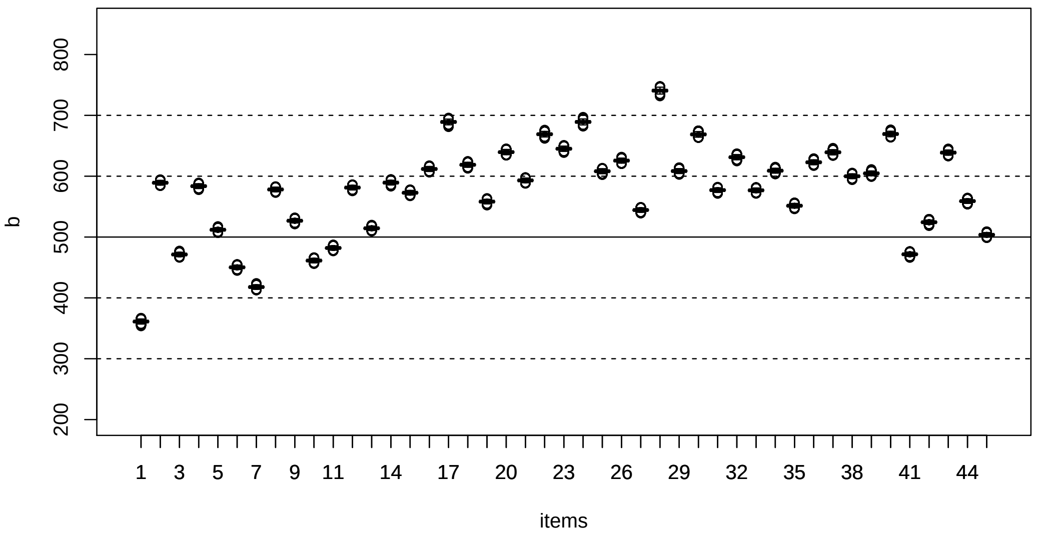

Figure 2: Boxplot of the estimates of parameter b's for 45 items of the ENEM Mathematics Exam.

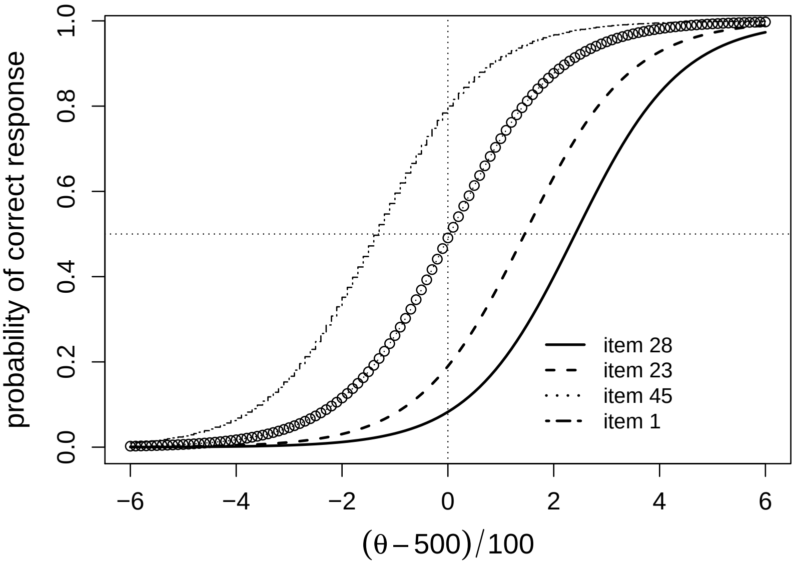

Figure 3: Item characteristic curves to some item choose of the ENEM Mathematics Exam: item 28 (solid line) is very difficult, item 23 (dashed line) is difficult, item 45 (dotted line) is intermediate, item 1 (twodash line) easy.