Measurement and Research Department Reports 2007-3

Accounting for Non-Normality in Latent Regression Models Using a Cumulative Normal Selection Function

Dylan Molenaar

Measurement and Research Department Reports 2007-3

Accounting for Non-Normality in Latent Regression Models Using a Cumulative Normal Selection Function

Masters Thesis in Psychology

Dylan Molenaar

Department of Psychology
University of Amsterdam
and
Psychometric Research Center
Cito

Supervisors
Dr. Norman D. Verhelst
Prof. dr. Han L.J. van der Maas

Cito
Arnhem, 2007

Accounting for non-normality in latent regression models using a cumulative normal selection function.

Dylan Molenaar
University of Amsterdam

Abstract

In educational measurement, group differences in ability are investigated using latent regression models. Commonly, these models assume a normal distribution of the latent ability, conditional on group membership. However, this assumption could be problematic. In the present paper, a new model is discussed in which the latent ability is modelled as if it has arisen from a latent selection process. This will enable the analysis of group differences under a more general family of distributions than the commonly used normal distribution.

1 Introduction

In the field of Educational Measurement, inferences are made about the ability of individuals or groups of individuals. 'Ability' is a broad term that refers to the educational and vocational potential of an individual (Thorndike, 1971. p. 3). Inferences about ability are based on the performances of a subject on a test that purport to measure the ability of interest. To get from the observed test performances (i.e., item scores) to inferences about the ability, psychometric models come into play. In the measurement model, the observed item responses are linked to a latent variable, which represents the ability. Parameters in the measurement model are *item parameters.* They provide information about the difficulty of the items and the degree in which the items can discriminate between individuals that are high in ability and individuals that are low in ability. Next, in the structural model, the latent variable (i.e., the ability) is linked to certain background variables. These are population characteristics that are of interest to the researcher, such as gender, age group, or ethnicity. Parameters in the structural model are *population parameters.* They provide information about the distribution of the ability in the (sub-)population(s). As there is a wide variety of psychometric models available, each with its own measurement and structural model, we focus on a class of models that is embedded within the framework of Item Response Theory (IRT). IRT is commonly used in Educational Measurement (see for instance the fourth edition of Educational Measurement, edited by Brennan, 2006). Specifically, we discuss the measurement model OPLM (One Parameter Logistic Model; Verhelst, Glas, & Verstralen, 1994; Verhelst & Glas, 1995) and the structural model SAUL (Structural Analysis of a Univariate Latent variable; Verhelst & Verstralen, 2002).

1.1 Measurement model: One Parameter Logistic Model

In OPLM, the probability of a specific item response is linked to a latent variable using a logistic regression function. The item parameters describe the shape and location of this so called *item response function.* The underlying ability is considered as a continuous latent variable which could account for the individual differences in the probability of the item scores. Let $P(X_{vi} = x_{vi}|\theta_v)$ denote the probability that respondent v answers x_{vi} to item i given his or her unobserved ability, θ_v. Now, for dichotomous items, the OPLM is defined as follows:

$$P(X_{vi} = x_{vi}|\theta_v) = \frac{\exp[\alpha_i(\theta_v - \beta_i)]}{1 + \exp[\alpha_i(\theta_v - \beta_i)]}, \qquad \text{for } i = 1, ..., k, \tag{1}$$

where α_i denotes the item-discrimination index ($\alpha_i \in \mathbb{N}_0$) and β_i denotes the item-difficulty parameter ($\beta_i \in \mathbb{R}$) of item i. In Eq. (1), the discrimination indices are imputed by hypothesis, leaving the item-difficulty parameters, β, and nuisance parameters, θ, as unknowns. Both sets of parameters could be estimated using Joint Maximum Likelihood (JML; see for instance Verhelst, 1993). However, at this stage, we are not interested in all individual ability estimates. The purpose of the measurement model is to firmly establish item

parameters to get an idea about the measurement properties of the items. Later on, we can use these properties to make inferences about θ. Therefore, we are only interested in estimating β. JML is thus inefficient, as all individual θ have to be estimated, while they are of no interest. Moreover, JML estimates are known to be biased and inconsistent (see Goldstein, 1980; Lord, 1984). A procedure that is able to estimate β independent of θ, is known as Conditional Maximum Likelihood (CML; Andersen, 1972; Verhelst & Glas, 1995). CML estimation in the OPLM is conducted conditional on the observed weighted test score (s_v), i.e.,

$$s_v = \sum_{i=1}^{k} \alpha_i x_{vi}, \qquad \text{for } v = 1, ..., n, \tag{2}$$

Thereby, θ disappears from the estimation formula (see for instance Verhelst, 1993). Using CML has the major advantage that item parameters are independent of θ. This enables estimation of the item parameters of the measurement model separately from the population parameters of the structural model. Separation of measurement model and structural model is desirable, see Verhelst & Verstralen (2002) and Hoijtink (1995). Integrating the measurement model and the structural model (as for instance in Zwinderman, 1991) greatly hampers the statistical testing of the model as a whole. If some restrictions -in either the measurement model or the structural model- do not hold, it is difficult to trace the misspecifications, as both models are totally intertwined. Another advantage is that separating the measurement model from the structural model enables calibration of the latter in a different sample than that of the former. A disadvantage is present as well. Separation of the measurement model and the structural model requires the estimated β_i's from the measurement model to be imputed in the structural model as if they were true parameter values (instead of *estimates* of the true parameters). This causes the standard errors of the estimated β_i's to be neglected, which will result in underestimation of the standard errors in the structural model. This problem is inevitable. However, we assume that the items and the calibration sample are both of such a quality that the standard errors of the β_i's are sufficiently small, so that the effects of neglecting them are minimal.

1.2 Structural model: Structural Analysis of an Univariate Latent variable

We proceed by discussing the structural model. It is supposed that the parameters in the measurement model are firmly established. That is, the model fit is statistically tested to be adequate, see for instance, Verhelst & Glas (1995). As the measurement model is accepted, the parameters are considered as constants. Therefore, we replace β_i with $\widehat{\beta}_i$ in what follows. The researcher is now able to evaluate and test a specific structural model. In the structural model, θ is related to a number of background variables by means of latent regression. At this point, we consider dichotomous background variables only, for the

ease of presentation. A structural model for θ, with r dichotomous background variables, and t interaction effects equals:

$$\boldsymbol{\theta} = \mathbf{Yb} + \mathbf{e} \tag{3}$$

in which $\boldsymbol{\theta}$ is a $(n \times 1)$ column vector of latent ability scores, $\mathbf{b}$ is a $[(r+t+1) \times 1]$ column vector of regression coefficients, $\mathbf{e}$ is a $(n \times 1)$ column vector of residuals, and $\mathbf{Y}$ is the $[n \times (r + t + 1)]$ design matrix. We now consider the case of two dichotomous background variables with interaction (e.g., $r = 2$, and $t = 1$). Notice that in the case of dichotomous background variables, $\mathbf{Y}$ only contains 1's and 0's. Mean θ of the subjects having a 1 on the background variable is estimated relatively to the subjects with a 0 on the background variable (i.e., the reference group). Notice further that for r dichotomous background variables there are r dummy variables for the main effects (i.e., y_{n1} to y_{nr}). Now we can specify $\mathbf{Y}$ and $\mathbf{b}$:

$$\mathbf{Y} = \begin{bmatrix} 1 & y_{11} & y_{12} & y_{11}y_{12} \\ 1 & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots \\ 1 & y_{n1} & y_{n2} & y_{n1}y_{n2} \end{bmatrix}, \tag{4}$$

$$\mathbf{b} = \begin{bmatrix} \upsilon & \delta_1 & \delta_2 & \gamma \end{bmatrix}, \tag{5}$$

where y_{ij} denotes the value of respondent i on background variable j, $y_{n1}y_{n2}$ is the product between y_{n1} and y_{n2}, υ denotes the mean of the reference group, δ_1 is the effect of the first background variable, δ_2 is the effect of the second background variable, and γ is the interaction effect. Models like Eq. (3) could be fitted with the computer program SAUL (Verhelst & Verstralen, 2002). In SAUL it is commonly assumed that

$$\mathbf{e} \sim N(0, \sigma_p^2), \tag{6}$$

that is, the residuals are normally distributed with mean 0, and variance σ_p^2. As a result of Eq. (6), $\boldsymbol{\theta}$, conditional on the background variables, is assumed to be normally distributed, with mean depending on the elements of $\mathbf{b}$, and variance σ_p^2.

The elements of vector $\mathbf{b}$ from Eq. (3)**,** and σ_p^2 from Eq. (6) are the structural parameters, define $\eta = (\mathbf{b}, \sigma_p^2)$. Estimates of these parameters, $\boldsymbol{\eta}$, are obtained by maximizing the likelihood of the data, i.e.

$$L(\boldsymbol{\eta}; \alpha_i, \hat{\beta}_i, \mathbf{X}, \mathbf{Y}) = \sum_{v=1}^{n} \int_{-\infty}^{\infty} P(\mathbf{x}_v|\theta) g(\theta|\mathbf{y}_v; \boldsymbol{\eta}) d\theta, \tag{7}$$

in which $g(\theta|\mathbf{y}_v; \eta)$ denotes the probability density function (pdf) of θ, with distributional parameter vector, $\boldsymbol{\eta}$, conditional on the vector of background values, $\mathbf{y}_v$. In Eq. (7), $\mathbf{X}$ is a matrix that contains the item scores and $P(\mathbf{x}_v|\theta)$ denotes the distribution of the item scores under the measurement model. The

procedure of maximizing Eq. (7) is known as Marginal Maximum Likelihood estimation (MML; e.g., Bock & Aitkin, 1981; Thissen, 1982; Verhelst & Eggen, 1989). As Zwinderman (1991) states, the advantage of MML is that we can regress θ on one or more background variables without the necessity of estimating all individual θ_v. However, we have to specify $g(\theta)$, the distribution of θ, which is not straightforward. We will elaborate on this in the next section.

1.3 The ability distribution

Estimating population parameters using MML requires the specification of an ability distribution. Commonly, this distribution is chosen to be the normal distribution (see among many: Bock & Aitkin, 1981; Bock & Lieberman, 1970; Mislevy, 1984; Rigdon & Tsutakawa, 1983; Thissen, 1982; Zwinderman, 1991). Undoubtedly, this will be an accurate decision in a wide range of settings. However, the possibility of alternative distributions should not be dismissed.

1.3.1 Problems with the normal distribution

The use of a normal distribution within MML is already questioned in many papers (e.g., van den Oord, 2005; Woods, 2006, Zwinderman & van der Wollenberg, 1990; Nandakumar & Yu, 1996; Stone, 1992). It has been shown that when the distribution of the latent variable deviates from a normal one, noticeable parameter differences result with MML (Zwinderman & van der Wollenberg, 1990; Nandakumar & Yu, 1996; Stone, 1992). In our approach of separating the measurement model form the structural model, the item parameters could not be biased by false assumptions about the ability distribution. However, bias in the population parameters could be present. In the light of Educational Measurement, this has major implications, as individual ability inferences could be systematically wrong. In addition, other research fields are benefitted with the possibility to fit non-normal latent variables as well. Consider the research into the so called Flynn effect (i.e., intergenerational increase in mean IQ). This effect is hypothesized as originating from a change in the general intelligence distribution over years. Other fields that deal with non-normal latent variables are psychopathology and personality. In general, all fields of research that use statistical models to explore or confirm theoretical knowledge can benefit from the possibility of non-normal latent variable(s). The normal distribution is a basic assumption in several statistical models. Testing whether this assumption is statisfied requires a more general model than the normal one. However, these models are often not (yet) available. Therefore, models incorporating a more general family of distributions, including the normal ones, is valuable in the field of statistical testing.

1.3.2 Alternatives

There are already some possibilities to model a non-normal ability distribution. Anderson and Madsen (1977) argued that the ability distribution could be any

existing distribution, and illustrated this by specifying a log-beta ability distribution. However, the disadvantage of using an existing distribution, is that -as the ability is not directly observed- no clues are available about which one to choose. A class of distributions that does not suffer from this problem is known as the non-parametric distribution family (De Leeuw & Verhelst, 1986; Follman, 1988; Bock & Aitkin; 1981). In this approach, the ability distribution is estimated from the data. The approach suffers from a couple of shortcomings. First, many parameters are needed to model the distribution, and second, sample fluctuations will make the ability distribution sample specific (van den Oord, 2005). Another problem is that the resulting distribution is not unique. Multiple solutions are possible, which result in different distributions with the same likelihood of the observed data. Alternative methods to approximate the ability distribution use B-splines (Woods, 2006) and Johnson Curves (van den Oord, 2005). However, both methods suffer from large standard errors of the parameter estimates, and have difficulties with highly non-normal ability distributions.

In the present paper, it is explored whether an alternative approach is better able to approximate the ability distribution. We will treat the ability distribution as if it has arisen from a so-called *latent selection* process. In latent selection, a selection function is used that models the probability to be selected from a population in which the ability is normally distributed. A flexible class of distributions arises which includes the normal distribution as a special case. How a selection function is used, and how the different distributions arise, is explained in Section 2. In Section 3 and 4, parameter estimation is discussed. We use both the Expectation Maximization algorithm (Section 3) and the Newton-Raphson algorithm (Section 4) to come to MML estimates of the parameters. In Section 5, SAUL and the latent selection model are compared, and in Section 6 the model is discussed.

