UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

A Bayesian Skew Mixture Item Response Model

Juliane Venturelli

Orientador: Flávio Bambirra Gonçalves

Co-Orientadora: Rosangela Helena Loschi

Belo Horizonte, April 5, 2015

Abstract

Under the Item Response Theory, the two most common link functions used to model dichotomous data are the symmetric probit and logit. However, some authors have emphasized that these symmetric links do not always provide the best fit for some data sets. To overcome this issue, asymmetric links have been proposed. This work aims at introducing a flexible Item Response Model able to accommodate both symmetric and asymmetric link. The c.d.f. of a centered skew normal distribution is assumed as the link function and, additionally, we consider a finite mixture of Beta distributions and a point mass distribution at zero to describe the uncertainty about the skewness parameter, so not all items need to be assumed asymmetric *a priori*. Therefore, the proposed model embraces symmetric and asymmetric normal models in one also performing an intrinsic model selection. We offer the full condition distribution of ability, discrimination and difficulty parameters. We also introduce efficient algorithms to sample from the posterior distributions.

vaidade, vaidade, tudo é vaidade

À nova geração: Aninha, Arthuzinho, Benjamin, Bernardo e Mateus,

(que não sejam peixes de aquário!)

e claro, à Biga, Frida, Emmy e Onix

(que nem sabem sobre aquários!)

Agradecimentos

Você sabe por que como raízes? Porque raízes são importantes [1].

Sou muito grata pela constante ajuda, suporte e orientação do Professor Flávio Bambirra, um dos professores de conhecimento mais versátil que conheci na Estatística, de ideias geniosas e ousadas. Tenho certeza de que em poucos anos seu nome estará no hall dos estatísticos de destaque. Sou muito grata à Professora Rosangela Loschi, que aceitou o desafio de me co-orientar em TRI. Muito obrigada por deixar meu trabalho mais colorido!

Sou imensamente grata ao papy e à mamy, que não mediram esforços para investir em minha educação acadêmica. Especialmente, agradeço meu pai por sempre se dispor a me ajudar com o meu "software". Mais ainda, agradeço por serem exemplos de esforço e dedicação. Definitivamente não sei como meu pai conseguiu fazer Engenharia e minha mãe Medicina, tendo que trabalhar e ainda com três filhas...!

Minha sincera gratidão ao Bráulio pela indescritível disposição em me ajudar nos problemas mais diversos. Agradeço à Larissa, aluna destaque do mestrado, Lívia, sempre neutra, e Rachel, com quem não tem como discutir, pela presença no café, no bandejão, nos choros, e por aguentarem minhas rabugentices. Agradeço à Gabi por sempre oferecer ajuda, e à Raquel Borges pelas ajudas com o *ggplot*. Agradeço à minha querida amiga Marina Flor Muniz, por quem eu tenho enorme admiração não só pela excelência acadêmica, mas por conseguir viver fora do aquário. Sou muito grata pela minha amizade com Vivi Paixão, quem conheci na luta, e que viveu comigo os mesmos dramas e desesperos de uma aluna de mestrado. Agradeço à Bruna, pelo constante apoio, carinho e compreensão, à Esther pelas "caronas"; ao Jaider -que não ajudou em nada!-, e à Tamires pelos ensaios sobre as ilusões da Ciência. Agradeço ao Gustavo Souza por ter parado suas balbúrdias quânticas para (tentar) me ensinar Análise.

Agradeço à Lay pelas muitas amostras grátis e muitos acolhimentos. Agradeço à Letícia Malta, talvez a pessoa de maior inteligência interpessoal que já conheci, a quem sou eternamente grata por me ajudar a ser mais leve, centrada e forte. Ao Bup, que nesses dias de conclusão de trabalho me colocou para cima.

[1] A Grande Beleza

Agradeço à Bruna e Carol, simplesmente por estarem sempre lá, ainda que longe.

Agradeço à Vó Nice, à Vó Irene e aos avôs Olívio e Abelardo (ainda que não os tenha conhecido pessoalmente) por serem boas raízes.

Acknowledgement

The authors would like to thank all of the people who contributed in some way to the work described in this dissertation.

This work was accomplished with the help and support of fellow lab-mates and collaborators. We would like to especially thank Bráulio Veloso and Professor Marcos Prates for being very helpful in finding a clever solution to a numerical problem, which had cost us considerable time and headache. We would like to thank Douglas Mesquita and Luis Gustavo Silva for their constant contributions in troubleshooting in R and Latex. We also would like thank the contributions of Professor Dani Gamerman, and James Olson for the English support.

This work could not have been accomplished without the financial support of CAPES.

Contents

1 Skew Distributions with Normal Kernels **12**

1.1 Skew Normal Distribution 12

1.2 Centered Skew Normal Distribution 16

1.3 A General Class of Multivariate skew Normal Distributions 20

2 Binary Item Response Models **22**

2.1 Symmetric IRM 23

2.1.1 The Rasch Model 23

2.1.2 Two-Parameter Models 25

2.1.3 Three-Parameter Models 26

2.2 Skew IRM 28

2.2.1 Skew Logit Model 28

2.2.2 BBB Skew Probit Model 30

2.2.3 Skew Normal model IRT under the Centered Parametrization 31

3 A Flexible Class of Centered Skew Probit IRT Model **33**

3.1 Proposed Model 34

3.2 Prior Specifications 35

3.3 Posterior Distribution 37

3.4 MCMC 37

3.4.1 Full Conditional Distribution for X 39

3.4.2 Full Conditional Distribution for $\boldsymbol{\theta}$ 39

3.4.3 Full Conditional Distribution for (a, b) . . . 41
3.4.4 Full Conditional Distribution for p . . . 44
3.5 Full Conditional Distribution for Z . . . 45
3.6 Full Conditional Distribution for $\boldsymbol{\gamma}_-$ and $\boldsymbol{\gamma}_+$ . . . 46
3.7 Computational Aspects . . . 47
3.7.1 Sampling of X . . . 47
3.7.2 Sampling of $\boldsymbol{\theta}$ and (a, b) . . . 48
3.7.3 Sampling of Z . . . 49

4 Simulations **51**

5 Final Remarks **56**
5.1 Full Conditional Distribution for θ . . . 57
5.2 Full Conditional Distribution for (a, b) . . . 59
5.3 Generalized Beta Distribution . . . 64

Introduction

Item Response Theory (IRT) is a psychometric theory commonly used in educational assessments and cognitive psychology. It aims to model variables which are by nature unobservable, such as abilities, intelligence and depression. As these variables are constructs rather than physical attributes, they are often called latent traits. Although these traits are not directly measured, an individual's responses to a test provide information from which traits can be inferred. In this work, the terms "ability", "latent traits", or simply "traits", are going to be used indiscriminately with the same meaning. As our focus is on the use of IRT in the education field, the main interest is to model the relationship between responses and abilities of J individuals submitted to a test comprised by I items.

Under the latent theory, it is customary to assume that only one trait influences, or determines, a person's performance when taking a test. Although this assumption is not reasonable in practice, it is enough to assume that a dominant factor influences on a test performance. This dominant factor is referred to as the ability measured by the item. Models that consider only one ability are called unidimensional models. However, more complex models are available to take into account a multidimensional vector of abilities (Van Der Linden & Hambleton, 1997).

In IRT, models are built taking into account that the relationship between an examinee's[2] item performance and the dominant trait is described by a monotonically increasing function denominated item characteristic curves (ICC). The choice of the ICC and its parameters is a crucial part of IRT. The well known item response models (IRM) for dichotomous responses, for instance, adjust response data for item characteristics such as difficulty, discriminating power and liability to guessing. Some most well known IRM's are presented and discussed in Chapter 2.

Early work in the IRT field was first addressed in the works of Richardson (1936), Lawley (1943) and Tucker (1946). However, most contemporary concepts of IRT were formulated around the 1950s and 1960s, mainly by Lord (1952), Rasch (1960), Birnbaum (1957, 1958), Lazarsfeld & Henry (1968) and Wright & Panchapakesan (1969). Also, a major contribution

[2]examinee is a term commonly used in the literature to refer to the person subjected to a test

to IRT was given by Samejima (1969, 1972) who introduced a new IRM able to handle both polychotomous and continuous response data. Samejima (1972) also extended unidimensional models to the multidimensional ones.

A startup and yet revealing work on IRT can be found in the book by Baker (2001) entitled *The Basics of Item Response Theory*. In Portuguese, the reader can resort to de Andrade et al. (2000) for a brief and comprehensive work on basic concepts. Some more detailed work can be found in Baker (1977), who provided a comprehensive review of parameter estimation methods. Van Der Linden & Hambleton (1997) offered a book with a collecting of complex IRM, such as models for multiple abilities or cognitive components and nonmonotone items, and nonparametric models. Also, some important estimation methods (EM and Bayesian approach) for dichotomous and polichotomous items were discussed by Azevedo (2003), and for a more complete work on the Bayesian approach to modern test theory, the reader can resort to the book by Fox (2010).

A common assumption in modelling academic ability and other latent traits associated with human behaviour is to assume that these traits follow a standard normal distribution. Such an assumption asserts that one believes the abilities population has a normal shape, and the J students taking the test are a random sample of this population. However, this assumption is not always observed in psychometric data, as noticed by Micceri (1989). Samejima (1997) also questioned the indiscriminate use of normality assumption without even checking its adequacy, and Azevedo et al. (2012) showed data sets where normality and symmetry did not entirely hold. To overcome this issue and to obtain better estimations for traits, some flexible approaches have been considered. Mislevy (1984), for instance, considered a mixture of normal distributions and a nonparametric estimation based on empirical histograms. Bazan (2005) proposed a flexible item response model by using the skew normal distribution (SN) (Azzalini, 1985) to model the behaviour of the ability parameter. However, according to Azevedo et al. (2011), Bazan (2005) did not consider the estimation of the skewness parameter concomitantly with the model estimation, neither addressed issues concerning to the model identifiability. To overcome the identifiability problem, Azevedo et al. (2011) considered the Centered Skew Normal distribution (CSN) introduced by Azzalini (1985). The centered parametrization (CP) approach brings some advantages to the model inference and the estimation of item's parameters. A brief presentation of the SN and CSN families of distributions is given in Chapter 1.

Another common feature of IRT is the usage of symmetric probit and logit link functions for ICC. However, it has been emphasized by some authors that these symmetric ICCs are not always appropriate for describing the relationship between examinee's ability and the probability of success (correct answer). According to Chen et al. (2000), the commonly used links for binary response data do not always fit the dataset properly, which can lead to biased estimation of the model parameters. Many authors have proposed possible solutions to overcome this issue. For instance, Samejima (2000) proposed a family of models, the so called Logistic Positive Exponent Family, which provides asymmetric ICCs and has the logistic model as a special case. In this class of models, the ICC is given by $L(\cdot)^{\epsilon_i}$, where $L(\cdot)$ is the cumulative distribution function (c.d.f.) of the standard logistic distribution, and $\epsilon_i > 0$ is the skewness parameter associated with the i^{th} item. The logistic link is recovered when $\epsilon_i = 1$. Santos (2009) presented an extension of the proposed model by Samejima (2000), allowing the guessing parameter to be different from zero, and using a prior specification to detect asymmetric items. For the skewness parameter ϵ, Santos (2009) considered a finite mixture type distribution with a point mass at one. Additionally, another asymmetric link function was obtained by Bazán et al. (2006), who considered the Skew Normal (SN) distribution (Azzalini, 1985) to model the ICC, thus, obtaining an asymmetric IRM that admits asymmetric items. Since the SN distribution has the normal one as a special case, the probit IRT model is included in the skew probit class of IRMs. These models are briefly reviewed in Section 2.2.

In this work, we focus on dichotomous unidimensional IRMs and inference is made under the Bayesian approach. Our main interest is to build a flexible IRT model able to accommodate both symmetric and asymmetric ICC. Differently from what is presented in Bazán et al. (2006), we introduce a skewed ICC based on the CSN distribution. One of the most important contributions of this work is to consider a finite mixture of Beta distributions and a point mass at zero to describe the uncertainty about the skewness parameter. Therefore, the proposed model embraces both probit and skew probit models which, under our approach, have a non null prior probability of occurrence. Consequently, such a strategy also provides an intrinsic methodology for model selection. We offer the full condition distribution of ability, discrimination and difficulty parameters. We also propose efficient algorithms, based on Markov Chain Monte Carlo (MCMC) methods, to sample from the posterior distribution.

In regard to its structure, this work is organized as follows. In Chapter 1 we review the SN and the CSN families of distribution (Azzalini, 1985) and some of their properties, since skewed distributions with normal kernels play an important role in our approach. We also review and make some adaptations for the family of distribution with Normal kernel introduced by Gonçalves & Gamerman (2015), which is based on the unified class of the SN (SUN) distribution presented by Arellano-Valle et al. (2006). In Chapter 2, we review some of the main adopted unidimensional symmetric IRM for dichotomous responses, and we also review the asymmetric IRM proposed by Bazán et al. (2006). Additionally, we briefly present the model by Azevedo et al. (2011), which is based on adopt the centered parametrization of the SN distribution. In Chapter 3, the main contributions of this work are presented. The proposed model is introduced and an algorithm to sample from the posterior distributions is developed. In Chapter 4, we present some simulation results to evaluate the efficiency of the proposed model. Finally, Chapter 5 presents the final remarks and some discussions about open problems.

Chapter 1

Skew Distributions with Normal Kernels

In recent years, the construction of new distributions that are able to accommodate skewness, multimodality and tails that are heavier than the normal distribution has received considerable attention. It is not feasible to mention all development in the area. Recent surveys of flexible distributions can be found in Azzalini (2005), Wang et al. (2004) and Arnold et al. (2002).

We are interested in unimodal skew distributions with normal kernels. Several distributions with these characteristics have been proposed since the introduction of the Skew Normal (SN) distribution in the seminal paper by Azzalini (1985). Most of such families belongs to the unified class of SN distributions introduced by Arellano-Valle & Azzalini (2006), the so-called SUN family of distributions.

In the following, we are going to briefly present the univariate Skew Normal (SN) and the Centered Skew Normal (CSN) families, both introduced by Azzalini (1985). We also review the skew distribution with normal kernel proposed by Gonçalves & Gamerman (2015). All these distributions are going to be considered in Chapter 3 to define a new skew IRT model.

1.1 Skew Normal Distribution

As defined by Azzalini (1985), a random variable X has a Skew Normal (SN) distribution, with location parameter $\xi \in \mathbb{R}$, scale parameter $\omega^2 \in \mathbb{R}^+$, and skewness parameter $\lambda \in \mathbb{R}$, which is denoted by $X \sim SN(\xi, \omega^2, \lambda)$, if its probability density function (p.d.f.) is given by

$$f_{SN}(x; \xi, \omega^2, \lambda) = 2\omega^{-1}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\left(\frac{x-\xi}{\omega}\right)\right), x \in \mathbb{R}, \tag{1.1}$$

where $\phi(.)$ and $\Phi(.)$ denote, respectively, the p.d.f. and the cumulative distribution function (c.d.f.) of the standard normal distribution. The parameter λ controls the degree of skewness of the distribution. A p.d.f. with negative asymmetry is obtained when $\lambda < 0$, and positive asymmetry when $\lambda > 0$. An important characteristic of the SN family is that the normal family is recovered when $\lambda = 0$. Moreover, it preserves some properties of the normal family, such as linearity.

If $X \sim SN(\xi, \omega^2, \lambda)$, its mean and variance are given, respectively, by

$$E(X) = \mu_X = \xi + r\delta\omega, \tag{1.2}$$

$$Var(X) = \sigma_X^2 = \omega^2\left(1 - r^2\delta^2\right), \tag{1.3}$$

where $r = \sqrt{2/\pi}$ and δ is an alternative parametrization of the skewness parameter λ given by

$$\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}, \quad \delta \in [-1, 1]. \tag{1.4}$$

The transformation $Y = (X - \xi)\omega^{-1}$ leads to the standard skew normal distribution, denoted by $Y \sim SN(0, 1, \lambda)$, which p.d.f. is

$$f_{SN}(y; 0, 1, \lambda) = 2\phi(y)\Phi(\lambda y),\ y \in \mathbb{R}. \tag{1.5}$$

Figure 1.1 shows the behaviour of λ as a function of δ. It can be noticed that δ approximate to 1(-1) as λ tends to $\infty(-\infty)$.

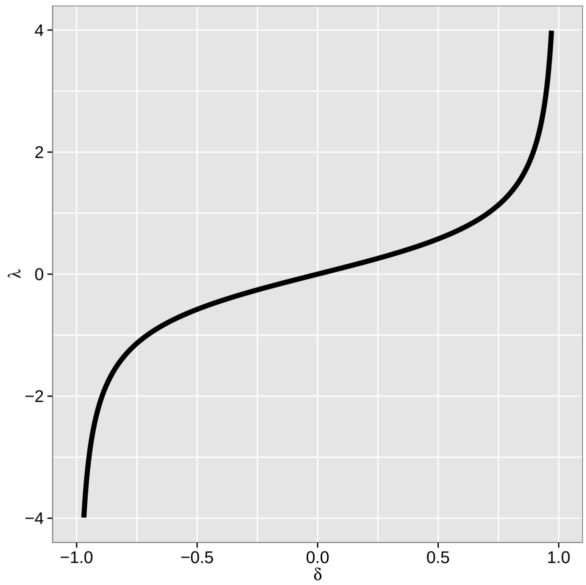

Figure 1.1: λ in function of δ.

It follows straightforward from equations (1.2) and (1.3) that the mean and variance of the standard SN distribution are, respectively,

$$\mu_Y = r\delta, \tag{1.6}$$

$$\sigma_Y^2 = 1 - r^2\delta^2 = 1 - \mu_Y^2. \tag{1.7}$$

The c.d.f. of the standard skew normal is denoted by $\Phi_{SN}(y;\lambda)$, and it is given by

$$\Phi_{SN}(y;\lambda) = \int_{-\infty}^{y} 2\phi(t)\Phi(\lambda t)dt = 2\Phi_2((y,0)^T;\mathbf{0},\mathbf{\Omega}), \tag{1.8}$$

where $\Phi_2(.;\mathbf{0},\mathbf{\Omega})$ denotes the c.d.f. of the bivariate normal distribution with mean vector $\mathbf{0} = (0 \quad 0)^T$, and covariance matrix $\mathbf{\Omega} = \begin{pmatrix} 1 & -\delta \\ -\delta & 1 \end{pmatrix}$.

Henze (1986) proposed a standard stochastic representation of a SN variable with density described by equation (1.5) given by

$$Y \stackrel{d}{=} \delta V + (1-\delta^2)^{1/2} W, \tag{1.9}$$

where $V \sim HN(0,1)$ and $W \sim N(0,1)$, for $V \perp W$, where HN denotes the half normal distribution. The representation in terms of convolution plays an important role in inference procedures that involves the SN family.

Figures 1.2*a* and 1.2*b* show the p.d.f. of the standard skew normal distribution defined in (1.5) for different values of the skewness parameter λ. The higher the value of λ, the higher the asymmetry of the p.d.f.. Also, positive values of λ induce a positive asymmetry in the p.d.f., and negative values of λ induce a negative asymmetry. It can be noticed that when $\lambda \to -\infty(+\infty)$, the asymmetric distribution tends to put most of its positive mass on negative (positive) values. In fact, the half normal distribution is a limit case of (1.5) if $\lambda \to +\infty$. Figures 1.2*c* and 1.2*d* show the c.d.f. of the standard skew normal for different values of λ.

