Bayesian smoothing and inference in longitudinal models with applications to pediatric developmental scores

Author: Kevin L.W. Duisters

Supervisors: dr. E.M.L. Dusseldorp  
Prof. dr. A.W. van der Vaart  
Mathematical Institute, Leiden University

External Supervisor:  
Prof. dr. S. van Buuren  
Utrecht University, TNO Leiden

September, 2015
## Contents

Preface 3

PART I: Model

1 Introduction 5

2 Sequential semiparametric Bayesian estimation of latent variables 5
   2.1 Statistical Model 5
   2.2 Related work 6
   2.3 Auxiliary variable regression 9
   2.4 Posterior distributions 9
   2.5 Sequential Bayesian shrinkage 10

PART II: Methodology and Computational Statistics

3 Truncated mixture normal distributions 12
   3.1 Probit, logit and the mixture normal approximation 12
   3.2 Truncated random variables 13
   3.3 Sampling 13
   3.4 Link versus scale invariance 14

4 Empirical Bayesian reference estimation 15
   4.1 Marginal likelihood 15
   4.2 Related work 16
   4.3 MCEM: Empirical Gibbs 17
   4.4 Full conditionals 19
   4.5 Parameter identification 21

5 Simulation study 22
   5.1 Simulation set-up 22
   5.2 Initialization and perfect response profiles 23
   5.3 Parameter recovery and convergence 27

6 Sequential Bayesian estimation and inference 29
   6.1 Gibbs sampler using reference parameters 29
   6.2 Inference 30

PART III: Applications to Developmental score data TNO

7 Introduction to the D-score and SMOCC data 32
   7.1 Description and background 32
   7.2 Descriptive statistics SMOCC 32
   7.3 Rasch Model 33
   7.4 D-score estimation 34
   7.5 Inference in the reference diagram 34
8 Reference estimation on the SMOCC data 35
  8.1 Initialization .......................................................... 35
  8.2 Model set-ups: polynomial degree and covariates .................. 36
  8.3 Sensitivity analysis \( w_0 \) ......................................... 40
  8.4 Probit vs logit and the link scale .................................. 40
  8.5 Summary of reference procedure .................................. 41

9 Subject specific updating on the SMOCC data 42
  9.1 Sequential estimation and shrinkage ................................ 42
  9.2 Covariate corrected reference diagrams .......................... 43
  9.3 Subject specific inference .......................................... 44

10 Comparison of methods 44
  10.1 Reference parameters ............................................. 44
  10.2 Cross-validation .................................................. 44
  10.3 Summary of comparison ........................................... 46

PART IV: Discussion and Conclusion
Discussion 48

Conclusion 49

APPENDICIES
A Appendix Part I 53
  A.1 Formal definition Frequentist and Bayesian Statistics .......... 53
  A.2 Basic definitions .................................................. 54

B Appendix Part II: Methodology and Computational Statistics 56
  B.1 Posterior distribution multivariate normal conjugate update .... 56
  B.2 Mixture normal approximation to logistic density ............... 56
  B.3 Sampling from a truncated normal ................................ 57
  B.4 LMS Likelihood details ........................................... 59

C Appendix Part III: Application D-score TNO 61
  C.1 Elapsed time ...................................................... 61
  C.2 Reproduction prior mcurve Van Buuren (2014) ................... 61
  C.3 Summary of reference procedure \( q = 2 \), without covariates, \( w_0 = 0 \) .... 62
Preface

This work is written as thesis in partial fulfilment of the degree M.Sc. Mathematics: Statistical Science for the Life and Behavioural Sciences, and represents its final requirements. Over the past six months, I have spent a considerable amount of time at TNO Life Sciences in Leiden. The applied statistical problems they encounter regarding child health monitoring have proven to be both substantially interesting as well as theoretically challenging.

It has been a privilege to have the author of one of the most important underlying studies for my thesis as a supervisor. In addition to his guidance and expertise, he has given me the freedom to develop my own ideas irrespective of whether this would validate or contradict his earlier work. For this I am highly grateful to Stef van Buuren.

Elise Dusseldorp has made great effort in helping me to understand the field and the specifics of the data application. Her insights and regular advice (at non-regular office hours) have significantly improved this thesis. As supervisor at TNO Leiden she has fulfilled an important role in embedding my work within the organization and introducing me to several experts. Moreover, she has allowed my thoughts to move into theoretical explorations of which the practical usability or planning was not always well defined beforehand. I am thankful for her trust and guidance.

Furthermore, I appreciate the time Aad van der Vaart has taken to listen to and read the more technical details of my work during several appointments. His questions were thought provoking and have contributed to the proposed methods and my academic development.

As this thesis represents a milestone, time and place feel appropriate to thank some persons that have been important over the past years. With their dedication to the field and its students, Jacqueline Meulman and Willem Heiser have turned my interest for statistical science into a career choice. I look forward to tapping into their knowledge and motivation for many years to come. On a more personal level, my gratitude towards MM, JD and ID is beyond words.

Leiden, September 2015
Part I: Model
1 Introduction

Van Buuren (2014) recently promoted new techniques in monitoring human development, including that of cognitive nature. With respect to conventional growth charts of continuous measures like height and weight, displaying a development score (“D-score”) is more challenging since ability is latent. Jacobusse, Van Buuren, and Verkerk (2006) have been successful in fitting a psychometric model to imply such abilities on a dataset of 2151 Dutch infants (“SMOCC” data, Herngreen et al., 1992), which Van Buuren (2014) later formalized to a D-score diagram including reference population curves.

Although practitioners were fast to recognize the potential of such diagrams in early detection of abnormal child development, some issues remained. Firstly, individual curves tended to be less smooth than expected of human development. Secondly, the method did not take the longitudinal character of the data into account, causing intra-subject correlations of the D-score to be low. Thirdly, the estimation procedure relied on a subjective prior ability distribution on each time point (irrespective of subject) to overcome low data availability at certain visits, to which its outcome was rather sensitive. Fourthly, it was hypothesized that background information could assist in modelling development more accurately.

In this paper an attempt is made to resolve these issues by casting the D-score estimation in a Bayesian longitudinal framework. A hierarchical model inspired by Azevedo, Fox, and Andrade (2015) and Fox and Glas (2001) is extended for use in longitudinal analysis with non-equidistant visit times by including a Gaussian Process (GP). Furthermore, attention is paid to modelling children using subject specific information. The estimation procedure now consists of two phases. Firstly, an Empirical Bayesian procedure determines a reference growth diagram, taking into account the longitudinal nature of the data. Secondly, inference is made for new children on a subject specific basis using a fully Bayesian sequential updating paradigm. In case of low information availability, estimates are shrunk towards a reference curve adjusted for covariates. As information becomes increasingly available over time, the reference is mixed with subject specific latent ability estimates.

The GP has a long standing history in machine learning, where it is widely used for both classification as well as regression (Rasmussen and Williams, 2006). Furthermore, the GP is often seen in spatio-temporal statistics, where it is known as kriging (Krige, 1951). More recently, it has found its way into applied longitudinal analysis in the medical field. Shi et al. (2012) describe the advantages of combining a parametric mixed model specification with a non-parametric GP to model subject specific longitudinal patterns (in their case, dose-response curves) flexibly. The major benefit of the GP is that it accommodates patient based estimation and prediction using a low amount of parameters.

This paper is structured as follows. In section 2 of Part I, the statistical model and the Bayesian framework are outlined. It does so as general as possible, since this approach is not limited to data of human development. Part II includes details on the estimation procedures for both the reference as well as the subject specific phases. Being decidedly pragmatic in our Bayesian approach, a reasonable amount of attention is paid to describing to what extent the procedures are dependent on subjective input and to the freedom of methods researchers have in this respect. A simulation study is included to demonstrate the efficacy of the proposed reference procedure. In Part III, the model is applied to the SMOCC data and a comparison is made against the results of Van Buuren (2014). Finally, Part IV contains a discussion and conclusion. Mathematical details are included in the appendices.

2 Sequential semiparametric Bayesian estimation of latent variables

2.1 Statistical Model

We consider a hierarchical binary model for random variable $Y_{ijt} \in \{0, 1\}$ for subject $i$, item $j$ at time $t$. Assume we observe 0/1 scores for $N$ subjects, each associated with their own set of observation times $\tau_i := \{t_k : k \in T\}$ where $T$ represents the continuous time scale and $i \in [1, n]$ the subject. In case multiple
items are observed at the same time, they are grouped into notation $\mathcal{J}_{ij} := \{j : t \in \tau_i\}$, with $j \in [1, J]$.