2 The general model

2.1 Using a selection function

The ability distribution could have a thinner upper or lower tail than that of a normal distribution. To account for this non-normality in statistical models, a selection function is particularly useful. Literally, a selection function models the probability that a subject with a given ability is selected from an ideal population -in which the ability is normally distributed- into an actual population. This so called *latent selection process* results in a non-normal distribution in the actual population. Notice that latent selection is only a tool to account for non-normality, it is not postulated that any real selection process caused the non-normal nature of the distributions under consideration. In Section 6, we will address this in more detail. Using latent selection, we can account for a thinner lower tail of an ability distribution. In this case, the selection function

should be increasing over the range of the ability, with small values at the lower tail and high values at the upper tail of the normal distribution in the ideal population. A subject with a low ability has a small probability of being selected into the actual population, and a subject with a high ability has a large probability of being selected. As a result, relatively more high-ability subjects will be incorporated in the actual population, resulting in a distribution that is skewed to the right. Likewise, a distribution that is skewed to the left (i.e., a lower upper tail) is obtained by specifying a decreasing selection function.

Let us consider the probability density in the actual population:

$$p(\theta) \propto s(\theta) \times g(\theta), \tag{8}$$

where $s(.)$ denotes the selection function and $g(.)$ denotes the normal probability density function of the ideal population. Both are a function of the ability, θ. The right-hand side of Eq. (8) is not a probability density function itself. To make it a probability density function, Eq. (8) is normalized, resulting in:

$$p(\theta) = \frac{s(\theta) \times g(\theta)}{\int s(\theta) \times g(\theta) d\theta}. \tag{9}$$

Let us now consider $s(\theta)$ in more detail. In fact, $s(\theta)$ could be any function that satisfy the condition: $s(\theta) \in [0, 1]$. A logistic function seems a satisfactory choice therefore. Logistic functions are widely used to model probability in -for instance- IRT and logistic regression models. We choose, however, for a cumulative normal selection function, for the following reasons: Using a cumulative normal selection function, implies that $s(\theta)$ in Eq. (9) is substituted with the normal distribution function. This results in a distribution that is defined as the normalized product of a normal density and a normal distribution function. This family of distributions is known as the *skew-normal distribution* (Azzalini, 1985;1986). Using this class of distributions is thus appealing as it is theoretically and technically well grounded (e.g., Azzalini, 1985, 1986; Azzalini & Capitanio, 1999). Another advantage of the skew-normal distribution is that many elegant properties of the normal distribution still apply (Arnold & Beaver, 2002). For instance, if a joint density is defined by the product of a cumulative normal selection function and a normal distribution, it has marginal densities of the same type (Azzalini & Dalla Valle, 1996). In addition, the use of a cumulative normal selection function enables the generalization to multivariate latent selection, as literature has accumulated around the multivariate skew-normal distribution (Azzalini & Dalla Valle, 1996). A multivariate logistic selection function is less appealing.

The cumulative normal selection function that we will use is defined as follows:

$$s(\theta) = \Phi\left(\lambda_0 + \lambda_1 \theta\right), \tag{10}$$

where $\Phi(.)$ denote

$$\Phi\left(\lambda_0 + \lambda_1\theta\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda_0+\lambda_1\theta} \exp(-\frac{1}{2}z^2)dz. \tag{11}$$

From Eq. (11), two selection parameters arise. The first parameter, λ_1, is a scale parameter that determines the slope of the selection function. The algebraic sign determines whether the selection function is increasing or decreasing. Notice that an increasing selection function (that is, $\lambda_1 > 0$) results in a distribution that is skewed to the right, and that a decreasing selection function ($\lambda_1 < 0$) results in a distribution that is skewed to the left. In the case that $\lambda_1 = 0$, the selection function is constant, and every subject has an equal probability (0.5) to be selected into the actual population. In that case, the resulting distribution equals the normal distribution of the ideal population (i.e., there is no selection). The second selection parameter, λ_0, is an intercept. To gain more insight into this parameter, we write for $\lambda_1 \neq 0$:

$$\Phi\left(\lambda_0 + \lambda_1\theta\right) = \Phi\left[\lambda_1 \times \left(\frac{\theta}{\lambda_1} + \frac{\lambda_0}{\lambda_1}\right)\right].$$

Now it is clear that a subject with ability $\theta = -\frac{\lambda_0}{\lambda_1}$, has an probability of 0.5 of being selected into the actual population. Notice that, in case of an increasing selection function, the higher λ_0, the less low-ability subjects are selected. See Figure (1) for some examples.

In Eq. (9), $s(\theta)$ is substituted with Eq. (10):

$$p(\theta) = \frac{\Phi\left(\lambda_0 + \lambda_1\theta\right) \times g(\theta)}{\int_{-\infty}^{\infty} \Phi\left(\lambda_0 + \lambda_1\theta\right) \times g(\theta)d\theta}, \tag{12}$$

where

$$g(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(\theta - \mu_p\right)^2}{2\sigma_p^2}\right]. \tag{13}$$

From Eq. (13), two more parameters arise. These concern μ_p and σ_p^2, which denote the mean and variance of the ability distribution in the ideal population respectively. These parameters should be interpreted with care, as they do not provide information about the mean and variance of the ability distribution *in the actual population*. Measures for these characteristics are discussed in chapter 5.

2.2 The normalizing constant

The denominator of Eq. (12) is the normalizing constant which we will study in more detail in this section. It can be shown that for the denominator, it holds that

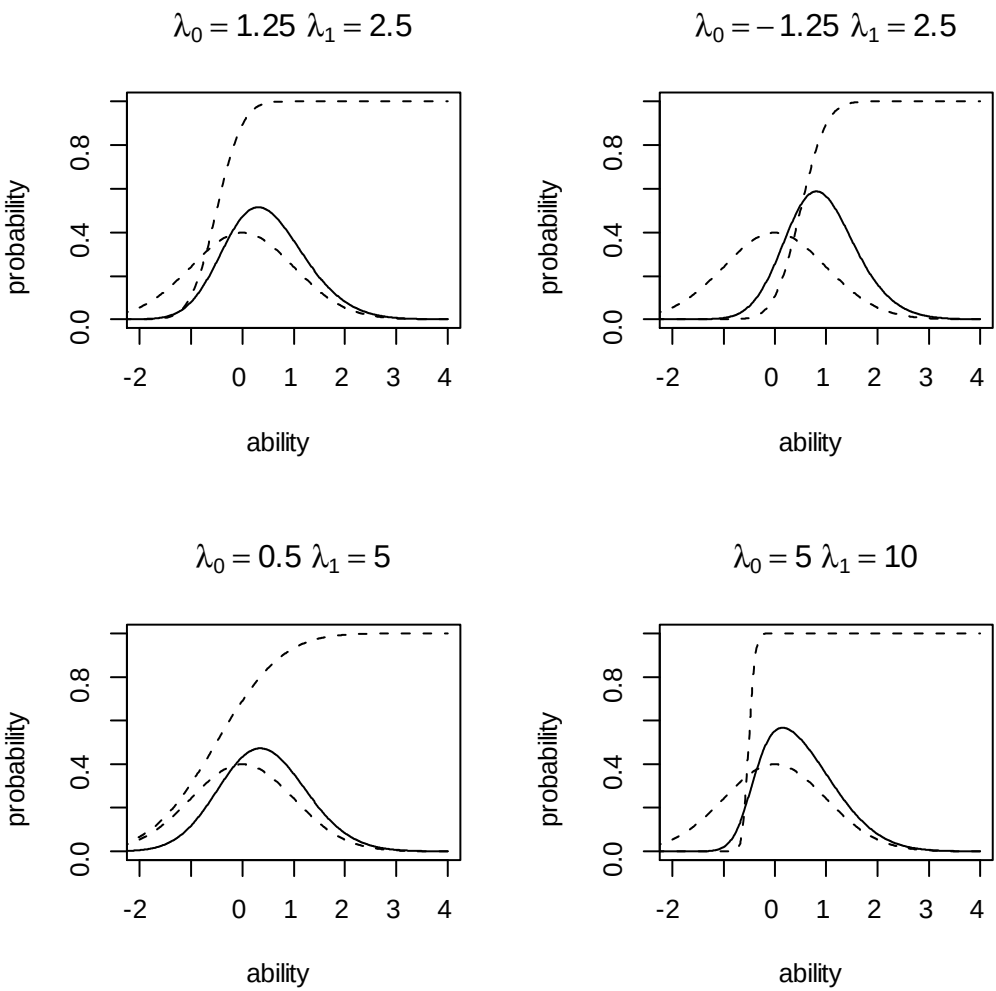

Figure 1: Some skew-normal distribution functions for varying values of the selection parameters. The dotted lines represent the selection function and the normal distribution of the target population. The solid line represents the resulting distribution in the working sample. Population parameters are not varied, and fixed to $\mu_p = 0$ and $\sigma_p^2 = 1$.

$$\int_{-\infty}^{\infty} \Phi\left[\lambda_0 + \lambda_1\theta\right] \times g(\theta)d\theta = \Phi\left(\frac{\lambda_0 + \lambda_1\mu_p}{\sqrt{1+\lambda_1^2\sigma_p^2}}\right). \tag{14}$$

Although the equality of Eq. (14) is widely used, we did not encounter any formal proof in the literature. Therefore, we provide it here. We start by a transformation

$$t = \frac{\theta - \mu_p}{\sigma_p}$$

Then, the left-hand side of Eq. (14) can be written as:

$$\int_{-\infty}^{\infty} \Phi\left[\lambda_0 + \lambda_1\theta\right] \times g(\theta)d\theta = \int_{-\infty}^{\infty} \Phi\left[\lambda_0 + \lambda_1\left(t\sigma_p + \mu_p\right)\right] \times \varphi(t)dt, \tag{15}$$

where $\varphi(.)$ denotes the standard normal pdf. Now we specify

$$a = \lambda_0 + \lambda_1\mu_p \text{ and } b = \lambda_1\sigma_p.$$

Thus we show that

$$\int_{-\infty}^{\infty} \Phi\left(a + bt\right) \times \varphi(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{a+bt} \exp(-\frac{1}{2}z^2)\exp\left[-\frac{1}{2}t^2\right]dzdt. \tag{16}$$

We rotate the z and t axes over an angle α, in such a way that the new t axis (which we call q) is parallel to the line $z = a + bt$. See Figure (2).

We use the following rotation:

$$\begin{bmatrix} q \\ r \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \times \begin{bmatrix} t \\ z \end{bmatrix}.$$

From this, it follows that:

$$z = q\sin\alpha + r\cos\alpha, \tag{17}$$

$$t = q\cos\alpha - r\sin\alpha. \tag{18}$$

Notice that

$$b = \tan\alpha = \frac{\sin\alpha}{\cos\alpha}, \tag{19}$$

$$\cos\alpha = \frac{1}{\sqrt{1+b^2}}. \tag{20}$$

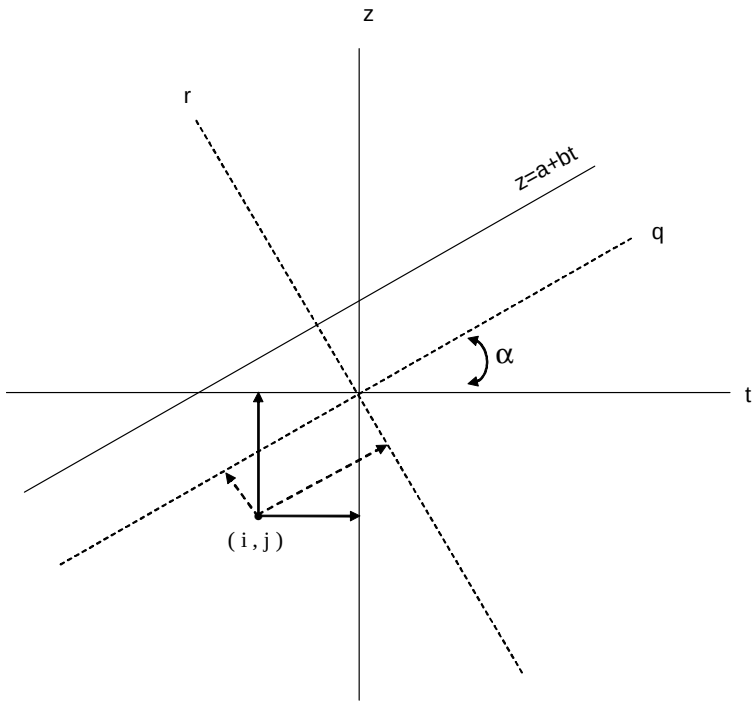

Figure 2: Orthogonal rotation of the z and t-axis. The line $z = a + bt$ is parallel to the q-axis, and (i, j) is an arbitrary point in respectively the (t, z)-plane (solid lines) and the (q, r)-plane (dotted lines).