Despite its flexibility, inference for the shape parameter based on likelihood methods in the SN family is not always possible. For instance, the likelihood can have local maximum (besides the global maximum) and the maximum likelihood estimator can be infinity (Azzalini, 1985).

Figure 1.3 shows the log-likelihood function based on an i.i.d. sample $\mathbf{y}$ of size n = 1000, of $Y \sim SN(0,1,1)$. The log-likelihood exhibits a non-quadratic shape and, as stated by Arellano-Valle & Azzalini (2008), there is a stationary point at $\lambda = 0$. According to Azzalini (1985), the stationary point occurs for any sample. To overcome such a problem, Azzalini (1985) also proposed an alternative parametrization of the skew normal family distribution. Such family is going to be briefly presented in next section.

1.2 Centered Skew Normal Distribution

In order to avoid some of the inference problems discussed in Section 1.1, such as the smooth behaviour of the likelihood function in the neighbourhood of $\lambda = 0$, Azzalini (1985) proposed a centered parametrization for the SN distribution, the so called Centered Skew Normal family (CSN). Under this parametrization the Pearson's skewness coefficient is considered instead of

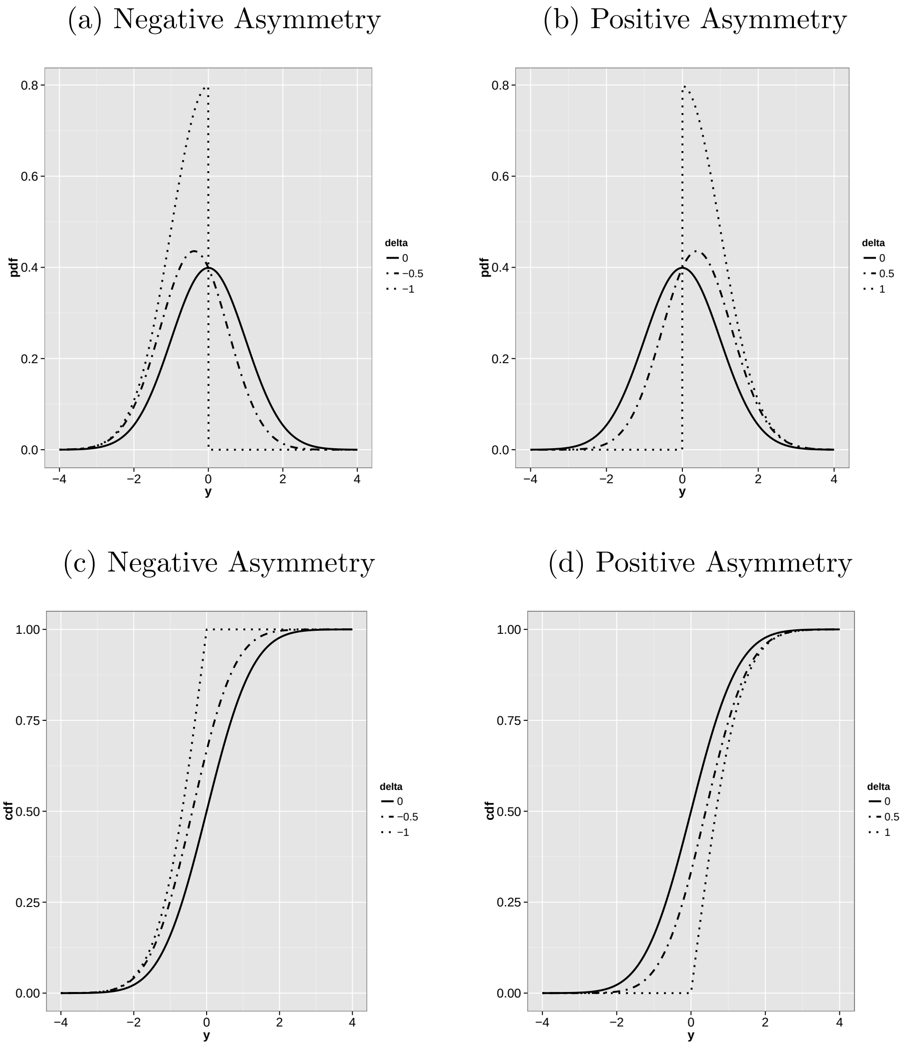

Figure 1.2: P.d.f. and c.d.f. of the SN(0,1,λ) for negative and positive skewness.

λ. Also, the location and scale parameters are the mean and variance of the distribution, respectively. Azevedo et al. (2009) presented very helpful results and properties of the CSN.

To obtain the p.d.f. of a random variable X^c with distribution in the CSN family, let us assume that $X \sim SN(0, 1, \lambda)$ and the following transformation

$$X^c = \psi \left(\frac{X - \mu_X}{\sigma_X} \right) + \varsigma, \tag{1.10}$$

where ς and ψ^2 are the mean and the variance of X^c, respectively, and μ_X and σ_X are defined

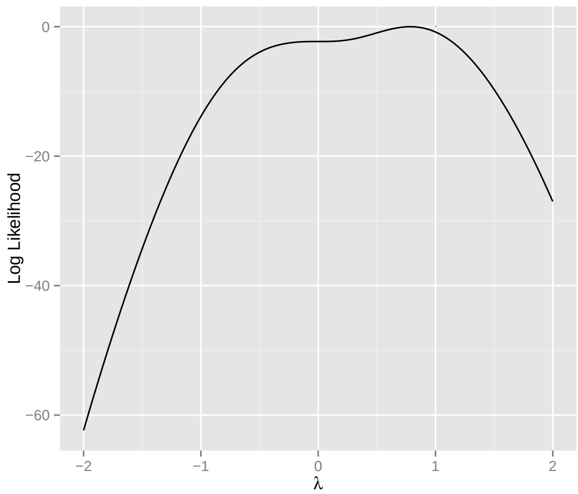

Figure 1.3: Skew Normal log Likelihood for λ.

in (1.6) and (1.7), respectively. Assuming the transformation given by (1.10) and considering the Jacobian method, it follows from equation (1.5) that the p.d.f of X^c is given by

$$f_{CSN}(x^c;\varsigma,\psi^2,\lambda) = \frac{2\sigma_X}{\psi}\phi\left(\mu_X + \sigma_X\left(\frac{x^c-\varsigma}{\psi}\right)\right)\Phi\left(\lambda\left(\mu_X + \sigma_X\left(\frac{x^c-\varsigma}{\psi}\right)\right)\right), \quad x^c \in \mathbb{R}. \tag{1.11}$$

In order to rewrite the p.d.f of X^c as a function of the Pearson's skewness coefficient of X, γ, the following results are considered.

The Pearson's skewness coefficient for a random variable X is defined by:

$$\gamma = \frac{E(X-E(X))^3}{Var(X)^{2/3}}. \tag{1.12}$$

Thus, if $X \sim SN(0,1,\lambda)$, Henze (1986) proved that the Pearson's skewness coefficient of X is

$$\gamma = r\delta^3(2r^2-1)(1-r^2\delta^2)^{-3/2}, \qquad \gamma \in (-0.99527, 0.99527), \tag{1.13}$$

where δ is defined in (1.4), $r = \sqrt{2/\pi}$ and $s = \left(\frac{2}{4-\pi}\right)^{1/3}$.

After some algebraic calculations we obtain that

$$\delta = \frac{s\gamma^{1/3}}{r\sqrt{1+s^2\gamma^{2/3}}}, \tag{1.14}$$

$$\lambda = \frac{s\gamma^{1/3}}{\sqrt{r^2+s^2\gamma^{2/3}(r^2-1)}}. \tag{1.15}$$

Consequently, it follows that the mean and variance of X can be rewritten as functions of γ and are given, respectively by

$$\mu_X = \frac{s\gamma^{1/3}}{\sqrt{1+s^2\gamma^{2/3}}}, \tag{1.16}$$

$$\sigma_X^2 = \frac{1}{1+s^2\gamma^{2/3}}. \tag{1.17}$$

Replacing (1.16) and (1.17) in (1.11) we obtain the p.d.f of $X^c \sim CSN(\varsigma, \psi^2, \gamma)$ as

$$f_{CSN}(x^c;\varsigma,\psi^2,\gamma) = \left(\frac{2}{\sqrt{\psi^2(1+s^2\gamma^{2/3})}}\right)\phi\left(\frac{x^c-(\varsigma-s\gamma^{1/3})}{\sqrt{\psi^2(1+s^2\gamma^{2/3})}}\right)\Phi\left(g(\gamma)\left(\frac{x^c-(\varsigma-s\gamma^{1/3})}{\sqrt{\psi^2(1+s^2\gamma^{2/3})}}\right)\right), \tag{1.18}$$

where $g(\gamma)$ is given by (1.15). To simplify the presentation of the p.d.f. in (1.18), define

$$\begin{aligned} \varsigma^* &= \varsigma - s\gamma^{1/3}, \\ \psi^* &= \sqrt{\psi^2(1+s^2\gamma^{2/3})}. \end{aligned}$$

Thus, (1.18) is rewritten as

$$f_{CSN}(x^c;\varsigma^*,\psi^{*2},\gamma) = \frac{2}{\psi^*}\phi\left(\frac{x^c-\varsigma^*}{\psi^*}\right)\Phi\left(g(\gamma)\left(\frac{x^c-\varsigma^*}{\psi^*}\right)\right). \tag{1.19}$$

The c.d.f. of a standard CSN variable is obtained by the integration of the density in (1.19)

$$\Phi_{CSN}(x^c;\gamma) = \int_{-\infty}^{x^c} \frac{2}{\psi^*}\phi\left(\frac{t-s\gamma^{1/3}}{\psi}\right)\Phi\left(g(\gamma)\left(\frac{t-s\gamma^{1/3}}{\psi}\right)\right)dt. \tag{1.20}$$

We can also obtain the c.d.f. of a CSN variable in terms of the standard SN c.d.f.

$$\Phi_{CSN}(x^c;\gamma) = \Phi_{SN}(x\sigma_X + \mu_X;\delta),$$

where μ_X and σ_X are defined in expressions (1.6) and (1.7), respectively. This equivalence can be very convenient, as there is a SN package in R (Azzalini, 2014).

From equations (1.10), (1.16), and (1.17), the stochastic representation for $X^c \sim CSN(\varsigma^*, \psi^{2*}, \gamma)$ becomes

$$X^c \stackrel{d}{=} (\delta V + (1-\delta^2)^{1/2} W)\psi^* + \varsigma^*, \tag{1.21}$$

where V and W are defined in (1.9), and δ is a function of γ as defined in (1.14).

Figure 1.4 shows the relationship of delta as a function of gamma. Notice that a slight change on γ around zero represents a strong change on δ.

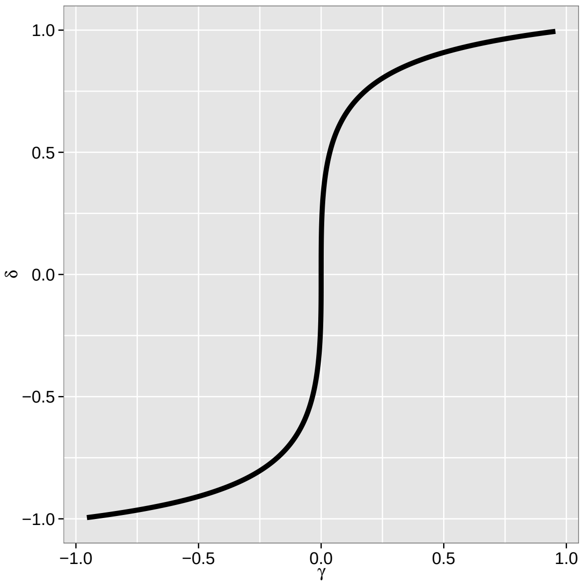

Figure 1.4: δ as a function of γ.

As stated by Azzalini (1985), Arellano-Valle & Azzalini (2008) and Azevedo et al. (2009), the p.d.f. defined in (1.19) is more appropriate for inference purposes if compared to the non centered given in (1.5). According to Azzalini & Capitanio (1999), under the centered parametrization in (1.19) the shape of the likelihood function well behave around zero, which provides better inference.

To observe this, consider a random sample $\mathbf{x} = (x_1, x_2, ..., x_n)$ of $X^c \sim CSN(0,1,\gamma)$. For such sample the log-likelihood function is

$$f(x|\varphi,\gamma) = -n\ln(\varphi) + \sum_{i=1}^{n}\left(\ln\left(\phi\left(\frac{x_i^c + s\gamma^{1/3}}{\varphi}\right)\right) + \ln\left(\Phi\left(g(\gamma)\left(\frac{x_i^c + s\gamma^{1/3}}{\varphi}\right)\right)\right)\right). \quad (1.22)$$

Figure 1.5 displays the plot based on the same sample information considered in Figure 1.3.

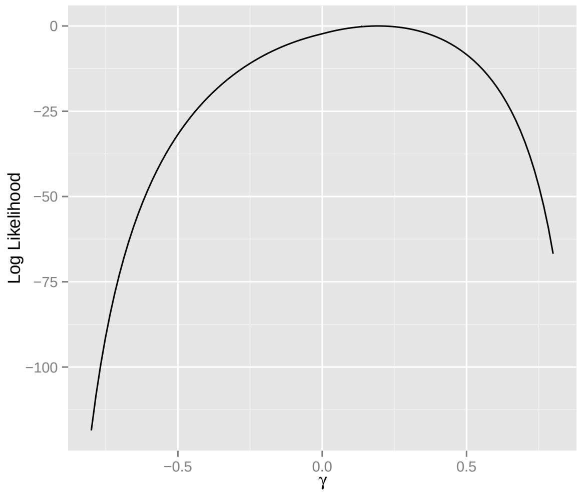

Figure 1.5: Centered Skew Normal Log Likelihood for γ.

Compared to the log-likelihood plot displayed in Figure 1.3, it is noticeable that the new plot exhibits a curve with a regular behaviour, much closer to quadratic functions and without a stationary point at $\gamma = 0$ ($\lambda = 0$).

1.3 A General Class of Multivariate skew Normal Distributions

We now define a more general distribution with Skew Normal kernel which appears in the derivation of the conditional distributions of the MCMC algorithm proposed in Chapter 3.

This distribution was introduced, in the form presented here, by Gonçalves & Gamerman (2015), and is based on the Unified Skew Normal (SUN) family of distributions, defined by

Arellano-Valle & Azzalini (2006). However, to adapt it to our model structure, we introduce a new parameter to the model proposed by Gonçalves & Gamerman (2015). As we do not use the parametrization of Arellano-Valle & Azzalini (2006), we call this new distribution Adapted SUN (ASUN).

Consider a d-dimensional column vector $\boldsymbol{\xi}$, an $m \times d$ matrix $\mathbf{W}$, an m-dimensional column vector $\boldsymbol{\eta}$, and a $d \times d$ matrix $\boldsymbol{\Sigma}$. Let us also assume column vectors $\mathbf{U}_0$ and $\mathbf{U}_1$ of dimension m and d, respectively, such that

$$\begin{pmatrix} \boldsymbol{U_0} \\ \boldsymbol{U_1} \end{pmatrix} \quad \sim \quad N_{m+d}(\mathbf{0}, \boldsymbol{\Sigma}^*), \tag{1.23}$$

where $\boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\Delta}^T \\ \boldsymbol{\Delta} & \boldsymbol{\Sigma} \end{pmatrix}$, $\boldsymbol{\Gamma} = \boldsymbol{I_m} + \boldsymbol{W\Sigma W^T}$, $\boldsymbol{I_m}$ denotes the identity matrix of order m and $\boldsymbol{\Delta}^T = \mathbf{W}\boldsymbol{\Sigma}$. We say that $(\boldsymbol{U_1}+\boldsymbol{\xi}|\boldsymbol{U_0}+\boldsymbol{\varepsilon}+\boldsymbol{\eta} > 0) \sim ASUN_{d,m}(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \mathbf{W}, \boldsymbol{\eta})$, where $\boldsymbol{\varepsilon} = \boldsymbol{\Delta^T \Sigma^{-1} \xi}$.

Lemma 1.3.1. *The density of* $(\boldsymbol{U_1} + \boldsymbol{\xi}|\boldsymbol{U_0} + \boldsymbol{\varepsilon} + \boldsymbol{\eta} > 0)$ *is given by*

$$f(z) = \frac{1}{\Phi_m(\boldsymbol{\varepsilon}; \boldsymbol{\Gamma})} \phi_d(z - \boldsymbol{\xi}; \boldsymbol{\Sigma}) \Phi_m(\boldsymbol{W} z - \boldsymbol{\eta}; \boldsymbol{I_m}), \tag{1.24}$$

where $\phi_k(.; \boldsymbol{A})$ *and* $\Phi_k(.; \boldsymbol{A})$ *are the p.d.f. and the c.d.f., respectively, of the k-dimensional Gaussian distribution with mean vector zero and covariance matrix* $\boldsymbol{A}$*.*

The proof of Lemma 1.3.1 can be found in the Appendix in Gonçalves & Gamerman (2015).

Simulation from the density (1.24) is not straightforward. Gonçalves & Gamerman (2015) propose the following algorithm to efficiently sample from this distribution.

Define $\mathbf{U}_0^* = \mathbf{A}^{-1}\mathbf{U}_0$, where A is obtained from the Cholesky decomposition of $\boldsymbol{\Gamma}$, i.e, $\boldsymbol{\Gamma} = \boldsymbol{A}\boldsymbol{A}^T$. This implies that $\mathbf{U}_0^* \sim N_m(\mathbf{0}, \mathbf{I}_M)$ and $\mathbf{U}_0 = \mathbf{A}\mathbf{U}_0^*$. Let B be the region in the $\mathbf{U}_0*$ domain where $\mathbf{U}_0 > -(\boldsymbol{\varepsilon} + \boldsymbol{\eta})$. This region defines linear constraints in this domain. The proposed algorithm is as follows:

Algorithm 1 GG Algorithm

1: Obtain u_i^* from $(U_{0i}^*|U_0^* \in B)$;

2: Obtain $u = Au^*$;

3: Obtain z^* from $(U_1|U_0 = u) \sim N(\Delta\Gamma^{-1}u, \Sigma - \Gamma^{-1}\Delta^T)$;

4: Obtain $z = z^* + \xi$;

5: **return** z

Step 3 in Algorithm 1 performed via MCMC, more specifically, using the Gibbs sampler, which means that only Monte Carlo error is involved. For a detailed explanation on this algorithm, see Gonçalves & Gamerman (2015).

Chapter 2

Binary Item Response Models

The Item Response Theory assumes that a change in the latent variable leads to a change in the probability of success (correct answer) in a specified item. This behaviour is described by the Item Characteristic Curve (ICC), which specifies how the probability of an item response changes due to changes in the ability level. The ICC is such that the probability of success will be small for examinees with low ability, and large for examinees with high ability. It is also important to recall that each item has its own ICC in a test.

