Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 1, 1–23
DOI: 10.1214/12-BJPS191

Extensions of the skew-normal ogive item response model

Jorge Luis Bazán[a], **Márcia D. Branco**[b] **and Heleno Bolfarine**[b]

[a]*Pontificia Universidad Católica del Perú*
[b]*Universidade de São Paulo*

Abstract. We introduce new applications of the skew-probit IRT model by considering a flexible skew-normal distribution for the latent variables and by extending this model to include an additional random effect for modeling dependence between items within the same testlet. A Bayesian hierarchical structure is presented using a double data augmentation approach. This can be easily implemented in WinBUGS or SAS by considering MCMC algorithms. Several Bayesian model selection criteria, such as DIC, EAIC and EBIC, have been considered; in addition, we use posterior sum of squares of the latent residuals as a global discrepancy measure to model comparison. Two applications illustrate the methodology, one data set related to a mathematical test and another related to reading comprehension, both applied to Peruvian students. Results indicate better performance of the more flexible models proposed in this paper.

1 Introduction

Typically, Item Response Theory (IRT) for multivariate dichotomous responses resulting from n individuals evaluated in a test with k items considers a unidimensional latent variable θ associated with individual abilities. Moreover, statistical models for IRT consider a set of parameters associated with the items under consideration that are related with the probability that the ith examinee is able to answer the jth item correctly.

Usually a symmetric Item Characteristic Curve (ICC) is considered for IRT models. The supposition of symmetry for the ICC implies that, the probability to answer correctly an item approaches zero with the same rate as it approaches one. Therefore, individuals with high or low abilities are discriminated in a similar fashion. However, when the interest of the test is more directed to discriminate among individuals with high abilities and not as much to discriminate among those with low abilities (or vice-versa) an asymmetric ICC can be more appropriate.

Samejima (1997, 2000) exhibits inconsistencies of symmetric ICCs under some philosophical principles. For example, when the goal is to detect student's brightness it is reasonable to consider two principles: (1) more credit (reward) should be given to a person who is successful with difficult items, and (2) a person who fails

Key words and phrases. Bayesian estimation, item response models, latent variables, skew-normal ogive model, skew-normal distribution, skew-probit link, testlet.
Received June 2011; accepted March 2012.

in an easy item should be penalized more. Since an asymmetric ICC deals better with these two principles, Samejima proposed a family of asymmetric ICCs. This family, termed *logistic positive exponential*, has the logit link as a special case. A new item parameter associated with skewness is introduced with the intention of adjusting the above principle. Estimation of this model together with a new IRT model, namely the *reciprocal logistic positive exponential*, is presented in Bolfarine and Bazán (2010).

Bazán, Branco and Bolfarine (2006) proposed another skew item response model, namely here the *skew-normal ogive model*. In this model, a new asymmetric ICC curve is assumed by considering the cumulative distribution function of the standard skew-normal distribution (Azzalini, 1985). Thus, a new skewness item parameter is introduced over normal curves and the symmetric normal ogive model (Albert, 1992) is a particular case. In this paper, we study extensions and additional properties of the skew-normal ogive model (SNO model).

Statistical assumptions in modeling academic achievement and other latent variables associated with human behavior are based on the normality assumption for these variables. Several authors have questioned this assumption (see Samejima, 1997; Micceri, 1989) since it is somewhat restrictive for modeling human behavior. Micceri (1989) presents many examples of situations where the latent variables can be assumed not to be normally distributed. For instance, when the focus is on item calibration, it is important to cover all levels of abilities in the population, in this case it is reasonable to consider a uniform distribution for the abilities.

For some authors, the parametric distributional assumptions for the latent variable is not part of the Item Response modeling and also a nonparametric model can be assumed for the latent variable (see, e.g., Duncan and MacEachern, 2008). However from a Bayesian perspective, which is assumed in this paper, the complete specification of the IRT model is given by the specification of the likelihood function and all prior distributions, in particular the prior distribution for the latent variable θ. It is a common practice to consider the normal distribution as a prior for latent variables (see, e.g., Albert, 1992; Patz and Junker, 1999). However, we find it important to explore the possibility of using flexible distributional assumptions for the latent variables, such as the skew-normal distribution. This flexible assumption can be an alternative between the restrictive normal model and a full nonparametric model. As we will show here the procedure to estimate the parameters using that distribution is a natural extension of the normal assumption for the latent variables.

On the other hand, in reading comprehension tests the design of testlets or item bundles has been adopted recently in educational and psychological tests (see, e.g., Wainer, Bradlow and Wang, 2001). Additionally, new applications are being considered as, for example, in Wainer et al. (2001) and Wang et al. (2010).

Fitting standard item response models to testlet responses ignores the possible dependence between the items within a testlet. As indicated by Wang and Wilson (2005), not considering this dependence tends to overestimate the precision

of measures obtained from testlets and yields biased estimation for item difficulty and discrimination parameters. Overstatement of precision and biased estimation lead to inaccurate inferences about the parameters (Wainer and Wang, 2000). Thus, Bradlow, Wainer and Wang (1999) extend the ogive normal model to include an additional random effect to model the dependence between items within the same testlet. The variances of the random testlet effects were assumed to be constant across testlets. This model has been successfully used in multiple applications (see Wainer, Bradlow and Wang, 2001). However, to the best of our knowledge we are no ware of studies on the use of asymmetric links to item characteristic curves in the context of tests with testlets.

The paper is organized as follows. In Section 2, we present a review of the skew-normal ogive IRT model and the data augmentation approach. In the third section, we extend the SNO IRT model by considering asymmetrically distributed latent variables. In Section 4, we present the skew-normal ogive testlet model. Section 5 presents two applications to illustrate the good performance of the approach developed here by considering different criteria for model comparison. The paper ends with a discussion in Section 6 and two appendices. In Appendix A.1, we present properties of the skew-normal distribution, and in Appendix A.2, we present the program code to implement the new model proposed.

2 The skew-normal ogive model

Let Y_{ij} denotes the dichotomous variable corresponding to the response of the ith individual, $i = 1, \ldots, n$, on the jth item, $j = 1, \ldots, k$. Y_{ij} takes the value 1 if the response is correct and 0, otherwise. The skew-normal ogive IRT model proposed by Bazán, Branco and Bolfarine (2006) follows by considering:

$$Y_{ij}|\theta_i, \eta_j \sim \text{Bern}(p_{ij}), \tag{2.1}$$

$$p_{ij} = F_{d_j}(m_{ij}), \tag{2.2}$$

$$m_{ij} = \alpha_j(\theta_i - \beta_j) \tag{2.3}$$

with $\alpha_j > 0$, $-\infty < \beta_j < \infty$, $-1 < d_j < 1$, and $-\infty < \theta < \infty$ and Bern$(\cdot)$ denoting a Bernoulli distribution. The probability $p_{ij} = P(Y_{ij} = 1|\theta_i, \eta_j)$ is the *conditional probability of correct response* given the ith ability value θ_i and the jth item parameter $\eta_j = (\alpha_j, \beta_j, d_j)$. Moreover, $F_{d_j}(\cdot)$ denotes the cumulative distribution function (c.d.f.) of the standard skew-normal distribution (see Appendix A.1) and m_{ij} is the *latent linear component* relating θ_i and the item parameter η_j.

Note that $F_{d_j}(\cdot)$ is an asymmetric ICC and $F_{d_j}^{-1}(\cdot)$ is the BBB skew-probit link (see Bazán, Bolfarine and Branco, 2010). The SNO model satisfies the *latent monotonicity property* (Holland and Rosenbaum, 1986), that is, it is a monotone increasing function of the unidimensional quantity θ_i. Thus, the SNO model is a unidimensional monotone latent variable model (Junker and Ellis, 1997). Moreover,

it satisfies the *latent conditional independence principle* or the *local independence* principle, which considers that for the ith examinee, $\{Y_{ij} : j \geq 1\}$ are conditionally independent given θ_i. It is *assumed* that responses from different individuals are also independent.