Now, we use these results to rewrite the double integral in the right-hand side of Eq. (16). First, we rewrite the upper limit of the second integral, using Eq. (17), Eq. (18) and Eq. (19). Thus, $z = a + bt$ is rewritten as

$$q \sin\alpha + r\cos\alpha = a + \frac{\sin\alpha}{\cos\alpha}(q\cos\alpha - r\sin\alpha).$$

This simplifies to:

$$\begin{aligned} r\cos\alpha &= a + \frac{\sin\alpha}{\cos\alpha}(q\cos\alpha - r\sin\alpha) - q\sin\alpha, \\ r\cos\alpha &= a - \frac{r\sin^2\alpha}{\cos\alpha}, \\ r\left(\cos\alpha + \frac{\sin^2\alpha}{\cos\alpha}\right) &= a, \\ r(\cos^2\alpha + \sin^2\alpha) &= a\cos\alpha, \\ r &= a\cos\alpha. \end{aligned}$$

Using Eq. (20) we can write:

$$a\cos\alpha = \frac{a}{\sqrt{1+b^2}}.$$

As a result, the right-hand side of Eq. (16) changes to:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\frac{a}{\sqrt{1+b^2}}} \exp(-\frac{1}{2}r^2)\exp\left[-\frac{1}{2}q^2\right] drdq.$$

Notice that both integrals are now independent:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}q^2\right]dq \times \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{a}{\sqrt{1+b^2}}}\exp(-\frac{1}{2}r^2)dr.$$

The infinite integral equals 1 as it is the area under a standard normal probability density function. We thus have

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{a}{\sqrt{1+b^2}}}\exp(-\frac{1}{2}r^2)dr = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right).$$

Transforming back to λ_0 and λ_1, we get

$$\Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\lambda_0 + \lambda_1\mu_p}{\sqrt{1+\lambda_1^2\sigma_p^2}}\right). \tag{21}$$

2.3 A reparameterisation

The model defined by (12) and (11) is used in many scientific work, e.g., in finite mixture modelling (Lin, Lee & Yen, preprint), in truncation models (Arnold, Beaver, Groeneveld & Meeker, 1993; Arnold & Beaver, 2002) and in Bayesian statistics (Liseo & Loperfidob, 2003). We will adopt, however, a somewhat different parameterisation, to improve the interpretation of the selection parameters, and to embellish the formulas. To obtain the new parameterisation we define

$$\lambda_1 = \frac{1}{\sigma_s},$$

$$\lambda_0 = -\frac{\mu_p}{\sigma_s}.$$

Substituting in Eq. (12) and Eq. (21) we get

$$p(\theta) = \frac{\Phi\left[\frac{\theta-\mu_s}{\sigma_s}\right] \times g(\theta)}{\Phi\left(\frac{\mu_p-\mu_s}{\sqrt{\sigma_p^2+\sigma_s^2}}\right).} \tag{22}$$

The re-parameterisation has the major advantage that the selection parameters are explicitly defined as the mean and standard deviation of the normal probability function. However, there seems to be a disadvantage of the present model, as the parameter σ_s has parameter space, $[0, \longrightarrow)$. As a result, decreasing selection functions could strictly not be fitted. However, some minor alteration of the data will evade this shortcoming, as is discussed in chapter 5.

3 Expectation Maximization

3.1 Introduction to the EM-algorithm

As is discussed in Section 1, parameters from the ability distribution are acquired by maximizing the marginal maximum likelihood function [see Eq. (7)]. Maximizing this function is equivalent to maximizing the logarithm of this function. Maximizing the loglikelihood is, however, not straightforward, as θ is unobservable. Expectation Maximization (EM) is an iterative procedure that maximizes the loglikelihood in the presence of missing observations (Demster, Laird & Rubin, 1977). Treating the individual θ's as unobserved, EM is a particular useful procedure to estimate the parameters from the ability distribution.

In EM, two sets of parameters are used. We denoted these as $\widetilde{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}$. A single EM-iteration consist of two steps: the E-step and the M-step. In the E-step, the first set of parameters, $\widetilde{\eta}$, is used to calculate the posterior

distribution of θ given the data. In the M-step, the Q-function, which will be specified later on, is maximized with respect to the second set of parameters, η. In the next EM–iteration, $\widetilde{\boldsymbol{\eta}}$ is replaced with $\boldsymbol{\eta}$, and new values for $\boldsymbol{\eta}$ are estimated. These two steps are repeated until a convergence criterion is met. The speed with which the EM-algorithm converges depends on the amount of information in the missing data. Demster, Laird & Rubin state that: "If the information loss due to incompleteness is small, then the algorithm converges rapidly " (p.11). As there are as many $\theta's$ missing as there are subjects, the loss of information can be expected to be large. In general, the EM-algorithm converges particularly slow as it gets near the solution. Therefore, the number of iterations is no indication of the stability of the solution. Even after hundreds of EM-iterations, the solution could still be far away. It is thus not a good idea to use EM to acquire Maximum Likelihood estimates, as a convergence criterion is hard to choose. Therefore, we use the EM-algorithm only to acquire starting values for the Newton-Raphson algorithm (which is discussed in Section 4).

As stated above, the EM-algorithm maximizes the loglikelihood in the presence of missing observations. As the θ's are considered as missing data, EM could be used to maximize the loglikelihood of the response vector x_v of person v. For sake of simplicity, we write the loglikelihood for a single observation as follows:

$$\ln L_v = \ln \int_{-\infty}^{\infty} f(x_v|\theta)k(\theta)d\theta, \tag{23}$$

where $f(x_v|\theta)$ denotes the probability distribution function under the measurement model, and $k(\theta)$ denotes the probability density function under the structural model. To obtain the loglikelihood for the whole sample, the right-hand side of Eq. (23) is summed over all subjects:

$$\ln L = \sum_v \ln \int_{-\infty}^{\infty} f(x_v|\theta)k(\theta)d\theta. \tag{24}$$

From now, we denote functions that are evaluated at the current parameter estimates ($\widetilde{\boldsymbol{\eta}}$), by the function symbol with a tilde.

Maximizing Eq.(24) is equivalent to maximizing the ratio

$$\ln \frac{L}{\widetilde{L}}, \tag{25}$$

because $\widetilde{L}$ is a constant as it is evaluated at the current parameter estimates. If we substitute the likelihood function we obtain

$$\ln \frac{L}{\widetilde{L}} = \sum_v \ln \frac{\int_{-\infty}^{\infty} f(x_v|\theta)k(\theta)d\theta}{\int_{-\infty}^{\infty} f(x_v|\theta)\widetilde{k}(\theta)d\theta}.$$

Now, we multiply the numerator by $\widetilde{k}(\theta)/\widetilde{k}(\theta)$:

$$\ln \frac{L}{\widetilde{L}} = \sum_v \ln \frac{\int_{-\infty}^{\infty} \frac{k(\theta)}{\widetilde{k}(\theta)} f(x_v|\theta)\widetilde{k}(\theta)d\theta}{\int_{-\infty}^{\infty} f(x_v|\theta)\widetilde{k}(\theta)d\theta}. \tag{26}$$

The posterior density of θ given x_v is defined by

$$\widetilde{h}(\theta|x_v) = \frac{f(x_v|\theta)\widetilde{k}(\theta)}{\int_{-\infty}^{\infty} f(x_v|\theta)\widetilde{k}(\theta)d\theta}, \tag{27}$$

leaving

$$\ln \frac{L}{\widetilde{L}} = \sum_v \ln \int_{-\infty}^{\infty} \frac{k(\theta)}{\widetilde{k}(\theta)} \widetilde{h}(\theta|x_v)d\theta. \tag{28}$$

Important to notice is that the right-hand side of Eq. (28) equals a conditional expected value in the posterior distribution of θ. Using Jensens inequality, we can write:

$$\ln \frac{L}{\widetilde{L}} \geq \sum_v \int_{-\infty}^{\infty} \ln \frac{k(\theta)}{\widetilde{k}(\theta)} \widetilde{h}(\theta|x_v)d\theta. \tag{29}$$

As a result, we can force the loglikelihood function to increase, by maximizing the right-hand side of Eq. (29), which can be written as

$$\sum_v \int_{-\infty}^{\infty} \ln \frac{k(\theta)}{\widetilde{k}(\theta)} \widetilde{h}(\theta|x_v)d\theta = \sum_v \int_{-\infty}^{\infty} \ln k(\theta)\widetilde{h}(\theta|x_v)d\theta - \sum_v \int_{-\infty}^{\infty} \ln \widetilde{k}(\theta)\widetilde{h}(\theta|x_v)d\theta. \tag{30}$$

As the second term at the right-hand side of Eq. (30) is only function of the current estimates, it is considered as constant. Therefore, the EM-algorithm is about maximizing the Q-function:

$$Q(\boldsymbol{\eta}) = \sum_v \int_{-\infty}^{\infty} \ln k(\theta)\widetilde{h}(\theta|x_v)d\theta. \tag{31}$$

Herein, only $\ln k(\theta)$ is function of the parameters, as $\widetilde{h}(\theta|x_v)$ is evaluated at the know parameters.

3.2 EM for skew-normal latent distributions

Let us specify $\ln k(\theta)$ in the latent selection model. The probability density function $k(\theta)$ is given by Eq. (22), thus

$$\ln k(\theta) = -\ln \Phi(\Lambda) + \ln \Phi\left[\lambda(\theta)\right] + \ln g(\theta), \tag{32}$$

$$= -\ln \Phi(\Lambda) + \ln \Phi[\lambda(\theta)] - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_p^2) - \frac{1}{2}\frac{(\theta-\mu_p)^2}{\sigma_p^2},$$

where

$$\Lambda = \frac{\mu_p - \mu_s}{\sqrt{\sigma_p^2 + \sigma_s^2}},$$

and

$$\lambda(\theta) = \frac{\theta - \mu_s}{\sigma_s}.$$

3.2.1 Estimating the parameters from the model

To obtain parameter estimates, the Q-function [see Eq. (31)] is maximized by setting its first order derivatives with respect to η equal to 0, and solving the resulting equations. As $\widetilde{h}(\theta|x_v)$ is evaluated at $\widetilde{\eta}$, it is a constant. Derivatives of the Q-function are thus given by

$$\frac{d}{d\eta}Q(\boldsymbol{\eta}) = \sum_v \int_{-\infty}^{\infty} \frac{d\ln k(\theta)}{d\eta}\widetilde{h}(\theta|x_v)d\theta. \tag{33}$$

Only derivatives of $\ln k(\theta)$ [see Eq. (32)] are needed. At this point, it is important to note that we take derivatives to the logarithm of the variances. The first reason is to omit computational problems, i.e., zero or negative variances or extremely small numbers in the denominator of a fraction. A second reason is that ML assumes the parameter estimates to be asymptotically normal distributed. This is very unlikely for variances as they are bounded by zero. Taking derivatives to the logarithm of a variance corrects for this skewed distribution.

First order derivatives of $\ln k(\theta)$ are now given by:

$$\frac{\partial \ln k(\theta)}{\partial \mu_s} = \frac{1}{N^{\frac{1}{2}}}R(\Lambda) - R[\lambda(\theta)]\frac{1}{\sigma_s}, \tag{34}$$

$$\frac{\partial \ln k(\theta)}{\partial \ln \sigma_s^2} = \frac{1}{2}(1-\rho)\Lambda R(\Lambda) - \frac{1}{2}R[\lambda(\theta)]\frac{1}{\sigma_s}\left(\theta - \mu_s\right), \tag{35}$$

$$\frac{\partial \ln k(\theta)}{\partial \ln \sigma_p^2} = \frac{1}{2}\rho\Lambda R(\Lambda) - \frac{1}{2} + \frac{1}{2}\frac{(\theta-\mu_p)^2}{\sigma_p^2}, \tag{36}$$

$$\frac{\partial \ln k(\theta)}{\partial \mu_p} = -\frac{1}{N^{\frac{1}{2}}} R(\Lambda) + \frac{(\theta - \mu_p)}{\sigma_p^2}, \tag{37}$$

where

$$R(x) = \frac{\varphi(x)}{\Phi(x)}, \tag{38}$$

$$N = \sigma_p^2 + \sigma_s^2,$$

and ρ is defined as the interclass correlation:

$$\rho = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_s^2}.$$

Notice that it is sufficient to take derivatives to the four basic parameters only. The regression parameters [i.e., $\mathbf{b}$ from Eq. (3)] are additive, meaning that they all have derivatives equal to that of μ_p.