A very important assumption on item response models (IRM) is known as Lararsfeld's assumption of *local independence*, which states that the examinee's responses to different items are independent, given the latent variables. That is, the performance of an examinee in an item cannot affect his/her performance in any other items in the test, given ability θ. This is equivalent to say that, given ability θ, two items are uncorrelated. Also, the local independence assures that the order of presentation of the test items must not affect the examinee's performance. In the case these assumptions are not held, a special model should be considered to take testlet into account.

There are different mathematical forms of the ICC, and each of them leads to different IRM. In this chapter some well known and employed IRM are reviewed. We focus on IRM for binary response, and consequently models presented are meant to be applied to the analyses of multiple choice items, corrected as right or wrong, or to the analyses of open items, when these are corrected dichotomously.

This chapter presents a brief review of some IRM. Section 2.1 describes the most common IRM for binary data based on symmetric item characteristic curve. The reader more acquainted can feel comfortable to skip this section. Section 2.2 describes some skew IRM.

2.1 Symmetric IRM

The most used models for dichotomous items are the logistic and the probit models. These models are symmetric and they can be divided mainly into three types according to the number of parameters that describe the item. These parameters represent features of the item such as the discrimination power, difficulty and guessing. Another very important parameter in the IRT, if not the most important one, is the individual's ability (or trait), denoted by θ. We start presenting the most simple IRM, the Rasch model.

To establish notation, through all this work, Y_{ij} is a dichotomous variable that takes 1, when the j^{th} examinee answers correctly to the i^{th} item, and 0 otherwise.

2.1.1 The Rasch Model

The simplest item response model is the Rasch model (Rasch, 1960), also known as the one-parameter logistic model (1-PL). This model involves parameters θ_j, which denotes the ability of the j^{th} examinee, and the item difficulty β_i of the i^{th} item. On the Rasch model, the parameter β_i is the point on the ability scale in which the probability of a correct response is 0.5. That is, if the ability of the examinee is higher than the item difficulty, he/she has more chance to correctly answer the item and vice versa. This parameter can be seen as a location parameter. An item i with β_i, is said to be easier then an item k with β_k, when the probability of success at a fixed ability is higher for i in comparison to k, for $\beta_i < \beta_k$.

In the 1-PL model, the probability of a correct response in the i^{th} item by the j^{th} examinee is given by

$$P(Y_{ij} = 1|\theta_j, \beta_i) = \frac{1}{1 + e^{-(\theta_j - \beta_i)}}, \tag{2.1}$$

where $-\infty < \theta_j < \infty$, for $j = 1, \ldots, J$, and item difficulty parameter $-\infty < \beta_i < \infty$, for

$i = 1, \ldots, I$. In Figure 2.1, the ICCs corresponding to model in equation (2.1) are plotted for different item difficulty levels. Notice that β_i has impact only in the location, and not in the shape of the curve. Also, notice that an examinee j with ability level 0 has $P(Y_{ij} = 1|\theta_j = 0, \beta_i = -1) = 0.84$ (solid line) and the same examinee has $P(Y_{ij} = 1|\theta_j = 0, \beta_i = 1) = 0.16$ (dashed line).

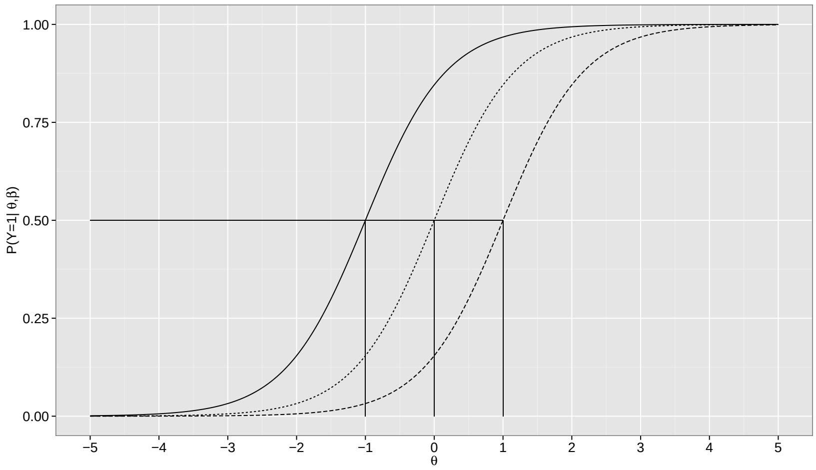

Figure 2.1: ICC of the Rasch model for $\beta = -1$ (solid line), $\beta = 0$ (dotted line) and $\beta = 1$ (dashed line).

The model presented in (2.1) is not identifiable, since the same probability of success can be obtained for two different levels of ability and item difficulty parameter. Notice that if $\theta_j^* = \theta_j + \Delta$ and $\beta_i^* = \beta_i - \Delta$, where Δ is a constant, we have that

$$\begin{aligned}
P(Y_{ij} = 1|\theta_j, \beta_i) &= [1 + e^{-(\theta_j - \beta_i)}]^{-1} \\
&= [1 + e^{-(\theta_j + \Delta - \beta_i - \Delta)}]^{-1} \\
&= [1 + e^{-(\theta_j^* - \beta_i^*)}]^{-1} = P(Y_{ij} = 1|\theta_j^*, \beta_i^*).
\end{aligned}$$

Therefore, for different set of parameters β and θ we can obtain the same likelihood, which brings some inference troubles. This identification problem can be solved, for instance, by adding a restriction on the sum of the difficulty parameters. Another way to identify the model is by assuming a prior distribution for the abilities, which solves the identification issue by setting a metric for θ_j.

The Rasch model is very simple, as it does not take into account more complex, but factual scenarios. In order to turn Rasch's model more flexible, a two-parameter model was developed by Lord (1952), and it is presented in the next subsection.

2.1.2 Two-Parameter Models

The two-parameter models (2P) are obtained by including a discrimination parameter to the model described in expression 2.1. The 2P model was originally developed by psychometrist Lord (1952), who built the ICC in terms of the normal ogive curve, which is formally defined by

$$P(Y_{ij} = 1|\theta_j, a_i, \beta_i) = \Phi(\alpha_i(\theta_j - \beta_i)), \tag{2.2}$$

where $\Phi(\cdot)$ denotes the standard normal cumulated density function (c.d.f..), and $\alpha_i > 0$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The constraint on $\alpha_i > 0$ assures that an examinee with a higher ability has higher probability of success on any item. The discrimination parameter is directly proportional to the slope of the curve at its maximum inclination point $\theta = \beta$. Items with high discrimination parameters are better at differentiating examinees around the location point (difficulty), such that an small change in the latent trait lead to large changes in probability.

Figure 2.2 shows three items under the 2P model, all curves with the same difficulty parameter $\beta_i = 0$, and with different discrimination values. The higher (lower) the discrimination parameter, the better (less) the item is to discriminate low and high ability around β_i. For instance, when $a = 0.2$ (solid line) $P(Y_{ij} = 1|\theta_j = -0.1, \beta_i = 0, \alpha_i = 0.2) = 0.49$, and $P(Y_{ij} = 1|\theta_j = 0.1, \beta_i = 0, \alpha_i = 0.2) = 0.51$. That is, both probabilities of success are very close. However, for $a = 1$ (dotted line), $P(Y_{ij} = 1|\theta_j = -0.1, \beta_i = 0, \alpha_i = 1) = 0.46$ $P(Y_{ij} = 1|\theta_j = 0.1, \beta_i = 0, \alpha_i = 1) = 0.54$. In an extreme scenario, for instance $a = 100$ (dashed line), the item perfectly discriminates students in the neighborhood of $\beta = 0$. According to Baker & Kim (2004), items whose $\alpha < 0.65$ have low discrimination power; $0.65 \leq a < 1.34$ have moderate discrimination; $1.35 \leq \alpha < 1.69$ have high discrimination, and $\alpha \geqslant 1.70$ have very high discrimination power.

Birnbaum (1968) modified Lord's model by using the logistic c.d.f. instead, creating the

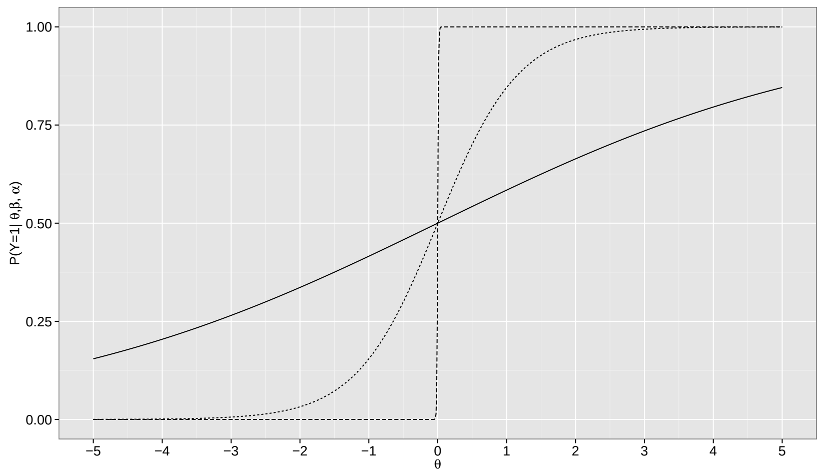

Figure 2.2: ICCs for the 2P IRT model for fixed difficulty parameter $\beta = 0$, $\alpha = 0.2$ (solid line), $\alpha = 1$ (dotted line) and $\alpha = 100$ (dashed line).

two-parameter logit model (2-PL), which is formally defined by

$$P(Y_{ij} = 1|\theta_j, \alpha_i, \beta_i) = \frac{1}{1 + e^{-D\alpha_i(\theta_j - \beta_i)}}, \tag{2.3}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where D is a scale factor. If D = 1.7, this model provides a reasonable approximation for the probit model defined by equation (2.2).

The quantity $\alpha_i(\theta_j - \beta_i)$ is often presented as $(a_i\theta_j - b_i)$ by many authors. According to Baker & Kim (2004) and Fox (2010), this parametrization may result in more stable computational procedures. Furthermore, under the Bayesian approach, this reparametrization leads to conjugated full conditional distributions for $(\mathbf{a} \quad \mathbf{b})$. This latter condition is very attractive in terms of sampling from the joint posterior distribution, and for this reason, we adopt this parametrization all through this work.

2.1.3 Three-Parameter Models

It is noteworthy from the model defined in (2.3) and from Figure 2.2 that the probability of a correct answer goes to zero when the ability goes to $-\infty$. However, it is very plausible that an examinee simply guesses an item, specially in educational tests. To take this into account, Birnbaum (1968) extended the 2P model by introducing a new parameter c, the guessing parameter, which is a nonzero lower asymptote for the ICC. In the three-parameter model (3P) the probability of correct response is explained by this additional factor of guessing. As in the case of the 2P models, the 3P models are mostly defined using both the logistic (3-PL) and the probit function (3-PP).

The three-parameter logit model (3-PL) is defined by

$$P(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + e^{-D(a_i\theta_j - b_i)}}, \tag{2.4}$$

where $1 < c_i < 0$, for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and the other quantities are defined as before. Although c is commonly known as the guessing parameter, it is an item parameter's. Therefore, it can also be interpreted as the probability of success of examinees with extremely low ability, being then the threshold probability of success.

Figure 2.3 shows two ICCs varying according to parameter c, with b = 0 and a = 1 fixed. Notice that, for $c = 0.25$ (dotted line), examinees with very low abilities have 0.25 of probability of success.

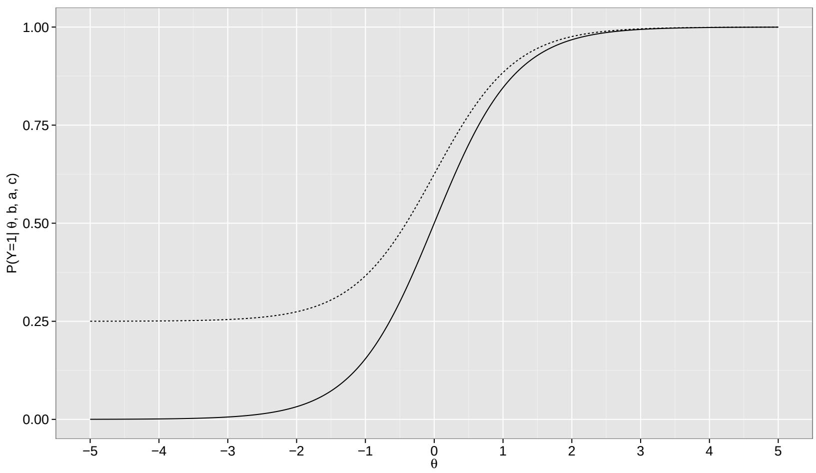

Figure 2.3: ICCs of the 3P logistic model for b = 0, a = 1, c = 0 (solid line) and c = 0.25 (dotted line).

The three-parameter probit model (3-PP) is expressed by

$$P(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\Phi(a_i\theta_j - b_i),$$

where a_i, b_i, c_i and θ_j are defined as in model 2.4.

A Bayesian approach for estimation and model selection for the 3PP models can be found in Sahu (2002).

2.2 Skew IRM

All the previous models presented in Section 2.1 are based on symmetric curves. That is an appropriated assumption whenever it is reasonable to assume that the probability of a correct answer to an item approaches zero at the same rate as it approaches one. Under these models, individuals with low or high abilities are discriminated in a similar way. However, it has been emphasized by several authors that symmetric ICC's are not always suitable to describe the relationship between the abilities and the probability of success. To better fit ICC's with asymmetric behaviour, some asymmetric link functions have been proposed.

These new models take into account the skewness parameter δ, which controls the curve asymmetry. The parameter δ is interpreted by Bazán et al. (2014) as a *penalization parameter*, since it can be seen as a penalty or as a reward on the probability of a correct answer. Figure 2.4 shows the ICC for three different values of δ. At a fixed ability level (for instance, $\theta = 0$) when $\delta = 0$, the probability of a correct answer is equal to 0.5. At the same ability level, however, for $\delta = -0.5$, the probability of a correct answer is equal to 0.67, and for $\delta = -0.9$, this probability goes up to 0.86. The same happens for positive values of δ. However, in this scenario the skewness parameter works as a "penalty". Again, for $\delta = 0$ and $\theta = 0$, the probability of a correct answer is equal to 0.5. At the same ability level however, for $\delta = 0.5$, the probability of a correct answer is equal to 0.33, and for $\delta = 0.9$, this probability goes down to 0.14.

2.2.1 Skew Logit Model

We briefly describe the Logistic Positive Exponent Family proposed by Samejima (2000) and its extension model proposed by Santos (2009).

Let Y_{ij} be a Bernoulli random variable assuming 1 if the answer of the j^{th} examinee to the i^{th} item is correct, and 0 otherwise. The proposed model by Samejima (2000) is defined as

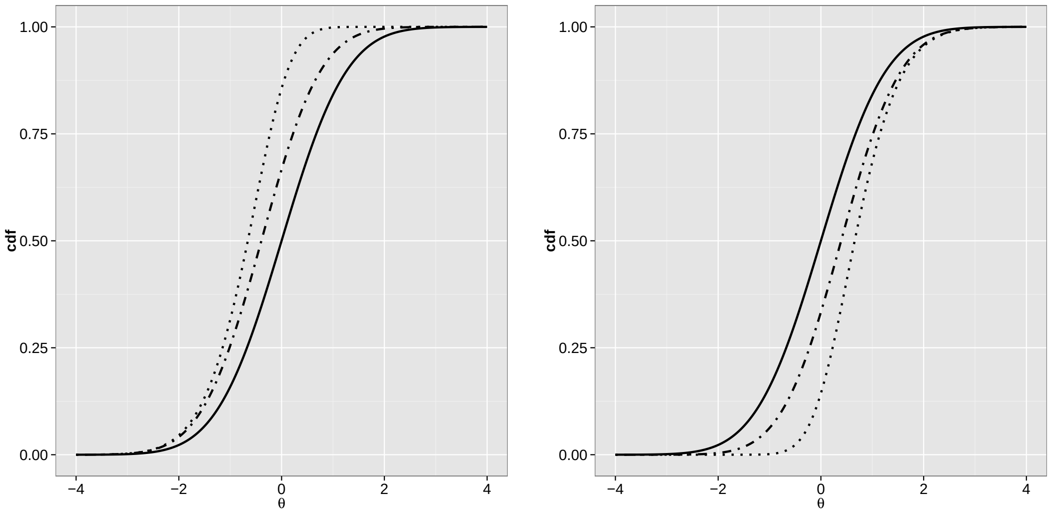

Figure 2.4: Cdf of standard skew normal for negative skewness (left) and positive skewness: $\delta = \pm 0$ (solide line), $\delta = \pm 0.5$ (dashed point line) and $\delta = \pm 0.9$ (pointed line).

follows

$$P(Y_{ij} = 1|\theta_j, a_i, b_i, \epsilon_i) = \left(\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}\right)^{\epsilon_i}, \tag{2.5}$$

for $i = 1, ..., I$ and $j = 1, ..., J$, with $a_i > 0, -\infty < b_i < \infty, \epsilon_i > 0$, and $-\infty < \theta_j < \infty$. Notice that when $\epsilon_i = 1$, the symmetric link is obtained. According to Samejima (2000), the skewness parameters ϵ_i represents the item complexity, which is distinct from item difficulty. This item parameter is related to the number sequential steps to successfully solve the complete problem. Figure (2.5) shows ICCs with parameters a = 1, b = 0 and different values for ϵ.

The proposed model by Santos (2009) extended model (2.5) by adding the pseudo-guessing parameter c. Thus, the proposed model by Santos (2009) is given by

$$P(Y_{ij} = 1|\theta_j, a_i, b_i, c_i, \epsilon_i) = c_i + \left(\frac{(1 - c_i)}{1 + e^{-Da_i(\theta_j - b_i)}}\right)^{\epsilon_i}, \tag{2.6}$$

for $i = 1, ..., I$ and $j = 1, ..., J$, with $a_i > 0, -\infty < b_i < \infty, 1 < c_i < 0, \epsilon_i > 0$, and $-\infty < \theta_j < \infty$. As in model (2.5), the symmetric three-parameter logit link is recovered when $\epsilon_i = 1$. Another contribution of Santos (2009) work is the modeling of the skewness parameters ϵ_i, by considering a finite mixture type with a point mass, given by

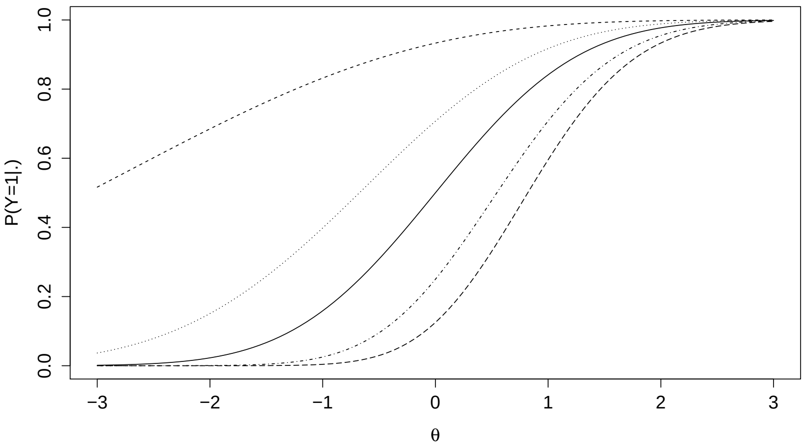

Figure 2.5: ICCs with parameters a = 1, b = 0, c = 0, and different values for ϵ: $\epsilon = 0.1$ (short dashed line), $\epsilon = 0.5$ (dotted line), $\epsilon = 1$ (solid line), $\epsilon = 2$ (dotted-dashed line), $\epsilon = 3$ (long dashed line).

$$(\epsilon_i|\pi_i) \sim \pi_i\delta_1 + (1-\pi_i)LN(\mu_{\epsilon_i}, \sigma^2_{\epsilon_i}), \tag{2.7}$$

where δ_1 is a distribution degenerated in 1, and π_i denotes the probability of item i be symmetric. Because of the complexity, the posterior distribution of all parameters are approximated via MCMC methods, more specifically, via Metropolis-Hastings.