Let $Y_i^T = (Y_{i1}, \ldots, Y_{ik})$, $\theta^T = (\theta_1, \ldots, \theta_n)$ and $\eta^T = (\eta_j, \ldots, \eta_k)$ and *suppose that the assumptions mentioned earlier hold*. The multivariate joint probability distribution of $Y = (Y_1^T, \ldots, Y_n^T)^T$, given the vector of latent variables θ^T and the item parameter vector η^T, can be written as

$$p(Y|\theta, \eta) = \prod_{i=1}^{n} \prod_{j=1}^{k} [p_{ij}]^{Y_{ij}} [1 - p_{ij}]^{1-Y_{ij}}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, k. \quad (2.4)$$

For $d_j = 0$ it follows that $p_{ij} = \Phi(m_{ij})$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. The symmetric ogive normal IRT model then follows. Additionally, an approximation for the logistic IRT model is obtained when $\Phi(\cdot/1.7)$ is considered (see Camilli, 1994).

Figure 1 depicts skew-probit ICCs for different values of the shape parameter d_j and fixed item parameters $\alpha_j = 1$ and $\beta_j = 0$. Six different ICCs are considered taking $d_j = -0.9, -0.7, -0.5, 0.5, 0.7, 0.9$ for comparison with $d_j = 0$ (the symmetric ICC). For $d_j > 0$ the ICC presents positive skewness, and for $d_j < 0$ it

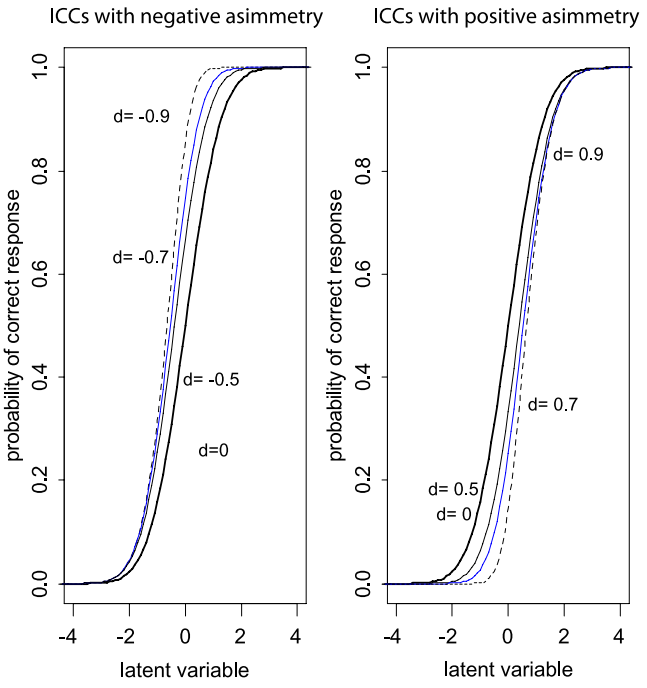

Figure 1 *Skew-probit ICCs for* $\alpha = 1$, $\beta = 0$ *and different values of the shape parameter* d.

presents negative skewness. As in the positive exponent logistic models considered by Samejima (1997, 2000), the skew-probit ICC also considers a new item parameter to control the asymmetry of the curve; however, now we have an extension for the probit ICC and not for the logistic ICC.

The item shape parameter can be psychometrically interpreted as a penalty (reward) on the probability of correct response. Hence, an item with negative shape parameter penalizes (rewards) students with larger (smaller) levels of the latent variable and an item with positive shape parameter rewards (penalizes) individuals with larger (smaller) levels of the latent variable (see Figure 1). We call the shape parameter d the *item penalization parameter*.

As in the usual symmetric IRT model, the parameters α_j and β_j are called discrimination (or slope) item parameter and difficulty (or intercept) item parameter, respectively. A steeper item response curve corresponds to an item that highly discriminates students of smaller and greater levels of the latent variable. An item with a small value of α_j is a relatively poor discriminator between students for changing values of the latent variable. Moreover, an item of a test with a highly negative value of β_j corresponds to an easy item in which individuals with smaller averages in the latent variable show relatively low probabilities of correct responses. In contrast, an item with a large value of β_j is a difficult item since individuals with larger levels of the latent variable show relatively low probabilities of correct responses.

Following different proposals in the literature, we reparameterize the model introduced by considering $a_j = \alpha_j$ and $b_j = -\alpha_j\beta_j$ such that $m_{ij} = a_j\theta_i - b_j$, with $\eta_j = (a_j, b_j)$ the item parameter corresponding to the jth item. According to Baker and Kim (2004), this parameterization may result in more stable computations. We use the notation $a = (a_1, \ldots, a_k)^T$ and $b = (b_1, \ldots, b_k)^T$.

2.1 Data augmentation approach

Denoting $\mathbf{D}_{obs} = \mathbf{y}$ the observed data, the likelihood function for the SNO can be written as

$$L(\theta, \eta, |\mathbf{D}_{obs}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[F_{d_j}(m_{ij})\right]^{y_{ij}}\left[1 - F_{d_j}(m_{ij})\right]^{1-y_{ij}}. \tag{2.5}$$

Following Albert (1992), for n examinees responding a test with k items, it is known that the probit link can be rewritten as

$$Y_{ij} = \begin{cases} 1, & Z_{ij} > 0, \\ 0, & Z_{ij} \leq 0, \end{cases}$$

where $Z_{ij} = m_{ij} + e_{ij}$ with $e_{ij} \sim N(0, 1)$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$. It follows that

$$p_{ij} = P(Y_{ij} = 1|\theta_i = u_i, \eta_j) = \Phi(m_{ij}).$$

This representation shows a linear *structure* in the auxiliary latent variable Z_{ij}, which produces an equivalent model with a probit link. Further, e_{ij}'s are latent residuals *which are* independent and identically distributed (see Albert and Chib, 1995) and this fact can be used for model checking.

Similarly, as shown in Bazán, Branco and Bolfarine (2006), it is possible to define a latent linear error structure for the skew-normal ogive model replacing the normality assumption of the error terms by the skew-normality assumption. Following notation from Appendix A.1, when $e_{ij} \sim \text{SN}(0, 1, -\lambda_j)$ we have

$$p_{ij} = P(Y_{ij} = 1|\theta_i, \eta_j) = F_{d_j}(m_{ij}).$$

Note that, the parameter $d_j = \frac{\lambda_j}{\sqrt{1+\lambda^2}}$ is a reparametrization of the shape parameter λ_j. In addition, using the stochastic representation (Henze, 1986) for a skew-normal distribution we can write

$$e_{ij} = -d_j V_{ij} - \left(1 - d_j^2\right)^{1/2} W_{ij},$$

with $W_{ij} \sim N(0, 1)$ and $V_{ij} \sim \text{HN}(0, 1)$ (the half-normal distribution). It follows that the conditional distribution of the e_{ij} given $V_{ij} = v_{ij}$ is a normal distribution with mean $-d_j v_{ij}$ and variance $1 - d_j^2$ (see Appendix A.1).

Simulation of the latent variables Z_{ij} should be considered in two steps. First simulate $V_{ij} \sim \text{HN}(0, 1)$ and then, simulate from the conditional distribution $Z_{ij}^* \equiv Z_{ij}|V_{ij} = v_{ij}$ where $Z_{ij}^* \sim N(m_{ij} d_j v_{ij}, 1 - dj)$. Additionally, the latent residuals e_{ij}'s (all independent) in the skew probit link can be used also for model checking.