**Definition 2.1** (Model I).

\[
Y_{ijt} | p_{ijt} \sim \text{Bin}(1, p_{ijt}) \tag{1}
\]

\[
g(p_{ijt}) = \theta_{it} - \delta_j \tag{2}
\]

\[
\theta_i = X_i \xi_i + f_i \tag{3}
\]

\[
\xi_i = W_i \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + b_i \tag{4}
\]

\[
f_i | v_0, w_0 \sim \text{GP}(0, K) \tag{5a}
\]

\[
b_i | \Sigma_b \sim N_{q+1}(0, \Sigma_b) \tag{5b}
\]

\[
K(t_k, t_l) = v_0^2 \cdot \exp(-\frac{1}{2w_0^2}|t_k - t_l|^2) \tag{6}
\]

At level (2), $g(\cdot)$ is a link function that maps the conditional mean onto a linear predictor, in case of the Bernoulli distribution $\mathbb{E}(Y_{ijt} | p_{ijt}) = p_{ijt}$. Notice that the model works with vectors $\theta_i \in \mathbb{R}^{|\tau_i|}$, such that correlations between visits can be introduced. Matrices $A_i \in \{0, 1\}^{Y_i \times |\tau_i|}$ and $\Delta_i \in \{0, 1\}^{Y_i \times J}$, used in notation later, transform $\theta_i$ respectively $\delta \in \mathbb{R}^J$ to their long formats for subject $i$ with length $|Y_i|$.

The latent variable is defined at level (3) and we consider its distribution to be the main object of interest. It is modelled through a regression function with errors $f_i$ distributed as GP at level (3). The GP has mean zero and covariance kernel $K$, which level (6) specifies as squared exponential (cf. e.g. Rasmussen and Williams, 2006) with parameters $v_0$ and $w_0$. The time input is later understood to be subject specific, but here the kernel is defined generally for scalars $t_k$ and $t_l$. Notice that the (continuous) time dependence of $f_i$ opens up the possibility to interpret $\theta_i$ as functional form (more on this in subsection 2.2). The regression parameters $\xi_i$ are modelled further at level (4), where the distribution of $b_i$ is used. We assume throughout that $f$ and $b$ are independent.

For the purposes of our application, we think of $X_i \in \mathbb{R}^{|\tau_i| \times (q+1)}$ as design matrix including an intercept and a time polynomial of degree $q$. Let $X_i(k,h) := (\tau_i)_{(h-1)}$ define the $(k,h)$ element of $X_i$, where $\tau_i$ is the $k$-th observation time in vector $\tau_i$. Furthermore, $W_i$ could be structured as $W_i = [I_{q+1} \Gamma_i]$, such that $\beta$ carries over to level (3) parametrizing polynomial of time. In case $\Gamma_i$ only has elements in its first row, $\gamma$ can be interpreted as additive effects (e.g. from covariates). Elements in other rows model interaction effects with time (or more generally, with variables in $X$).

To make this model fully Bayesian, priors must be placed on $\delta, \beta, \gamma, \Sigma_b, v_0$ and $w_0$. Unless mentioned otherwise, these are always non-informative Gaussian. In case of $\Sigma_b$, a non-informative choice would be to use an Inverse-Wishart distribution with the inverse sum of squares matrix about the sample mean as scale matrix (Gelman et al., 2014), although many alternatives exist (Chung et al., 2015).

### 2.2 Related work

#### 2.2.1 Psychometrics

When $\theta_i(t)$ is not considered as function but as time indexed latent variable $\theta_{it}$, a Generalized Linear Mixed Model (GLMM; Breslow and Clayton, 1993) is fit. Given our application, the most relevant related field is that of psychometrics, in which $\theta$ is referred to as latent ability of subjects and $\delta$ as difficulty of items. In case $g(\cdot)$ is taken to be the logit link, the model specified by level (2) is formally equivalent to the Rasch
model (Rijmen et al., 2003). In that case, \( \theta \) is considered as random and \( \delta \) as fixed effects. When a probit link is used instead, definition 2.1 (again at the same level) is known as the normal ogive model. A generalization can be made by allowing non-zero elements in \( A \) to be different from 1 (the so called discrimination parameter). In case this is done under the logit link the model is known as 2PL IRT. Without the longitudinal perspective, definition 2.1 is close to the work of Fox and Glas (2001) in modelling latent ability across two levels (i.e., level (3) and (4)) in a Bayesian framework. There, \( g(\cdot) \) is taken to be the probit link, such that level (2) specifies a normal ogive model.

Reintroducing the longitudinal character of the data, Azevedo, Fox, and Andrade (2015) start from the normal ogive model and specify a multivariate normal distribution for \( \theta \) given mean vector and covariance matrix. The focus of their study is on assessing different covariance matrix structures (with varying amounts of parameters) while estimating in a Bayesian framework. Azevedo, Fox, and Andrade (2015) pay little attention to modelling the mean vector, although to make the approach fully Bayesian they assume it is multivariate normal with some prior mean and covariance. A downside of most covariance matrices reviewed by Azevedo, Fox, and Andrade (2015) is that observations must be equally timed over subjects and equally spaced within subjects.

### 2.2.2 Functional regression

One way to generalize to situations of continuously spaced observations is by moving from discrete time indexed latent variables \( \theta \) to functional forms \( \theta(t) \). Shi and Wang (2014) do so by introducing the GP called \( f \) in definition 2.1, as generalization based on earlier work for continuous outcome variables (Shi et al., 2012). In the latter, the authors have introduced the notion of combining a parametric mixed-effects part with a GP thus creating a semiparametric model which they refer to as ME-GPFR. Besides the ME fixed and random effects, Shi et al. (2012) include scalar covariates with time-varying parameters. For instance, a spline model is fit on the differential gender effect. Notice these are "fixed" effects too, the difference being that in the fixed ME-part, the covariates are time-varying and the parameters fixed, whereas the reverse holds in the scalar covariate part. In definition 2.1, both types of fixed effects are grouped into \( \gamma \), which can be extended towards time-dependence. Although suggesting the possibility, Shi and Wang (2014) use an Empirical Bayes method instead of a fully Bayesian approach, in contrast to Azevedo, Fox, and Andrade (2015). Bayesian estimation is discussed further in section 4.

### 2.2.3 LMM versus GPFR

Given its importance to formalize the equivalence of our work to that of Azevedo, Fox, and Andrade (2015), a brief intermezzo on the relationship between covariance specification in LMM and the GP kernel is outlined here.

**Definition 2.2 (Gaussian Process).** A real valued stochastic process \( X := \{X_t : t \in T\} \) is Gaussian if \( \forall a_1, \ldots, a_n \in \mathbb{R} \) and \( \forall t_1, \ldots, t_n \in T \) with \( n \in \mathbb{N} \), the linear combination \( Y := \sum_{i=1}^{n} a_i X_{t_i} \) has a (univariate) normal distribution.

Definition 2.2 can be alternatively formulated in terms of the multivariate normal distribution.

**Definition 2.3 (GP definition adapted from Rasmussen and Williams (2006, p.13)).** A real valued stochastic process \( X := \{X_t : t \in T\} \) is Gaussian if any subset of random variables \( X_s \subseteq X \) has a (joint multivariate) normal distribution.

In the context of model definition 2.1, observations of the GP are discretized on level (3) towards a multivariate normal distribution by observing \( \theta \) on a finite set of times \( \tau_i \). Define the subject specific covariance matrix by \( \Psi_{\theta,j} \) where its \((k,l)\) element equals \( K(\tau_{ij}, \tau_{lj}) \) and let \( \mu_{\theta,j} := X_i \xi_i \). Then,

\[
\theta_i | \mu_{\theta,j}, \Psi_{\theta,j} \sim N(\mu_{\theta,j}, \Psi_{\theta,j})
\]

---

\[\text{To the extent one is willing to assume the Rasch model is equivalent to a 1PL IRT model}\]

\[\text{Mixed-effects Gaussian Process Functional Regression}\]
A notable difference with Azevedo, Fox, and Andrade (2015) is that we provide all subjects with their own mean and covariance matrix, which becomes visible in the subscript \(i\). To be clear on notation with respect to the GP, \(\Psi_{\theta,i}\) represents a covariance matrix and \(K\) is used to specify the covariance kernel through which it is generated. Since many parameters depend on specific observation times \(\tau_i\), the subscript \(i\) is occasionally dropped from notation throughout the remainder of this work.

Since the mixed effects distinction “fixed” and “random” is unfortunate in the Bayesian setting, we adopt the “non-varying” and “varying” terminology of Chung et al. (2015). Although the distinction between (3) and (4) is important for estimation, rewriting the latent ability in model 2.1 leads to a well-known interpretation.