2.2.2 BBB Skew Probit Model

We now define the BBB Skew Probit model proposed in Bazán et al. (2006), with some extensions proposed in Bazán et al. (2014). Let Y_{ij} be a Bernoulli random variable assuming 1 if the answer of the j^{th} examinee to the i^{th} item is correct, and 0 otherwise. The BBB Skew Probit model is defined as follows

$$\begin{aligned} Y_{ij}|\theta_j, a_i, b_i, \delta_i &\sim Bernoulli(p_{ij}), &(2.8)\\ p_{ij} &= \Phi_{SN}(m_{ij}; \delta_i), &(2.9)\\ m_{ij} &= a_i\theta_j - b_i, &(2.10)\end{aligned}$$

for $i = 1, ..., I$ and $j = 1, ..., J$, with $a_i > 0, -\infty < b_i < \infty, -1 < d_i < 1$, and $-\infty < \theta_j < \infty$, and where $\Phi_{SN}(m_{ij}; \delta_i)$ denotes the c.d.f. of the standard SN distribution defined in 1.8. As in the usual symmetric IRM, a_i and b_i denote the discrimination and difficulty item parameters, respectively.

To avoid a Bernoulli type likelihood, the authors used the data augmentation strategy proposed by Albert (1992), which equivalently represents the skew probit IRM defined in (2.8) by

$$Y_{ij} = \begin{cases} 1, & \text{if} \quad X_{ij} > 0, \\ 0, & \text{if} \quad X_{ij}, \leq 0, \end{cases} \tag{2.11}$$

where $X_{ij} = m_{ij} + e_{ij}$, and $e_{ij} \sim SN(0, 1, -\delta_i)$. For sampling purposes, the authors considered the stochastic representation of a SN variable (Henze, 1986) assuming

$$e_{ij} \stackrel{d}{=} \delta_i V_{ij} + (1 - \delta_i^2)^{1/2} W_{ij}, \tag{2.12}$$

where $V \sim HN(0, 1)$, $W \sim N(0, 1)$, and $V \perp W$. As priors specification, they elicited $a_i \sim N(\mu_a, \sigma_a^2)\mathbb{I}(a_i > 0)$, $b_i \sim N(\mu_b, \sigma_b^2)$ and $d_i \sim Uniform(-1, 1)$. For θ_j the authors used $\theta_j \sim N(0, 1)$, but also extended the SN IRM by considering asymmetrically distributed latent variables, assuming, $\theta_j \sim SN(\mu, \sigma^2, \omega)$, where $-\infty < \mu < \infty, \sigma^2 > 0$, and $-1 < \omega < 1$. As noticed by Azevedo et al. (2011), the model considering asymmetric distribution for the latent variable is not identifiable. To overcome this issue, Bazán et al. (2014) assumed priors distributions for θ hyperparameters.

2.2.3 Skew Normal model IRT under the Centered Parametrization

As mentioned in Section 1.2 of Chapter 1, because of the behaviour of the likelihood function in the neighborhood of $\lambda = 0$ (under the original parametrization), Azzalini (1985) proposed a centered parametrization for the SN distribution. Under this parametrization, Azevedo et al. (2011) proposed a model based on the centered skew distribution (CSN). However, it is important to notice that the authors did not apply the CSN to the ICC, but only to the latent trait, that is, they assumed $\theta \sim CSN(0, 1, \gamma)$. Their proposal comes from the fact that many works in literature suggest the lack of normality in the latent traits, although it is questionable if there is a significant gain in doing so. As cited in Section 1.2, the CSN is parametrized by the Pearson's skewness coefficient γ, and it preserves the asymmetric behavior of the SN distribution. An important contribution in Azevedo et al. (2011) is that the proposed model can take into account omitted responses. For estimation purpose the authors consider the Metropolis-Hasting within Gibbs sampling algorithm.

A great advantage of the CSN if compared to the SN in item response theory is the role of the discrimination and difficulty parameters of the item. Under the CSN distribution, the parameters "a" and "b" play the same role in the symmetric and in the skewed models, since the expected value and the variance of the latent distribution are invariant with respect to skewness parameter γ. Taking into account

$$X_{ij} = m_{ij} + e_{ij}, \tag{2.13}$$

$E[X_{ij}] = m_{ij}$ and $Var[X_{ij}] = 1$, as $E[e_{ij}] = 0$ and $Var[e_{ij}] = 1$, for both $e_{ij} \sim CSN(0,1,-\gamma_i)$ and for $e_{ij} \sim N(0,1)$. The same does not happen when $e_{ij} \sim SN(0,1,-\delta_i)$, as $E[X_{ij}] = m_{ij} + \sqrt{\frac{2}{\pi}}\delta_i$, and $Var[X_{ij}] = 1 - \frac{2}{\pi}\delta_i^2$, since $E[e_{ij}] = \sqrt{\frac{2}{\pi}}\delta_i$ and $Var[e_{ij}] = 1 - \frac{2}{\pi}\delta_i^2$.

That means that the posterior of **a** and **b** will be very similar when considering the symmetric model or the skewed model under the centered parametrization, and significantly different when considering the skewed model under the non-centered parameter. A consequence of that is that considering the non-centered parametrization could have a negative impact on the MCMC due to potential multimodes of the posterior.

Chapter 3

A Flexible Class of Centered Skew Probit IRT Model

As noticed by many authors, in many situations it is more appropriated to use an asymmetric link for the ICC. Bazán et al. (2014) introduced a skewness parameter associated to each item, in order to build a more flexible model, as it allows the use of the symmetric probit as well as the use of the asymmetric probit ICC. The authors considered a Uniform(-1,1) prior for the skewness parameter δ. However, by eliciting this prior, the prior probability of having a symmetric probit link, that is, $\delta = 0$, is zero. Even more, the choice of this prior assumes that all items are asymmetric with positive probability. It would be desirable, if not ideal, to assume the asymmetric structure only for items that require it. Nevertheless, that information is not available *a priori* and a naive model selection procedure would require 2^I models to be fitted and compared. A way to overcome this issue is to consider a mixture component for the skewness parameter γ, so that all items may or may not be asymmetric, and in that way, the data point out which model is more likely.

In this work, we introduce a new skew probit IRT model, in which a mixture component on the skewness parameter is introduced, so not all items need to be assumed asymmetric *a priori*. This makes the proposed model flexible enough to embrace all possible models in one (symmetric and asymmetric probit with different degrees of skewness). An advantage of the proposed approach is that an intrinsic model selection is performed through the mixture posterior probabilities. Our asymmetric ICC is built under the Centered Skew Normal (CSN) distribution since centered parametrization shows some advantages over the direct parametrization (see Section 1.2). Another contribution of our work is that we present different algorithms to sample from the posterior distribution (Section 3.4).

3.1 Proposed Model

Let Y_{ij} be a dichotomous variable assuming 1, if the answer of the $j^{th}, j = 1, \ldots, J$, examinee to the $i^{th}, i = 1, \ldots, I$, item is correct and 0 otherwise. The CSN probit model is given by

$$\begin{aligned} Y_{ij}|p_{ij} &\sim Bernoulli(p_{ij}), \\ p_{ij} &= \Phi_{CSN}(m_{ij}; \gamma_i), \\ m_{ij} &= a_i\theta_j - b_i, \end{aligned} \tag{3.1}$$

where a_i, b_i and θ_j are, respectively, the discriminant, difficulty and ability parameters defined in Subsection 2.1.2, γ_i is the skewness parameter defined in 1.13, and $\Phi_{CSN}(.)$ is the c.d.f of the CSN distribution, defined in expression (1.20).

To establish notation hereafter, let $\boldsymbol{\theta} = (\theta_1; \ldots; \theta_J)$, $\mathbf{a} = (a_1; \ldots; a_I)^T$, $\mathbf{b} = (b_1; \ldots; b_I)^T$, $\boldsymbol{\gamma} = (\gamma_1; \ldots; \gamma_I)^T$, and, matrix $\boldsymbol{y} = (y_{ij})_{I \times J}$ the observed data.

The likelihood function based on an independent sample of the centered skew probit IRM in 3.1 is given by

$$L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}|\mathbf{y}) = \prod_{i=1}^{I}\prod_{j=1}^{J}[\Phi_{CSN}(m_{ij}; \gamma_i)]^{y_{ij}}[1 - \Phi_{CSN}(m_{ij}; \gamma_i)]^{1-y_{ij}}. \tag{3.2}$$

To avoid working with a Bernoulli type likelihood, it is more convenient to work with the stochastic representation proposed by Albert (1992), where the model presented in (3.1) can be rewritten as

$$Y_{ij} = \begin{cases} 1, & \text{if} \quad X_{ij} > 0, \\ 0, & \text{if} \quad X_{ij}, \leq 0, \end{cases} \tag{3.3}$$

where $X_{ij} = m_{ij} + e_{ij}$, $e_{ij} \sim CSN(0, 1, -\gamma_i)$ and, hence, $X_{ij} \sim CSN(m_{ij}, 1, -\gamma)$. Notice that the skewness parameter of X_{ij} is the opposite of the skewness parameter of the ICC as

$$\begin{aligned}
P(Y_{ij} = 1) &= P(X_{ij} > 0) = P(m_{ij} + e_{ij} > 0) && (3.4)\\
&= P(e_{ij} > -m_{ij}) = 1 - P(e_{ij} < -m_{ij}) && (3.5)\\
&= 1 - \Phi_{CSN}(-m_{ij}; \gamma_i) && (3.6)\\
&= \Phi_{CSN}(m_{ij}; -\gamma_i), && (3.7)
\end{aligned}$$

where the last equality is justified by the fact that $\Phi_{CSN}(-x; \gamma) = 1 - \Phi_{CSN}(x; -\gamma)$. It is important to notice that the representation given by 3.3 preserves the probability model of Y_{ij}. For further notation reference, let $\boldsymbol{X} = (X_{ij})_{I\times J}$ and $\boldsymbol{x} = (x_{ij})_{I\times J}$ the observed realization of $\mathbf{X}$.

3.2 Prior Specifications

In order to build a more robust centered skew probit item response model (IRM) that permits inference about both symmetric and asymmetric items, the following finite mixture prior distribution for $\gamma_i's$ it is assumed

$$\begin{aligned}
\gamma_i &= Z_{i0}W_{i0} + Z_{i1}W_{i1} + Z_{i2}W_{i2},\\
Z_i &\sim Mult(1, p_{i0}, p_{i1}, p_{i2}),\\
W_{i0} &\sim \delta_0,\\
-W_{i1} &\sim Beta(\alpha_w, \beta_w, l_l, l_u),\\
W_{i2} &\sim Beta(\alpha_w, \beta_w, l_l, l_u),\\
p_i &\sim Dirichlet(\alpha_0, \alpha_1, \alpha_2),
\end{aligned} \tag{3.8}$$

where δ_0 is a point-mass at zero and Beta(.) stands for the General Beta distribution defined in Appendix 5.3 with support (-0.99527, 0.99527). The random variables W_{i0}, W_{i1} and W_{i2} are associated with the symmetric, negative asymmetric and positive asymmetric models, respectively. The vector of probability $\boldsymbol{p_i} = (p_{i0}, p_{i1}, p_{i2})$ is such that $\sum_{k=0}^{2} p_{ik} = 1$, and hyperparameters $(\alpha_0, \alpha_1, \alpha_2)$ are assumed to be known. For further notation reference, let $\mathbf{Z} = (Z_{ic})_{I\times 3}$, where $c = 0, 1$ and 2.

We assume independence among the components $\boldsymbol{\theta}$, $\mathbf{a}$, and $\mathbf{b}$ such that the joint prior distribution is given by

$$\pi(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) = \left[\prod_{i=1}^{I} \pi(a_i)\pi(b_i)\right] \left[\prod_{j=1}^{J} \pi(\theta_j)\right], \tag{3.9}$$

where $\theta_j \overset{iid}{\sim} N(\mu^*_{\theta_j}, \sigma^{*2}_{\theta_j})$, $a_i \overset{iid}{\sim} N(\mu^*_{a_i}, \sigma^{2*}_{a_i})\mathbb{I}(a_i > 0)$, and $b_i \overset{iid}{\sim} N(\mu^*_{b_i}, \sigma^{*2}_{b_i})$. The hyperparameters $\mu^*_{a_i}, m^*_{b_i}, \sigma^{2*}_{a_i}$ and $\sigma^{2*}_{b_i}$ are assumed to be known.

The values for α_w and β_w need to be carefully chosen. Firstly, the prior density cannot concentrate too much probability mass around zero, which refers to the symmetric model and, therefore, could lead to identifiability problems. Secondly, it cannot be strongly informative by concentrating most of its mass in high values, which could overestimate γ. The shape of the γ prior is shown in Figure 3.1.

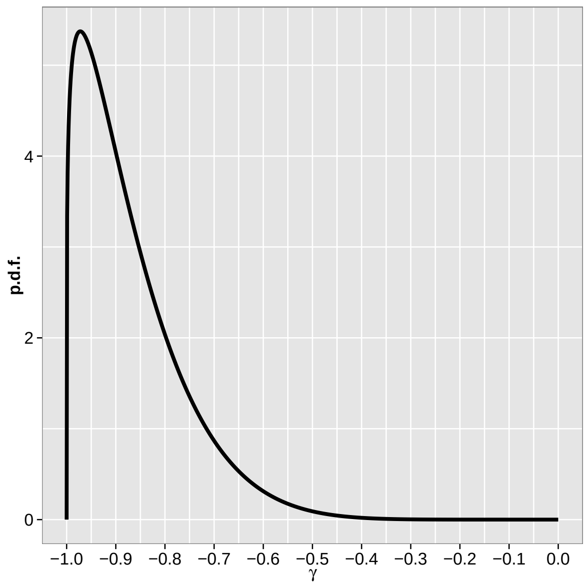

Figure 3.1: Shape of γ prior.

3.3 Posterior Distribution

The main goal of the inference process is to obtain the posterior distributions of all unknown quantities in the model, which in our proposed model is $\mathbf{\Psi} = (\mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}, \mathbf{p}, \mathbf{Z}, \boldsymbol{\gamma})$. Considering the prior dependence structure given in 3.8, the joint density of $(\mathbf{Y}, \mathbf{\Psi})$ may be factorized as

$$\pi(\mathbf{Y}, \mathbf{\Psi}) = \left[\prod_{i=1}^{I}\prod_{j=1}^{J} \pi(Y_{ij}|X_{ij})\pi(X_{ij}|\theta_j, a_i, b_i, \gamma_i)\pi(\gamma_i|Z_i)\pi(Z_i|p_i)\pi(a_i)\pi(b_i)\pi(p_i)\right]\left[\prod_{j=1}^{J}\pi(\theta_j)\right]. \tag{3.10}$$

Considering the prior specifications, the joint distribution is given by

$$\begin{aligned}
\pi(\mathbf{Y}, \mathbf{\Psi}) = \Bigg[\prod_{i=1}^{I}\prod_{j=1}^{J} & \Big(\mathbb{I}(Y_{ij}=1)\mathbb{I}(X_{ij}>0) + \mathbb{I}(Y_{ij}=0)\mathbb{I}(X_{ij}\le 0)\Big) f_{CSN}(X_{ij}; m_{ij}, 1, -\gamma_i) \\
& \times \Big(\mathbb{I}(Z_{i0}=1)\mathbb{I}(\gamma_i=0) + \mathbb{I}(Z_{i1}=1) f_B(-\gamma_i; \alpha_w, \beta_w, l_l, l_u) + \mathbb{I}(Z_{i2}=1) f_B(\gamma_i; \alpha_w, \beta_w, l_l, l_u)\Big) \\
& \times p_{i0}^{Z_{i0}} p_{i1}^{Z_{i1}} p_{i2}^{Z_{i2}} \frac{\Gamma(\alpha_0+\alpha_1+\alpha_2)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)} p_{i0}^{\alpha_0-1} p_{i1}^{\alpha_1-1} p_{i2}^{\alpha_2-1} f_N(a_i; \mu^*_{a_i}, \sigma^{*2}_{a_i})\mathbb{I}(a_i>0) f_N(b_i; \mu^*_{b_i}, \sigma^{*2}_{b_i})\Bigg] \\
& \times \left[\prod_{j=1}^{J} f_N(\theta_j; \mu^*_{\theta_j}, \sigma^{*2}_{\theta_j})\right],
\end{aligned} \tag{3.11}$$

where $f_B(.; \alpha_w, \beta_w, l_l, l_u)$ denotes the p.d.f. of the Beta distribution defined in Appendix 5.3, and $f_N(.; \mu, \sigma^2)$ stands for a Normal density with mean μ and variance σ^2.

In the next section we introduce an MCMC algorithm to sample from the posterior (3.11).

3.4 MCMC

The model presented in (3.11) has a complex high dimensional distribution which is very hard to be explored analytically. Despite the complexity of the proposed model, the full conditional distributions of $\mathbf{X}$, $\boldsymbol{\theta}$, $(\mathbf{a} \quad \mathbf{b})$, $\mathbf{p}$, $\mathbf{Z}$ and $\boldsymbol{\gamma}$ are tractable, which permit us to develop an efficient algorithm to sample from the posterior distribution. In order to obtain these samples, we propose an MCMC scheme based on a Gibbs sampler with Metropolis Hastings steps.