We consider now the complete data likelihood function involving the conditional auxiliary latent variables $\mathbf{Z}^* = (Z_1^{*T}, \ldots, Z_n^{*T})^T$, with $Z_i^{*T} = (Z_{i1}^*, \ldots, Z_{ik}^*)$, $i = 1, \ldots, n$, and $\mathbf{V} = (V_1^T, \ldots, V_n^T)^T$, with $V_i^T = (V_{i1}, \ldots, V_{ik})$, $i = 1, \ldots, n$. The complete-data likelihood function for the SNO IRT model with $\mathbf{D} = (\mathbf{Z}^*, \mathbf{V}, \mathbf{y})$ is given by

$$L(\mu, \eta, \lambda|D) = \prod_{i=1}^{n} \prod_{j=1}^{k} \phi^*\left(Z_{ij}^*\right) I\left(Z_{ij}^*, Y_{ij}\right) \phi(V_{ij}) I(V_{ij} > 0), \tag{2.6}$$

where $\phi^*(\cdot)$ denotes the probability density function of the normal distribution with mean $m_{ij} - d_j v_{ij}$, variance $1 - d_j^2$ and $I(Z_{ij}^*, Y_{ij}) = I(Z_{ij}^* > 0) I(Y_{ij} = 1) + I(Z_{ij}^* \leq 0) I(Y_{ij} = 0)$. Note that, when $d_j = 0$ the likelihood function above is the one given in Albert (1992).

2.2 Prior specification

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

$$\pi(\theta, \eta) = \prod_{i=1}^{n} g_1(\theta_i) \prod_{j=1}^{k} g_2(\eta_j), \tag{2.7}$$

where g_1 and g_2 are specified probability *density functions* for θ_i and η_j, respectively, $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Additionally, for simplicity we assume independence between a_j, b_j, d_j, so that

$$g_2(\eta_j) = g_{21}(a_j)g_{22}(b_j)g_{23}(d_j).$$

For the normal ogive model, Ghosh et al. (2000) and Albert and Ghosh (2000) showed that g_{21} should be proper in order to guarantee a proper posterior distribution, however g_{22} can be improper. We consider proper prior densities of probabilities for all item parameters to avoid possible problems. Following Rupp, Dey and Zumbo (1992) and Sahu (2002), we take a normal with positive values for g_{21} denoted by $N(\mu_a, \sigma_a^2)I(\cdot, \cdot)$ and $N(\mu_b, \sigma_b^2)$ for g_{22}. For the new item parameters d_j, we consider a noninformative prior uniform on $[-1, 1]$. The prior specification for the parameters associated with the individuals θ_i will be discuss in the next section.

3 Extending the SNO IRT model by considering asymmetrically distributed latent variables

In the classical formulation of the IRT models one assumption typically added to the model, although not widely accepted, is that $\theta_i \sim N(\mu, \sigma^2)$. This establishes that the latent variables associated with the individuals taking the test are *well behaved* and that their abilities are a random sample from this distribution. We can specify, for μ and σ^2, $\mu = 0$ and $\sigma^2 = 1$ (as in Albert, 1992), or specify probability distributions for μ and σ^2 (as Patz and Junker, 1999) to solve the identifiability problem of the IRT model.

From a Bayesian perspective, the specification of a probability distribution for the latent variables θ_i is the same as the specification of a prior distribution. We propose here an asymmetric class of prior distributions for the ith individual latent variable θ_i, given by

$$\theta_i \sim \mathrm{SN}(\mu, \sigma^2, \gamma), \qquad i = 1, \ldots, n \tag{3.1}$$

with $-\infty < \mu < \infty, \sigma^2 > 0, -\infty < \gamma < \infty$.

This skew-normal probability density function (p.d.f.) will be denoted by $g_1(\theta_i)$.

Notice that, considering a priori equal asymmetry for all the abilities, does not imply that, a posteriori, equal asymmetry will result. That is, after we observe the results of the tests, the posterior distributions of the asymmetry for the abilities can take different values for each individual.

As before, we prefer to use the alternative parametrization for the shape parameter given by $\omega = \frac{\gamma}{\sqrt{1+\gamma^2}}$, which takes values in $[-1, 1]$.

Figure 2 depicts the density functions of the latent variables for different values of the shape parameter γ. The three curves on the right-hand side are examples with positive shape parameter γ modeling latent variables concentrated on lower

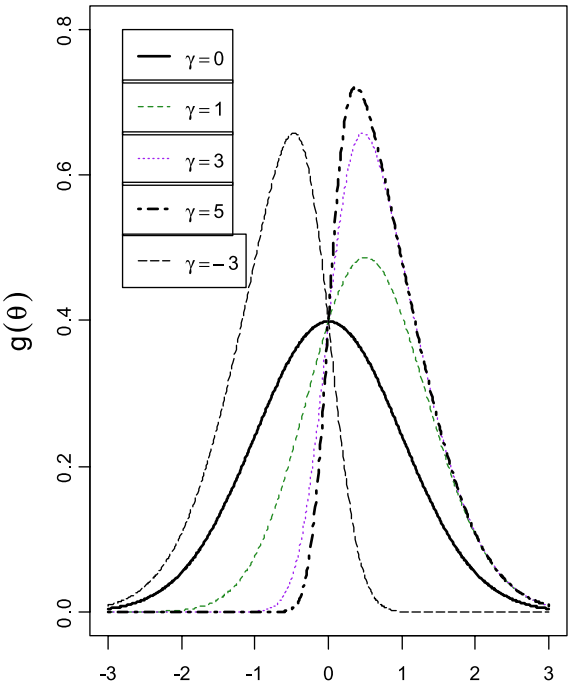

Figure 2 *Different skew-normal density functions, with* $\mu = 0$ *and* $\sigma^2 = 1$.

values. The three curves on the left side are examples with a negative shape parameter γ modeling latent variables concentrated on higher values. As a reference, in all cases the $N(0, 1)$ curve is also presented.

The curves shown in Figure 2 are reasonable assumptions for the distribution of the latent variables associated with human behavior in different contexts, as observed in Micceri (1989). Examples of such behavior are anxiety (see Zaider et al., 2001) and depression (see Riddle, Blais and Hess, 2002), where certain skewness is expected considering a non-clinic population. Moreover, as noted in Hashimoto (2002), in educational contexts several predictor variables related to school proficiency can be asymmetrically distributed.

Another situation to motivate the assumption of asymmetry for latent variables is obtained by a selection process where a bias is induced in the sample process which induces skewness in the original distribution as discussed in Arellano-Valle, Branco and Genton (2006).

Hence, the specification in (3.1) is flexible and also accommodates the normal distribution as a special case. Note that, the asymmetry considered here is different from the ICC asymmetry, which is related to a latent error for the definition of threshold to correct response.

The idea of the formulation considered here for the latent variable in (3.1) was originally presented in Arellano-Valle, Branco and Genton (2006). As noted by

Azevedo, Bolfarine and Andrade (2010) the model is not identifiable. However, differently from their formulation which considers a centered parameterization to overcome this issue, we consider the use of priors distributions for the hyperparameter.

A particular specification for the abilities is to consider the $\mathrm{SN}(0, 1, \gamma)$. In this case, the usual normal specification is a particular case ($\gamma = 0$). In that case, the mean and variance are not centered, that is, they are not 0 and 1, respectively. A centered parametrization can be obtained by considering $\mu = \frac{-0.7978846\gamma}{\sqrt{(1-0.6366198\gamma^2)}}$ and $\sigma^2 = \frac{1}{1-0.6366198\gamma^2}$, then $E(\theta_i) = 0$ and $V(\theta_i) = 1$. It is equivalent to the use of a centered parametrization (see Azevedo, Bolfarine and Andrade, 2010).