**Definition 2.4** (Ability as LMM).

\[
\begin{align*}
\theta_i &= X_i(\beta + b_i) + u_i \gamma + f_i \\
\text{Cov}(\theta_i|\beta, \gamma) &= X_i\Sigma_bX_i' + \Psi_{\theta,i}
\end{align*}
\]

Hence we see that without priors on \(\beta, \gamma, \Sigma_b, \nu_0\) and \(w_0\), the model reduces to an LMM. In case \(\Sigma_b\) is diagonal, this is a random intercept random slope model, where \(\beta, \gamma\) are considered as non-varying, and \(b_i\) as varying parameters. The subject specific variables \(u_i\) can be thought of as background information that does not change over time (e.g. gender, duration of pregnancy), which fill the \(\Gamma_i\) design matrix. In this case they enter as additive effects, but the model can be extended by introducing a \(\gamma\beta\) interaction. The conditional variance specification makes the role of \(\Psi_{\theta,i}\) as residual variance in the LMM apparent. In its most simple form, this covariance matrix would reduce to a diagonal \(\sigma^2_{\epsilon|\tau_i}\). Another structure with only one parameter is the continuous AR1. Furthermore, the GP can be parametrized more heavily by introducing heterogeneous \(w\) instead of a single \(w_0\).

### 2.2.4 Growth curve modelling

From the perspective of the application, our work is related to that of modelling growth curves. As noted in the introduction, model definition 2.1 is an attempt to describe latent ability (human development) and interpret it in a similar fashion as is currently done with physical growth. A general overview of methodologies in this field can for instance be found in Johnston, Roche, and Susanne (1980). A common method to describe growth curves is the Jenns-Bayley (JB) non-linear mixed effects model (Jenss and Bayley, 1937). For simplicity without covariates \(\gamma\), this model would specify:

**Definition 2.5** (Ability as JB model). Let \(t\) denote a time for subject \(i\), i.e. \(t \in \tau_i\).

\[
\theta_{it} = (\beta_0 + b_{i,0}) + (\beta_1 + b_{i,1})t - \exp((\beta_2 + b_{i,2}) + (\beta_3 + b_{i,3})t) + \epsilon_{it}
\]

Here, \(b_i\) are zero-mean multivariate normal distributed random effects and \(\epsilon_{it}\) i.i.d. Gaussian. Remark that the JB model specifies a random intercept-random slope structure corrected by a (subject specific) non-linear exponential term.

Although less common in the field, we argue that we can capture the same individual growth curves (of ability) using a polynomial of time. For the same amount of \(\beta\) parameters as the JB model (not taking into account variance specification), a subject specific cubic polynomial including intercept is possible. Using such a linear model instead of the non-linear JB method is necessary for the Bayesian estimation through Gibbs sampling described in section 4. Fortunately, this comes at little cost, since most growth curves can be mimicked well by a cubic (or even a quadratic) polynomial. Furthermore, note that our model is more general in (multivariate) modelling of the residual variance \(\epsilon_i\). The JB model is based on a diagonal covariance matrix, where we specify the GP.
2.2.5 Extensions

Further extensions that can be considered are those based on the choices made in definition 2.1. For instance, the degree of the polynomial \( \beta \) could be altered, as well as the (structure of the) covariance matrix \( \Sigma_0 \) on \( \xi_i \) at level (3). Furthermore, a different covariance kernel can be introduced at level (6). Finally, the level of subjectivism in the hyperparameters can be varied.

2.3 Auxiliary variable regression

Define for all combinations \( ij \) the auxiliary variable \( Z_{ij} = \theta_{ij} - \delta_j + \epsilon_{ij} \) such that \( Y_{ij} = 1_{Z_{ij} > 0} \). The distribution of \( Z_{ij} \) enters through a distribution on \( \epsilon_{ij} \). Retrieving model definition 2.1 using continuous \( \epsilon_{ij} \) that are easy to sample from instead of linking levels (1) and (2) has shown to be useful in Bayesian estimation (e.g. Gelman et al., 2014, p.408).

**Proposition 2.1** (Auxiliary variable probit regression). Let \( \epsilon_{ij} \sim N(0,1) \). Then model definition 2.1 is retrieved under the probit link.

**Proof.** Consider definition 2.1 where conditional on \( \theta \) and \( \delta \) we have \( p_{ij} = g^{-1}(g(p_{ij})) = \Phi(\theta_{ij} - \delta_j) \) the standard normal CDF, i.e. the probit model. Let \( m := \theta_{ij} - \delta_j \). The proof follows by a change of variable \( x' := x - m \) in the integral.

\[
P(Y = 1|m,p) = P(Z > 0|p,m) = 1 - P(Z \leq 0|p,m) = 1 - \int_{-\infty}^{m} \Phi(x') \, dx' = 1 - \Phi(-m) = \Phi(m)
\]

Similarly, Holmes and Held (2006) and Albert and Chib (1993) describe auxiliary variable methods that are equivalent to model levels (1) and (2) in case of a logit link.

**Proposition 2.2** (Auxiliary variable logit regression). Let \( \epsilon_{ij} \sim Log(0,1) \), the logistic distribution with mean zero and scale one. Then model definition 2.1 is retrieved under the logit link.

**Proof.** Consider definition 2.1 where conditional on \( \theta \) and \( \delta \) we have \( p_{ij} = g^{-1}(g(p_{ij})) = \text{logit}^{-1}(\theta_{ij} - \delta_j) \), i.e. the logit model. Let \( m := \theta_{ij} - \delta_j \). Denote the logistic density with location \( m \) and scale 1 evaluated at \( x \) as \( \pi^L(x;m,1) \) and its distribution ("logistic") function as \( \Pi^L(x;m,1) \).

\[
\pi^L(x;m,1) = \frac{\exp(-(x-m))}{1 + \exp(-(x-m))^2} \quad \Pi^L(x;m,1) = \frac{1}{1 + \exp(-(x-m))}
\]

\[
P(Y = 1|p,m) = P(Z > 0|p,m) = 1 - P(Z \leq 0|p,m) = 1 - \int_{-\infty}^{0} \pi^L(x;m,1) \, dx = 1 - \Pi^L(0;m,1) = 1 - \frac{1}{1 + \exp(m)} = \frac{\exp(m)}{1 + \exp(m)}
\]

2.4 Posterior distributions

To end this section, we present an expression for the posterior distribution of the parameters given the data \( Y \) (through auxiliary variable \( Z \)) and all design matrices (excluded from notation). We rely on Bayes’ Rule A.1 and specify the posterior up to proportionality. Remark that we slightly abuse notation and denote all densities by \( \pi \). Details on Bayesian notation and theory can be found in appendix A.1.2.
with subject specific data as it becomes available sequentially (i.e. longitudinally). As time progresses, the
which Azevedo, Fox, and Andrade (2015) later adapted for use in their paper on longitudinal covariance
θ variable, discussed further in section 3. Independence of items given
the Empirical Bayes estimation to be discussed in full generality. Concerning (sampling from) truncated normal mixture distributions is presented as this is a prerequisite for
sampler becomes available under the logit link as well. Before going into estimation details, an intermezzo
conjugacy. However, by using a mixture normal approximation to the logistic distribution, the Gibbs
normal variable (definition 2.1), which is particularly easy to sample from and does well in terms of
probabilistic ogive model, i.e. \( g(\cdot) \) in definition 2.1 is the probit link, as this has computational advantages as opposed to the logit link (e.g. Gelman et al., 2014, p. 408). The reason is that probit regression depends on a
normal variable (definition 2.1), which is particularly easy to sample from and does well in terms of
conditioned on \( \cdot \) implies that the continuous Gaussian process is discretized into a multivariate normal distribution
\( \pi(\theta) \times \pi(\beta) \times \pi(\gamma) \times \pi(\Sigma_b) \times \pi(v_0, w_0) \)

Remark that (7) assumes latent values \( z_{ijt} \) are conditionally independent and a product measure is formed over \( i, J_i \) and \( \tau_i \). Depending on the probit or logit specification, this is either a normal or logistic truncated variable, discussed further in section 3. Independence of items given \( \theta \) and \( \delta \) is known as "local stochastic independence" in the IRT literature. The subject independence is a natural assumption and the conditional time independence, which implies that given \( \theta_i(t) \) items are exchangeable across time, is standard in practice as well. The conditional independence of subjects is repeated in (8) and (9). Conditioning on observation times \( \tau_i \) implies that the continuous Gaussian process is discretized into a multivariate normal distribution as in definition 2.4. Extensions (e.g. specifying priors on \( \delta, \gamma \) and/or hyperparameters) can be introduced by adding detail to the distributions in the product of proposition 2.3 at (10).