To facilitate the MCMC construction, we reparametrize the model (3.10) taking into account the mixture structure for γ_i, presented in (3.8), in the following way

$$\gamma_i = Z_{i0}0 + Z_{i1}\gamma_{i-} + Z_{i2}\gamma_{i+}, \tag{3.12}$$

where γ_{i-} and γ_{i+} denote, respectively, the negative and positive values generated for γ_i. By doing this, we construct a Markov chain in γ_{i-} and γ_{i+} instead of constructing a single chain for γ_i, that is, we treat the negative and positive parts of γ_i separately. Expression (3.12) shows that, given Z_i, the model is symmetric when $Z_{i0} = 1$, negative asymmetric when $Z_{i1} = 1$, and positive asymmetric when $Z_{i2} = 1$. Therefore, model (3.10) is rewritten as

$$\begin{aligned}
\pi(\mathbf{Y}, \mathbf{\Psi}) =& \left[\prod_{i=1}^{I}\prod_{j=1}^{J} \pi(Y_{ij}|X_{ij})\pi(X_{ij}|\theta_j, a_i, b_i, Z_i, \gamma_{i-}, \gamma_{i+})\pi(Z_i|p_i)\pi(\gamma_{i-})\pi(\gamma_{i+})\pi(a_i)\pi(b_i)\pi(p_i)\right]\left[\prod_{j=1}^{J}\pi(\theta_j)\right], \\
=& \left[\prod_{i=1}^{I}\prod_{j=1}^{J}\Big(\mathbb{I}(Y_{ij}=1)\mathbb{I}(X_{ij}>0) + \mathbb{I}(Y_{ij}=0)\mathbb{I}(X_{ij}\leq 0)\Big)\right. \\
& \times f_{CSN}(X_{ij}; a_i\theta_j - b_i, 1, -(Z_{i0}0 + Z_{i1}\gamma_{i-} + Z_{i2}\gamma_{i+}))f_B(-\gamma_{i-}; \alpha_w, \beta_w, l_l, l_u)f_B(\gamma_{i+}; \alpha_w, \beta_w, l_l, l_u) \\
& \left.\times \frac{\Gamma(\alpha_0+\alpha_2+\alpha_2)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)} p_{i0}^{\alpha_0-1}p_{i1}^{\alpha_1-1}p_{i2}^{\alpha_2-1}p_{i0}^{Z_{i0}}p_{i1}^{Z_{i1}}p_{i2}^{Z_{i2}} f_N(a_i; \mu^*_{a_i}, \sigma^{*2}_{a_i})\mathbb{I}(a_i>0)f_N(b_i; \mu^*_{b_i}, \sigma^{*2}_{b_i})\right] \\
& \times \left[\prod_{j=1}^{J} f_N(\theta_j; \mu^*_{\theta_j}, \sigma^{*2}_{\theta_j})\right].
\end{aligned}$$

To improve the MCMC algorithm, an efficient strategy is to use a blocking scheme, where strongly correlated parameters must be jointly sampled. For the proposed model, we consider the following blocking scheme

$$(\mathbf{X}) \quad (\boldsymbol{\theta}) \quad (\mathbf{a}, \mathbf{b}) \quad (\mathbf{p}) \quad (\mathbf{Z}) \quad (\boldsymbol{\gamma}_-, \boldsymbol{\gamma}_+). \tag{3.13}$$

The next subsections show the full conditional distribution for blocks in (3.13).

3.4.1 Full Conditional Distribution for X

Given parameters $\boldsymbol{\theta}$, $\mathbf{a}$, $\mathbf{b}$ and $\boldsymbol{\gamma}$, it follows from (3.13) that the components of $\mathbf{X}$ are independent and such that

$$(X_{ij}|\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{y}) \sim \begin{cases} CSN(m_{ij}, 1, -\gamma_i)\mathbb{I}(x_{ij} > 0), & \text{if} \quad y_{ij} = 1, \\ CSN(m_{ij}, 1, -\gamma_i)\mathbb{I}(x_{ij} < 0), & \text{if} \quad y_{ij} = 0. \end{cases} \tag{3.14}$$

That is, the full conditional distribution of X_{ij} is a truncated CSN distribution putting positive mass in positive values, when $y_{ij} = 1$, and a truncated CSN distribution putting positive mass in negative values, when $y_{ij} = 0$. If $\gamma_i = 0$, then the full conditional distribution of X_{ij} is equivalent to the Truncated Normal distribution.

It is not possible to sample directly from (3.14). To do so we consider two algorithms: a Rejection Sampling (RS) and an embedded Gibbs Sampler.

3.4.2 Full Conditional Distribution for $\boldsymbol{\theta}$

For parameter vector $\boldsymbol{\theta}$, the full conditional distribution can be factored for each examinee. Such strategy does not compromise the algorithm convergence, because given the item parameters, examinees are independent. Therefore, it is equivalent to simulate jointly all vector $\boldsymbol{\theta}$, which has a J-dimension distribution, or to simulate each θ_j separately from its marginal full conditional distribution, which is unidimensional. The latter is more attractive because of computational costs.

Each ability parameter $\theta_1, \ldots, \theta_J$, given $\mathbf{X}$, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{Z}$ and $\boldsymbol{\gamma}$, has full conditional density given by

$$f_N(\theta_j; \mu_\theta^*, \sigma_\theta^{*2}) \propto \prod_{i=1}^{I} f_{CSN}(x_{ij}; a_i\theta_j - b_i, 1, -\gamma_i) f_N(\theta_j; 0, \sigma_\theta^{*2}), \tag{3.15}$$

where μ_θ^* and σ_θ^{*2} are the prior mean and variance of $\boldsymbol{\theta}$, respectively. After some algebraic manipulation (see details on Appendix 5.1), it follows that the full conditional distribution in

(3.15) for each θ_j is

$$f(\theta_j|\mathbf{X},\mathbf{a},\mathbf{b},\boldsymbol{\gamma}) \propto \prod_{i=1}^{I} \phi\left(\frac{x_{ij} + s\gamma_i^{1/3} - (a_i\theta_j - b_i)}{\varphi_i} + \frac{\theta_j}{\sigma_\theta^{*2}}\right) \Phi\left(g(\gamma_i)\left(\frac{x_{ij} + s\gamma_i^{1/3} - (a_i\theta_j - b_i)}{\varphi_i}\right)\right), \tag{3.16}$$

where

$$g(\gamma_i) = \frac{s\gamma_i^{1/3}}{\sqrt{r^2 + s^2\gamma_i^{2/3}(r^2 - 1)}}, \tag{3.17}$$

and

$$\varphi_i = \sqrt{(1 + s^2\gamma_i^{2/3})}. \tag{3.18}$$

From expression derived in (3.16), each θ_j has full conditional distribution of the form

$$f(\theta_j|\mathbf{X},\mathbf{a},\mathbf{b},\mathbf{Z},\boldsymbol{\gamma}) \propto \phi(\theta_j - \xi_{\theta_j};\Sigma_\theta)\Phi_I(\mathbf{W}\theta_j - \boldsymbol{\eta}_j,\mathbf{I}_I), \tag{3.19}$$

where $\Phi_I(.)$ is a normal c.d.f of dimension I, that is, given $\mathbf{X}$, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{Z}$, and $\boldsymbol{\gamma}$, each θ_j has an ASUN distribution, defined in (1.24), denoted by

$$(\theta_j|\mathbf{X},\mathbf{a},\mathbf{b},\mathbf{Z},\boldsymbol{\gamma}) \sim ASUN(\xi_{\theta_j},\Sigma_\theta,\mathbf{W},\boldsymbol{\eta}_j), \tag{3.20}$$

where Σ_θ and ξ_{θ_j} are given by

$$\Sigma_{\theta_j} = \frac{\sigma_{\theta_j}^{2*}}{\sigma_{\theta_j}^{2*}\sum_{i=1}^{I}\frac{a_i^2}{\varphi_i^2} + 1}, \tag{3.21a}$$

$$\xi_{\theta_j} = \Sigma_{\theta_j}\left(\sum_{i=1}^{I}\frac{a_i(x_{ij} + b_i + s\gamma_i^{1/3})}{\varphi_i} + \frac{\mu_{\theta_j}^*}{\sigma_{\theta_j}^{*2}}\right), \tag{3.21b}$$

$$\mathbf{W} = \begin{pmatrix} -a_1\frac{g(\gamma_1)}{\varphi_1} & -a_2\frac{g(\gamma_2)}{\varphi_2} & \dots & -a_I\frac{g(\gamma_I)}{\varphi_I}\end{pmatrix}^T,$$

$$\boldsymbol{\eta_j} = \left(\left(\frac{g(\gamma_1)}{\varphi_1}(x_{1j} + b_1 + s\gamma_1^{1/3})\right) \quad \left(\frac{g(\gamma_2)}{\varphi_2}(x_{2j} + b_2 + s\gamma_2^{1/3})\right) \quad \dots \quad \left(\frac{g(\gamma_I)}{\varphi_I}(x_{Ij} + b_I + s\gamma_I^{1/3})\right)\right)^T,$$

where $\Sigma_{\theta_j} \in \mathbb{R}_+$, $\xi_{\theta_j} \in \mathbb{R}$, and $\mathbf{W}$ and $\boldsymbol{\eta}_j$ column vectors of order I.

Notice that if $\gamma_i = 0$ (or equivalently if $Z_{i0} = 1$), the full conditional distribution in 3.20 becomes

$$(\theta_j|\mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}) \sim N(\xi_{\theta_j}, \Sigma_{\theta_j}), \tag{3.22}$$

where Σ_θ and ξ_{θ_j} are given, respectively, by

$$\Sigma_{\theta_j} = \frac{\sigma^{2*}_{\theta_j}}{\sigma^{2*}_{\theta_j} \sum_{i=1}^{I} \frac{a_i^2}{\varphi_i^2} + 1},$$
$$\xi_{\theta_j} = \Sigma_{\theta_j} \left(\sum_{i=1}^{I} \frac{a_i(x_{ij} + b_i + s\gamma_i^{1/3})}{\varphi_i} + \frac{\mu^*_{\theta_j}}{\sigma^{*2}_{\theta_j}} \right).$$

Simulation from (3.20) is performed, for each fixed j, by using Algorithm 1. However, this simulation can be computationally expensive as, at each iteration k of the MCMC, it involves a Cholesky decomposition and the inversion of a matrix of order $I \times I$. Another possible way to sample from distribution in (3.16) is to consider the Metropolis-Hastings algorithm.

The target density is given by (3.16), and as the proposal distribution we consider a random walk given by

$$q(\theta_j^*|\theta_j^{(k)}) = f_N(\theta_j^*; \theta_j^{(k)}, \tau^2_{\theta_j}) \tag{3.23}$$

where θ_j^* is the candidate value for θ_j, $\theta_j^{(k)}$ is the current value of the chain, and $\tau^2_{\theta_j}$ is the variance (tuning), which value is chosen such that the acceptance ratio is about 0.44. The acceptance probability is given by

$$\alpha(\theta^*, \theta^{(k)}) = min \left\{ 1, \frac{\prod_{i=1}^{I} f_{CSN}(x_{ij}; a_i\theta_j^* - b_i, 1, -\gamma_i) f_N(\theta_j^*; \mu^*_{\theta_j}, \sigma^{*2}_{\theta_j})}{\prod_{i=1}^{I} f_{CSN}(x_{ij}; a_i\theta_j^{(k)} - b_i, 1, -\gamma_i) f_N(\theta_j^{(k)}; \mu^*_{\theta_j}, \sigma^{*2}_{\theta_j})} \right\}. \tag{3.24}$$

3.4.3 Full Conditional Distribution for (a, b)

Since parameters a_i and b_i are strongly correlated, a good strategy to improve the MCMC performance is to generate a_i and b_i from the joint full conditional distribution of ($\mathbf{a}$,$\mathbf{b}$). However, the full conditional distribution of $(\mathbf{a}, \mathbf{b})$ can be factored for each item, since the pairs (a_i, b_i) are conditionally independent. Therefore, it is equivalent to simulate jointly all vector (**a, b**), which has a distribution of order $2 \times I$ or to simulate each pair separately from its marginal full conditional distribution. The latter option is chosen here as it has lower computational cost.

The joint distribution for each pair (a_i, b_i) given $\mathbf{X}$, $\mathbf{Z}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, is given by

$$
\begin{aligned}
f(a_i, b_i|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &\propto \prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_i) f_N(a_i; \mu_{a_i}, \sigma_{a_i}^{*2}) \mathbb{I}(a_i > 0) f_N(b_i; \mu_{b_i}, \sigma_{b_i}^{*2}) \\
&\propto \prod_{j=1}^{J} \phi \left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i} + \frac{a_i - \mu_{a_i}^*}{\sigma_{a_i}^*} + \frac{b_i - \mu_{b_i}^*}{\sigma_{b_i}^*} \right) \mathbb{I}(a_i > 0) \times \\
&\times \Phi \left(g(\gamma_i) \left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i} \right) \right), \qquad (3.25)
\end{aligned}
$$

where $\sigma_{a_i}^{*2}$ and $\sigma_{b_i}^{*2}$ are the prior variances of a_i and b_i, respectively, $g(\gamma_i)$ and φ_i are defined in (3.17) and in (3.18), respectively.

The joint full conditional distribution of $(a_i, b_i|.)$ is the truncated ASUN distribution given by

$$
(a_i, b_i|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \sim ASUN(\boldsymbol{\xi}_i, \boldsymbol{\Sigma}_{ab_i}, \boldsymbol{W}_i, \boldsymbol{\eta}_i) \mathbb{I}((a_i, b_i) \in A), \qquad (3.26)
$$

where $A = \{(a_i, b_i) \in \mathbb{R} : a_i > 0\}$, $\xi_i = (\mu_{a_i} \quad \mu_{b_i})^T$, $\Sigma_{ab_i} = \begin{pmatrix} \sigma_{a_i}^2 & \rho\sigma_{a_i}\sigma_{b_i} \\ \rho\sigma_{a_i}\sigma_{b_i} & \sigma_{b_i}^2 \end{pmatrix}$,

$$
\begin{aligned}
\rho_i &= \frac{\sigma_{a_i}^* \sigma_{b_i}^* \sum_{j=1}^{J} \theta_j}{\left[\left(\sigma_{a_i}^{2*} \sum_{j=1}^{J} \theta_j^2 + \varphi_i^2 \right) \left(\sigma_{b_i}^{*2} J + \varphi_i^2 \right) \right]^{1/2}}, \\
\sigma_{b_i}^2 &= \left(\frac{\sigma_{b_i}^{*2} \varphi_i^2}{\sigma_{b_i}^{*2} J + \varphi_i^2} \right) \frac{1}{(1 - \rho_i^2)}, \\
\sigma_{a_i}^2 &= \left(\frac{\sigma_{a_i}^{*2} \varphi_i^2}{\sigma_{a_i}^{*2} \sum_{j=1}^{J} \theta_j^2 + \varphi_i^2} \right) \frac{1}{(1 - \rho_i^2)}, \\
\mu_{a_i} &= \sigma_{a_i}^2 \left(\frac{\sum_{j=1}^{J} x_{ij}\theta_j + s\gamma_i^{1/3} \sum_{j=1}^{J} \theta_j + \varphi_i^2 \mu_{a_i}^* \sigma_{a_i}^{*-2}}{\varphi_i^2} \right) - \sigma_{a_i}\sigma_{b_i}\rho_i \left(\frac{\sum_{j=1}^{J} x_{ij} + Js\gamma_i^{1/3} - \varphi_i^2 \mu_{b_i}^* \sigma_{b_i}^{*-2}}{\varphi_i^2} \right), \\
\mu_{b_i} &= \sigma_{a_i}\sigma_{b_i}\rho_i \left(\frac{\sum_{j=1}^{J} x_{ij}\theta_j + s\gamma_i^{1/3} \sum_{j=1}^{J} \theta_j + \varphi_i^2 \mu_{a_i}^* \sigma_{a_i}^{*-2}}{\varphi_i^2} \right) - \sigma_{b_i}^2 \left(\frac{\sum_{j=1}^{J} x_{ij} + Js\gamma_i^{1/3} - \varphi_i^2 \mu_{b_i}^* \sigma_{b_i}^{*-2}}{\varphi_i^2} \right), \\
W_i &= \frac{\gamma_i}{\varphi_i} \left(\boldsymbol{\theta}^T \quad \mathbf{1}_J \right),
\end{aligned}
$$

$$\eta_i = \frac{\gamma_i}{\varphi_i}\left(\boldsymbol{x_i}^T + s\gamma_i^{1/3}\right), \tag{3.27}$$

where $\mathbf{1}_J$ is a column vector of ones of order J, and $\boldsymbol{x_i} = \begin{pmatrix} x_{i1} & x_{i2} & \dots & x_{iJ} \end{pmatrix}$. The algebraic derivation for distribution (3.26) is given in Appendix 5.2. If $\gamma_i = 0$, it implies that $\varphi_i = 1$, and consequently distribution (3.26) becomes equivalent to a symmetric truncated bivariate normal distribution. in this case, the full distribution of each pair (a_i, b_i) is given by

$$(a_i, b_i \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \sim N_2(\mu_{ab_i}, \sigma^2_{ba_i}; A), \tag{3.28}$$

where A $= \{(a_i, b_i) \in \mathbb{R} : a_i > 0\}$, and

$$\rho_i = \frac{\sigma^*_{a_i}\sigma^*_{b_i}\sum_{j=1}^J \theta_j}{\left[\left(\sigma^{*2}_{a_i}\sum_{j=1}^J \theta_j^2 + 1\right)\left(\sigma^{*2}_{b_i}J + 1\right)\right]^{1/2}}, \tag{3.29}$$

$$\sigma^2_{a_i} = \left(\frac{\sigma^{*2}_{a_i}}{\sigma^{*2}_{a_i}\sum_{j=1}^J \theta_j^2 + 1}\right)\frac{1}{(1-\rho_i^2)}, \tag{3.30}$$

$$\sigma^2_{b_i} = \left(\frac{\sigma^{*2}_{b_i}}{\sigma^{*2}_{b_i}J + 1}\right)\frac{1}{(1-\rho_i^2)}, \tag{3.31}$$

$$\mu_{a_i} = \sigma^2_{a_i}\left(\sum_{j=1}^J x_{ij}\theta_j + \mu^*_{a_i}\sigma^{*-2}_{a_i}\right) - \sigma_{a_i}\sigma_{b_i}\rho\left(\sum_{j=1}^J x_{ij} - \mu^*_{b_i}\sigma^{*-2}_{b_i}\right), \tag{3.32}$$

$$\mu_{b_i} = \sigma_{a_i}\sigma_{b_i}\rho_i\left(\sum_{j=1}^J x_{ij}\theta_j + \mu^*_{a_i}\sigma^{*-2}_{a_i}\right) - \sigma^2_{b_i}\left(\sum_{j=1}^J x_{ij} - \mu^*_{b_i}\sigma^{*-2}_{b_i}\right), \tag{3.33}$$

where $\sigma^2_{a_i}$, $\sigma^2_{b_i} \in \mathbb{R}^+$, $\rho_i \in (-1, 1)$, and μ_{a_i}, $\mu_{b_i} \in \mathbb{R}$.

Simulation from (3.26) is performed for each item i by Algorithm 1. The simulation from (3.26) can be computationally expensive as, at each iteration k of the MCMC, it involves the

Cholesky decomposition and the inversion of a matrix of order $J \times J$. Hence, we consider as well the Metropolis-Hastings algorithm to sample each pair (a_i, b_i).

The target density is given by (3.25), and the proposal distribution considered is a random walk which is given by

$$q((a_i \quad b_i)^{*T} | (a_i \quad b_i)^{(k)T}) = f_{N_2}((a_i \quad b_i)^{*T}; (a_i \quad b_i)^{(k)T}, \mathbf{\Omega}_{ab_i}), \tag{3.34}$$

where $f_{N_2}(.; (\mu_a \quad \mu_b)^T, \mathbf{\Omega})$ stands for the bivariate normal density with mean vector $(\mu_1 \quad \mu_2)^T$ and covariance matrix $\mathbf{\Omega} = \begin{pmatrix} \tau_{a_i}^2 & \rho_i \tau_{a_i} \tau_{b_i} \\ \rho_i \tau_{a_i} \tau_{b_i} & \tau_{b_i}^2 \end{pmatrix}$. The quantities $\tau_{a_i}^2$, $\tau_{b_i}^2$ and ρ_i are chosen such that the acceptance ratio is about 0.44.