By considering two or just one type of asymmetry on the specification of the ICC curve or on the specification of the distribution of the latent variable, four scenarios are possible:

(a) skew-normal distribution for ICC and skew-normal distribution for abilities, namely *Skew-Probit Skew-Normal (SPSN) model*;
(b) skew-normal distribution for ICC and normal distribution for abilities, namely *Skew-Probit Normal (SPN) model*;
(c) normal distribution for ICC and skew-normal distribution for abilities, namely *Probit Skew-normal (PSN) model* and
(d) normal distribution for ICC and normal distribution for abilities, namely *Probit normal (PN) model.*

Notice that, (a) and (b) are skew-normal ogive models and (c) and (d) are normal ogive models where (d) is the usual model (Albert, 1992). We call these four scenarios the SN-IRT family (see also Bazán, Bolfarine and Branco, 2004).

An interesting aspect of the models formulated above is the flexibility in detecting items specified according to ICCs with skew-probit links and items specified according to ICCs with probit links. Thus, the SN-IRT family is a very flexible model.

The model in (a) is more general and it involves a total of $n + 3 + 3k$ unknown parameters. In contrast, the model in (d) is simpler and it involves $n + 2K$ unknown parameters. As a consequence of the introduction of new parameters, these models become overparameterized and then nonidentifiable. In such cases, a potential advantage of the Bayesian analysis over likelihood-based one is that if informative priors are available, as is the case here, proper inferences can be obtained despite of having an overparameterized (Rannala, 2007) model. In addition, as indicated in Poleto et al. (2011), the use of proper prior distributions in the Bayesian framework readily allows us to obtain valid inferences even for nonidentifiable models.

3.1 Fully Bayesian specification

Considering the augmented likelihood given in (2.6) and the prior specification discussed before, the fully Bayesian set-up for the most general ogive skew-normal

IRT model is given by the following hierarchical structure:

$$
\begin{aligned}
Z^*|v_{ij}, y_{ij}, \theta_i, a_j, b_j, d_j &\sim N\big(m_{ij} - d_j v_{ij}, 1 - d_j^2\big) I\big(Z_{ij}^*, Y_{ij}\big), \\
v_{ij} &\sim \mathrm{HN}(0, 1), \\
\theta_i &\sim \mathrm{SN}\big(\mu, \sigma^2, \omega\big), \\
a_j &\sim N\big(\mu_a, \sigma_a^2\big) I(\cdot, \cdot), \qquad b_j \sim N\big(\mu_b, \sigma_b^2\big), \\
d_j &\sim U(-1, 1), \\
\mu &\sim N(0, 1), \qquad 1/\sigma^2 \sim \mathrm{Gamma}(0.01, 0.01), \\
\omega &\sim U(-1, 1).
\end{aligned}
$$

By considering the hierarchical structure, this formulation can be easily implemented in WinBUGS (Lunn et al., 2000). Alternatively, one can use the MCMC procedure in SAS (SAS Institute Inc., 2009).

Particular cases of this more general SNO model can be derived by eliminating some lines in the hierarchical structure.

4 Extending the SNO IRT model by considering testlets

Wainer and Kiely (1987) introduced the testlet terminology and defined that as a group of items related to a single content area, developed as a unit. It contains a fixed number of predetermined paths that an examinee may follow. Examples of testlets include a set of reading items associated with a common passage, a group of social studies items referring to a map, or a collection of mathematics items based on a graph or a table. Thus, items that are part of a testlet are not statistically independent. Item responses within a testlet are not locally independent because they are related through a common stimulus. Therefore, in this case, the usual IRT models can lead to an inaccurate estimation of examinees' and items' parameters, and also overestimate the precision of these parameters (Tuerlinckx and De Boeck, 2001). Bradlow, Wainer and Wang (1999) propose to retain the item as the unit of measurement and add a person-specific random effect parameter, to account for the shared variance among items within a testlet. This parameter is called the testlet effect parameter and is denoted γ_{il}.

In order to extend the model by considering the testlets, we consider the hierarchical structural specification of the most general SNO IRT model later in this section. The following modification of the latent linear component presented in (2.3) is considered, that is,

$$
m_{ij} = \alpha_j(\theta_i - \beta_j + \gamma_{il}), \qquad i = 1, \ldots, n,\ j = 1, \ldots, k, l = 1, \ldots, t, \tag{4.1}
$$

where γ_{il} is a person-specific testlet effect or the random effect for person i on testlet l. Thus, γ_{il} describes the interaction between persons and items (local item

dependence) within the testlet independent of the ability and item parameters. The prior specification for these parameters is given by

$$\gamma_{il} \sim N\big(0, \sigma^2_{\gamma_l}\big). \tag{4.2}$$

The testlet effects were centered at zero to emphasize their status as deviations from the SNO IRT model and to identify the model and $\sigma^2_{\gamma_l}$ indicates the amount of the testlet effect for testlet l.

4.1 Fully Bayesian specification

The hierarchical structure specification of the SNO testlet model is given by

$$\begin{aligned}
Z^*|v_{ij}, y_{ij}, \theta_i, a_j, b_j, d_j &\sim N\big(m_{ij} - d_j v_{ij}, 1 - d_j^2\big) I(Zs_{ij}, Y_{ij}),\\
v_{ij} &\sim \mathrm{HN}(0,1),\\
\theta_i &\sim N(0,1),\\
\gamma_{il} &\sim N\big(0, \sigma^2_{\gamma_l}\big),\\
a_j \sim N\big(\mu_a, \sigma^2_a\big), &\qquad b_j \sim N\big(0, \sigma^2_b\big), \qquad d_j \sim U(-1,1),\\
1/\sigma^2_{\gamma_k} &\sim \chi(s),
\end{aligned}$$

where $\chi(s)$ corresponds to a chi-square distribution with s degrees of freedom.

This formulation can be easily implemented in WinBUGS or with the MCMC procedure in SAS.

When $\sigma^2_{\gamma_l} = \sigma^2_\gamma$ (the same variance for all testlets) the model reduces to the two-parameter testlet model proposed by Bradlow, Wainer and Wang (1999). When $a_i = 0$ the model reduces to the one-parameter Rasch testlet model proposed by Wang and Wilson (2005). If a skew-normal distribution is considered for the abilities, then corresponding prior must be considered for the hyperparameters.

5 Applications

We illustrate the methodology developed in the earlier sections using two data sets from Peruvian students. The first data set shows results from a mathematical test applied in rural schools in Peru. The second data set, is related to reading comprehension test applied to Peruvian students in some cities in the jungle region of Peru. Prior specification, starting values for the MCMC algorithm and convergence diagnostics are discussed. The MCMC procedure is based on the data augmentation approach discussed in Sections 3.1 and 4.1, respectively, and it is implemented using the WinBUGS software. Model comparison between symmetrical and asymmetrical IRT models are developed by using the Deviance Information Criterion (DIC) described in Spiegelhalter et al. (2002), the Expected Akaike Information Criterion (EAIC) and the Expected Bayesian Information Criterion (EBIC). Moreover, the

sum-of-squared-latent residuals ($SSR = \sum_{i=1}^{n} \sum_{j=1}^{k} e_{ij}^2$) introduced in Section 2 is also considered.

EAIC and EBIC are criteria proposed in Brooks (2002) and Carlin and Louis (2000) and were used by Bolfarine and Bazán (2010) in the context of TRI. These criteria penalize the *posterior expected deviance* 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. On the other hand, DIC penalizes the posterior expected deviance by using $2\rho_D$, where ρ_D is a complexity measure associated with the *effective number of parameters in a model*. This is given by the difference between the posterior mean of the deviance function and the deviance at the posterior estimates of the parameter of interest. In fact, the *posterior expected deviance* or *Dbar* can also be considered as a model comparison criterion. For all criteria, the smaller values indicate better fit.