The posterior in proposition 2.3 is analytically intractable. Fox and Glas (2001) have suggested a specific Gibbs sampling scheme for Bayesian estimation of multilevel IRT models based on work by Albert (1992), which Azevedo, Fox, and Andrade (2015) later adapted for use in their paper on longitudinal covariance structures (cf. section 2.2). Computational details of the Gibbs sampling scheme as well as hyperparameter estimation using Empirical Bayes can be found in section 4.

2.5 Sequential Bayesian shrinkage

Our main goal is to find a posterior distribution on \( \theta_{ij} \), based on observations at times \( \tau_i \) of subject specific history shrunk towards population averages at times where data availability is low. The approach is therefore split in two, firstly estimating a reference latent curve (once) and secondly mixing this information with subject specific data as it becomes available sequentially (i.e. longitudinally). As time progresses, the balance moves increasingly towards subject specific estimation.

In the next part, a(n Empirical) Bayesian estimation procedure is proposed for the reference parameters and a fully Bayesian procedure for the subject specific estimation. The reference procedure is accompanied by a simulation study to show its efficacy. In Part III, both the reference estimation as well as the subject specific procedure are demonstrated on a real dataset, followed by a discussion on sequential mixing of information.

Both Bayesian estimations rely heavily on the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984). Here, we initially follow Albert (1992) and Fox and Glas (2001) in parametrizing a normal ogive model, i.e. \( g(\cdot) \) in definition 2.1 is the probit link, as this has computational advantages as opposed to the logit link (e.g. Gelman et al., 2014, p. 408). The reason is that probit regression depends on a normal variable (definition 2.1), which is particularly easy to sample from and does well in terms of
conjugacy. However, by using a mixture normal approximation to the logistic distribution, the Gibbs
sampler becomes available under the logit link as well. Before going into estimation details, an intermezzo
concerning (sampling from) truncated normal mixture distributions is presented as this is a prerequisite for
the Empirical Bayes estimation to be discussed in full generality.

\[
\pi(Z, \delta, \theta, \xi, \beta, \gamma, \Sigma_b, v_0, w_0 | Y) \propto \pi(Z | Y, \delta, \theta) \times \pi(\delta) \times \pi(\theta | \xi, v_0, w_0) \times \pi(\xi | \beta, \gamma, \Sigma_b)
\]

\[
\times \pi(\beta) \times \pi(\gamma) \times \pi(\Sigma_b) \times \pi(v_0, w_0)
\]

\[
\propto \left( \prod_{i=1}^{n} \prod_{j \in J_i} \prod_{t \leq \tau_i} \pi(Z_{ijt} | \theta_{ij}, \delta_{ij}, Y_{ijt})(1(z_{ijt} > 0)(Y_{ijt} = 1) + 1(z_{ijt} \leq 0)(Y_{ijt} = 0)) \right)
\]

\[
\times \left( \prod_{i=1}^{n} \frac{1}{\sqrt{\Sigma_b}} \exp\left( -\frac{1}{2}(\xi_i - X_i \xi_i) \Sigma_b^{-1} (\xi_i - X_i \xi_i) \right) \right)
\]

\[
\times \left( \prod_{i=1}^{n} \frac{1}{\sqrt{\Sigma_b}} \exp\left( -\frac{1}{2}(\xi_i - W_i \xi_i) \Sigma_b^{-1} (\xi_i - W_i \xi_i) \right) \right)
\]

\[
\times \pi(\delta) \times \pi(\beta) \times \pi(\gamma) \times \pi(\Sigma_b) \times \pi(v_0, w_0)
\]
Part II:
Methodology and Computational Statistics
3 Truncated mixture normal distributions

Before discussing truncation, we revisit the probit and logit link, and suggest an approximation to auxiliary logit regression of definition 2.2.

3.1 Probit, logit and the mixture normal approximation

In case of a latent probit regression, the full conditionals are easy to sample from and the Gibbs procedure presented in section 4.3.3 will generally be fast. However, a latent logistic regression may be preferred for two reasons. Firstly, since some models inherently rely on the logit link (e.g. the Rasch model), which has an interpretation that is preferred over that of the probit link by many in practice. Secondly, being canonical to the Bernoulli model, using a logit link has theoretical advantages in some estimation procedures. To be as general as possible at this point, we discuss the Gibbs scheme (Algorithm 2) for both links. Since the Gibbs sampler relies heavily on conjugate updates of the multivariate normal, we choose to approximate the logistic density by a mixture of normal densities.

When only one normal density is included in the mixture, a well-known scaling factor of 1.64 is obtained, i.e. \( \log \left( \frac{p}{1-p} \right) \approx 1.64 \cdot \Phi^{-1}(p) \) for \( p \in (0,1) \). Factors of 1.6 or 1.7 are also often seen in practice, depending on the wish to match on density or distribution and the matching criteria used. By including more normal densities in the mix, the approximation becomes better. In appendix B.2, worked out details are included. There we argue a normal mix of 5 densities to be best for our purposes, with the following characteristics.

**Definition 3.1** (normal mixture approximation, \( k = 5 \)). Let \( \pi^L(m, 1) \) denote the logistic density with location \( m \) and scale \( 1 \) as function of \( z \) on support \( (-\infty, \infty) \). We define the following normal mixture \( \phi^M(z; w, m, s^2) \) around mean \( m \) with weights \( w \) and standard deviations \( s \). Let \( \phi(z; m, s^2) \) represent the normal density with mean \( m \) and variance \( s^2 \) evaluated in the point \( z \). Appendix B.2 establishes a choice for \( k = 5 \) with the following weights.

\[
\pi^L(x; m, 1) \approx \phi^M(z; w, m, s^2) := \sum_{h=1}^{k} w_h \cdot \phi(z; m, s_h^2)
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w )</td>
<td>0.436</td>
<td>0.377</td>
<td>0.017</td>
<td>0.109</td>
<td>0.061</td>
</tr>
<tr>
<td>( s )</td>
<td>1.935</td>
<td>1.337</td>
<td>1.107</td>
<td>2.879</td>
<td>0.920</td>
</tr>
</tbody>
</table>

Figure 1 includes a qq-plot and a plot of distribution functions of this mixture density \( \phi^M(z; w, m, s^2) \) and a standard normal (probit) against the true logistic density.
3.2 Truncated random variables

When conditioned on the outcome $Y_{ij}$, the density of $Z$ is truncated. We define the density of a truncated normal mixture with weights $w$ and standard deviations $s$, all translated by mean $m$. Truncation is defined such that no probability mass is located outside of the interval $[a, b]$. In case $Y = 1$, $Z$ is left truncated hence $a = 0$ and $b = \infty$. In case $Y = 0$, right truncation implies $a = -\infty$ and $b = 0$.

Definition 3.2 (Truncated mixture normal variable).

$$
\pi(Z|Y, m, w, s) = \mathbb{1}_{z \in [a, b]} \cdot \left( \frac{\sum_{h=1}^{k} w_h \cdot \frac{1}{s_h} \phi\left( \frac{z-m}{s_h} \right)}{\sum_{h=1}^{k} \left( w_h \left( \Phi\left( \frac{b-m}{s_h} \right) - \Phi\left( \frac{a-m}{s_h} \right) \right) \right)} \right)
$$

It follows that $\pi$ is a density as it integrates to 1 and is non-negative on its support.

3.3 Sampling

Sampling from a truncated mixture can be done by sampling from single truncated variables under an adjusted weight vector, which we refer to as truncation weights $wt$.

$$
w_{th} := \frac{\Phi\left( \frac{b-m}{s_h} \right) - \Phi\left( \frac{a-m}{s_h} \right)}{\sum_{h=1}^{k} \left( \Phi\left( \frac{b-m}{s_h} \right) - \Phi\left( \frac{a-m}{s_h} \right) \right)} \in [0, 1] \quad \forall h \in [1, k]
$$

To sample an element of the mixture distribution, select an index $h \in [1, k]$ distributed according to $h \sim \text{Bin}(k, wt)$ and sample from a truncated normal variable with mean $m$ and standard deviation $s_h$.

The fastest way to sample from a single truncated normal variable is through the inverse CDF method using the probability integral transform. In case the mean $m$ is at risk of being far from the truncation point in a zero mass region, this method becomes affected by computational accuracy in the tail of the normal quantile and distribution function. Robert (1995) and Damien and Walker (2001) have suggested algorithms that can deal with these issues at the cost of computational efficiency. More information can be found in appendix B.3. Since in our practical situations the mean $m$ will be close to the truncation point or at least in a positive probability region, we suggest the use of the inverse CDF method.