After substituting Eq. (34) to (37) in Eq. (33) and performing some algebraic manipulation, we obtain the derivatives of the Q-function:

$$\frac{\partial Q}{\partial \mu_s} = R(\Lambda)\frac{1}{N^{\frac{1}{2}}} - \frac{1}{\sigma_s} A = 0, \tag{39}$$

$$\frac{\partial Q}{\partial \ln \sigma_s^2} = R(\Lambda)\frac{1}{2}(1-\rho)\Lambda - \frac{1}{2}\frac{1}{\sigma_s} B + \frac{1}{2}\frac{\mu_s}{\sigma_s} A = 0, \tag{40}$$

$$\frac{\partial Q}{\partial \ln \sigma_p^2} = R(\Lambda)\frac{1}{2}\rho\Lambda - \frac{1}{2} + \frac{1}{2\sigma_p^2} D = 0, \tag{41}$$

$$\frac{\partial Q}{\partial \mu_p} = -R(\Lambda)\frac{1}{N^{\frac{1}{2}}} + \frac{1}{\sigma_p^2} C - \frac{\mu_p}{\sigma_p^2} = 0, \tag{42}$$

where

$$A \equiv A(\sigma_s^2, \mu_s) = \frac{1}{n}\sum_v \int_{-\infty}^{\infty} R[\lambda(\theta)]\widetilde{h}(\theta|x_v)d\theta = \frac{1}{n}\sum_v \widetilde{E}[R(\lambda(\theta))|x_v], \tag{43}$$

$$B \equiv B(\sigma_s^2, \mu_s) = \frac{1}{n}\sum_v \int_{-\infty}^{\infty} \theta R[\lambda(\theta)]\widetilde{h}(\theta|x_v)d\theta = \frac{1}{n}\sum \widetilde{E}[\theta R[\lambda(\theta)]|x_v] \tag{44}$$

$$C = \frac{1}{n}\sum_v \int_{-\infty}^{\infty} \theta\widetilde{h}(\theta|x_v)d\theta = \frac{1}{n}\sum_v \widetilde{E}(\theta|x_v), \tag{45}$$

$$D \equiv D(\mu_p) = \frac{1}{n}\sum_v \int_{-\infty}^{\infty} (\theta - \mu_p)^2\widetilde{h}(\theta|x_v)d\theta = \frac{1}{n}\sum_v \widetilde{E}\left[(\theta - \mu_p)^2|x_v\right], \tag{46}$$

in which n denotes the number of subjects, and $\widetilde{E}(.)$ denotes expected value in the posterior distribution of θ given the data and the $\widetilde{\eta}$ parameters. Notice that A, B, and D are function of the η parameters which are unknown. This causes major problem when the system of equations given by Eq. (39) to (42) is solved. We want to maximize Q with respect to these η parameters, but it is impossible to find explicit expressions for these parameters as the selection parameters are embedded in $R[\lambda(\theta)]$ within the integrals of Eq. (43) and Eq. (44). A possible solution would be to use the Newton-Raphson algorithm [see Section (4.1)]. However, as Netwon-Raphson is a very sensitive procedure that requires well chosen starting values, we strongly doubt whether it will find a solution. Therefore, we operate as follows. We keep σ_s^2 and μ_s constant, and start by estimating σ_p^2 and μ_p. We rewrite Eq. (39) to find an expression for $R(\Lambda)/N^{\frac{1}{2}}$:

$$\frac{R(\Lambda)}{N^{\frac{1}{2}}} = \frac{1}{\sigma_s} A.$$

Using this result, Eq. (42) could be written as

$$\mu_p = C - \frac{\sigma_p^2}{\sigma_s} A. \tag{47}$$

Applying Eq. (47) leads to an update of μ_p. From Eq. (41) we solve σ_p^2:

$$\sigma_p^2 = \frac{D}{1 - R(\Lambda)\rho\Lambda}. \tag{48}$$

Note that, when evaluating the right-hand side of Eq. (48), Λ and D are re-

computed using the updated value of μ_p. After estimating σ_p^2, Eq. (47) will not longer hold, as this equation is a function of σ_p^2. Therefore, we iteratively estimate μ_p and σ_p^2 until the estimates stabilize. In practise, two iterations will suffice. After every update in every iteration we checked whether the loglikelihood increased.

Now, we continue by keeping μ_p and σ_p^2 constant, and estimating the selection parameters. We do this for both parameters separately. We seek a value for $\Delta\mu_s$ that, when added to $\widetilde{\mu}_s$, increases the loglikelihood. At the first M-step, we start with $\Delta\mu_s = 0.1$. Then the new estimate of μ_s is given by

$$\mu_s = \widetilde{\mu}_s + \Delta\mu_s.$$

If the loglikelihood increases at the new estimate, we accept it. If it does not, we proceed by calculating

$$\mu_s = \widetilde{\mu}_s - \Delta\mu_s.$$

If the loglikelihood neither increases at the new estimate, we take $\Delta\mu_{s,new} = \frac{1}{2} \times \Delta\mu_s$ and start again. This procedure is continued until the loglikelihood increases, or until $\Delta\mu_{s,new} = 1e-6$. In case of the latter, the estimate of μ_s

from the previous iteration is not changed.For σ_s^2, we prevent negative variances by seeking a value $\delta\sigma_s^2$ that, when *multiplied* with $\widetilde{\sigma}_s^2$ increases the likelihood. At the first M-step, we start with $\delta\sigma_s^2 = 0.9$. The new estimate of σ_s^2 is then given by

$$\sigma_s^2 = \widetilde{\sigma}_s^2 \times (2 - \delta\sigma_s^2).$$

If the loglikelihood increases at the new estimate, we accept it. If it does not, we proceed by calculating

$$\sigma_s^2 = \widetilde{\sigma}_s^2 \times \delta\sigma_s^2.$$

If the loglikelihood neither increases at the new estimate, we take $\delta\sigma_{s,new}^2 = \delta\sigma_s^2 + \frac{1}{4} \times (1 - \delta\sigma_s^2)$ and start again. This procedure is continued until the loglikelihood increases, or until $\delta\sigma_s^2 = 0.9995$. In case of the latter, the estimate of σ_s^2 from the previous iteration is not changed.

Notice that we do not maximize the Q-function. However, we do make the Q-function to increase in each iteration. Therefore, the procedure described above is a Generalized Expectation Maximization algorithm (i.e., a special case of EM which does not require the Q function to be maximized; Demster, Laird & Rubin, 1977).

3.3 Numerical tools

To fit the latent distribution from: Eq. (22) using the EM-algorithm, some numerical tools are needed. First, the common method of Gauss-Hermite Quadrature is used to approximate the integrals in Eq. (43), Eq. (45), and Eq. (46). Second, and far less common, an approximation of $R(x)$ from Eq. (38) is needed to prevent numerical problems.

3.3.1 Approximation of $R(x)$

Notice that $R(x)$ from Eq. (38) equals the ratio between the standard normal density function and the cumulative normal density function, both evaluated at x. See Figure (3) for a graph of $R(x)$. In the figure, the line $y = -x$ is draw to show that

$$\lim_{x \to -\infty} \frac{R(x)}{x} = 1,$$

and

$$\lim_{x \to \infty} R(x) = 0.$$

As a result, it holds by definition that $R(x) \in (0, \infty)$. In other words, $R(x)$ is strictly positive and nonzero. However, the widely used statistical software package R (R Development Core Team, 2006) returns $\varphi = 0$ for $x < -38.5$ and $\Phi(x) = 0$ for $x < -37.5$. This causes $R(x)$ to be undefined for $x < -37.5$, because the denominator of $R(x)$ equals zero for these values of x. This problem

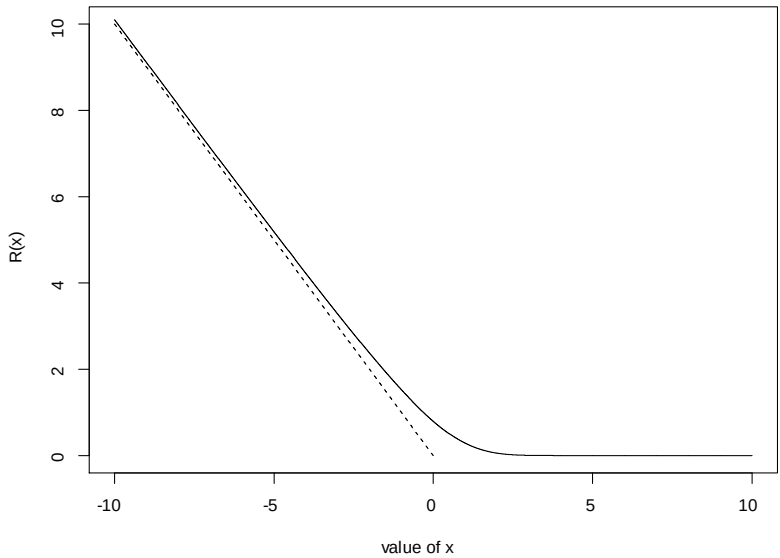

Figure 3: Graph of $R(x)$. Solid line: $y = R(x)$, dotted line: $z = -x$.

is not associated with software package R alone. In numerical computation, $\Phi(x)$ from the denominator of Eq. (38), is generally piecewise approximated with a set of rational functions, which are accurate up to a certain decimal place, if $\Phi(x) > 10^{-308}$. When $\Phi(x) < 10^{-308}$, the software will return $\Phi(x) = 0$, while in fact, $\Phi(x)$ is nonzero for all x's. This error imposes major problems in our model. Consider A from Eq. (43), which represents the conditional expectation of $R[\lambda(\theta)]$ in the skew-normal density. The use of Hermite- Gauss Quadratures causes $\lambda(\theta)$ to take values that are much smaller than -37.5, which will result in errors when evaluating $R[\lambda(\theta)]$ at the quadrature points.

The problem is solved by recognizing that we are not interested in the numerator and denominator of Eq. (38) separately. It is sufficient to have $R(x)$ available as a whole. We will thus use an approximation for $R(x)$ to get rid of the numerical limitations. As can be seen from Figure (3), two intervals could be distinguished for x in which $R(x)$ is monotome decreasing. These concern: $x \in (-\infty, -4]$ and $x \in [4, \infty)$. For both intervals a separate approximation is used. In the interval $(-4, 4)$ the ordinary normal density and distribution functions are used.

3.3.2 Approximation for $x \in (-\infty, -4]$

In this interval, we use Mill's ratio (Johnson & Kotz, 1970, p. 279) to approximate $R(x)$. Mill's ratio is commonly used in survival analysis to model hazard rates. Mill's ratio is given by

$$M(x) = \frac{1 - \Phi(x)}{\varphi(x)}. \tag{49}$$

We will thus need the inverse Mill's ratio:

$$[M(x)]^{-1} = \frac{\varphi(x)}{1 - \Phi(x)} = \frac{\varphi(-x)}{\Phi(-x)} = R(-x).$$

Johson & Kotz (p. 279) provide an approximation for Eq. (49) that is given by

$$M(x) = x^{-1}[1 - (x^2 + 3)^{-1} - 6(x^6 + 13x^4 + 25x^2 + 145)^{-1} + O(x^{-14})]. \tag{50}$$

We use the inverse of Eq. (50) to approximate $R(x)$ in the interval $x \in (-\infty, -4]$.

3.3.3 Approximation for $x \in [4, \infty)$

Strictly, for $x > 4$ no major numerical problems arise, as the denominator of $R(x)$ is close to 1. However, to enhance numerical accuracy, we will use an approximation for $R(x)$ in this interval nevertheless. Feller (1968, p.175)

provides the following asymptotic expression: as x approaches infinity, the ratio of the sides given in expression (51) tends to one:

$$1 - \Phi(x) \sim x^{-1}\varphi(x). \tag{51}$$

We will thus use that

$$R(x) \approx \frac{\varphi(x)}{1 - x^{-1}\varphi(x)}, \tag{52}$$

hereby, we omit $\Phi(x)$ from the denominator of $R(x)$.

3.4 Input for the model

To fit the skew-normal regression model, data is needed. When n subjects complete k items, a $(n \times k)$ matrix of item scores is obtained. However, it is not necessary to input the whole matrix of observed data in the model. Only the item discrimination indices, the estimated item-difficulty parameters and the test scores are needed. See $\widetilde{h}(\theta|x_v)$ from Eq. (27). This is the only part of the model that needs data input. Specifically $f(x_v|\theta)$ contains the measurement model, with all observed data in x_v. However, $\widetilde{h}(\theta|x_v)$ could be rewritten, in such a way that the individual item scores disappear from the equation. Consider $\widetilde{h}(\theta|x_v)$:

$$\widetilde{h}(\theta|x_v) = \frac{f(x_v|\theta)\widetilde{k}(\theta)}{\int_{-\infty}^{\infty} f(x_v|\theta)\widetilde{k}(\theta)d\theta},$$

where, $f(x_v|\theta)$ is defined by

$$f(x_v|\theta) = \prod_i \frac{\exp[x_{vi}\alpha_i(\theta - \beta_i)]}{1 + \exp[\alpha_i(\theta - \beta_i)]}. \tag{53}$$

and x_{vi} is the observed score of subject v on item i. Eq. (53) can be rewritten as

$$
\begin{aligned}
f(x_v|\theta) &= \prod_i \frac{\exp[x_{vi}\alpha_i(\theta-\beta_i)]}{1+\exp[\alpha_i(\theta-\beta_i)]}, \\
&= \frac{\prod_i \exp(x_{vi}\alpha_i\theta) \times \prod_i \exp(x_{vi}\alpha_i\beta_i)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}, \\
&= \frac{\exp\sum_i x_{vi}\alpha_i\theta \times \prod_i \exp(x_{vi}\alpha\beta_i)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}, \\
&= \frac{\exp(s_v\theta) \times \prod_i \exp(x_{vi}\alpha\beta_i)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]},
\end{aligned}
$$

Inputting in Eq. (27):

$$\widetilde{h}(\theta|x_v) = \frac{\frac{\exp(s_v\theta)\times\prod_i \exp(x_{vi}\alpha\beta_i)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)}{\int_{-\infty}^{\infty}\frac{\exp(s_v\theta)\times\prod_i \exp(x_{vi}\alpha\beta_i)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)d\theta},$$

$$= \frac{\prod_i \exp(x_{vi}\alpha\beta_i)\times\frac{\exp(s_v\theta)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)}{\prod_i \exp(x_{vi}\alpha\beta_i)\times\int_{-\infty}^{\infty}\frac{\exp(s_v\theta)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)d\theta},$$

$$= \frac{\frac{\exp(s_v\theta)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)}{\int_{-\infty}^{\infty}\frac{\exp(s_v\theta)}{\prod_i 1+\exp[\alpha_i(\theta-\beta_i)]}\widetilde{k}(\theta)d\theta}.$$

Thus, the item scores (x_{vi}) disappear from the equation. The test scores (s_v), the item discrimination indices (α_i), and the item difficulty parameters (β_i) are left. These are sufficient as input for the model.