5.1 Math data set

We consider an analysis on the response pattern obtained by the application of a Mathematical Test for fourth grade students of rural Peruvian Elementary Schools. Item response vectors are available from the authors upon request and correspond to the response of 974 students to 18 items qualified as binary response (correct and incorrect). The scores have mean equal to 8.27, median 8 and standard deviation 4.20. The sample skewness and kurtosis indexes are -0.075 and -0.836, respectively. The test presents a regular reliability index given by Cronbach's alpha equal to 0.83, and the mean proportion for the items equal to 0.449. The Mathematical Test is formed with independent items corresponding to different tasks with different definitions. Given the latent ability θ_i, it is considered that the correct responses to the items are independent. Furthermore, the autocorrelations within individual responses seem to be low, which provides additional support for the assumption of local independence.

As it has been mentioned, proper priors for a_j and b_j guarantee that the complete posterior for the model is proper. Further, informative prior distributions placed on a_j and b_j can be used to reflect the prior belief that the values of the item parameters are not extreme (in the boundary of the parametric space). In the common situations where little prior information is available on the difficulty parameters, we can choose S_b^2 to be large. This choice will have a modest effect on the posterior distribution for non-extreme data, and it will result in a proper posterior distribution in the case of extreme data. Extreme data occurs when students get correct (or incorrect) answers for all items. However, Sahu (2002) states that larger values of the variance led to unstable estimates. We consider here the same priors given in Sahu (2002) and Bazán, Branco and Bolfarine (2006), that is $a_j \sim N(1; 0.5)I(0; \cdot)$ and $b_j \sim N(0; 2)$, $j = 1, \ldots, k$. For the skew models, a uniform prior distribution on $(-1, 1)$ is specified for each d_j, $j = 1, \ldots, k$. Finally, we consider $\theta_i \sim \mathrm{SN}(\mu, \sigma^2, \gamma)$, $i = 1, \ldots, n$, where it is assumed that $w \sim U(-1, 1)$, $1/\sigma^2 \sim \mathrm{Gamma}(0.01, 0.01)$ and $\mu \sim N(0, 1)$.

Table 1 *Comparing models using different criteria for the Math data*

Criterion	PN	SPN	PSN	SPSN
Number of parameters	1010	1028	1013	1049
Deviance of the posterior means	16,865	15,139	16,861	15,168
Posterior expected deviance	16,012	14,096	15,999	15,994
ρ_D effective number of parameters	853	1042	862	−826
DIC	17,718	16,181	17,723	14,341
Expected AIC	18,885	17,195	18,887	17,266
Expected BIC	26,734	25,184	26,760	25,418
SSR posterior mean	17,570	12,800	17,550	12,470

The model considering asymmetry for the ICC and the abilities, denoted by SPSN, is the more general one. The Math data set involves 54 item parameters and 974 individual traits for the individuals in the sample. Additionally, in the SPSN model the hyperparameters (μ, σ^2, w) were estimated from the data set. The others models considered are particular cases and have fewer parameters.

Table 1 on page 13 shows the number of parameters in each model.

The MCMC procedure is somewhat slow because of the great number of chains that must be generated. For example, the PSN model takes about 150 seconds to run 1000 iterations on a Intel Core2 Duo E8400 Processor 3.003 GHz with 3.2 GB RAM. For the SPSN model it takes twice this time, under the SPN model it takes about 1.5 times and under the PN model it takes about three times the PSN times. The time to run the Markov Chains for each model is related to the presence of dependence structures on the latent variables (Chen, Shao and Ibrahim, 2000), with the sample size (Sahu, 2002) and also with the prior specification.

We consider 1 and 0 as initial values for the item parameters a_j and b_j, $j = 1, \ldots, k$, respectively. We propose zero as initial values for the skewness parameters w and d_j's because it corresponds to the mean of the uniform distribution on $(-1, 1)$. Initial values for the latent variable θ_i and auxiliary latent variables corresponding to the different models (as V_{ij} and W_{ij}) can be considered as generated from distributions specified in Section 3.1 but we prefer fixing this value in 0.5 to improve the performance and stability of the developed software.

When using MCMC, the sampled values for initial iterations of the chains are discarded because of their dependence on starting states. Also, the presence of autocorrelations between values of the chain is expected when latent variables are introduced (Chen, Shao and Ibrahim, 2000). Therefore, thinning values up to 100 are recommended. For example, Jackman (1992) consider for the PN model, running half million iterations and retaining only every thousandth iteration so as to produce an approximately independent sequence of sampled values from the joint posterior density.

Chen, Shao and Ibrahim (2000) reported the slow convergence of the Albert–Chib algorithm. In the mathematics data set, the convergence for the shape pa-

rameter was found to be slow. Some algorithms to improve the convergence of the Gibbs sampler in the second data augmentation approach are proposed in the literature (see Holmes and Held, 2006). A simpler alternate considered here was to generate a great number of iterations and use large thinning values. Here we considered a total of 204,000 iterations. Starting with a burn-in of 4000 iterations and them using thinning $= 100$, a sample of size 2000 is obtained. Several criteria computed using the CODA package in the WinBUGS program were used for the convergence analysis. Results are showed in Table 1.

From Table 1 on page 13, we see that all models in the skew-normal IRT family improve the corresponding symmetric probit model. Moreover, the SPSN and SPN models present the best fit for the data set by considering several criteria. Spiegelhalter et al. (2002) mention the possibility of negative values of the effective number of parameters, as we can see in Table 1 for the SPSN model. Some possible explanations for the negative values are that, the posterior distribution is extremely asymmetric or symmetric and bimodal, where the posterior mean is a poor summary statistics. Negative ρ_D can also indicate conflicting information between prior and data. Informative prior elicitation using historical data and model sensitivity to different prior choices will be explored in subsequent studies.

One of the advantages of the proposed model is to be able to extract more information from the items, by considering a new item parameter, and more information from abilities, by considering the hyperparameter in the new specification of the prior distribution for the latent variables. Figure 3 illustrates the behavior of the item parameters estimates and Table 2 on page 15 the behavior of hyperparameters of abilities estimates for the SPSN model.

Given the high skewness observed in the posterior distribution for the shape parameters d_j, we prefer median to mean as summary measures. Moreover, a 95% credible interval for the shape parameter γ results in negative endpoints and, hence, not containing zero, clearly indicating γ to be different from zero, so that the PN model is not adequate for fitting this data set. Thus, while perhaps asymmetry in ICC is not justified and Normal ICC can be sufficient, asymmetry in abilities was found.

A high and positive value for the penalty parameter corresponding to item 14 (see Table 2 and Figure 3) is observed.