Proposition 3.1 (Sampling a single truncated normal $N(m,s)$). Let the truncated density evaluated at point $z \in \mathbb{R}$ be denoted as follows.

$$
\pi(z; m, s, a, b) = \mathbb{1}_{z \in [a, b]} \cdot \frac{\frac{1}{z} \phi\left( \frac{z-m}{s} \right)}{\frac{1}{s} \Phi\left( \frac{b-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right)}
$$

Let $U \sim \text{UN}(0,1)$ and sample a realization $u$ from it. Then, a sample $z$ from $\pi$ is obtained by the following equation.

$$
z = m + s \cdot \Phi^{-1}\left( \Phi\left( \frac{a-m}{s} \right) + u \cdot \left( \Phi\left( \frac{b-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right) \right) \right)
$$

Proof. Inverse CDF method with probability integral transform:

$$
\Pi(z; m, s, a, b) := \int_{-\infty}^{z} \pi(z')dz' = \frac{1}{\Phi\left( \frac{b-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right)} \cdot \left( \Phi\left( \frac{z-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right) \right)
$$

$$
= u 
\iff 
\Phi\left( \frac{z-m}{s} \right) = \Phi\left( \frac{a-m}{s} \right) + u \cdot \left( \Phi\left( \frac{b-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right) \right)
\iff 
z = m + s \cdot \Phi^{-1}\left( \Phi\left( \frac{a-m}{s} \right) + u \cdot \left( \Phi\left( \frac{b-m}{s} \right) - \Phi\left( \frac{a-m}{s} \right) \right) \right)
$$

\qed
The following simulation study shows a sample of size $10^6$ from the probit (no mix) and logit (mixture normal) links left truncated at zero with $m = 1.5$ and $w$ and $s$ from definition 3.1.

![Probit and Logit Densities](image)

*Figure 2: Sampling $10^6$ values from truncated normal (probit) and mixture truncated normal $k = 5$ (logit) densities. Lines drawn in are theoretical probabilities for probit and logit (dotted) under the sampled values shown.*

Firstly, remark in the plot on the right hand side that the normal mixture approximation does a decent job (compare the histogram to the dotted line). Notice that the approximation is slightly less successful than suggested by figure 1, because the truncation places more mass in the tails (where the approximation becomes weaker). Secondly, by including the true density (for the respective values) of the other link as well, we observe in both plots that the fatter tails of the logistic density become increasingly apparent after truncation. The differences between the links are larger than one might expect on the basis of figure 1.

### 3.4 Link versus scale invariance

Subsection 4.5 elaborates on identification of $\theta$ and $\delta$ in model definition 2.1. Notwithstanding the conclusion that only the difference between ability and difficulty is identified, thereby making the model scale invariant at the latent level, such invariance does not carry over to the estimation procedure at the link level.

The distribution of the error term in the auxiliary variable regression causes the differences (essentially represented by $Z$) to be inherently scaled with high probability. This characteristic affects the relative positioning of item parameters in the longitudinal framework (and resulting placement of abilities). Invariance of the latent scale is preserved in the sense that the entire set of parameter estimates can be translated vertically (i.e. only differences identified), but once a single estimate is fixed, the link function affects relative positioning of all other parameters. This will become relevant in subsection 4.5.

A poor alignment of latent scale and link function eliminates discriminative power of items and expresses itself at all levels of the model. For example, imagine we are in a low scale situation in which the difficulties are close together in absolute value. Then, answering a low difficulty question correctly is likely to lead to a strong overestimation of ability (far in excess of more difficult items answered incorrectly), since the ability-difficulty difference is generated by the (poorly aligned) link function. Similarly, ability will be
strongly underestimated for items answered incorrectly. Although this may still produce mean estimates that are roughly correct, it causes variance parameters, e.g. \( v_0 \), to be insensible. A detailed discussion is included in subsection 8.4.

4 Empirical Bayesian reference estimation

In section 2, it was outlined that the proposed model (definition 2.1) rewritten as latent probit regression (definition 2.1) has a posterior density whose full conditionals are particularly easy to sample from using a Gibbs procedure (Geman and Geman, 1984; Gelfand and Smith, 1990). Here, we extend work of Albert (1992), Fox and Glas (2001) and Azevedo, Fox, and Andrade (2015) for use with the discretized GP, which directs towards multivariate normal conjugate updating. Furthermore, as section 3 points out, the benefits of latent probit regression are preserved for the logit link under a mixture normal approximation.

4.1 Marginal likelihood

The Gibbs procedure of Fox and Glas (2001) and Azevedo, Fox, and Andrade (2015) is a fully Bayesian Markov Chain Monte Carlo (MCMC) framework. However, Casella (2001) describes situations in which the researcher does not want or cannot specify priors on hyperparameters deep in the hierarchy. In our model, examples are the kernel parameters introducing structure to the ability covariance, whose full conditionals are non-conjugate and hence difficult to sample in the Gibbs framework. For consistency, we consider the unstructured \( \Sigma_b \) matrix to be such a type of hyperparameter as well.

Interest goes out to point estimation of parameters in the collection \( C := \{ (C_i := (\Sigma_{\theta_i}, \Psi_{\theta_i}) : i \in [1, n] \} \) based on data \( Y = \{ Y_i : i \in [1, n] \} \). Let all underlying parameters that make up the collection \( C \) be denoted by \( \Xi \), such that we can write \( C(\Xi) \) to signify the underlying parametrizations. \( \Sigma_{\theta} \) is a \((q+1) \times (q+1)\) symmetric matrix with \( \frac{1}{2}(q+1)(q+2) \) unique parameters. \( \Psi_{\theta,j} \) is parametrized by a limited amount of kernel parameters. Note, in a fully Bayesian framework one places priors on \( C \) and computes the following posterior:

\[
\pi(C|Y) = \frac{\pi(Y|C)\pi(C)}{\pi(Y)} \tag{11}
\]

Since we are interested in point estimates of \( C \) (after which \( \Xi \) can be estimated), we could look at EAP or MAP estimates of this posterior. Particularly for the MAP, the reasoning follows through that it is unnecessary to compute the normalizing constant. Furthermore, in the situation of a non-informative prior \( \pi(C) \) that we consider ourselves to be in, this problem is often reduced to maximizing \( \pi(Y|C) \). This entity is known as the marginal likelihood or ML-II (e.g. Rasmussen and Williams, 2006). The resulting estimate as the Empirical Bayes or MML estimate.

**Definition 4.1 (Marginal Likelihood).** The marginal likelihood of the data given the parameters is defined as \( \pi(Y|C) \) and sometimes denoted by \( L(C|Y) \) to express the fact this function will be maximized over the parameters. See further definition A.1 in appendix A.1.

**Definition 4.2 (Empirical Bayes estimate).** The EB estimate \( \hat{C} \) is defined as:

\[
\hat{C}(\Xi) := \arg\max_{C} \pi(Y|C)
\]

And consequently, the EB estimate for \( \Xi \) is defined as:

\[
\hat{\Xi} := \arg\max_{\Xi} \hat{C}(\Xi)
\]

Under conditional independence of subjects \( \pi(Y|C) \) can be factorized into the usual product of individual likelihoods. Then, we integrate out what is unknown (i.e. the varying effects \( b_i \) and \( f_j \)), a concept similar to that in GLMM.
Proposition 4.1 (Computing the marginal likelihood). Let $b_i \subseteq \mathbb{R}^{d+1}$ and $f_i \subseteq \mathbb{R}^{|\tau|}$.

$$
\pi(Y|C) = \prod_{i=1}^{n} \pi(Y_i|C_i) = \prod_{i=1}^{n} \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^{|\tau|}} \pi(Y_i|C_i, b_i, f_i) \pi(b_i, f_i|C_i) \ db_i \ df_i \\
= \prod_{i=1}^{n} \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^{|\tau|}} \pi(Y_i|b_i, f_i) \pi(b_i|\Sigma_b) \pi(f_i|\nu_0, \nu_0) \ db_i \ df_i
$$

Where in case of the conditionally Bernoulli outcome with logit (canonical) link we have $\pi(Y_i|b_i, f_i) = \exp(l_i)$ with:

$$
l_i := \sum_{j \in J_i} \sum_{t \in \tau_i} (y_{ijt} \cdot \log(\frac{p_{ijt}}{1-p_{ijt}}) + \log(1-p_{ijt}))
$$