4 Newton-Raphson

4.1 Introduction to the Newton-Raphson algorithm

The Newton-Raphson algorithm is an iterative procedure that finds the root(s) of a (system of) equation(s). Suppose that one knows that for x_0 the function $f(x)$ is nearly zero. The logic behind NR is that $f(x)$ is replaced by the first two terms of its Taylor expansion round x_0, to obtain a better approximation of the root of the function $f(x)$. This expansion is a polynomial of the first degree and equals

$$p_{x_0}(x) = f(x_0) + \frac{f\,'(x_0)}{1}(x - x_0),$$

where $f\,'(x_0)$ is the first order derivative of $f(x)$ evaluated at x_0. Now, the root of $f(x)$ is approximated by the root of $p_{x_0}(x)$; Meaning that

$$0 = f(x_0) + \frac{f\,'(x_0)}{1}(x - x_0), \tag{54}$$

resulting in

$$x = x_0 - \frac{f(x_0)}{f\,'(x_0)}. \tag{55}$$

In the next iteration, x is substituted with x_0, and a new approximation for x is determined. If the original x_0 is sufficiently close to the root of $f(x)$, one gets closer and closer to the root of $f(x)$ every iteration. This procedure is repeated until some convergence criterion is met. When starting values (i.e., x_0 in the first iteration) are badly chosen, the root of $p_{x_0}(x)$ is too far away from the root of $f(x)$. As a result, the NR-algorithm will diverge. Another consequence of bad starting values is that x can take values in a domain in which $f(x)$ is not defined. It is thus important to have reasonable starting values available. However, we have good starting values at our disposal from the EM-algorithm (as is discussed in Section 3).

Eq. (55) generalizes in the multivariate case as follows:

$$\mathbf{x} = \mathbf{x}_0 - [f'(\mathbf{x}_0)]^{-1} f(\mathbf{x}_0), \tag{56}$$

where $f'(\mathbf{x}_0)$ is a matrix with first order derivatives and $f(\mathbf{x}_0)$ denotes a vector which contains the values of $f(x)$ evaluated at $\mathbf{x}_0$. We will use this result in the next section.

4.2 NR for skew-normal latent distributions

To obtain parameter estimates for η, the method of maximum likelihood is used. Maximum likelihood involves the maximization of the likelihood function, which equals for the skew-normal latent distribution:

$$L = \prod_v \frac{\int_{-\infty}^{\infty} f(x_v|\theta)\Phi(\theta)g(\theta)d\theta}{\Phi(\Lambda)}. \tag{57}$$

Maximizing Eq. (57) is equivalent to maximizing the logarithm of Eq. (57):

$$\ln L = -n \times \ln \Phi(\Lambda) + \sum_v \ln \int_{-\infty}^{\infty} f(x_v|\theta)\Phi(\theta)g(\theta)d\theta. \tag{58}$$

At the maximum of Eq. (58) first order derivatives to η are equal to zero. We use the NR-algorithm to find the zeros of these first order derivatives. Therefore, $f(\mathbf{x}_0)$ from Eq. (56) changes to the gradient, $\mathbf{G}_{\boldsymbol{\eta}_t}$, a vector with first order derivatives of the loglikelihood function evaluated at $\boldsymbol{\eta}_t$, where $\boldsymbol{\eta}_t$ is a vector

with estimates of η in the t-th NR-iteration. $f'(\mathbf{x}_0)$ of Eq. (56) changes to $\mathbf{H}_{\boldsymbol{\eta}_t}$, the Hessian, a matrix with second order derivatives evaluated at η_t. Thus we have

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - [\mathbf{H}_{\boldsymbol{\eta}_t}]^{-1}\mathbf{G}_{\boldsymbol{\eta}_t}. \tag{59}$$

For the derivatives of the loglikelihood function, see appendix A.

In section (3.4) it is shown that EM only needs test scores as data input. The same can be shown for the NR-algorithm. As $f(x_v|\theta)$ appears in both the numerator and the denominator of the moments shown in Appendix A, the observed test scores (x_{vi}) disappear from the model. The test scores (s_v), the item discrimination indices (α_i), and the item difficulty parameters (β_i) are thus sufficient as input for the model.

As NR is converged at the maximum likelihood estimates, $\widehat{\eta}$, $\mathbf{H}_{\widehat{\boldsymbol{\eta}}}$ from Eq. (59) can be used to obtain an estimation of the variance-covariance matrix of the parameter estimates using the following:

$$\widehat{\Sigma} = \left(-\mathbf{H}_{\widehat{\eta}}\right)^{-1},$$

where the diagonal elements of $\widehat{\Sigma}$ contain the squared standard errors of the parameter estimates, and the off-diagonal elements contain their covariance.

5 Applications

5.1 An example

In this section we demonstrate the latent regression model with a skew-normal ability distribution. We first analyze a single data set to see how parameters are estimated, and to see how the loglikelihood function is changing during the estimation process.

We generated a data set with 5000 subjects who completed 40 dichotomous items. For each subject, a θ_v is drawn from the skew-normal distribution with $\mu_p = 0$, $\sigma_p^2 = 1$, $\mu_s = -0.5$, and $\sigma_s = 0.5$. The generation of the θ-values was accomplished in the following way: First, a θ_v is drawn from the normal distribution with $\mu_p = 0$ and $\sigma_p^2 = 1$. Then, using this θ_v, the probability that this subject is selected into the actual population is calculated using a selection function with $\mu_s = -0.5$ and $\sigma_s = 0.5$. This probability is compared to a uniformly distributed random number, U in $[0, 1)$. If the probability to be selected exceeded U, the θ_v was used in the data simulation. Otherwise, it was rejected. This procedure was repeated until the specified number of subjects was obtained (i.e., 5000 in this case). The number of subjects that were rejected equalled 2377. This is a rejection rate of 32.2%.

Item discrimination indices are all set to 1. True item-difficulty parameters are uniformly chosen between -2.5 and 2.5. As the skew-normal regression

model is a structural model, we estimated the item-difficulty parameters from the measurement model using CML, and imputted these in the structural model. To estimate the parameter from the structural model, 25 EM-iterations are conducted followed by 8 NR-iterations. From the 7th to the 8th NR-iteration, parameter estimates did not change in the first 7 decimals, thus, the procedure seems to have converged.

In Figure (4), the value of the loglikelihood function is plotted at each iteration. Notice that the likelihood increases rapidly when switched from EM to NR.

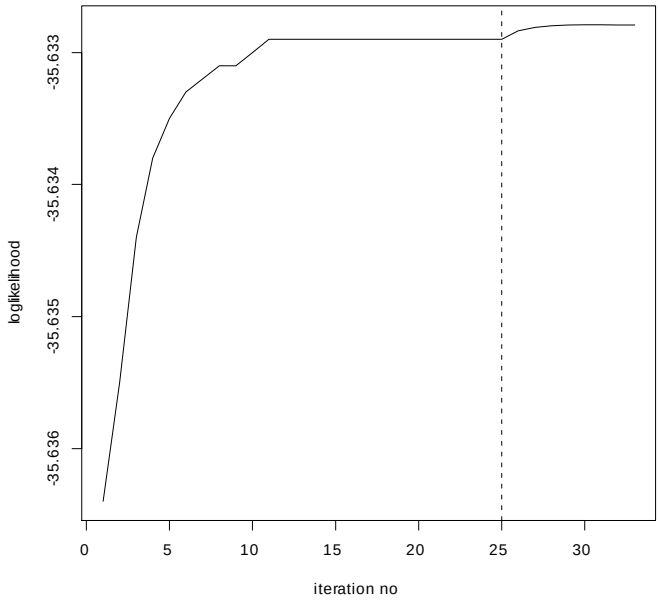

Figure 4: The value of the loglikelihood function at each iteration (devided by n). The dotted line indicates the switch from EM to NR.

In Figure (5), parameter estimates are plotted at each iteration. Notice that after 6 iterations, parameter change is already small. Therefore, we could have switched to NR at this point as well. It is also evident from the figure that, when switched to NR, parameter estimates make a jump. Hereafter, the solution is reached within a few iterations. When only the EM-algorithm was used to estimate the parameters from the model, an enormous number of iterations would have been needed. After 25 iteration the solution is still far away. In

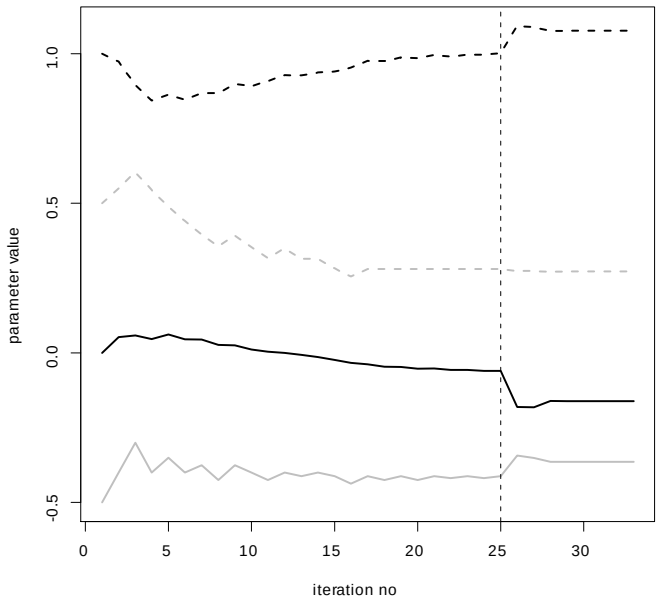

Figure 5: Parameter estimates at each iteration. Black lines: population parameters; grey lines: selection parameters. Solid lines: variance parameters; dotted lines: mean parameters. The vertical line indicates the switch from EM to NR.

Table (1), parameter estimates of the EM and NR-algorithm are displayed, as well as the standard errors of the NR-estimates (notice that no standard errors are involved in EM). As the standard errors of σ_p^2 and σ_s^2 concerned the standard errors of their logarithm, we transformed these using the Univariate delta method (Bishop, Fienberg & Holland, 1975, Section 14):

$$var(\sigma^2) = var(\ln \sigma^2) \times \left(\frac{d\sigma^2}{d \ln \sigma^2} \right)^2 = var(\ln \sigma^2) \times \sigma^4. \tag{60}$$

where Eq. (60) σ^2 is either σ_p^2 or σ_s^2.

Table 1: Parameter estimates of EM and NR

	μ_p	σ_p^2	μ_s	σ_s^2
True values	0	1	-0.5	0.25
Estimated values EM[1]	-0.060	1.002	-0.413	0.281
Estimated values NR	-0.162	1.077	-0.363	0.273
Standard errors NR	0.364	0.242	0.239	0.017

[1] note that these concern estimates after 25 iterations, the solution is not converged yet.

Notice the large standard errors of the parameter estimates. Except for σ_s^2, which has a reasonable standard error, the parameter estimates are unstable. In Table (2) a possible reason is shown: very large correlations between the parameter estimates.

Table 2: Correlations between the NR parameter estimates

	μ_p	$\ln \sigma_p^2$	μ_s	$\ln \sigma_s^2$
μ_p				
$\ln \sigma_p^2$	-0.985			
μ_s	0.617	-0.701		
$\ln \sigma_s^2$	0.942	-0.981	0.804	

It could be concluded from the table that the skew-normal regression model is nearly unidentifiable. Parameter estimates are highly correlated, therefore, the different parameters from the model are hard to distinguish statistically. In Section 6, we will address this problem in greater detail.

5.2 A decreasing selection function.

To account for a thinner upper tail in the ability distribution, a decreasing selection function has to be fitted. The present parameterisation of the model [see Eq.(22)] does not permit this directly. However, a minor manipulation of the data enables negative selection functions as well.