Figure 4(a) compares item 14 ICCs under the PN and SPSN models, clearly showing differences among them. Figure 4(b) shows that for low ability levels, the probability of a correct response with the PN model is greater than with the SPSN model and it is the opposite for high ability levels. The positive value of the d_{14} implies that, for low ability levels the probability of a correct response increases faster with the SPSN than with the PN model. For instance, when $-1.5 \leq \theta \leq -0.5$, the change on the probability of a correct response with the SPSN model is 0.28 and with the PN model is 0.23. The opposite is true, because with high levels of abilities the change in the probability of a correct response increases slower with the PN

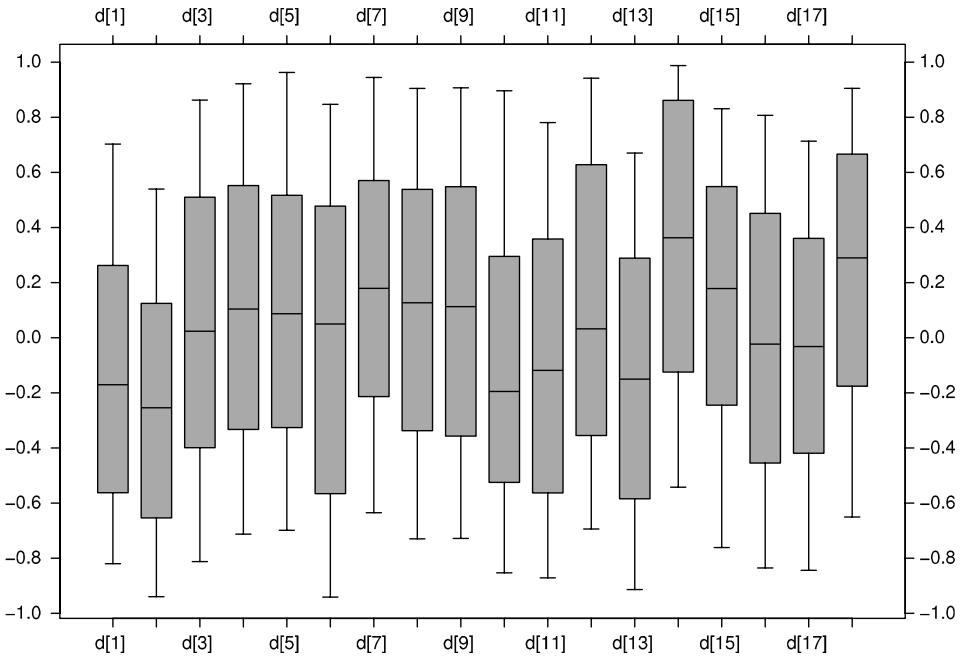

Figure 3 *Box-plots for the d parameters for the* 18 *items of Math test under the SPSN IRT model.*

Table 2 *Posterior statistics for latent hyperparameters under the SPSN IRT model for the Math data set*

Parameter	Mean	SD	Median	Percentil 2.5	Percentil 97.5
γ	−6.420	2.737	−5.886	−13.300	−2.644
μ	0.554	0.250	0.557	0.088	1.044
σ^2	0.739	0.189	0.721	0.427	1.169
$E(\theta)$	−1.139	0.3325	−1.111	−1.733	−0.5968
$V(\theta)$	0.3513	0.07661	0.3438	0.2273	0.5101

model than with the SPSN model. For instance, when $1.5 \leq \theta \leq 2.0$ the change in the probability of a correct response with the PN model is 0.053 and with the SPSN model is 0.031. Hence, we note that the PN model overestimates the probability of correct response for lower levels of mathematical ability and underestimates it for higher levels of mathematical ability. Therefore, considering the information provided by the SPSN model, item 14 rewards students with lower levels of mathematical ability and penalizes students with higher levels of mathematical ability.

By considering the estimates of the hyperparameter to the Mathematical ability in Table 2 on page 15, we found that the posterior mean and variance for the mean mathematical ability are −1.144 and 0.354, respectively, substantially differing

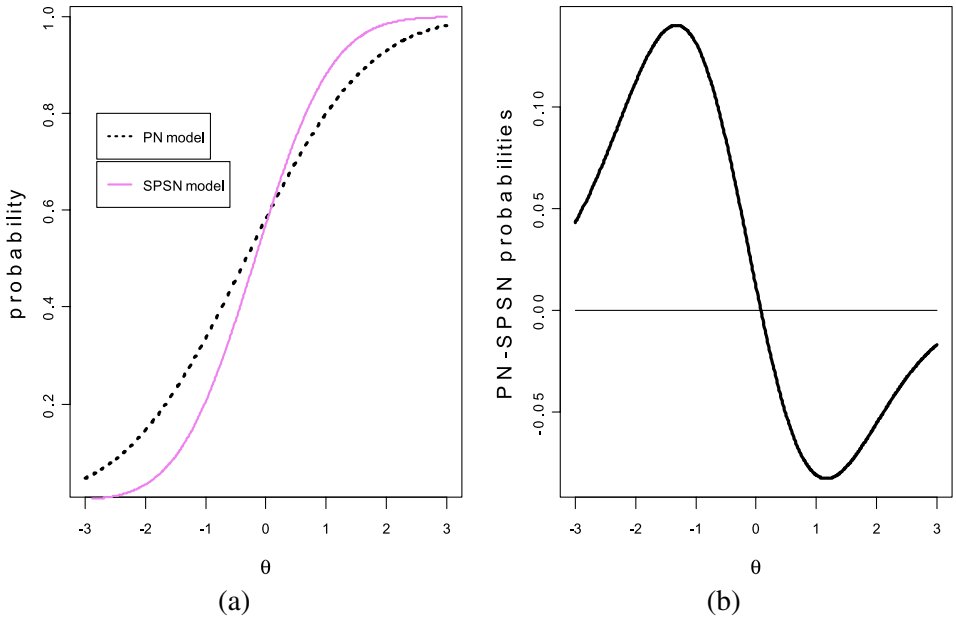

Figure 4 (a) *ICCs for item* 14 *for the Math data set, under PN and SPSN-IRT models*, (b) *Differences between the PN and SPSN probabilities estimative.*

from the values when considering the PN model which are assumed as 0 and 1, respectively. The values found are more in accordance with the distribution for the scores that presents a proportion mean of 0.4595 (-0.0405 with respect to an ideal proportion mean of 0.5) and a standard deviation of 0.2335.

5.2 Reading comprehension data set

We consider an analysis on the response pattern obtained by the application of a Reading Comprehension Test in a group of seventh grade students from Peruvian elementary schools. Item response data are available from authors upon request and correspond to the response of the 297 students to 14 items qualified as binary response (correct and incorrect). The mean score is 10.28, the median 11 and the standard deviation 2.41. The sample skewness and kurtosis indexes are 0.141 and 0.537, respectively. The test presents a median reliability index given by Cronbach's alpha equal to 0.65, and the mean proportion for the items equal to 0.734.

The original Reading Comprehension Test was to read four comprehension passages by 1535 students (see Chincaro, 2010). In our analysis, only items from the three first passages and students from the cities in Peruvian jungle region are considered. Items in testlet corresponds to different task with different definitions. The first testlet have 3 items, the second testlet have 6 items and, finally, the third testlet have 5 items.

Table 3 *Comparing testlet models using different criteria for the Reading Comprehension data*

Criteria	PN	SPN	PN testlet	SPN testlet
Number of parameters	325	339	1513	1527
Deviance of the posterior means	3434.44	3166.29	3335.6	3138.88
Posterior expected deviance	3236.27	2825.55	3064.98	2852.31
ρ_D effective number of parameters	198.17	340.739	270.63	286.57
DIC	3632.61	3507.03	3606.23	3425.45
Expected AIC	4084.44	3844.29	6361.6	6192.88
Expected BIC	6142.60	5991.11	15,943.11	15,863.05
SSR posterior mean	4172	2992	4167	3044

The priors considered are $u_i \sim N(0,1)$, $\gamma_{il} \sim N(0, \sigma^2_{\gamma_l})$, $a_j \sim N(1; 0.5)I(0;\cdot)$, $b_j \sim N(0; 2)$, $d_j \sim U[-1, 1]$, $1/\sigma^2_{\gamma_k} \sim \chi(0.5)$.

In the SNO testlet model, 42 item parameters, 297 individual traits, 891 parameters associated with testlets and 3 hyperparameters corresponding to $\sigma^2_{\gamma_1}, \sigma^2_{\gamma_2}$ and $\sigma^2_{\gamma_3}$ were estimated from the data set. The WinBUGS code is in Appendix A.2. Table 3 on page 17 shows the number of parameters in each model.