And because $b_i$ and $f_i$ multivariate normal:

$$
\pi(Y|C) \propto \prod_{i=1}^{n} \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^{|\tau_i|}} \exp(l_i) \times \exp(-\frac{1}{2} b_i' \Sigma_b^{-1} b_i) \times \exp(-\frac{1}{2} f_i' \Phi_{\beta, \nu}^{-1} f_i) \ db_i \ df_i \tag{12}
$$

$$
=: \prod_{i=1}^{n} \int_{\mathbb{R}^d} \exp(h(v_i)) \ dv_i \tag{13}
$$

We go from (12) to (13) by writing vector $v_i := (b_i, f_i) \in \mathbb{R}^d$, integer $d := q + 1 + |\tau_i|$ and introducing the function $h: v \mapsto \mathbb{R}$.

$$
h(v_i) := l_i - \frac{1}{2} (b_i' \Sigma_b^{-1} b_i - f_i' \Phi_{\beta, \nu}^{-1} f_i)
$$

4.2 Related work

As e.g. McCullough, Searle, and Neuhaus (2008) or Raudenbush, Yang, and Yosef (2000) describe, an approximation of $\pi(Y|C)$ is made typically by analyzing a tractable quasi-likelihood, numerical integration (either a Gauss-Hermite quadrature on (12), a Laplace approximation on (13)) or MCMC (c.q. MC Expectation Maximization) methods.

The Laplace approximation uses a Taylor expansion of the function $h$ around its maximum $\tilde{v}$. This expansion is given for multivariate $v \in \mathbb{R}^d$ and general order $p$ in definition A.7, appendix A.2.

Definition 4.3 (Laplace approximation of order $p$). Let $d$ denote the dimension of the integral of interest, and $h(\cdot)$ a $p$-times differentiable function, with partial derivatives continuous in a neighbourhood of $\tilde{v}$. Then, $\forall v \in \mathbb{R}^d$

$$
\int_{\mathbb{R}^d} \exp(h(v)) \ dv \approx (2\pi)^{d/2} |V|^{1/2} \exp(h(\tilde{v})) \times \mathbb{E}(\exp(\sum_{k=3}^{p} \frac{1}{k!} \times \times (v - \tilde{v})^T h^{(k)}(\tilde{v})(v - \tilde{v})))
$$

With $V := (-h^{(2)}(\tilde{v}))^{-1}$ and $h^{(k)}$ denotes the $k$th derivative of $h$ with respect to $v$. Furthermore, $\mathbb{E}$ is the expectation with respect to a zero-mean MVN with covariance matrix $V$.

For $p = 2$ (as originally suggested by Laplace) and $d$ small, Tierney and Kadane (1986) have shown the error to be $O(n^{-5/2})$. However, when the dimension $d$ of the integral in (13) grows faster than $O(n^{1/3})$, second-order Laplace approximations are generally under risk of being (asymptotically) insufficiently accurate (Shun and McCullagh, 1995). In case of the GP, for fixed dimension $q$ of $b$, the dimension of the integral equals $q + 1 + |\tau_i|$, where the possibility that $|\tau_i|$ grows one-to-one with $n_i$ cannot be excluded.

Raudenbush, Yang, and Yosef (2000) handle this problem by using approximations based on $p = 6$. There results look promising as mentioned by McCullough, Searle, and Neuhaus (2008), but are insufficiently tested for high dimensions and thereby run the same (albeit lower) risk of inaccuracy.

In resolving this matter, Shi and Wang (2014) resort to a method that bears resemblance to Integrated Nested Laplace Approximation (INLA; Rue, Martino, and Chopin, 2009), which was already discussed
by its predecessor (Rue and Martino, 2007). Main topic of both works is the approximation of posterior marginals, e.g. \( \pi(C|\mathcal{Y}) \) and \( \pi(f_i|\mathcal{Y}) \), the latter of which they find via approximations to \( \pi(f_i|\mathcal{Y}, X, C) \) and integrating out numerically. In Rue and Martino (2007), a Gaussian approximation to \( \pi(f_i|\mathcal{Y}, X, C) \) is made, which was later improved to a (simplified) Laplace approximation hence creating INLA (Rue, Martino, and Chopin, 2009). As a matter of fact, this strategy would be an alternative to our EB approach in case one would be willing to specify a prior \( \pi(C) \), after which INLA approximates equation (11).

Instead, we share an interest with Shi and Wang (2014) in finding \( \pi(\mathcal{Y}|C) \) as outlined above. Let \( \pi_G \) indicate the Gaussian approximation. Using notation equivalent to proposition 4.1, Shi and Wang (2014) propose an approximation of the following form\(^\ddagger\):

\[
\pi(\mathcal{Y}|C) \approx \left. \frac{\pi(f_i, \mathcal{Y}|C)}{\pi_G(f_i|\mathcal{Y}, C)} \right|_{f_i = \tilde{f}_i(C)}
\]

Although there is no particular reason to suspect the results of Shi and Wang (2014) to be incorrect, we have two concerns about this approximation. Firstly, notwithstanding the rather satisfying results obtained by Rue, Martino, and Chopin (2009) in approximating \( \pi(C|\mathcal{Y}) \) by such a form, Shi and Wang (2014) do not test nor refer to any literature backing the applicability of this approximation to \( \pi(\mathcal{Y}|C) \). Secondly, Rue and Martino (2007) emphasize the insufficient ability of \( \pi_G(f_i|\mathcal{Y}, C) \) in capturing location and skewness of \( \pi(f_i|\mathcal{Y}, C) \), which may have effects on the \( \pi(\mathcal{Y}|C) \) approximation that are hard to anticipate.

### 4.3 MCEM: Empirical Gibbs

McCulloch (1997) has developed a Monte Carlo Expectation Maximization (MCEM) for Generalized Linear Mixed Models. In particular, the MCEM algorithm can use Gibbs samples as input, in which case one may speak of MCMC-EM. Casella (2001) has studied a similar idea and labels it Empirical Gibbs, which is closest to our implementation. The major benefit of Empirical Gibbs over the related work discussed in subsection 4.2 is that no integration is required, although this comes at computational cost.

The MCEM algorithm is based on maximizing the marginal likelihood of equation (11) after rewriting:

\[
\pi(Y|C) = \frac{\pi(Y, \theta, \delta, \zeta, \beta, \gamma|C)}{\pi(\theta, \delta, \zeta, \beta, \gamma|Y, C)} \tag{14}
\]

**Definition 4.4** (EM algorithm (Dempster, Laird, and Rubin, 1977)). Let \( E \) denote the expectation operator with respect to a random variable with density \( \pi(\theta, \delta, \zeta, \beta, \gamma|Y, C) \), i.e. the posterior distribution conditioned on \( C \). Let \( m \) and \( m \) denote indices (defined later).

\[
\begin{align*}
E-[\log \pi(Y|C)] &= E[\log \pi(Y, \theta, \delta, \zeta, \beta, \gamma|C)] - E[\log \pi(\theta, \delta, \zeta, \beta, \gamma|Y, C)] \\
E-[\log \pi(Y, \theta, \delta, \zeta, \beta, \gamma|Y, C)] &= \min_{c \in C} E[\log \pi((Y, \theta, \delta, \zeta, \beta, \gamma|C)|c^{m+1})]
\end{align*}
\]

With respect to the marginal likelihood computation based on integration, the difference here is the conditioning on extra parameters in the hierarchy. The trick is that these parameters become available through the Gibbs sampler. The expectation in the EM algorithm is taken with respect to the posterior \( \pi(\theta, \delta, \zeta, \beta, \gamma|Y, C) \), such that the E-step effectively cancels (or is "automated" (Casella, 2001)) when (an approximation to) this posterior is available. This bypasses the use of equation (14) in practice, but it is included for reference against proposition 4.3.