The acceptance probability is given by

$$\alpha((a_i, b_i)^*, (a_i, b_i)^{(k)}) =$$
$$min \left\{ 1, \frac{\prod_{j=1}^{J} f_{CSN}(x; a_i^* \theta_j - b_i^*, 1, -\gamma_i) f_N(a_i; \mu_{a_i}^*, \sigma_{a_i}^{*2}) \mathbb{I}\{a_i^* > 0\} f_N(b_i^*; \mu_{b_i}^*, \sigma_b^{*2})}{\prod_{j=1}^{J} f_{CSN}(x; a_i^{(k)} \theta_j - b_i^{(k)}, 1, -\gamma_i) f_N(a_i^{(k)}; \mu_{a_i}^*, \sigma_a^{*2}) \mathbb{I}\{a_i^{(k)} > 0\} f_N(b_i^{(k)}; \mu_{b_i}^*, \sigma_b^{*2})} \right\}. \tag{3.35}$$

Although the proposal distribution given in (3.34) will eventually draw negative values for a_i, these candidates are not going to be accepted, as the prior specification for a_i puts probability mass zero on negative values.

3.4.4 Full Conditional Distribution for p

The full conditional distribution of $\boldsymbol{p_i} = (p_{i0}, p_{i1}, p_{i2})$ is given by

$$\begin{aligned} f(p_i|\mathbf{X}, \mathbf{a}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{Z}) &\propto \prod_{i=1}^{I} p_{i0}^{\alpha_0 - 1} p_{i1}^{\alpha_1 - 1} p_{i2}^{\alpha_2 - 1} p_{0i}^{Z_{0i}} p_{1i}^{Z_{1i}} p_{2i}^{Z_{2i}}, \\ &\propto \prod_{i=1}^{I} p_{0i}^{Z_{0i} + \alpha_0 - 1} p_{1i}^{Z_{1i} + \alpha_1 - 1} p_{2i}^{Z_{2i} + \alpha_2 - 1}. \end{aligned} \tag{3.36}$$

Therefore, $\boldsymbol{p_i} \sim Dirichlet(\boldsymbol{\alpha_i})$, where $\boldsymbol{\alpha_i} = (Z_{0i} + \alpha_0, Z_{1i} + \alpha_1, Z_{2i} + \alpha_2)$, and $(\alpha_0, \alpha_1, \alpha_2)$ are hyperparameters assumed to be known.

3.5 Full Conditional Distribution for Z

As discussed in the introduction of Section 3.4, for computational advantages, we treat γ_i as two separate Markov chains, one related to its negative values, γ_{i-}, and one related to its positive values, γ_{i+}. Under this approach, the full conditional distribution of each Z_i is proportional to

$$f(Z_i|\mathbf{X}, \mathbf{a}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \propto f_{i0}^{Z_{0i}} f_{i1}^{Z_{1i}} f_{i2}^{Z_{2i}} p_{0i}^{Z_{0i}} p_{1i}^{Z_{1i}} p_{2i}^{Z_{2i}}, \tag{3.37}$$

where f_{i0}, f_{i1}, and f_{i2} are defined as

$$\prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_i) = \begin{cases} f_{i0}, & \text{for} \quad \gamma_i = 0, \\ f_{i1}, & \text{for} \quad \gamma_i = \gamma_{i-}, \\ f_{i2}, & \text{for} \quad \gamma_i = \gamma_{i+}. \end{cases} \tag{3.38}$$

Rewriting expression (3.37) we obtain

$$f(Z_i|\mathbf{X}, \mathbf{a}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \propto \left(\frac{f_{i0}p_{i0}}{f_s}\right)^{Z_{0i}} \left(\frac{f_{i1}p_{i1}}{f_s}\right)^{Z_{1i}} \left(\frac{f_{i2}p_{i2}}{f_s}\right)^{Z_{2i}}, \tag{3.39}$$

where $f_s = f_{i0}p_{i0} + f_{i1}p_{i1} + f_{i2}p_{i2}$. Therefore, the full conditional distribution of Z_i is

$$Z_i|. \sim Multinomial(1, p_{i0}^*, p_{i1}^*, p_{i2}^*), \tag{3.40}$$

where

$$p_{ik}^* = \frac{f_{ik}p_{ik}}{f_s}, \quad for \quad k = 0, 1, 2.$$

The computational issues related to the sampling of $\mathbf{Z}$ are dealt in Section 3.7.

3.6 Full Conditional Distribution for $\boldsymbol{\gamma_-}$ and $\boldsymbol{\gamma_+}$

We consider a mixture model with for γ_i such

$$\gamma_i = Z_{i0}0 + Z_{i1}\gamma_{i-} + Z_{i2}\gamma_{i+}, \tag{3.41}$$

where γ_{i-} and γ_{i+} represent, respectively, negative and positive values for γ_i. Thus, we construct a Markov chain for γ_{i-} and for γ_{i+} instead of a single chain for γ_i. Since the algebraic calculation for γ_{i-} and for γ_{i+} are suchlike, we define $\gamma_{i\pm}$ to represent both quantities.

It is important to note that by using such sampling scheme, if at iteration k of the MCMC, the model points to, say, a positive value γ_{i+}, the full conditional distribution for γ_{i-} is proportional to its prior distribution. For this reason, it is reasonable to update only the skewness parameter corresponding to the current model, that is, we update γ_{i-} if $Z_{i1} = 1$, and we update γ_{i+} if $Z_{i2} = 1$. If $Z_{i0} = 1$, we update the chain by assuming the symmetric model ($\gamma_i = 0$).

The full conditional distribution of $\gamma_{i\pm}$ is given by

$$\pi(\gamma_{i\pm}|.) = \begin{cases} \prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, 0) f_B(-\gamma_{i-}; \alpha_w, \beta_w, l_l, l_u) f_B(\gamma_{i+}; \alpha_w, \beta_w, l_l, l_u), & \text{if} \quad Z_{i0} = 1, \\ \prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_{i-}) f_B(-\gamma_{i-}; \alpha_w, \beta_w, l_l, l_u) f_B(\gamma_{i+}; \alpha_w, \beta_w, l_l, l_u), & \text{if} \quad Z_{i1} = 1, \\ \prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_{i+}) f_B(-\gamma_{i-}; \alpha_w, \beta_w, l_l, l_u) f_B(\gamma_{i+}; \alpha_w, \beta_w, l_l, l_u), & \text{if} \quad Z_{i2} = 1. \end{cases} \tag{3.42}$$

Since it is difficult to sample directly from this distribution, we use a MH step with Gaussian distribution for the proposal distribution with tuning parameters τ_- (for γ_{i-}) and τ_+ (for γ_{i+}). Thus, the proposal distribution for $\gamma_{i\pm}$ is given by

$$q(\gamma^*_{i\pm}|\gamma^{(k)}_{i\pm}) = \begin{cases} f_N(\gamma^*_{i-}; \gamma^{(k)}_{i-}, \tau^2_-), & \text{if} \quad Z^{(k)}_{i1} = 1, \\ f_N(\gamma^*_{i+}; \gamma^{(k)}_{i+}, \tau^2_+), & \text{if} \quad Z^{(k)}_{i2} = 1. \end{cases} \tag{3.43}$$

In equation 3.43, $\gamma_{i-}^{(k)}$ and $\gamma_{i+}^{(k)}$ denote, respectively, the current values of the chain of γ_{i-} and of γ_{i-}, and γ_{i-}^{*} and γ_{i+}^{*} denote the proposed values for γ_{i-} and γ_{i+}, respectively.

Thus, the acceptance probability for $\gamma_{i\pm}$ is given by

$$\alpha(\gamma_{i\pm}^{*}, \gamma_{i\pm}^{(k)}) = min\left\{1, \frac{\prod_{j=1}^{J} f_{CSN}(x; a_i\theta_j - b_i, 1, -\gamma_{i\pm}^{*})\pi(\gamma_{i\pm}^{*})}{\prod_{j=1}^{J} f_{CSN}(x; a_i\theta_j - b_i, 1, -\gamma_{i\pm}^{(k)})\pi(\gamma_{i\pm}^{(k)})}\right\}. \tag{3.44}$$

3.7 Computational Aspects

We now discuss some relevant computational aspects related to the implementation of the proposed algorithms.

3.7.1 Sampling of X

As described in Section 3.4.1, the full conditional distribution of X_{ij} is a truncated CSN, and it is not possible to sample directly from it. We consider two algorithms to sample from this distribution: a rejection sampling (RS) and an embedded Gibbs Sampler.

Notice that, when $\gamma_i = 0$, sampling from distribution 3.4.1 is equivalent to sample from a truncated normal distribution. This can be done effortless by the inverse c.d.f. method, unless the distribution is restricted to the tail of the density (outside $\mu - 6\sigma, \mu + 6\sigma$). In that case, we use the following accept-reject algorithm, derived by Robert (1995).

Let us denote by $N(\mu, \sigma^2; a^-)$ the normal distribution with left truncation point a^-, from which we wish to obtain samples from. Also, let $Exponential(\alpha, a^-)$ be a translated exponential distribution with density

$$g(z|\alpha, a^-) = \alpha \exp -\alpha(z - a^-)\mathbb{I}(z \geqslant a^-). \tag{3.45}$$

The algorithm is given by

To sample from a CSN distribution, we use the stochastic representation by Henze (1986) given by

Algorithm 2 Truncated Normal Distribution Sampling

1: **do**
2: Generate $\alpha = (a^- + (a^{2-} + 4)^{1/2})/2$;
3: **while** u > p
4: Generate $p = \exp(-(z - \alpha)^2/2)$;
5: Generate $z \sim Exponential(\alpha, a^-)$;
6: Generate $u \sim Uniform(0, 1)$;
7: **return** z;

$$X_{ij} = m_{ij} + \psi_i d_{ij} V_{ij} + \psi_i (1 - d_{ij}^2)^{1/2} W_{ij} - s\gamma_i^{1/3}. \tag{3.46}$$

The RS algorithm to sample from a truncated CSN consists of proposing a candidate from the non-truncated distribution and accepting the candidate if it falls into the region specified by the truncation.

This method may be inefficient as its computational cost depends on the probability of the region of interest, which is the global acceptance probability of the algorithm. It is likely that, for some X_{ij}, at some point of the MCMC, this probability will be very small due to a disparity between the parameter values and the data. For this reason, we also consider an embedded Gibbs Sampler. This method suggests a Gibbs sampler that alternates between the simulation of V and W given the truncation restrictions. The fact that this has dimension 2 and W and V are independent assures very fast convergence of the algorithm. The initial values are sampled from $V_{ij} \sim HN(0, 1)$ and then $W_{ij}|V_{ij} = v_{ij} \sim N(0, 1)\mathbb{I}(w_{ij} > ma_{ij})$ which guarantees that we already start in the region of interest. After a few iterations, say 10, the last simulated value is taken.

3.7.2 Sampling of $\boldsymbol{\theta}$ and (a, b)

As noticed in Section 3.4.2, we can sample from the full conditional distribution of $(\theta_j|.)$ and of $(a_i, b_i|.)$, which results in a fast convergence of the Markov chains. However, these simulations can be computationally expensive. For instance, the sampling of $(\theta_j|.)$, as at each iteration k, involves a Cholesky decomposition and the inversion a matrix of order $I \times I$. And the sampling of (a_i, b_i) involves a Cholesky decomposition and the inversion of a matrix of order $J \times J$. Hence, the cost of the sampling from the full conditional distribution depends on the number of items and students. A solution to optimize time is to resort to a Metropolis-Hastings step, which is very fast, although it demands more Monte Carlo iterations to converge.

To avoid numeric problems on the Metropolis-Hastings step, it is convenient to work with the sum of the logarithm of the densities on the product of the CSN p.d.f.

An important aspect when sampling (a_i, b_i) is the restriction $a_i > 0 \;\; \forall \;\; i$. To do so, we use a RS algorithm, that accepts the proposed value if $a_i > 0$. However, it is important to derive the exact conditional full distribution of (a_i, b_i) since nowadays parallel computing could easily overcome this timing issue.

3.7.3 Sampling of Z

The sampling from the full conditional distribution of Z can result in many numeric errors. The most important problem occurs on the product in J in the expression

$$\prod_{j=1}^{J} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_i) = \begin{cases} f_{i0}, & \text{for} \quad \gamma_i = 0, \\ f_{i1}, & \text{for} \quad \gamma_i = \gamma_{i-}, \\ f_{i2}, & \text{for} \quad \gamma_i = \gamma_{i+}. \end{cases} \tag{3.47}$$

To relieve notation, let us denote, without loss of generality, $f_{i0} = f_0$, $f_{i1} = f_1$, and $f_{i2} = f_2$.

Notice that we have a product of J densities. Thus, the values $f_c, c = 0, 1, 2$ go to zero quickly. The directly application of the logarithm function does not help, as we need that

$$1 = \frac{f_0}{f_0 + f_1 + f_2} + \frac{f_1}{f_0 + f_1 + f_2} + \frac{f_2}{f_0 + f_1 + f_2}, \tag{3.48}$$

in order to the full conditional distribution of Z be a Multinomial.

To solve this numeric issue, we use the identity

$$\frac{a}{a + b + c} = \frac{1}{1 + \frac{b}{a} + \frac{c}{a}}, \tag{3.49}$$

which implies that

$$\begin{aligned}\frac{f_0}{f_0+f_1+f_2} &= \frac{1}{1+\frac{f_1}{f_0}+\frac{f_2}{f_0}} = \frac{1}{1+\frac{\exp\{f_1^*\}}{\exp\{f_0^*\}}+\frac{\exp\{f_2^*\}}{\exp\{f_0^*\}}} \\ &= \frac{1}{1+\exp\{f_1^*-f_0^*\}+\exp\{f_2^*-f_0^*\}}.\end{aligned} \tag{3.50}$$

We then calculate each parcel $f_c, c = 0, 1, 2$ of (3.48) using the result in (3.50), and the numeric problem with (3.47) is solved for high values of J.

Chapter 4

Simulations

In this chapter we investigate the efficiency of the proposed MCMC methodology to estimate the parameters of the CSN-probit model. The simulated data sets vary according to the number of students and items, and also according to the skewness of the item. The abilities were generated from a standard Normal distribution, the discrimination parameters from a Uniform(0.7, 2), and the difficulty parameters from a Uniform(-3, 3). All the MCMC chains run for 20.000 iterations with a burn-in of 5.000. All codes were developed in the OX language.

To sample from the full conditional distribution of $\mathbf{X}$, we use the embedded Gibbs sampler described in Section 3.4.1. The sampling of $\boldsymbol{\theta}$, $(\mathbf{a}, \mathbf{b})$, $\boldsymbol{\gamma}_-$ and $\boldsymbol{\gamma}_+$ were performed via Metropolis-Hastings. Each Z_i is sampled from its full conditional distribution given by (3.40). For this work, we set $\mathbf{p}$ fixed at (1/3 , 1/3 , 1/3). Simulations indicate that the posterior of $\boldsymbol{Z_i}$ is very sensitive to the choice of $\mathbf{p}$.

Also, for all cases in which we sample from the full conditional distribution of $\boldsymbol{\gamma}_-$ and $\boldsymbol{\gamma}_+$, we set the prior Beta(4, 1.2), which puts significant probability mass for values outside the interval (-0.20, 0.20).

We first present simulation results for a set of symmetric items. The sample consists of 10.000 students and 30 items. For the initial values of vector $\boldsymbol{\theta}$, we use the standard score of each student, which consists of calculating the raw score (mean score) of each student, subtracting it by the group average mean, and then dividing it by the group standard deviation. For the initial values of vectors $\mathbf{a}$ and $\mathbf{b}$, we sampled from a Uniform(0.5, 2.5) and from a Uniform(-3,3) distribution, respectively. For prior distributions, we considered $\theta \sim N(0,1), a \sim N(0,2)\mathbb{I}(a > 0)$ and $b \sim N(0,2)$.

Figure 4.1 shows the real values of parameters versus the estimated values (posterior mean). As expected, the symmetric model estimates well ability, discrimination and difficulty parameters. Notice that there is some loss on the precision of estimation on extreme values of ability. This is expected as it would be needed many easy (difficult) items to better estimate these extreme values.

(a) Ability (b) Discrimination (c) Difficulty

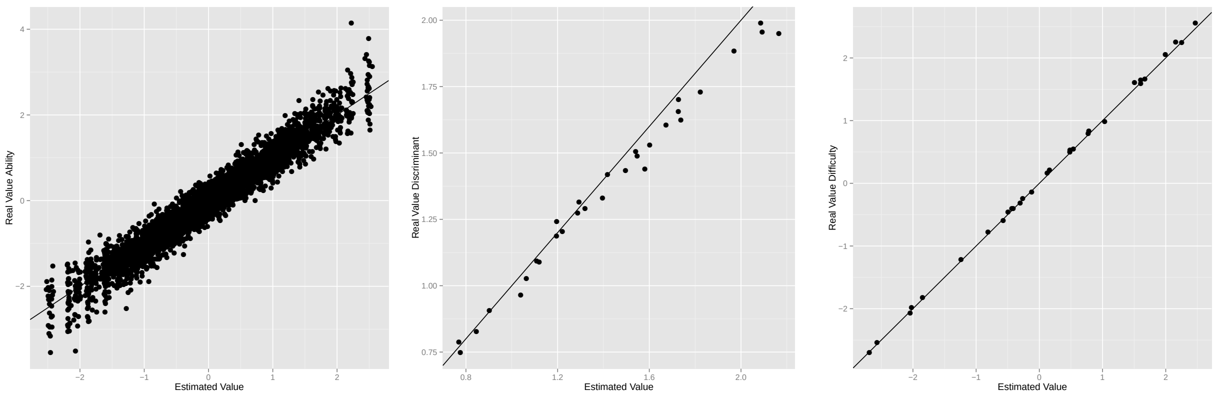

Figure 4.1: Real values versus estimated values (posterior mean). Black lines are $x = y$

Figure 4.2 shows the tracing plots of two items and one examinee. Most chains show fast convergence. Figure 4.3 shows interval estimation (credibility interval) for parameters "a" and "b". Notice that the model was very efficient in the estimation of the item parameters with all credible intervals containing the true value of the parameters, although the credibility interval of the discrimination parameter has higher variability.

Figure 4.4 shows three estimated (solid line) and real (dashed line) ICC. The estimated curves were very close to the real ones. Table 4.1 shows the real and estimated values for discrimination and difficulty parameters for these three ICCs.