We consider 54,000 iterations, starting with a burn-in of the 4000 iterations and them using thinning equal to 25. The MCMC final sample size is 2000. Several criteria, computed using the CODA package in the WinBUGS program, were used for the convergence analysis.

We consider initial values 1, 0 and 0 for the item parameters a_j, b_j and d_j, respectively. Initial values for the latent variable θ_i and auxiliary latent variables corresponding to the different models could be by generated from the distributions specified in Section 4, however, we prefer to fix this value in 0.5 to improve the performance and stability of the developed software. From Table 3 on page 17, all criteria seem to indicate that SPN fits better than the PN IRT model and hence asymmetrical ICCs are justified. However, we found that only item 9 has a slight asymmetry with the following posterior estimates for the shape parameter: $Median = 0.2975$, $Percentile\,5 = -0.1293$ and $Percentile\,95 = 0.5662$.

For the testlet models, considering DIC and SSR, we notice that PN and SPN models present improvement. The EAIC and EBIC criteria are discarded because the number of parameters in the models with and without testlet are not comparable and inadequately penalize the deviance of the posterior mean. The SPN testlet model presents better fit for the data set than PN testlet model, by considering the four criteria. However, we found that only item 9 has a significant asymmetry with the following posterior estimates for the shape parameter: $Median = 0.3016$, $Percentile\,5 = 0.03991$ and $Percentile\,95 = 0.5428$ (see Figure 5).

Finally, considering the posterior mean, the estimates for the variance of each testlet effect are 0.2992, 0.3184 and 0.3495. These results indicate a modest effect of testlet dependence for this particular data set.

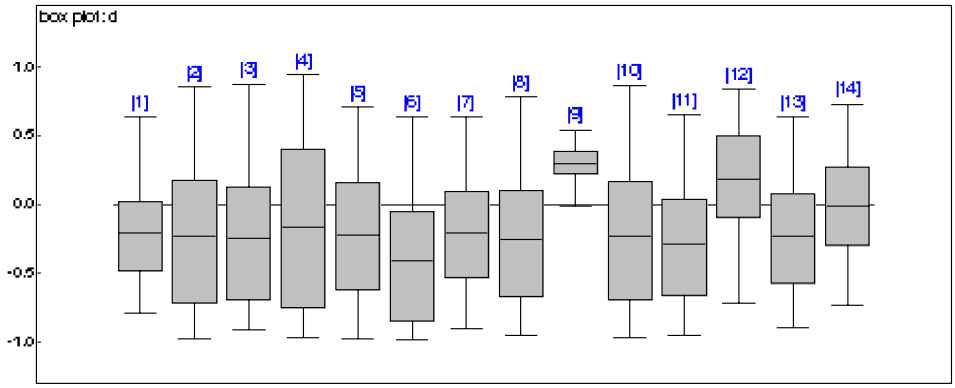

Figure 5 *Box-plots for the d parameters of the* 14 *items of Reading Comprehension test under the skew-probit testlet IRT model.*

6 Concluding remarks

This article introduces new applications of the skew-normal ogive IRT model proposed by Bazán, Branco and Bolfarine (2006), which extends the work of Albert (1992) for asymmetrical IRT models. Two extensions are considered for this model: the standard skew-normal distribution as prior distribution for the latent variables and the inclusion of an additional random effect for the dependence between items within the same testlet. The full Bayesian specification by considering the hierarchical structure can be easily implemented using MCMC methodology in WinBUGS or SAS.

In addition, several model comparisons procedures are used to compare the symmetrical and asymmetrical IRT models (DIC, EAIC and EBIC). We also introduce latent residuals for the models and global discrepancy measures as the posterior sum of squares of the latent residuals. All these measurements show that the SN-IRT model obtained considering combinations of both types of asymmetry presents better fit than the usual ogive normal IRT model for the observed data.

For the Math data set considered, there is clear indication that the shape parameter for the ability distributions is different from zero, indicating the usefulness of the SN-IRT model in explaining asymmetric abilities. Extensions to more general models such as SN-IRT multidimensional model, hierarchical SN-IRT model, SN-IRT multilevel models will be the subject of future work. Other extensions of the skew probit link for ordinal responses, as in Johnson (2003), will also be studied in future developments.

For the Reading Comprehension data set, we showed that the version testlet of the skew-normal model, which combine both the estimation of the penalty parameter as well as the random effects associated with the testlet, improves the model fit.

It may be interesting to study versions of skew-normal ogive models which consider Rasch type models or guessing parameter with and without testlet effect. Also extensions for Polychotomous and multidimensional Item Response Models can be studied.

Appendix

A.1 The skew-normal distribution (Azzalini, 1985)

A random variable X follows a skew-normal distribution with location parameter μ, scale parameter σ^2 and shape parameter λ, which controls skewness, if the density function of $X$is given by

$$f_\lambda(x) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\left(\frac{x-\mu}{\sigma}\right)\right).$$

$\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the density and distribution function of the standard normal distribution. We use the notation $X \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$. The parameter λ is also called skewness parameter, the asymmetry is positive when $\lambda > 0$ and negative when $\lambda < 0$. If $\lambda = 0$, then the skewness vanishes and the density above reduces to the density of the $N(\mu, \sigma^2)$.

An alternative parametrization of the skewness parameter is given by

$$d = \frac{\lambda}{(1+\lambda^2)^{1/2}}, \tag{A.1}$$

where d is in $[-1, 1]$.

The mean and variance are given, respectively, by $E(X) = \mu + \sqrt{\frac{2}{\pi}}\sigma d$ and $V(X) = (1 - \frac{2}{\pi}d^2)\sigma^2$. The special case with $\mu = 0$ and $\sigma^2 = 1$ is called standard skew-normal distribution. Moreover, the random variable $Z = (X - \mu)/\sigma$ is distributed according to the standard skew-normal distribution with density function given by

$$f_\lambda(z) = 2\phi(z)\Phi(\lambda z),$$

and cumulative distribution function (c.d.f.) give by

$$F_\lambda(z) = \int_{-\infty}^{z} 2\phi(t)\Phi(\lambda t)\,dt = 2\Phi_2\big((z,0)^T; 0, \Omega\big).$$

Straightforward algebraic manipulations yield the expression on the right (see Bazán, Branco and Bolfarine, 2006) with $\Phi_2(\cdot)$ denoting the distribution function of the bivariate standard normal distribution with mean vector $\mathbf{0} = (0,0)^T$ and correlation matrix $\Omega = \left(\begin{smallmatrix} 1 & -d \\ -d & 1 \end{smallmatrix}\right)$ where d is given by (A.1).

Considering the stochastic representation (see Henze, 1986), the conditional distribution $Z|V = v$ is normal with mean dv and variance $1 - d^2$, that is, $Z|V = v \sim N(dv, 1 - d^2)$. In addition, if $Z \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$ then $Zs = aZ + b \sim \mathrm{SN}(a\mu + b, a^2\sigma^2, \mathrm{sign}(a)\lambda)$.