By the present parameterisation, we can not invert the selection function (because σ_s will then be negative). However, a selection function that is decreasing over θ, is increasing over $-\theta$. Thus, as we are able to fit increasing selection functions, we should invert the θ-scale. However, as θ is unobservable, it could not be inverted directly. Therefore, we tackle the problem by inverting the test scores and the item-difficulty parameters estimates. First, the test scores are inverted using the following transformation:

$$s_{v,new} = \sum_i \alpha_i - s_v,$$

where s_v is defined in Eq. (2), and α_i are the item discrimination indices from Eq. (1). Now, for a test with 50 items and all α_i's set to 1, a subject with test score 39 gets an inverted test score 11. Second, $\widehat{\beta}_i$ is inverted, in this way:

$$\widehat{\beta}_{i,new} = -\widehat{\beta}_i + 2 \times \frac{\sum \widehat{\beta}_i}{k},$$

if the mean of the item-difficulty parameters equals zero, only the sign changes. Now, using both the inverted item-difficulty parameters and inverted test scores, the latent regression model could be fitted as it stands. Notice that the resulting parameter estimates have to be interpreted with care, as they describe the inverted θ.

5.3 Comparing the skew-normal regression model

In this section, 42 data-sets are analyzed using both the skew-normal regression model (SN henceforth) and SAUL. Data sets were simulated according to the procedure outlined in section (5.1), and differed systematically according to a $7 \times 3 \times 2$ design. The first manipulated factor was σ_s. True values for this parameter were chosen to be 0.7, 0.6,0.5,0.4,0.3,0.2 and 0.1. The second factor that was manipulated concerned the number of subjects, these were either 10000, 2500 or 1000. The final factor was the number of items, they were either 40 or 20. Other parameters were fixed: $\mu_p = 0$, $\sigma_p^2 = 1$ and $\mu_s = -0.5$. Notice that we did not manipulate μ_s, because it gave major problems to both EM and NR. Systematically manipulating μ_s together with σ_s causes the selection function to be of a useless kind in the majority of the cases. For example, selection functions that are far away from the normal distribution of the ideal population. This caused that much convergence problems, that we choose not to manipulate this parameter, but to fix it to a common value.

In Table (3), success of convergence is displayed for all 42 data sets. It is clear that in many cases, NR fails to converge. This occured troughout the design. When σ_s gets small, problems arise independent of the number of items. The combination: few subjects and few items also cause NR to fail.

Table 3: Succes of convergence

		n=10000		n=2500		n=1000	
		items		items		items	
		40	20	40	20	40	20
	0.7	ok[1]	ok	*	ok	*	*
	0.6	ok	ok	ok	*	ok	ok
	0.5	ok	ok	ok	*	ok	*
σ_s	0.4	ok	ok	ok	*	ok	*
	0.3	ok	*	*	*	ok	*
	0.2	*	*	*	*	*	*
	0.1	*	*	*	*	*	*

[1] ok: converged, *: failure

From the 7×3 × 2 design results of 9 succeeded SN and SAUL analyses are displayed in Table (4). Parameter estimates of the two models are hard to compare, as both sets of parameters have different meaning and interpretation. Therefore, the predicted test score distribution, the goodness of fit measure χ^2, and the estimated mean and standard deviation of θ [$E(\theta)$ and $SD(\theta)$ respectively] are compared.

It is clear that the predicted mean and standard deviation of the observed test score distribution of SAUL are very close to the observed distribution. The SN model is even closer, but this is not that surprising, as the data were generated according this model. The differences regarding predicted mean and standard deviation are thus considered as negligible. This holds for $E(\theta)$ and $SD(\theta)$ as well. No systematical differences are observed. However, some important differences are evident in predicted skewness of the observed distribution. Where SN predicts the skewness of the observed distribution sufficiently well, SAUL is (completely) wrong in all cases. This implies that the misfit of SAUL is particular evident in the tails of the distribution. This could clearly be seen in Figure (6) and Figure (7). From Figure (6) it is evident that the predicted distribution of the SN model highly resembles the observed distribution. While, from Figure (7) it seems that SAUL is systematically wrong, particular in the tails of the distribution. This misfit is also apparent from the χ^2-values in Table (4). These values are calculated using a formula familiar from loglinear analysis:

$$\chi^2 = \sum_i \frac{(E_i - O_i)^2}{E_i},$$

where E_i is the expected test score frequency, and O_i is the observed test score frequency. The summation runs over i which denotes all possible scores (in this case 0-40).

The values of χ^2 vary greatly between SN and SAUL, with the lower values systematically for SN, indicating that this model fits better. As the difference

Table 4: Observed and predicted test score distribution for SN and SAUL

				$\sigma_s = 0.6$			
n		Mean	SD	Skewness	E(θ)[1]	SD(θ)[1]	χ^2
10000	Observed	23.07	5.46	0.06	-	-	-
	Predicted SN	23.08	5.46	0.04	0.49	0.78	42.72
	Predicted SAUL	23.10	5.47	-0.13	0.49	0.77	115.01
2500	Observed	23.19	5.32	0.10	-	-	-
	Predicted SN	23.19	5.31	0.10	0.51	0.76	33.00
	Predicted SAUL	23.24	5.32	-0.14	0.51	0.76	67.68
1000	Observed	23.08	5.67	-0.05	-	-	-
	Predicted SN	23.08	5.67	-0.05	0.49	0.83	20.27
	Predicted SAUL	23.13	5.66	-0.14	0.50	0.83	23.09
				$\sigma_s = 0.5$			
n		Mean	SD	Skewness	E(θ)	SD(θ)	χ^2
10000	Observed	22.94	5.34	0.14	-	-	-
	Predicted SN	22.94	5.34	0.13	0.47	0.76	29.40
	Predicted SAUL	22.97	5.36	-0.12	0.47	0.76	209.16
2500	Observed	23.11	5.33	0.12	-	-	-
	Predicted SN	23.11	5.33	0.12	0.49	0.76	50.99
	Predicted SAUL	23.21	5.33	-0.14	0.50	0.77	75.81
1000	Observed	22.98	5.33	0.18	-	-	-
	Predicted SN	22.98	5.33	0.18	0.48	0.77	58.96
	Predicted SAUL	22.97	5.36	-0.12	0.48	0.77	90.93
				$\sigma_s = 0.4$			
n		Mean	SD	Skewness	E(θ)[1]	SD(θ)[1]	χ^2
10000	Observed	23.13	5.20	0.13	-	-	-
	Predicted SN	23.13	5.20	0.13	0.50	0.74	19.23
	Predicted SAUL	23.13	5.23	-0.12	0.49	0.72	184.20
2500	Observed	23.10	5.22	0.16	-	-	-
	Predicted SN	23.10	5.22	0.16	0.49	0.74	16.86
	Predicted SAUL	23.12	5.25	-0.12	0.49	0.74	78.86
1000	Observed	23.16	5.21	0.16	-	-	-
	Predicted SN	23.16	5.20	0.17	0.50	0.74	22.69
	Predicted SAUL	23.20	5.22	-0.13	0.50	0.74	47.81

[1]The expected value of θ, E(θ), and the standard deviation of θ, SD(θ), are estimated using respectively the first and second moment about the mean of he corresponding distribution (that is: the skew-normal density for SN, and the normal density for SAUL).

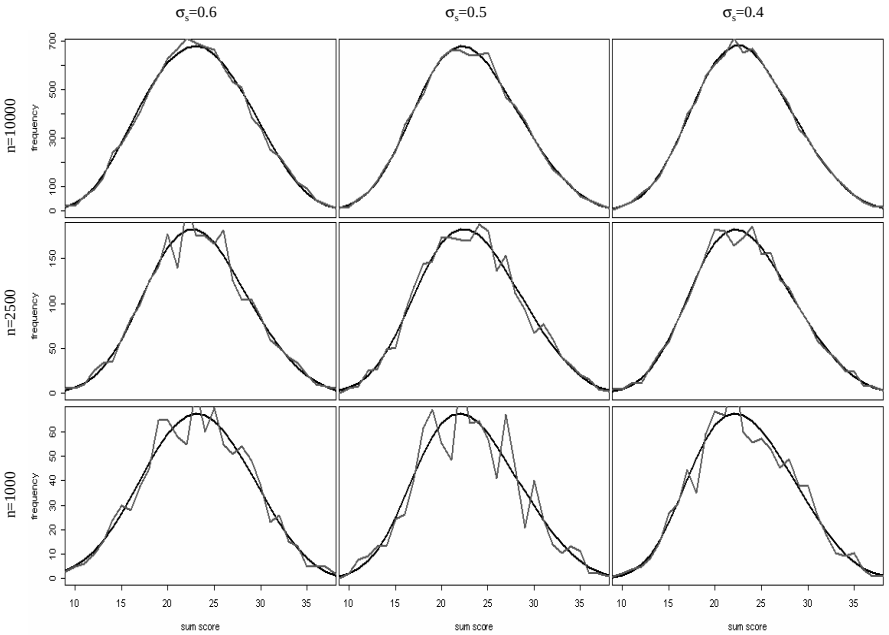

Figure 6: The observed and predicted test score distribution for varying n and σ_s. The grey line represents the observed test scores, and the black line represents the distribution predicted by the skew-normal regression model.

in degrees of freedom is 2, all the differences in χ^2 are significant (only one exception: the case $n = 1000$ and $\sigma_s = 0.6$).

It could be concluded that SN is better in predicting the observed test score distribution. However, the predicted mean and standard deviation of SAUL do not differ importantly from the observed mean and standard deviation. Therefore, if a researcher is only interested in these descriptives, no concerns are necessary about the ability distribution. Although the ability distribution is non-normal, using SAUL will yield approximately the same predicted mean and standard deviation of the observed distribution. However, the predicted skewness and model fit differ greatly from SN to SAUL. These aspects could thus bias statistical testing at individual level or group level when an ability distribution is wrongly specified. Thus, when model fit or skewness are of concern to the researcher, an ability distribution should be chosen with great care.

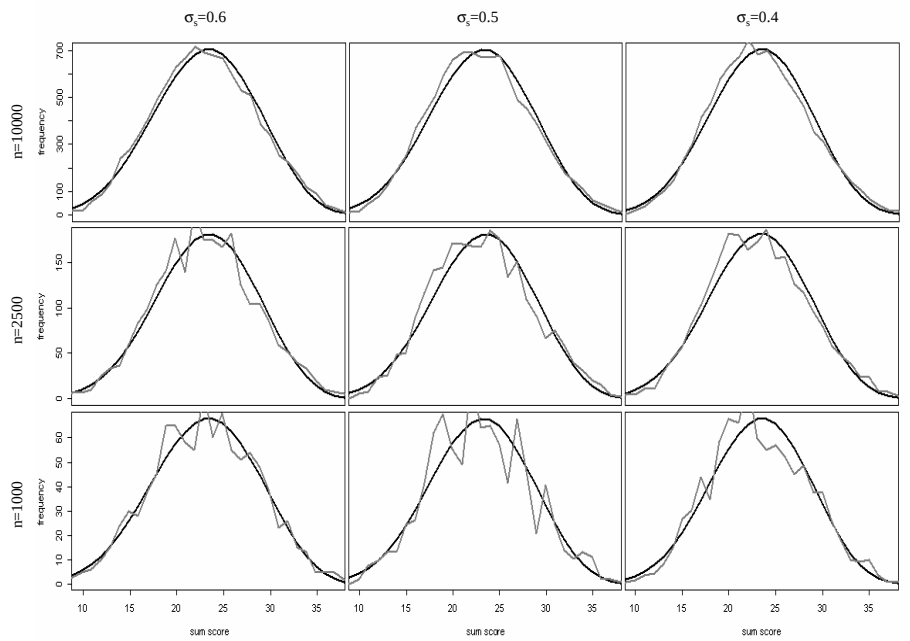

Figure 7: The observed and predicted test score distribution for varying n and σ_s. The grey line represents the observed test scores, and the black line represents the distribution predicted by SAUL.

5.4 Regression

Next, a regression model is specified, to see whether differences are present between SN and SAUL in making group comparisons. Three data-sets for two groups are simulated following the procedure as outlined in Section (5.1). In the present application, responses for 10000 subjects were generated, 5000 for each group. The same selection function ($\mu_s = -0.5$ and $\sigma_s = 0.5$) applied in both groups. In the first group, $\sigma_p^2 = 1, \mu_p = 0$, and in the second group: $\sigma_p^2 = 1, \mu_p = \alpha$, where $\alpha \in \{0, 0.25, -0.25\}$. The solution converged in all cases. Two SAUL models are fitted, first a model with equal variances in both groups (referred to as SAUL) and second, a model with unequal variances in both groups (referred to as SAUL-U). SAUL-U uses a multiplicative parameter to model group differences in variances (denoted β henceforth). This $\beta-$parameter represents the ratio between the variances of two groups. In the model, the logarithm of this ratio is modelled.

In Table (5), parameter estimates and standard errors are displayed. Notice that standard errors of the SN parameters are higher than those of SAUL and SAUL-U. Within SN, the standard errors of α and σ_s^2 are remarkably lower. It is particular evident that, when the true value of α deviated from zero, the α-parameters are not recovered. Specifically, α is systematically underestimated as its small standard errors cause the true parameter values to be outside a 95% confidence interval when $\alpha \neq 0$. Notice that SAUL-U detects a difference in σ_p^2 in the case that $\alpha = 0.25$ [that is, $\ln(\beta)$ deviates significantly from 0, as judged by its standard error].