\(^\ddagger\)Our treatment of \( \beta \) is not included in Shi and Wang (2014) and therefore excluded from this expression.
4.3.1 Monte Carlo approximation

Based on the output from the Gibbs sampler, the M-step is replaced by a Monte Carlo approximation (Casella, 2001).

\[
C^{m+1} \approx C^{m+1} := \arg \max_{C} \frac{1}{M-B} \sum_{m=B}^{M} \log \pi(Y, \theta^{m}, \delta^{m}, \xi^{m}, \beta^{m}, \gamma^{m}|C)
\]

Here, \(B := Bp \times M\) iterations are used as burn-in to approximately exclude the part of the MCMC (i.e. Gibbs) chain that has not converged yet, with burn percentage \(Bp \in [0, 1)\). In case of \(C := \{(C_{i} := (\Sigma_{b}, \Psi_{\theta,i})) : i \in [1, n]\}\) in the hierarchy of model definition 2.1, we have:

\[
\tilde{\Sigma}_{b} = \arg \max_{\Sigma_{b}} \frac{1}{M-B} \sum_{m=B}^{M} \frac{1}{n} \sum_{i=1}^{n} \pi(\Sigma_{b}|b_{i}^{m}) = \arg \max_{\Sigma_{b}} \frac{1}{M-B} \sum_{m=B}^{M} \frac{1}{n} \sum_{i=1}^{n} \left( -\frac{1}{2} b_{i}^{m} \Sigma_{b}^{-1} b_{i}^{m} - \frac{1}{2} \log(|\Sigma_{b}|) \right)
\]

\[
\left(1 - \frac{1}{2} \frac{\Sigma_{b}^{-1}}{\sum_{m=B}^{M} \frac{1}{n} b_{i}^{m} \Sigma_{b}^{-1} b_{i}^{m}} \right)
\]

Where, \(b := \left[ b_{1}^{m}, b_{2}^{m}, \ldots, b_{n}^{m} \right] \in \mathbb{R}^{n \times (q+1)}\) are implied by the Gibbs sampler (through \(\xi, \beta, \gamma\)), and the expression follows since \(\frac{1}{2} b_{i}^{m} \Sigma_{b}^{-1} b_{i}^{m}\) is the MLE of the unstructured covariance matrix at iteration \(m\) for multivariate normal \(\pi(\Sigma_{b}|b_{i}^{m})\). Similarly, we compute the MLE of the (temporarily unstructured) covariance matrix \(\Psi_{\theta,i}\), specific for subject \(i\), as \(\tilde{\Psi}_{\theta,i} = f_{i}^{m} f_{i}^{mT}\). Then, based on the structure imposed by the GP, optimal kernel parameters are found. Remark that this \(\Psi_{\theta}\) procedure implying structure is very close to that of Azevedo, Fox, and Andrade (2015), although the authors do not use an outer loop and the process is not subject specific.

\[
\tilde{\omega}_{0} = \frac{1}{M-B} \sum_{m=B}^{M} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{n_{i}} \sqrt{\text{tr}(\tilde{\Psi}_{\theta,i}^{m})} \right)
\]

(16a)

\[
\bar{\omega}(\tilde{\omega}_{0}) = \arg \max_{\omega_{0}} \frac{1}{M-B} \sum_{m=B}^{M} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{2} f_{i}^{mT} \tilde{\Psi}_{\theta,i}^{-1} f_{i}^{m} - \frac{1}{2} \log(|\Psi_{\theta,i}|) \right)
\]

(16b)

Notice that in (16a), the diagonal structure is imposed (based on ML) by selecting (the mean over \(m\), the mean over \(i\) and the mean of the trace of each MLE \(\tilde{\Psi}_{\theta,i}^{m}\). In other words, given the unstructured MLE matrix, \(\tilde{\omega}_{0}\) is the most likely parameter value for the structured matrix. Remark that Azevedo, Fox, and Andrade (2015) imply (off-diagonal) structured variance parameters, e.g. an AR1, directly from the unstructured matrices as well. However, we have found this to be a poor approximation in case of \(\omega_{0}\) and suggest a general purpose maximum likelihood optimization for (16b) instead. Although it comes at computational expense, it is rather quick after profiling \(\omega_{0}\) from the likelihood equation. Remark that \(\tilde{\omega}_{0}\) enters the \(\omega_{0}\) optimization through \(\Psi_{\theta,i}\), suppressed from notation in (16b).

4.3.2 Convergence

Define \(m1 \in [1, M1]\) to be the index for the outer loop and \(M2\) a vector of sample sizes for the inner loop, such that the \(m1\)-th outer loop runs over \(m \in [1, M]\) with \(M := M2[m1]\). Casella (2001) shows, given some conditions, that the posterior distribution from the Gibbs sampler based on \(C\) converges in total variation to the true posterior distribution given \(C\). These asymptotics involve \(n\), \(M1\) and \(M2\).

Booth and Hobert (1999) and Levine and Casella (2001) have suggested automated variants of these algorithms that determine simulation sizes \(M1\) and \(M2\). The easiest improvement is to place a convergence criterion on the parameters \(\Xi\) in \(C\), such that a stopping criterion is introduced in the outer loop. Once convergence in the outer loop is reached, a final inner loop of size \(M\) (the last element in \(M2\)) acts as Gibbs
procedure to sample from the posterior, where convergence is stochastic. Determining the other elements in M2 is even more intricate, since they may affect convergence in the outer loop. Despite the computational gain and theoretical justification made by Booth and Hobert (1999) and Levine and Casella (2001), we predetermine M1 and the vector M2 ad hoc at large values and verify convergence manually. This approach is similar to that of McCulloch (1997).

4.3.3 Algorithm

We discuss the Empirical Gibbs algorithm used for estimation of the reference procedure next. It combines many familiar Gibbs steps from Fox and Glas (2001) and Azevedo, Fox, and Andrade (2015) in the “inner” loop (m ∈ [1, M]) with a marginal maximum likelihood step in the “outer” loop (m1 ∈ [1, M1]) for the parameters C. In subsection 4.4, the individual steps of the Gibbs sampler are discussed in detail. The algorithm suppresses conditioning on data and hyperparameters (e.g. Y, A, w, s, Σδ, Σγ) from notation.

Algorithm 1: Gibbs sampler for \((Z, \delta, \theta, \zeta, \beta, \gamma)\) with empirical \((\Sigma_b, v_0, w_0)\) under posterior prop. 2.3

\[
\begin{cases}
\text{initialization} \quad (\delta^0, \theta^0, \zeta^0, \beta^0, \gamma^0) \quad \text{and} \quad (\Sigma_b^0, v_0^0, w_0^0) ; \\
\text{for } m1 \in [1 : M1] \text{ do} \\
\quad M=M2[m1], B=Bp*M ; \\
\quad \text{for } m \in [1 : M] \text{ do} \\
\quad \quad \text{Step 1} \quad \text{Sample } Z^m \text{ from } \pi(Z|\delta^{m-1}, \theta^{m-1}) ; \\
\quad \quad \text{Step 2} \quad \text{Sample } \delta^m \text{ from } \pi(\delta|Z^m, \theta^{m-1}) ; \\
\quad \quad \text{Step 3} \quad \text{Sample } \theta^m \text{ from } \pi(\theta|Z^m, \delta^m, \zeta^{m-1}, \tilde{v}_0, \tilde{w}_0) ; \\
\quad \quad \text{Step 4} \quad \text{Sample } \zeta^m \text{ from } \pi(\zeta|\theta^m, \beta^{m-1}, \gamma^{m-1}, \tilde{\Sigma}_b, \tilde{v}_0, \tilde{w}_0) ; \\
\quad \quad \text{Step 5} \quad \text{Sample } (\beta^m, \gamma^m) \text{ from } \pi(\beta, \gamma|\zeta^m, \tilde{\Sigma}_b) ; \\
\quad \quad \text{if } m>B \text{ then} \\
\quad \quad \quad \text{Step 6} \quad \text{Store variance information iteration } m \text{ for use in outer loop} \\
\quad \end{cases}
\]

Based on stored information, compute \((\hat{\Sigma}_b, \hat{v}_0, \hat{w}_0)\) from equations (15) and (16).

Starting from initial values of the parameter vector, the Gibbs sampler iterates M times through a cycle of fully conditional posteriors, sampling a single value from each and using those to sample from the others in the cycle. Each iteration delivers a new parameter vector, and upon convergence these are considered to be samples from the joint posterior distribution. We consider the Gibbs procedure to be converged after a burn-in period of B iterations, yielding a net sample size of \(\tilde{M} = M - B\). A prerequisite for the Gibbs sampler is that the fully conditional posterior of every step can be (easily) sampled from.