A.2 Code in WinBUGS for testlet model

```
 model{
    for (i in 1:n) {    for (j in 1:I) {
   m[i,j] <- alpha[j]*(theta[i] - beta[j]+gamma[i,id[j]])
# Probit
#       Zs[i,j] ~ dnorm(m[i,j],1)I(lo[y[i,j]+1],up[y[i,j]+1])
#       resid[i,j]<-Zs[i,j]-m[i,j]
#SProbit
      muz[i,j]<-m[i,j]-d[j]*V[i,j]
      Zs[i,j] ~ dnorm(muz[i,j],preczs[j])I(lo[y[i,j]+1],up[y[i,j]+1])
       V[i,j] ~ dnorm(0,1)I(0,)
     resid[i,j]<-Zs[i,j]-muz[i,j]
        res2[i,j]<-pow(resid[i,j],2)
                    }
                    }
#prioris latent variable
     for(i in 1:n){ u[i]~ dnorm(0,1)
                    for(k in 1:t){
                       gamma[i,k] ~ dnorm(0,pgamma[k])   }
                      gamma[i,t+1]<-0
                              }

for(k in 1:t){
   pgamma[k]  ~ dchisqr(0.5)
  sigma2gamma[k] <-1/pgamma[k]
                 }
    #priors for item parameters
     for (j in 1:I) {
    #priors due Sahu (2002)
          alpha[j] ~ dnorm(1,2)I(0,)
          beta[j] ~ dnorm(0,0.5)
          d[j]  ~ dunif(-1,1)
    preczs[j]<- 1/(1-pow(d[j],2))
    lambda[j]<-d[j]*sqrt(preczs[j])
           }
lo[1]<- -50; lo[2]<- 0;
## i.e.,Zs*|y=0~N(-delta*V+m,1-delta^2)I(-50,0)
up[1]<-  0;  up[2]<-50;
## i.e., Zs*|y=1~N(-delta*V+m,1-delta^2)I(0,50)

        for(j in 1:I) {     resmean[j]<-sum(res2[,j])}
           sse<-sum(resmean[])
           mu<-mean(u[])
           du<-sd(u[])
        }
```

Acknowledgments

Jorge Luis Bazán was supported by DGI PUCP 2011-0173 and FAPESP 2009/50105-8, for which we are duly grateful. This work was completed while the first author was a visiting research fellow at the University of Sao Paulo under supervision of Márcia Branco.

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 Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika* **82**, 747–759. MR1380812

Albert, J. H. and Ghosh, M. (2000). Item response modeling. In *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. K. Ghosh and B. F. Mallick, eds.) 173–193. New York: Marcel Dekker. MR1893789

Arellano-Valle, R. B., Branco, M. D. and Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *The Canadian Journal of Statistics* **34**, 581–601. MR2347047

Azzalini A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal Statistical* **12**, 171–178. MR0808153

Azevedo, C. L. N., Bolfarine, H. and Andrade, D. F. (2010). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics and Data Analysis* **55**, 353–365. MR2736560

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory—Parameter Estimation Techniques*, 2nd ed. New York: Marcel Dekker. MR2086862

Bazán, J. L., Branco, D. M. and Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis* **1**, 861–892. MR2282209

Bazán, J. L., Bolfarine, H. and Branco, D. M. (2004). A new family of asymmetric models for item response theory: A Skew-Normal IRT Family. Technical report RT-MAE-2004-17, Dept. Statistics, Univ. São Paulo.

Bazán, J. L., Bolfarine, H. and Branco, D. M. (2010). A framework for skew-probit links in Binary regression. *Communications in Statistics—Theory and Methods* **39**, 678–697. MR2745312

Bolfarine, H. and Bazán, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational Behavioral Statistics* **35**, 693–713.

Bradlow, E., Wainer, H. and X. Wang (1999). A Bayesian random effect model for testlets. *Psychometrika* **64**, 153–168.

Brooks, S. P. (2002). Discussion on "Bayesian measures of model complexity and fit" by Spiegelhalter, Best, Carlin and van der Linde (2002). *Journal of the Royal Statistical Society, Ser. B* **64**, 616–618.

Camilli G. (1994). Origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational and Behavioral Statistics* **19**, 293–295.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Boca Raton, FL: Chapman & Hall.

Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag. MR1742311

Chincaro, O. (2010). Dichotomous Rasch model with application to education. M.Sc. thesis, Pontificia Universidad Católica del Perú (in Spanish).

Duncan, K. and MacEachern, S. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modeling* **8**, 41–66. MR2750630

Ghosh, M., Ghosh, A., Chen, M.-H. and Agresti, A. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference* **88**, 99–115. MR1767562

Hashimoto, Y. (2002). Motivation and willingness to communicate as predictors of reported l2 use: The Japanese ESL context. *Second Language Studies* **20**, 29–70.

Holland, P. and Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics* **14**, 1523–1543. MR0868316

Henze, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scandinavian Journal Statistical* **13**, 271–275. MR0886466

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168. MR2227368

Jackman, S. (2004). Bayesian analysis for political research. *Annual Review of Political Science* **7**, 483–505.

Johnson, T. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika* **68**, 563–583. MR2272428

Junker, B. W. and Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics* **25**, 1327–1343. MR1447754

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* **105**, 156–166.

Patz, R. J. and Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics* **24**, 146–178.

Poleto, F. Z., Paulino, C. D., Molenberghs, G. and Singer, J. M. (2011). Inferential implications of over-parametrization: A case study in incomplete categorical data. *International Statistical Review* **79**, 92–113.

Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology* **51**, 754–760.

Riddle, A. S., Blais, M. R. and Hess, U. (2002). A multi-group investigation of the CES-D's measurement structure across adolescents, young adults and middle-aged adults. Scientific Series 2002s-36, Centre Interuniversitaire de recherche eit analysis des organizations, Montreal.

Rupp, A., Dey, D. K. and Zumbo, B. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to item response modeling. *Structural Equations Modeling* **11**, 424–451. MR2061898

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation* **72**, 217–232. MR1909259

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika* **62**, 471–493.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristics curves. *Psychometrika* **65**, 319–335.

SAS Institute Inc. (2009). *The MCMC Procedure, SAS/STAT Help Documentation*. Cary, NC: SAS Institute Inc.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Ser. B* **64**, 583–639. MR1979380

Tuerlinckx, F. and De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods* **6**, 181–195.

Wainer H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement* **24**, 185–201.

Wainer, H. and Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement* **37**, 203–220.

Wainer, H., Bradlow, E. T. and Wang, X., eds. (2007). *Testlet Response Theory and Its Applications*. New York: Cambridge Univ. Press.

Wainer, H., Brown, L. M., Bradlow, E. T., Wang, X., Skorupski, W. P., Boulet, J. and Mislevy, R. J. (2006). An application of testlet response theory in the scoring of a complex certification examination. In *Automated Scoring of Complex Tasks in Computer Based Testing* (D. M. Williamson, R. J. Mislevy and I. I. Bejar, eds.) Chapter 6, 169–200. Hillsadle, NJ: Lawrence Erlbaum Associates.

Wang, X., Baldwin, S., Wainer, H., Bradlow, E. T., Reeve, B. B., Smith, A. W., Bellizzi, K. M. and Baumgartner, K. B. (2010). Using testlet response theory to analyze data from a survey of attitude change among breast cancer survivors. *Statistics in Medicine* **29**, 2028–2044. MR2758445

Wang, W.-C. and Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement* **29**, 126–149. MR2113223

Zaider, T. I., Heimberg, R. G., Fresco, D. M., Schneier, F. R. and Liebowitz, M. R. (2003). Evaluation of the Clinical Global Impression Scale among individuals with social anxiety disorder. *Psychological Medicine* **33**, 611–622.

J. L. Bazán
Departamento de Ciencias
Pontificia Universidad Católica del Perú
Av. Universitaria 1801
San Miguel, Lima
Lima 32
Peru
URL: http://argos.pucp.edu.pe/~jlbazan/principal.html
E-mail: jlbazan@pucp.edu.pe

M. D. Branco
H. Bolfarine
Departamento de Estatística
Universidade de São Paulo
Rua do Matão, 1010 - Cidade Universitária - São Paulo
SP - Brasil CEP 05508-090
Brasil
E-mail: mbranco@ime.usp.br
hbolfar@ime.usp.br