Table (6) lists descriptive statistics of the observed test score distribution and the predicted test score distribution for SN, SAUL and SAUL-U. From the table it appears that findings are roughly the same as those in Section (5.3). That is: the mean and standard deviation are predicted well by both models, while skewness is predicted better by SN, and χ^2 is much smaller for SN. Notice that in the SN model, specifying a mean difference results in differences in the standard deviation of θ between groups as well. Recap that SAUL-U did detect this difference [see Table (5) again] in the case that $\alpha = 0.25$. The difference in $SD(\theta)$ when $\alpha = -0.25$ is thus undetected. This difference is thus either hard to detect for SAUL-U, or the power to detect it is low. However, given the number of subjects (10000) power should be reasonable. Apparently in detecting differences in $SD(\theta)$, the assumption of a normal distribution is problematic in this particular case.

6 Discussion

This paper presented a method to account for non-normality in ability distributions. Using a selection function, the commonly used normal distribution

Table 5: Parameter estimates and standard errors of SN and SAUL

$\alpha = 0$

		μ_p	α	σ_p^2	μ_s	σ_s^2	β
SN	estimate	0.070	-0.006	0.937	-0.515	0.235	-
	SE	0.164	0.028	0.103	0.153	0.012	-
SAUL	estimate	0.490	-0.004	0.553	-	-	-
	SE	0.0120	0.017	0.010	-	-	-
SAUL-U	estimate	0.490	-0.004	0.549	-	-	1.013 (0.012)[1]
	SE	0.012	0.017	0.015	-	-	0.037[2]

$\alpha = 0.25$

		μ_p	α	σ_p^2	μ_s	σ_s^2	β
SN	estimate	0.037	0.217	0.960	-0.519	0.224	
	SE	0.137	0.035	0.088	0.122	0.010	
SAUL	estimate	0.479	0.131	0.576	-	-	
	SE	0.012	0.017	0.0107	-	-	
SAUL-U		0.479	0.131	0.5533	-	-	1.0810 (0.0779)
		0.012	0.017	0.0146	-	-	0.0371

$\alpha = -0.25$

		μ_p	α	σ_p^2	μ_s	σ_s^2	β
SN	estimate	0.163	-0.182	0.906	-0.693	0.205	
	SE	0.124	0.033	0.086	0.144	0.012	
SAUL	estimate	0.472	-0.116	0.567	-	-	
	SE	0.012	0.017	0.011	-	-	
SAUL-U	estimate	0.472	-0.116	0.574	-	-	0.974 (0.026)
	SE	0.012	0.017	0.015	-	-	0.037

[1] As the logarithm of the β-parameter is modelled, $\ln(\beta)$ is displayed between brackets.

[2] The standard error that is displayed for β is the standard error of its logarithm.

Table 6: Observed and predicted test score distribution

$\alpha = 0$							
		mean	sd	skew	E(θ)	SD(θ)	χ^2
observed	group 1	23.09	5.26	0.11	-	-	-
	group 2	23.06	5.29	0.11	-	-	-
predicted SN	group 1	23.09	5.28	0.11	0.49	0.75	20.67
	group 2	23.07	5.28	0.11	0.49	0.75	21.17
predicted SAUL	group 1	23.10	5.30	-0.12	0.49	0.74	88.35
	group 2	23.10	5.30	-0.12	0.49	0.74	89.20
predict SAUL-U	group 1	23.10	5.29	-0.12	0.49	0.74	89.27
	group 2	23.08	5.31	-0.12	0.49	0.74	88.28
$\alpha = 0.25$							
		mean	sd	skew	E(θ)	SD(θ)	χ^2
observed	group 1	23.03	5.29	0.13	-	-	-
	group 2	23.85	5.39	0.07	-	-	-
predicted SN	group 1	23.03	5.30	0.12	0.48	0.75	32.12
	group 2	23.85	5.39	0.07	0.61	0.78	33.56
predicted SAUL	group 1	23.04	5.39	-0.12	0.48	0.76	102.86
	group 2	23.86	5.35	-0.16	0.61	0.76	101.78
predict SAUL-U	group 1	23.05	5.32	-0.12	0.48	0.74	105.68
	group 2	23.86	5.42	-0.16	0.61	0.77	94.65
$\alpha = -0.25$							
		mean	sd	skew	E(θ)	SD(θ)	χ^2
observed	group 1	22.99	5.35	0.11	-	-	-
	group 2	22.27	5.33	0.15	-	-	-
predicted SN	group 1	22.99	5.38	0.10	0.47	0.77	25.14
	group 2	22.27	5.30	0.15	0.36	0.75	29.41
predicted SAUL	group 1	23.00	5.34	-0.12	0.47	0.75	101.30
	group 2	22.27	5.37	-0.09	0.36	0.75	112.66
predict SAUL-U	group 1	23.002	5.37	-0.12	0.47	0.76	98.82
	group 2	22.275	5.35	-0.09	0.36	0.76	115.682

could be accommodated to account for a thinner upper or lower tail. We presented two estimation algorithm that, when combined, come to the maximum likelihood estimations of the parameters from the normal distribution and the selection function. It appeared that when the ability distribution was indeed non-normal, our model outperformed the normal model regarding 1) predicting skewness of the observed test scores, and 2) goodness of fit. The mean and standard deviation are predicted with about the same precision by both models. At first sight, the latent selection model seems a good choice when the ability distribution is (suspected to be) non-normally distributed. However, from Section 5, some major problems arise. These concern the high correlations between the parameter estimates, and the high standard errors of the parameter estimates. We see some explanations for this: The model, as it stands, is extremely flexible. Flexibility arises because many possible combinations of a selection function with a normal density function result in closely resembling skew-normal distributions [see Figure (8)]. For instance, if some constant is added to parameter μ_p, the resulting distribution changes. However, the original distribution could be closely regained, by adding some carefully chosen constant to another parameter (say σ_p^2). As a result, high correlation between the parameter estimates arise. These correlations are that high [see the example in Section (5.1)] that some parameters are nearly unidentifiable. Even worse, these correlations cause the Newton-Raphson algorithm (as presented in Section 4) to have difficulties with finding the solution, either when 1) the selection parameters are extreme (e.g., $\mu_s = -3$ or $\sigma_s = 0.01$) or, 2) when the number of subjects is too small (e.g., n<1000). These difficulties are expressed in either that NR diverges, or that NR needs more than 10 iterations, which is unusual for NR. Because of the high correlations between the parameter estimates, a solution to the problems of this model might be the use of parameter restrictions. As noticed, μ_s is suspected to be the malefactor. Two reasons are provided. First, σ_s is systematically associated with the smallest standard error, it is thus unlikely that this parameter causes the problems. Second, μ_s can cause the selection function to take very uncommon shapes as is noted in section (5.3). Restricting this parameter will thus hopefully reduce the flexibility of the model, and thereby the standard errors and intercorrelations. However, other restrictions should be considered as well (e.g., μ_p or σ_p^2).

Another problem seems that true parameter values are badly recovered in most applications. Particularly in the regression example, the regression parameter is systematically underestimated with a small standard error. As a result, the true parameter value is not in the 95% confidence interval. This raises questions about the use of this model in practice. Applying this model in the field of Educational Measurement causes mean differences to remain undetected, or some major differences to be underestimated.

Remarkable is that no literature is existing about these problems with parameter estimation in the skew-normal density. Azzalini (1985,1986) presents the model, but does not discuss parameter estimation. Neither do Arnold & Beaver (2002). Lin, Lee & Yin (2006) discuss parameter estimation of the skew-normal distribution in the light of mixture modelling. No problems are reported. How-

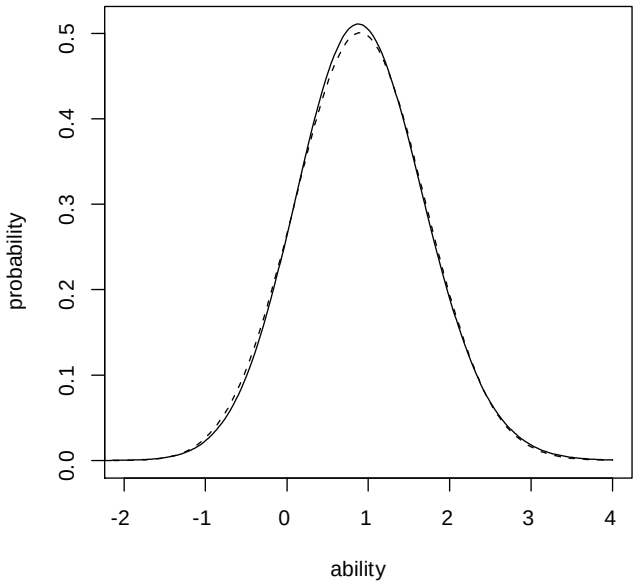

Figure 8: Two highly resembling skew-normal distributions, with completely different parameters. Solid line: $\mu_p = 0, \sigma_p^2 = 1, \mu_s = 1, \sigma_s = 1$. Dotted line: $\mu_p = 0.9, \sigma_p^2 = .8, \mu_s = -1.6, \sigma_s = .5$.

ever, the fitted distributions are not latent, as is the case in the present model. This applies to Azzalini and Capatanio (1999) as well, who discuss parameter estimation using maximum likelihood but do not use latent variables. Therefore, we presume that the difficulties are associated with latent distributions, and possibly with dichotomous data as well. As the model, for dichotomous data, needs the test scores as input only, the available degrees of freedom is small. Therefore, the model demands a lot of the data.

In Section 1, the possibility of a logistic selection function was mentioned. We discarded this option, and chose the cumulative normal density function. This choice could be questioned. Some may reason that -given the problems with the cumulative normal density function- a logistic function seems a better choice. However, taking into account the strong similarities between the logistic function and the cumulative normal density function (Lord & Novick, 1968, see Camilli, 1994, for a detailed prove), we expect that the logistic selection function will do no better. That is, with the model as presented, a logistic selection function will suffer from the same identification problems.

Finally, a remark is made about the nature of the model. Latent selection seems conceptually a practical tool to accommodate non-normality. We fitted some latent selection models to simulated data to explore the method. We have however no intention to claim that, when fitted to real data, this model is able to prove a real selection process. As presented, the model is a tool to describe interindividual differences in ability. To show that a real selection process caused the ability to be non-normally distributed, is not the purpose of this model. Therefore the parameters from the selection function, μ_s and σ_s, are not interpreted in terms of selection. They are meaningless, as are the parameters from the normal density function, μ_p and σ_p^2. It is *the product* that does matter, and that is useful in accounting for non-normality in latent ability distributions.

We showed that wrongly assuming a normal ability distribution could have negative consequences for the goodness of fit of the model and the prediction of test scores. Our model based on latent selection failed. Therefore, it is valuable to develop alternative skewed regression models, as an underlying normal distribution is not necessarily the best way to model group differences in ability.

References

Andersen, E. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the royal statistical society. Series B*, 34, 42–54.

Andersen, E. & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, 42, 357–374.

Arnold, B. & Beaver, R. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, 11, 7–54.

Arnold, B., Beaver, R., Groeneveld, R., & Meeker, W. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, 58, 471–488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–17.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46, 199–208.

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

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83, 715–726.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis. Theory and practice.* Cambridge: MIT Press.

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

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 189–197.

Brennan, R. L., Ed. (2006). *Educational Measurement. Fourth edition.* Washington, DC: ACE/Praeger Series on Higher Education.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, 19, 379–388.

de Leeuw, J. & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized rasch models. *Journal of Educational Statistics*, 11, 183–196.

Demster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 38, jan–38.

Feller, W. (1968). *Introduction to probability theory and its applications volume 1.* New York: Wiley.

Follman, D. (1988). Consistent estimation in the rasch model based on nonparametric margins. *Psychometrika*, 53, 553–562.

Goldstein, H. (1980). Dimensionality, bias, independence and measuremenstcale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 3, 234–260.

Hoijtink, H. (1995). Linear and repeated measures models for the person parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications.* New York: Springer-Verlag.

Johnson, N. L. & Kotz, S. (1970). *Continuous univariate distributions 2.* New York: Wiley.

Lin, T. I., Lee, J. C., & Yen, S. Y. (preprint). Finite mixture modeling using the skew normal distribution. *Statistica Sinica.*

Liseo, B. & Loperfidob, N. (2003). A bayesian interpretation of the multivariate skew-normal distribution. *Statistics and Probability Letters*, 61, 395–401.

Lord, F. M. (1984). *Maximum likelihood and bayesian parameter estimation in IRT.* Princeton NJ: Educational Testing Service.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Nandakumar, R. & Yu, F. (1996). Empirical validation of dimtest on nonnormal ability distributions. *Journal of Educational Measurement*, 33, 355–368.

R-Development-Core-Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rigdon, S. E. & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika*, 48, 567–574.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of multilog. *Applied Psychological Measurement*, 16, 1–6.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175–186.