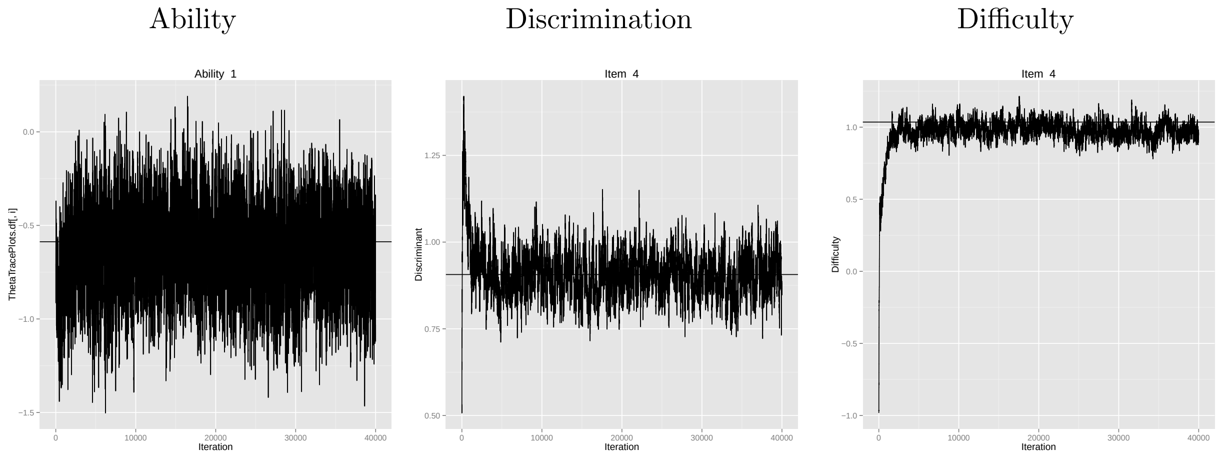

Figure 4.2: Trace Plots for some chains.

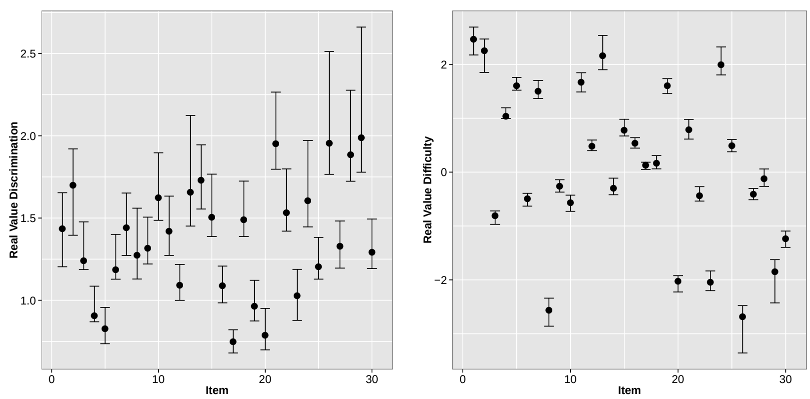

Figure 4.3: Interval estimation (credibility interval) for parameters “a”(left) and “b” (right). Points represent real parameter values and horizontal lines represent the interval estimation.

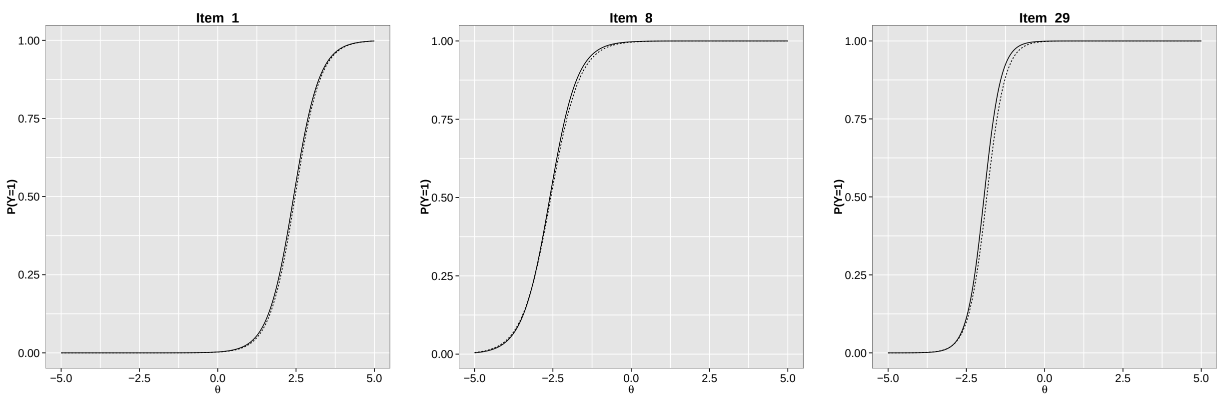

Figure 4.4: Estimated symmetric ICC from some items.

	Discrimination		Difficulty	
Item	**Real**	**Estimated**	**Real**	**Estimated**
1	1.43	1.42	2.47	2.42
8	1.27	1.34	-2.57	-2.60
29	1.99	2.16	-1.85	-1.93

Table 4.1: Real and estimated values for discrimination and difficulty parameters

Next results show the simulation outcome for a sample of 5.000 examinees and 50 items among which three are asymmetric *a prior*, such that items 1 ($\gamma_1 = 0.92$), item 2 ($\gamma_2 = 0.85$) and item 3 ($\gamma_1 = -0.78$). The parameters $\boldsymbol{\theta}$, $\mathbf{a}$ and $\mathbf{b}$ were fixed. Figure 4.5 shows the posterior distribution of $\mathbf{Z}$. The y-axis represents the posterior mean of $\boldsymbol{\gamma}_+$ and of $\boldsymbol{\gamma}_-$, and the size of each bubble represents the mixture posterior probability of Z (only probabilities higher then 0.5). The x-axis represents each item. For illustration purposes, let us take item 30. The item was generated as its skewness parameter being 0 (prior skewness). The negative mixture component had a posterior probability of 0.68, and the estimated value for the negative gamma was -0,23. Notice that the model detects the asymmetric items correctly. The results also indicate skewness behaviour in four other items. Table 4.2 shows the posterior probability of $\mathbf{Z}$ and posterior means of $\boldsymbol{\gamma}$ for the detected skew items.

Item	Posterior Probability	Posterior Mean	Prior Skewness
1	1.00	0.99	0.92
2	1.00	0.96	0.85
3	1.00	-0.99	-0.71
6	0.86	0.25	0
7	0.63	0.21	0
30	0.68	-0.23	0
45	0.89	-0.27	0

Table 4.2: Posterior distribution of $\mathbf{Z}$ and posterior mean of $\boldsymbol{\gamma}$ for detected asymmetric items.

Although some items (symmetric *a priori*) were estimated as asymmetric, the observed data set were better fitted by the asymmetric link, which is supported by the fact that the likelihood of the asymmetric model was higher than the likelihood of the symmetric model when evaluated with the posterior mean of the parameters. It is worthy noticing that the symmetric

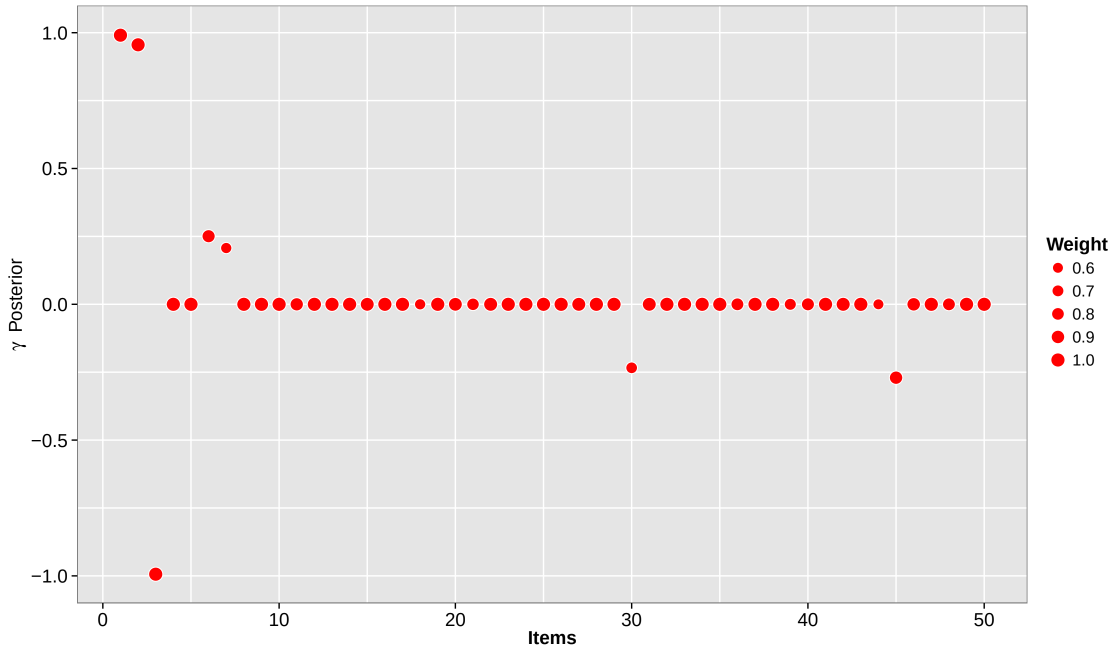

Figure 4.5: Posterior distribution of $\mathbf{Z}$ and posterior mean for $\boldsymbol{\gamma}$. Items asymmetric are: 1 ($\gamma_1 = 0.92$), 2 ($\gamma_2 = 0.85$) and 3 ($\gamma_1 = -0.78$).

and asymmetric probit ICC are similar even for moderated levels of skewness. Additionally, the skewness parameter is known to be hard to be estimated (Arellano-Valle & Azzalini, 2008, see), which is even worse in our case since the SN variables are not observed.

Because of timing issue, the simulation results considering the estimation of all parameter are going to be presented in future works. We acknowledge, however, that a more extensive analysis should be done in order to evaluate possible fragilities and advantages of the proposed methodology.

Chapter 5

Final Remarks

This work proposes a flexible item response model able to accommodate both symmetric and asymmetric item characteristic curves based on the centered skew normal distribution. One of the most important contributions of this work is to consider a finite mixture of Beta distributions and a point mass at zero to describe the uncertainty about the skewness parameter. Consequently, such a strategy also provides an intrinsic methodology for model selection. We offer the full condition distribution of ability, discrimination and difficulty parameters. We also propose efficient algorithms, based on MCMC methods, to sample from the posterior distribution. The presented methodology is as general as the one proposed in Bazán et al. (2006), as both cover symmetric and asymmetric links. However, our model is more parsimonious because items are not all assumed asymmetric *a priori*, which also leads to lower computational cost.

Simulation studies performed well in situations where some parameters were fixed. A more general study where all the parameters are estimated will be reported in future work. We acknowledge that reasonable results rely on properly choices for the prior distributions.

For future work, we intent to do a sensitive analysis of prior specifications of skewness parameter and $\mathbf{p}$. Also, we intent to do a sensitive analysis on the estimation impact of different levels of skewness based on different sample sizes (for both number of examinees and items). Additionally, we intent to sample $\boldsymbol{\theta}$ and ($\mathbf{a}$, $\mathbf{b}$) directly from its full condition distribution, which could be done by parallel computation.

5.1 Full Conditional Distribution for θ

The full model from which we wish to sample is given by expression (3.16) in Chapter 3. From this later expression, the full conditional distribution for each θ_j, conditional on $\mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}$, $\mathbf{a}$ and $\mathbf{b}$ is

$$f(\theta_j|\mathbf{X}, \mathbf{Z}, \mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}) \propto \left[\prod_{i=1}^{I} f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_i)\right] f_N(\theta_j; \mu^*_\theta, \sigma^{*2}_\theta), \tag{5.1}$$

where $m_{ij} = a_i\theta_j - b_i$, μ^*_θ and σ^{*2}_θ are the hyperparameters of the prior of $\boldsymbol{\theta}$, and

$$f_{CSN}(x_{ij}; a_i\theta_j - b_i, 1, -\gamma_i) \propto \phi\left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i}\right)\Phi\left(g(\gamma_i)\left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i}\right)\right). \tag{5.2}$$

where $g(\gamma_i) = \dfrac{s\gamma_i^{1/3}}{\sqrt{r^2 + s^2\gamma_i^{2/3}(r^2 - 1)}}$, $s = \left(\dfrac{2}{4-\pi}\right)^{1/3}$, $r = \sqrt{2/\pi}$, and $\varphi_i = \sqrt{1 + s^2\gamma_i^{2/3}}$.

In order to make the algebraic manipulations of (5.1) easier, let us first solve the product given by

$$\prod_{i=1}^{I} \phi\left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i}\right)\phi\left(\frac{\theta_j - \mu_{\theta^*}}{\sigma^{*2}_\theta}\right) \propto$$

$$\propto \exp\left\{\frac{-1}{2}\sum_{i=1}^{I}\left(\frac{x_{ij} - m_{ij} + s\gamma_i^{1/3}}{\varphi_i}\right)^2\right\} exp\left\{\frac{-1}{2}\frac{(\theta_j - \mu^*_\theta)^2}{\sigma^{*2}_\theta}\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\sum_{i=1}^{I}\frac{1}{\varphi_i^2}\left((x_{ij} - a_i\theta_j)^2 + 2(x_{ij} - a_i\theta_j)(b_i + s\gamma_i^{1/3}) + (b_i + s\gamma_i^{1/3})^2 + \frac{(\theta_j - \mu^*_\theta)^2}{\sigma^{*2}_\theta}\varphi_i^2\right)\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\sum_{i=1}^{I}\frac{1}{\varphi_i^2}\left(-2a_i\theta_j x_{ij} + a_i^2\theta_j^2 + 2(-a_ib_i\theta_j + a_i\theta_j s\gamma_i^{1/3}) + \frac{\theta_j^2}{\sigma^{*2}_\theta}\varphi_i^2 - \frac{2\theta_j\mu^*_\theta}{\sigma^{*2}_\theta}\varphi_i^2\right)\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\left(\theta_j^2\sum_{i=1}^{I}\frac{a_i^2}{\varphi_i^2} + \frac{\theta_j^2}{\sigma^{*2}_\theta} - 2\theta_j\sum_{i=1}^{I}\frac{a_i x_{ij}}{\varphi_i^2} - 2\theta_j\sum_{i=1}^{I}\frac{a_ib_i}{\varphi_i^2} - 2\theta_j\sum_{i=1}^{I}\frac{a_i s\gamma_i^{1/3}}{\varphi_i^2} - \frac{2\theta_j\mu^*_\theta}{\sigma^{*2}_\theta}\right)\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\left(\theta_j^2\left(\sum_{i=1}^{I}\frac{a_i^2}{\varphi_i^2}+\frac{1}{\sigma_\theta^{*2}}\right)-2\theta_j\left(\sum_{i=1}^{I}\frac{a_ix_{ij}+a_ib_i+a_is\gamma_i^{1/3}}{\varphi_i^2}+\frac{\mu_\theta^*}{\sigma_\theta^{*2}}\right)\right)\right\} \tag{5.3}$$

Equation (5.3) is proportional to the univariate normal distribution given by

$$f(\theta|\mu_\theta,\sigma_\theta^2)\propto exp\left\{\frac{-1}{2}\left(\frac{\theta^2}{\sigma_\theta^2}-\frac{2\theta\mu_\theta}{\sigma_\theta^2}+\frac{\mu_\theta^2}{\sigma_\theta^2}\right)\right\}. \tag{5.4}$$

Now, we equate quantities (5.3) and (5.4) such as

$$\begin{aligned}\frac{\theta_j^2}{\sigma_{\theta_j}^2} &= \theta_j^2\left(\sum_{i=1}^{I}\frac{a_i^2}{\varphi_i^2}+\frac{1}{\sigma_\theta^{*2}}\right),\\ \sigma_{\theta_j}^2 &= \frac{\sigma_\theta^{2*}}{\sigma_\theta^{2*}\sum_{i=1}^{I}\frac{a_i^2}{\varphi_i^2}+1}.\end{aligned} \tag{5.5}$$

To derive μ_θ, we equate equations

$$\begin{aligned}-2\theta\frac{\mu_{\theta_j}}{\sigma_\theta^2} &= -2\theta_j\left(\sum_{i=1}^{I}\frac{a_i(x_{ij}+b_i+s\gamma_i^{1/3})}{\varphi_i^2}+\frac{\mu_\theta^*}{\sigma_\theta^{*2}}\right),\\ \mu_{\theta_j} &= \sigma_\theta^2\left(\sum_{i=1}^{I}\frac{a_i(x_{ij}+b_i+s\gamma_i^{1/3})}{\varphi_i^2}+\frac{\mu_\theta^*}{\sigma_\theta^{*2}}\right).\end{aligned} \tag{5.6}$$

Now, for the product of the I c.d.f. in (5.3) we have

$$\begin{aligned}\prod_{i=1}^{I}\Phi\left(g(\gamma_i)\left(\frac{x_{ij}+s\gamma_i^{1/3}-a_i\theta_j+b_i}{\varphi_i}\right)\right) &= \prod_{i=1}^{I}\Phi\left(-\theta_ja_i\frac{g(\gamma_i)}{\varphi_i}+(x_{ij}+s\gamma_i^{1/3}+b_i)\frac{g(\gamma_i)}{\varphi_i}\right)\\ &= \Phi_I(\mathbf{W}\theta_j-\boldsymbol{\eta_j};\mathbf{1}_I).\end{aligned} \tag{5.7}$$

Column vectors $\mathbf{W}$ and $\boldsymbol{\eta_j}$ are

$$\mathbf{W}=\begin{pmatrix}-a_1c_1\\-a_2c_2\\\vdots\\-a_Ic_I\end{pmatrix},\boldsymbol{\eta_j}=\begin{pmatrix}c_i(x_{1j}+s\gamma_1^{1/3}+b_1)\\c_2(x_{2j}+s\gamma_1^{1/3}+b_2)\\\vdots\\c_I(x_{Ij}+s\gamma_I^{1/3}+b_I)\end{pmatrix}, \tag{5.8}$$

where $c=\dfrac{g(\gamma_i)}{\varphi_i}$.

Remark: When $\gamma_i=0$, it follows that $\varphi_i^2=1+s^2\gamma^{2/3}=1$, and we go back in the symmetric case. In this case, the full conditional distribution of each θ_j is

$$\theta_j|\mathbf{X},\mathbf{Z},\mathbf{a},\mathbf{b}\sim N(\mu_{\theta_j},\sigma_\theta^2), \tag{5.9}$$

where

$$\sigma_\theta^2=\frac{\sigma_\theta^{2*}}{\sigma_\theta^{2*}\sum_{i=1}^I a_i^2+1},$$
$$\mu_{\theta_j}=\sigma_\theta^2\left(\sum_{i=1}^I a_i(x_{ij}+b_i)+\frac{\mu_\theta^*}{\sigma_\theta^{*2}}\right).$$

5.2 Full Conditional Distribution for (a, b)

To obtain the full conditional distribution for ($\mathbf{a}$, $\mathbf{b}$), we must notice that, given $\mathbf{X}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$, the pairs (a_i,b_i) are independent. Therefore, we can simulate from each pair (a_i,b_i) separately, and it is equivalent of simulate from the whole vector ($\mathbf{a}$, $\mathbf{b}$).