Thorndike, R. L. (1971). Educational measurement for the seventies. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed).* Washington, DC: American Council on Eduation.

van den Oord, E. J. C. G. (2005). Estimating johnson curve population distributions in multilog. *Applied Psychological Measurement*, 29, 45–64.

Verhelst, N. D. (1993). Itemresponstheorie [item response theory]. In T. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk. [Psychometrics in practice.].* Arnhem: Cito.

Verhelst, N. D. & Eggen, T. H. M. (1989). *Psychometrische en statistische aspecten van peilingonderzoek [Psychometric and statistical aspects of survey research].* Arnhem: Cito.

Verhelst, N. D. & Glas, C. A. W. (1995). The one parameter logistic model. In G. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications.* New-York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1994). *OPLM: Computer program and manual*. Arnhem: Cito.

Verhelst, N. D. & Verstralen, H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL): Theory and a computer program.* Arnhem: Cito.

Woods, C. M. (2006). Ramsay-curve item response theory (rc-irt) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11, 253–27.

Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, 56, 589–600.

Zwinderman, A. H. & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the rasch model. *Applied Psychological Measurement*, 14, 73–81.

A Formula of the NR-algorithm

For the NR-algorithm, we need first and second order derivatives of the loglikelhihood function. The loglikelihood function for a single subject is given by

$$\begin{aligned} \ln L_v &= -\ln \Phi(\Lambda) + \ln \int_{-\infty}^{\infty} f(x_v|\theta)\Phi(\theta)g(\theta)d\theta, \\ &= \ell_1 + \ell_{2v}. \end{aligned}$$

Notice that ℓ_1 is only function of the parameters, and not of the data. The loglikelihood for the whole sample is thus calculated as follows:

$$\ln L = n \times \ell_1 + \sum_v \ell_{2v}$$

A.1 Taking the first-order derivatives of ℓ_1

A.1.1 Taking the first-order derivative w.r.t. a parameter in η_i :

$$\frac{\partial}{\partial \eta_i}\ell_1 = -R(\Lambda)\frac{\partial \Lambda}{\partial \eta_i}, \qquad \eta_i \in \{\mu_s, \ln \sigma_s^2, \mu_p, \ln \sigma_p^2\}. \tag{61}$$

A.1.2 Taking the second-order derivative with respect to a parameter in η_i:

$$\frac{\partial^2}{\partial\eta_i\partial\eta_j}\ell_1 = \left(\frac{\partial\Lambda}{\partial\eta_i}\frac{\partial\Lambda}{\partial\eta_j}\right)Z - R(\Lambda)\frac{\partial^2\Lambda}{\partial\eta_i\partial\eta_j}, \qquad \eta_i \in \eta_j \in \{\mu_s, \ln\sigma_s^2, \mu_p, \ln\sigma_p^2\}, \tag{62}$$

where

$$Z = \Lambda R(\Lambda) + R(\Lambda)^2.$$

A.1.3 First order derivatives

Using (61), the first order derivatives equal:

$$\begin{aligned}
\frac{\partial}{\partial\mu_s}\ell_1 &= \frac{R(\Lambda)}{N^{\frac{1}{2}}},\\
\frac{\partial}{\partial\ln\sigma_s^2}\ell_1 &= \frac{1}{2}R(\Lambda)(1-\rho)\Lambda,\\
\frac{\partial}{\partial\ln\sigma_p^2}\ell_1 &= \frac{1}{2}R(\Lambda)\rho\Lambda,\\
\frac{\partial}{\partial\mu_p}\ell_1 &= -\frac{R(\Lambda)}{N^{\frac{1}{2}}}.
\end{aligned}$$

A.1.4 Second order derivatives

Using (62), the second order derivatives equal:

$$\frac{\partial^2}{\partial\mu_s^2}\ell_1 = \frac{Z}{N},$$

$$\frac{\partial^2}{\partial\left(\ln\sigma_s^2\right)^2}\ell_1 = \left(\frac{1}{4}(1-\rho)^2\Lambda^2\right)Z - R(\Lambda)\left(\frac{3}{4}(1-\rho)^2\Lambda - \frac{1}{2}(1-\rho)\Lambda\right),$$

$$\frac{\partial^2}{\partial\left(\ln\sigma_p^2\right)^2}\ell_1 = \left(\frac{1}{4}\rho^2\Lambda^2\right)Z - R(\Lambda)\left(\frac{3}{4}\rho^2\Lambda - \frac{1}{2}\rho\Lambda\right),$$

$$\frac{\partial^2}{\partial\mu_p^2}\ell_1 = \frac{Z}{N}.$$

A.1.5 Mixed derivatives

Using (62), the mixed derivatives equal

$$\frac{\partial^2}{\partial\mu_s\partial\mu_p}\ell_1 = -\frac{Z}{N},$$

$$\frac{\partial^2}{\partial\mu_s\partial\ln\sigma_s^2}\ell_1 = \left(\frac{1}{2N^{\frac{1}{2}}}(1-\rho)\Lambda\right)Z - R(\Lambda)\left(\frac{1}{2}(1-\rho)\frac{1}{N^{\frac{1}{2}}}\right),$$

$$\frac{\partial^2}{\partial\mu_s\partial\ln\sigma_p^2}\ell_1 = \left(\frac{1}{2N^{\frac{1}{2}}}\rho\Lambda\right)Z - R(\Lambda)\left(\frac{1}{2}\rho\frac{1}{N^{\frac{1}{2}}}\right),$$

$$\frac{\partial^2}{\partial\mu_p\partial\ln\sigma_s^2}\ell_1 = \left(-\frac{1}{2N^{\frac{1}{2}}}(1-\rho)\Lambda\right)Z - R(\Lambda)\left(-\frac{1}{2}(1-\rho)\frac{1}{N^{\frac{1}{2}}}\right),$$

$$\frac{\partial^2}{\partial\mu_p\partial\ln\sigma_p^2}\ell_1 = \left(-\frac{1}{2N^{\frac{1}{2}}}\rho\Lambda\right)Z - R(\Lambda)\left(-\frac{1}{2}\rho\frac{1}{N^{\frac{1}{2}}}\right),$$

$$\frac{\partial^2}{\partial\ln\sigma_p^2\partial\ln\sigma_s^2}\ell_1 = \left(\frac{1}{4}\rho(1-\rho)\Lambda^2\right)Z - R(\Lambda)\left(\frac{3}{4}\rho(1-\rho)\Lambda\right).$$

A.2 ℓ_{2v}: The "integral-part"

ℓ_{2v} equals:

$$\ell_{2v} = \ln\int f(x|\theta)\Phi[\lambda(\theta)]g(\theta)d\theta = \ln\kappa_0.$$

Next, we present a general rule to calculate first- and second-order derivatives of ℓ_{2v}

A.2.1 Taking the first-order derivative w.r.t. a parameter in η:

$$\frac{\partial}{\partial\eta_i}\ell_{2v} = \frac{\frac{\partial\kappa_0}{\partial\eta_i}}{\kappa_0}, \qquad \eta_i \in \{\mu_p, \ln\sigma_p^2, \mu_s, \ln\sigma_s^2\}. \tag{63}$$

A.2.2 Taking the second-order derivative with respect to a parameter in η:

$$\frac{\partial^2}{\partial\eta_i\partial\eta_j}\ell_{2v} = \frac{\frac{\partial\kappa_0}{\partial\eta_i\partial\eta_j}}{\kappa_0} - \frac{\frac{\partial\kappa_0}{\partial\eta_i}}{\kappa_0}\frac{\frac{\partial\kappa_0}{\partial\eta_j}}{\kappa_0}, \qquad \eta_i, \eta_j \in \{\mu_p, \ln\sigma_p^2, \mu_s, \ln\sigma_s^2\}. \tag{64}$$

Using Eq. (63) and (64) and performing some algebraic manipulation we obtain

the first and second order derivatives of ℓ_{2v}. They are a function of the following eleven conditional expectations[1]:

[1] Note that B does not appear in the list.

$$
\begin{aligned}
A &= E_h\left(R|x_v\right), \\
C &= E_h[(\theta-\mu_p)|x_v], \\
D &= E_h[(\theta-\mu_p)^2|x_v], \\
E &= E_h[(\theta-\mu_p)^3|x_v], \\
F &= E_h[(\theta-\mu_p)^4|x_v], \\
G &= E_h[(\theta-\mu_s)R|x_v], \\
H &= E_h[(\theta-\mu_s)^2R|x_v], \\
I &= E_h[(\theta-\mu_s)^3R|x_v], \\
J &= E_h[(\theta-\mu_s)(\theta-\mu_p)R|x_v], \\
K &= E_h[(\theta-\mu_s)(\theta-\mu_p)^2R|x_v], \\
L &= E_h[(\theta-\mu_p)R|x_v], \\
M &= E_h[(\theta-\mu_p)^2R|x_v],
\end{aligned}
$$

where $E_h(.)$ denotes conditional expected value in the posterior density of θ given the data, and R is a shorthand notation for $R[\lambda(\theta)]$. For instance A is written fully as

$$
A = \frac{\int R[\lambda(\theta)]f(x|\theta)\Phi[\lambda(\theta)]g(\theta)d\theta}{\int f(x|\theta)\Phi[\lambda(\theta)]g(\theta)d\theta}.
$$

Now, for a single subject, first and second order derivatives are presented in the next subsections.

A.3 First order derivatives

$$
\begin{aligned}
\frac{\partial}{\partial\mu_p}\ell_{2v} &= \frac{1}{\sigma_p^2}C, \\
\frac{\partial}{\partial\ln\sigma_p^2}\ell_{2v} &= -\frac{1}{2}+\frac{1}{2\sigma_p^2}D, \\
\frac{\partial}{\partial\mu_s}\ell_{2v} &= -\frac{1}{\sigma_s}A, \\
\frac{\partial}{\partial\ln\sigma_s^2}\ell_{2v} &= -\frac{1}{2\sigma_s}G.
\end{aligned}
$$

A.4 Second order derivatives

$$\frac{\partial}{\partial \mu_p^2}\ell_{2v} = -\frac{1}{\sigma_p^2} + \frac{1}{\sigma_p^4}D - \frac{\partial^2}{\partial \mu_p^2}\ell_{2v},$$

$$\frac{\partial}{\partial \left(\ln \sigma_p^2\right)^2}\ell_{2v} = -\frac{1}{\sigma_p^2}D + \frac{1}{4} + \frac{1}{4\sigma_p^4}F - \frac{\partial^2}{\partial \left(\ln \sigma_p^2\right)^2}\ell_{2v},$$

$$\frac{\partial^2}{\partial \left(\mu_s\right)^2}\ell_{2v} = -\frac{1}{\sigma_s^3}G - \frac{\partial^2}{\partial \left(\mu_s\right)^2}\ell_{2v},$$

$$\frac{\partial^2}{\partial \left(\ln \sigma_s^2\right)^2}\ell_{2v} = \frac{1}{4\sigma_s}G - \frac{1}{4\sigma_s^3}I - \frac{\partial^2}{\partial \left(\ln \sigma_s^2\right)^2}\ell_{2v}.$$

A.5 Mixed derivatives

$$\frac{\partial}{\partial \mu_p \ln \sigma_p^2}\ell_{2v} = -\frac{3}{2\sigma_p^2}C + \frac{1}{2\sigma_p^4}E - \frac{\partial}{\partial \mu_p}\ell_{2v}\frac{\partial}{\partial \ln \sigma_p^2}\ell_{2v},$$

$$\frac{\partial^2}{\partial \mu_s \ln \sigma_s^2}\ell_{2v} = \frac{1}{2\sigma_s}A - \frac{1}{2\sigma_s^3}H - \frac{\partial}{\partial \mu_s}\ell_{2v}\frac{\partial}{\partial \ln \sigma_s^2}\ell_{2v},$$

$$\frac{\partial}{\partial \mu_p \ln \sigma_s^2}\ell_{2v} = -\frac{1}{2\sigma_s}\frac{1}{\sigma_p^2}J - \frac{\partial}{\partial \mu_p}\ell_{2v}\frac{\partial}{\partial \ln \sigma_s^2}\ell_{2v},$$

$$\frac{\partial}{\partial \mu_p \mu_s}\ell_{2v} = -\frac{1}{\sigma_s \sigma_p^2}L - \frac{\partial}{\partial \mu_p}\ell_{2v}\frac{\partial}{\partial \mu_s}\ell_{2v},$$

$$\frac{\partial}{\partial \ln \sigma_p^2 \mu_s}\ell_{2v} = \frac{1}{2\sigma_s}A - \frac{1}{2\sigma_s}\frac{1}{\sigma_p^2}M - \frac{\partial}{\partial_p \ln \sigma_p^2}\ell_{2v}\frac{\partial}{\partial \mu_s}\ell_{2v},$$

$$\frac{\partial}{\partial \ln \sigma_p^2 \ln \sigma_s^2}\ell_{2v} = \frac{1}{4\sigma_s}G - \frac{1}{4\sigma_s}\frac{1}{\sigma_p^2}K - \frac{\partial}{\partial_p \ln \sigma_p^2}\ell_{2v}\frac{\partial}{\partial \ln \sigma_s^2}\ell_{2v}.$$