It follows from (3.25), in Chapter 3, that the full conditional distribution for each (a_i,b_i), conditional on $\mathbf{X},\mathbf{Z},\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ is

$$f(a_i,b_i|\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\propto\prod_{j=1}^J f_{CSN}(x_{ij};m_{ij},1,-\gamma_i)f_N(a_i;\mu_a^*,\sigma_a^{*2})\mathbb{I}\{a_i>0\}f_N(b_i;\mu_b^*,\sigma_b^{2*}), \tag{5.10}$$

where the CSN density is given by

$$f_{CSN}(x_{ij}; m_{ij}, 1, -\gamma_i) \propto \phi\left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i}\right)\Phi\left(g(\gamma_i)\left(\frac{x_{ij} + s\gamma^{1/3} - m_{ij}}{\varphi_i}\right)\right), \qquad (5.11)$$

where $(\gamma_i) = \dfrac{s\gamma_i^{1/3}}{\sqrt{r^2 + s^2\gamma_i^{2/3}(r^2-1)}}$, $s = \left(\dfrac{2}{4-\pi}\right)^{1/3}$, $r = \sqrt{2/\pi}$, and $\varphi_i = \sqrt{1 + s^2\gamma_i^{2/3}}$.

We first derive the product of the J normal p.d.f. in equation (5.10) which is given by

$$\begin{aligned}
&\prod_{i=1}^{I}\phi\left(\frac{x_{ij} + s\gamma_i^{1/3} - m_{ij}}{\varphi_i}\right)\phi\left(\frac{a_i - \mu_a^*}{\sigma_a^{*2}}\right)\phi\left(\frac{b_i - \mu_b^*}{\sigma_b^{*2}}\right) \propto \\
&\propto \exp\left\{\frac{-1}{2\varphi_i^2}\left(\sum_{j=1}^{J}(x_{ij} + s\gamma_i^{1/3} - (a_i\theta_j - b_i))^2\right)\right\}\exp\left\{\frac{-1}{2}\left(\frac{(a_i - \mu_a^*)^2}{\sigma_a^{*2}} + \frac{(b_i - \mu_b^*)^2}{\sigma_b^{2*}}\right)\right\}\mathbb{I}\{a_i > 0\} \\
&\propto \exp\left\{\frac{-1}{2\varphi_i^2}\left(\sum_{j=1}^{J}((x_{ij} - a_i\theta_j) + (b_i + s\gamma_i^{1/3}))^2 + \frac{(a_i - \mu_a^*)^2}{\sigma_a^{*2}}\varphi_i^2 + \frac{(b_i - \mu_b^*)^2}{\sigma_b^{2*}}\varphi_i^2\right)\right\}\mathbb{I}\{a_i > 0\}, \\
&\propto \exp\Bigg\{\frac{-1}{2\varphi_i^2}\Bigg(\sum_{j=1}^{J}(-2x_{ij}a_i\theta_j + 2x_{ij}b_i + a_i^2\theta_j^2 - 2a_ib_i\theta_j - 2a_is\gamma_i^{1/3}\theta_j + b_i^2 + 2b_is\gamma_i^{1/3}) \\
&\qquad\qquad + \frac{(a_i - \mu_a^*)^2}{\sigma_a^{*2}}\varphi_i^2 + \frac{(b_i - \mu_b^*)^2}{\sigma_b^{2*}}\varphi_i^2\Bigg)\Bigg\}\mathbb{I}\{a_i > 0\}.
\end{aligned} \qquad (5.12)$$

After some arrangements, equation (5.12) becomes

$$\begin{aligned}
&\propto \exp\Bigg\{\frac{-1}{2}\Bigg(a_i^2\left(\frac{\sum_{j=1}^{J}\theta_j^2 + \varphi_i^2\sigma_a^{-2*}}{\varphi_i^2}\right) + b_i^2\left(\frac{J + \varphi_i^2\sigma_b^{-2*}}{\varphi_i^2}\right) \\
&- 2a_i\left(\frac{\sum_{j=1}^{J}x_{ij}\theta_j + s\gamma_i^{1/3}\sum_{j=1}^{J}\theta_j + \varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right) - 2b_i\left(\frac{-\sum_{j=1}^{J}x_{ij} - Js\gamma_i^{1/3} + \varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right) \\
&\qquad - 2a_ib_i\left(\frac{\sum_{j=1}^{J}\theta_j}{\varphi_i^2}\right)\Bigg)\Bigg\}\mathbb{I}\{a_i > 0\}.
\end{aligned} \qquad (5.13)$$

Equation (5.13) is proportional to a bivariate normal distribution of the form

$$f(a_i, b_i) \propto \exp\Bigg\{\frac{-1}{2}\Bigg(a^2\Bigg(\frac{1}{\sigma_a^2(1-\rho^2)}\Bigg) + b^2\Bigg(\frac{1}{\sigma_b^2(1-\rho^2)}\Bigg) - 2a\Bigg(\frac{\mu_a}{\sigma_a^2(1-\rho^2)} - \frac{\rho\mu_b}{\sigma_a\sigma_b(1-\rho^2)}\Bigg)$$

$$- 2b\Bigg(\frac{\mu_b}{\sigma_b^2(1-\rho^2)} - \frac{\rho\mu_a}{\sigma_a\sigma_b(1-\rho^2)}\Bigg) - 2ab\Bigg(\frac{\rho}{\sigma_a\sigma_b(1-\rho^2)}\Bigg)\Bigg)\Bigg\}, \quad (5.14)$$

From expressions and (5.13) and (5.14) and equating the proper parts we can derive the quantities ρ, σ_{a_i}, σ_{b_i}, μ_{a_i}, μ_{b_i}

$$\begin{aligned} \frac{a_i^2}{\sigma_{a_i}^2(1-\rho^2)} &= a_i^2\left(\frac{\sum_{j=1}^J \theta_j^2 + \varphi_i^2\sigma_a^{-2*}}{\varphi_i^2}\right) \\ \sigma_{a_i}^2 &= \left(\frac{\sigma_a^{*2}\varphi_i^2}{\sigma_a^{*2}\sum_{j=1}^J \theta_j^2 + \varphi_i^2}\right)\frac{1}{(1-\rho^2)}, \end{aligned} \quad (5.15)$$

$$\begin{aligned} \frac{b_i^2}{\sigma_{b_i}^2(1-\rho^2)} &= b_i^2\left(\frac{J + \varphi_i^2\sigma_b^{-2*}}{\varphi_i^2}\right) \\ \sigma_{b_i}^2 &= \left(\frac{\sigma_b^{2*}\varphi_i^2}{\sigma_b^{2*}J + \varphi_i^2}\right)\frac{1}{(1-\rho^2)}, \end{aligned} \quad (5.16)$$

$$-2a_ib_i\frac{\rho}{(1-\rho^2)\sigma_{a_i}\sigma_{b_i}} = -2a_ib_i\frac{\sum_{j=1}^J \theta_j}{\varphi_i^2}. \quad (5.17)$$

Solving the system of equations (5.15), (5.16), and (5.17) we have that

$$\rho = \frac{\sigma_{a_i}^*\sigma_{b_i}^*\sum_{j=1}^J \theta_j}{\left[\left(\sigma_a^{*2}\sum_{j=1}^J \theta_j^2 + \varphi_i^2\right)(\sigma_b^{2*}J + \varphi_i^2)\right]^{1/2}}. \quad (5.18)$$

The means μ_{a_i} and μ_{b_i} are obtained by solving the following equations:

$$-2a_i\left(\frac{\mu_{a_i}}{\sigma_{a_i}^2(1-\rho^2)}-\frac{\rho}{(1-\rho^2)}\frac{\mu_{b_i}}{\sigma_{a_i}\sigma_{b_i}}\right)=-2a_i\left(\frac{\sum_{j=1}^J x_{ij}\theta_j+s\gamma_i^{1/3}\sum_{j=1}^J\theta_j+\varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right) \tag{5.19}$$

$$-2b_i\left(\frac{\mu_{b_i}}{\sigma_{b_i}^2(1-\rho^2)}-\frac{\rho}{(1-\rho^2)}\frac{\mu_{a_i}}{\sigma_{b_i}\sigma_{a_i}}\right)=-2b_i\left(\frac{-\sum_{j=1}^J x_{ij}-Js\gamma_i^{1/3}+\varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right) \tag{5.20}$$

Equation (5.19) leads to

$$\mu_{a_i}=\frac{\sigma_{a_i}}{\sigma_{b_i}}\rho\mu_{b_i}+\sigma_{a_i}^2(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}\theta_j+s\gamma_i^{1/3}\sum_{j=1}^J\theta_j+\varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right). \tag{5.21}$$

Replacing (5.21) in (5.20), we obtain

$$\begin{aligned}
\mu_{b_i}&=\frac{\sigma_{b_i}}{\sigma_{a_i}}\rho\mu_{a_i}-\sigma_{b_i}^2(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}+Js\gamma_i^{1/3}-\varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right)\\
&=\frac{\sigma_{b_i}}{\sigma_{a_i}}\rho\left(\frac{\sigma_{a_i}}{\sigma_{b_i}}\rho\mu_{b_i}+\sigma_{a_i}^2(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}\theta_j+s\gamma_i^{1/3}\sum_{j=1}^J\theta_j+\varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right)\right)\\
&\qquad-\sigma_{b_i}^2(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}+Js\gamma_i^{1/3}-\varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right)\\
&=\rho^2\mu_{b_i}+\sigma_{a_i}\sigma_{b_i}\rho(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}\theta_j+s\gamma_i^{1/3}\sum_{j=1}^J\theta_j+\varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right)\\
&\qquad-\sigma_{b_i}^2(1-\rho^2)\left(\frac{\sum_{j=1}^J x_{ij}+Js\gamma_i^{1/3}-\varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right)\\
&=\sigma_{a_i}\sigma_{b_i}\rho\left(\frac{\sum_{j=1}^J x_{ij}\theta_j+s\gamma_i^{1/3}\sum_{j=1}^J\theta_j+\varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right)-\sigma_{b_i}^2\left(\frac{\sum_{j=1}^J x_{ij}+Js\gamma_i^{1/3}-\varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right).
\end{aligned} \tag{5.22}$$

Finally, replacing (5.22) in (5.21), we obtain the final expression for μ_{a_i}, which is given by

$$\mu_{a_i} = \frac{\sigma_{a_i}}{\sigma_{b_i}}\rho\Bigg(\sigma_{a_i}\sigma_{b_i}\rho\Bigg(\frac{\sum_{j=1}^{J} x_{ij}\theta_j - s\gamma_i^{1/3}\sum_{j=1}^{J}\theta_j + \varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\Bigg)$$

$$-\sigma_{b_i}^2\Bigg(\frac{\sum_{j=1}^{J} x_{ij} + Js\gamma_i^{1/3} - \varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\Bigg)\Bigg) + \sigma_{a_i}^2(1-\rho^2)\Bigg(\frac{\sum_{j=1}^{J} x_{ij}\theta_j + s\gamma_i^{1/3}\sum_{j=1}^{J}\theta_j + \varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\Bigg)$$

$$= \sigma_{a_i}^2\left(\frac{\sum_{j=1}^{J} x_{ij}\theta_j + s\gamma_i^{1/3}\sum_{j=1}^{J}\theta_j + \varphi_i^2\mu_a^*\sigma_a^{*-2}}{\varphi_i^2}\right) - \sigma_{a_i}\sigma_{b_i}\rho\left(\frac{\sum_{j=1}^{J} x_{ij} + Js\gamma_i^{1/3} - \varphi_i^2\mu_b^*\sigma_b^{*-2}}{\varphi_i^2}\right). \tag{5.23}$$

Now, for the product of the J c.d.f. in (5.10) we have

$$\begin{aligned}\prod_{i=1}^{J}\Phi\left(g(\gamma_i)\left(\frac{x_{ij} + s\gamma_i^{1/3} - a_i\theta_j + b_i}{\varphi_i}\right)\right) &= \prod_{i=1}^{J}\Phi\left(-a_i\theta_j\frac{g(\gamma_i)}{\varphi_i} + b_i\frac{g(\gamma_i)}{\varphi_i} + (x_{ij} + s\gamma_i^{1/3})\frac{g(\gamma_i)}{\varphi_i}\right)\\ &= \Phi_J((a_i \quad b_i)\mathbf{W}_i - \boldsymbol{\eta_i}).\end{aligned} \tag{5.24}$$

The matrix $\boldsymbol{W_i}$ of order $2\times J$ and column vector $\boldsymbol{\eta}_i$ are then

$$\boldsymbol{W_i} = \begin{pmatrix} -\theta_1 c_i & c_i \\ -\theta_2 c_i & c_i \\ \vdots & \vdots \\ -\theta_J c_i & c_i \end{pmatrix}, \boldsymbol{\eta}_i = \begin{pmatrix} c_i(x_{i1} + s\gamma_i^{1/3}) \\ c_2(x_{i2} + s\gamma_i^{1/3}) \\ \vdots \\ c_i(x_{iJ} + s\gamma_i^{1/3}) \end{pmatrix}, \tag{5.25}$$

where $c_i = \dfrac{g(\gamma_i)}{\varphi_i}$.

Remark: When $\gamma_i = 0$, it follows that $\varphi_i^2 = 1 + s^2\gamma_i^{2/3} = 1$, and we turn back to the symmetric case. In this case, the full conditional distribution of each pair (a_i, b_i) is the truncated bivariate normal distribution given by

$$(a_i, b_i) \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta} \sim N_2((a_i \quad b_i)^T; (\mu_{a_i} \quad \mu_{b_i})^T, \Sigma_i; A), \tag{5.26}$$

where $A = (a_i, b_i) \in \mathbb{R} : a_i > 0$,

$$\rho = \frac{\sigma_a^* \sigma_b^* \sum_{j=1}^J \theta_j}{\left[\left(\sigma_a^{*2} \sum_{j=1}^J \theta_j^2 + 1\right)\left(\sigma_b^{2*} J + 1\right)\right]^{1/2}},$$

$$\sigma_a^2 = \left(\frac{\sigma_a^{*2}}{\sigma_a^{*2} \sum_{j=1}^J \theta_j^2 + 1}\right) \frac{1}{(1-\rho^2)},$$

$$\sigma_b^2 = \left(\frac{\sigma_b^{2*}}{\sigma_b^{2*} J + 1}\right) \frac{1}{(1-\rho^2)},$$

$$\mu_{a_i} = \sigma_{a_i}^2 \left(\sum_{j=1}^J x_{ij}\theta_j + \mu_a^* \sigma_a^{*-2}\right) - \sigma_{a_i}\sigma_{b_i}\rho \left(\sum_{j=1}^J x_{ij} - \mu_b^* \sigma_b^{*-2}\right),$$

$$\mu_{b_i} = \sigma_a \sigma_b \rho \left(\sum_{j=1}^J x_{ij}\theta_j + \mu_a^* \sigma_a^{*-2}\right) - \sigma_b^2 \left(\sum_{j=1}^J x_{ij} - \mu_b^* \sigma_b^{*-2}\right).$$

5.3 Generalized Beta Distribution

The skewness parameter γ has its support on (-0.99527, 0.99527). Therefore, we define a more general Beta distribution, which does not restrict its support to the interval (0, 1).

The probability density function of the beta distribution, for $a \leq x \leq b$, and shape parameters α, $\beta > 0$,is given by

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta-1}}, \qquad \alpha, \beta > 0, \tag{5.27}$$

where B(p,q) is the beta function defined by

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt. \tag{5.28}$$

The variance of X is given by

$$Var[X] = (b-a)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{5.29}$$

and the mean is given by

$$E[X] = a + (b-a)\frac{\alpha}{(\alpha+\beta)}. \tag{5.30}$$

The case where a = 0 and b = 1 is called the standard beta distribution. The equation for the standard beta distribution is

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad 0 \le x \le 1; \quad \alpha, \beta > 0. \tag{5.31}$$

For $\alpha, \beta > 1$ it is valid:

$$f(x;\alpha,\beta) = f(1-x;\beta,\alpha). \tag{5.32}$$

Simulation of a generalized beta distribution can be done in terms of the standard beta distribution $F(x)$. First, we generate x from the standard beta distribution. Then we do the linear transformation $y = x(b-a)+a$, where a and b are the lower and upper bounds, respectively, as before.

Bibliography

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

Arellano-Valle, R. B., & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, *33*(3), 561–574.

Arellano-Valle, R. B., & Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis*, *99*(7), 1362–1382.

Arellano-Valle, R. B., Branco, M. D., & Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics*, *34*(4), 581–601.

Arnold, B. C., Beaver, R. J., Azzalini, A., Balakrishnan, N., Bhaumik, A., Dey, D., Cuadras, C., & Sarabia, J. M. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, *11*(1), 7–54.

Azevedo, C. L. (2003). *Métodos de Estimação na Teoria de Resposta ao Item*. Master's thesis, USP.

Azevedo, C. L., Andrade, D. F., & Fox, J.-P. (2012). A bayesian generalized multiple group irt model with model-fit assessment tools. *Computational statistics & data analysis*, *56*(12), 4399 – 4412.
URL http://doc.utwente.nl/83848/

Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2009). A note on identification and metric issues for skew irt models. Tech. rep., USP.

Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal irt model under the centred parameterization. *Computational Statistics & Data Analysis*, *55*(1), 353–365.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, (pp. 171–178).

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*(2), 159–188.

Azzalini, A. (2014). *The R package* `sn`*: The skew-normal and skew-t distributions (version 1.1-2)*. Università di Padova, Italia.
URL `http://azzalini.stat.unipd.it/SN`

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 579–602.

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research*, (pp. 151–178).

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bazan, J. L. (2005). *Uma família de modelos de resposta ao item normal assimétrica*. Ph.D. thesis, Universidade de São Paulo.

Bazán, J. L., Branco, M. D., Bolfarine, H., et al. (2006). A skew item response model. *Bayesian analysis*, *1*(4), 861–892.

Bazán, J. L., Branco, M. D., Bolfarine, H., et al. (2014). Extensions of the skew-normal ogive item response model. *Brazilian Journal of Probability and Statistics*, *28*(1), 1–23.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. *Randolph Air Force Base, Texas: Air University, School of Aviation Medicine*, *26*.

Birnbaum, A. (1958). On the estimation of mental ability. *Randolph Air Force Base, Texas: Air University, School of Aviation Medicine*, *15*.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, (pp. 395–479).

Chen, M., Shao, Q., & Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

de Andrade, D. F., Tavares, H. R., & da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Gonçalves, F. B., & Gamerman, D. (2015). Exact bayesian inference in spatiotemporal cox processes driven by multivariate gaussian processes.

Henze, N. (1986). A probabilistic representation of the'skew-normal'distribution. *Scandinavian journal of statistics*, (pp. 271–275).

Lawley, D. N. (1943). Xxiii.—on problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, *61*(03), 273–287.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin Boston.

Lord, F. (1952). A theory of test scores. *Psychometric monographs*.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, *1*(2), 33–49.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and computing*, *5*(2), 121–125.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, *72*(3), 217–232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*.

Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, *62*(4), 471–493.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*(3), 319–335.

Santos, V. L. F. (2009). *Teoria de Resposta ao Item: uma abordagem generalizada das Curvas Características dos Itens*. Ph.D. thesis, Universidade Federal do Rio de Janeiro.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*(1), 1–13.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory*, (pp. 1–28). Springer.

Wang, J., Boyer, J., & Genton, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica*, *14*(4), 1259–1270.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological measurement*, *29*(1), 23–48.