4.4 Full conditionals

This subsection describes the steps of algorithm 1 in detail. It is important to realize that \(\beta, \gamma\) and \(f\) are sampled subject specifically \(\forall i \in [1, n]\), since the covariance matrix of \(f\) is discretized at idiosyncratic observation times. Similar to the notation for data matrices introduced before, the subscript \(i\) is used when referring to such a specific parameter. When this subscript is suppressed, the total information across all subjects is included.
Step 1

For probit, similar to Albert (1992), step 1 involves sampling \( Z_i \) from a truncated normal distribution as described in definition 2.1. For logit, we use the normal mixture approximation to the logistic distribution as described in subsection 3.1. For both links a sample of (a mixture of) truncated normal variates must be obtained, which is therefore generalized. Truncation occurs left or right of zero depending on \( Y \) being 1 respectively 0. The mean of the truncated mixture is positioned at \( \theta_{it} - \delta_{jt} \). Let \( tN \) denote the truncated mixture distribution with density as defined in definition 3.2. In short, we have:

\[
Y_{ijt} = \begin{cases} 
1 & Z_{ijt} > 0 \\
0 & Z_{ijt} \leq 0
\end{cases} \quad \Leftrightarrow \quad Z_{ijt|\theta_{it}, \delta_{jt}, Y_{ijt}} \sim \begin{cases} 
tN(\theta_{it} - \delta_{jt}, w, s) & \text{truncated left of zero if } Y_{ijt} = 1 \\
tN(\theta_{it} - \delta_{jt}, w, s) & \text{truncated right of zero if } Y_{ijt} = 0
\end{cases}
\]

In the next step, by proportionality, we may assume the truncated normal realization of vector \( Z_i \) is in fact distributed as (a mixture of) a multivariate normal variable (i.e. non-truncated).

Step 2

Sample \( \delta \) from the following full conditional:

\[
\pi(\delta|Z, \theta) \propto \pi(Z|\delta, \theta) \times \pi(\delta)
\]

Firstly, by conditional independence of subjects, \( \pi(Z|...) \) is a product measure of Gaussian form. Secondly, let \( \pi(\delta) \) be conjugate multivariate normal with \( \mu_\delta \) and \( \Sigma_\delta = \sigma_\delta^2 \cdot I_{|\delta|} \).

At this stage, we recognize a Generalized Least Squares problem:

\[
Z - \theta = -\Delta \delta + \epsilon \quad \text{with } \Delta := [\Delta_1 \quad \Delta_2 \quad ... \quad \Delta_n]' \in \{0, 1\}^{|Y| \times |J|} \text{ and } \delta \in \mathbb{R}^{|J|}
\]

Furthermore, \( \theta \in \mathbb{R}^{|Y|} \) is the long vector of appropriate \( \theta_{it} \) elements across all \( i \) and \( t \). In this regression, \( \Delta \) selects the appropriate difficulties and \( Z \) contains all information there is available (latent through \( Y \)). In case of probit, \( \epsilon \sim N_{|Y|}(0, 1) \). In case of logit (normal mixture) \( \epsilon \sim \sum_{h=1}^k w_h \cdot N_{|Y|}(0, s_h \cdot I_{|Y|}) \). Then, it is straightforward to show the following conjugate update.

\[
\pi(\delta|Z, \theta) \propto \sum_{h=1}^k w_h \cdot \phi_{|J|}(\mu_{\delta h}, \Sigma_{\delta h})
\]

With the following updated parameters \( \forall h \in [1, k] \):

\[
\mu_{\delta h} = \Sigma_{\delta h}^{-1} (\Delta'(s_h \cdot I_{|Y|})^{-1}(Z - \theta) + \Sigma_\delta^{-1} \mu_\delta) \\
\Sigma_{\delta h} = (\Delta'(s_h \cdot I_{|Y|})^{-1} \Delta + \Sigma_\delta^{-1})^{-1}
\]

Since this GLS solution is important (and analogous) in most steps, a proof (using parameter name \( \beta \)) is provided in appendix B.1.

Step 3

\( \forall i \in [1, n] \) sample \( \theta_i \) as follows:

\[
\pi(\theta_i|Z_i, \delta, \xi_i, \nu_0, \nu_0) \propto \pi(Z_i|\delta, \theta_i) \times \pi(\theta_i|\xi_i, \nu_0) \times \pi(\xi_i|\nu_0) \times \pi(\nu_0) \\
\propto \sum_{h=1}^k w_h \cdot \phi_{|J|}(\mu_{\theta_{hi}}, \Sigma_{\theta_{hi}})
\]

With the following updated parameters \( \forall i \in [1, n] \) and \( \forall h \in [1, k] \):

\[
\mu_{\theta_{hi}} = \Sigma_{\theta_{hi}}^{-1}(A_i'(s_h \cdot I_{|\tau_i|})^{-1}(Z_i + \Delta \delta)' + \Psi_{\theta_{hi}}^{-1} X_i \xi_i) \\
\Sigma_{\theta_{hi}} = (A_i'(s_h \cdot I_{|\tau_i|})^{-1} A_i + \Psi_{\theta_{hi}}^{-1})^{-1}
\]
Step 4

∀ \in [1, n] sample ζ_i as follows:

$$
\pi(ζ_i | θ_i, β, γ, \hat{Σ}_b, \hat{v}_0, \hat{w}_0) \propto \pi(θ_i | ζ_i, \hat{Ψ}_{θ,i}) \times \pi(ζ_i | β, γ, \hat{Σ}_b) \\
\propto φ_{q+1}(\tilde{µ}_{ζ,i}, \tilde{Σ}_{ζ,i})
$$

With the following updated parameters ∀ i ∈ [1, n]:

$$
\tilde{µ}_{ζ,i} = \tilde{Σ}_{ζ,i}(X_i' \hat{Ψ}_{θ,i}^{-1}θ_i + \hat{Σ}_b^{-1}W_i [β] [γ])
$$

$$
\tilde{Σ}_{ζ,i} = (X_i' \hat{Ψ}_{θ,i}^{-1}X_i + \hat{Σ}_b^{-1})^{-1}
$$

Step 5

Remember that the β and γ separation was pure notation and these parameters can be blocked without loss of generality. Assume β and γ have multivariate normal priors such that (β, γ) ~ N(µ, Σ). Accumulating information across individuals, sample (β, γ) as follows:

$$
\pi(β, γ | ζ, \hat{Σ}_b) \propto \pi(ζ | β, γ, \hat{Σ}_b) \times \pi(β, γ | µ, Σ) \\
\propto φ_{q+1}(\tilde{µ}, \tilde{Σ})
$$

With the following updated parameters:

$$
\tilde{µ} = \tilde{Σ}((\sum_{i=1}^{n} W_i^{'} \hat{Σ}_b^{-1}ζ_i) + Σ^{-1}µ)
$$

$$
\tilde{Σ} = ((\sum_{i=1}^{n} W_i^{'} \hat{Σ}_b^{-1}W_i) + Σ^{-1})^{-1}
$$

4.5 Parameter identification

Remark that IRT models in general, and model definition 2.1 in particular, classify Y_{ijt} based on the difference in subject ability and item difficulty. This implies that these entities are not identified separately. In early work on IRT computation, Bock and Aitkin (1981) handled this by fixing the origin and scale of the latent ability distribution when estimating difficulties by Marginal Maximum Likelihood. Typically, such identification occurs using a standard normal distribution.

In case of auxiliary variable regression, the identification problem becomes evident by an unbounded variance of the conditional Z distribution, since θ and δ move (hand-in-hand) freely. McCulloch, Polson, and Rossi (2000) discuss a general method of identification in multilevel models by restricting the multivariate normal ability distribution to have mean zero at the first time point and top-left element of the covariance matrix equal to 1. In that sense, it is a direct generalization of Bock and Aitkin (1981). This multivariate identification is also used by Azevedo, Fox, and Andrade (2015). A downside of this method is that the variance parameters become difficult to interpret; they are scale dependent themselves.

Furthermore, in case of model definition 2.1, such a method of parameter identification is hampered by the fact the covariance matrix of the latent ability distribution is specified indirectly through Σ_b and Ψ_{θ,i}. In order to allow the variance parameters to move freely, an alternative method of identification is suggested, which is moreover straightforward in use. The latent ability becomes identifiable in the longitudinal setting when a sufficient amount of difficulties are fixed, depending on the study design. Imagine the items are linked over time (all items are asked on at least two visits) and are positioned on a straight line with positive slope. Then fixing two would be enough to identify all difficulty and ability parameters. With different scenarios come different identification requirements.
Since we are interested in comparing latent abilities under the proposed model with sequential Bayesian updating versus results obtained by Van Buuren (2014), fixating all difficulties to the values obtained by Van Buuren (2014) seems the best choice. This implies that we bypass step 2 of the Gibbs sampler in algorithm 1.

5 Simulation study

In this section, a simulation study of the method presented in section 4 is made. In addition to the efficacy of the Gibbs sampler in longitudinal psychometric analysis (Azevedo, Fox, and Andrade, 2015), we intend to show the ability of our method to recover kernel estimates of the GP. More generally, this simulation is designed to be close to the real data application of TNO.

We follow Fox and Glas (2001) in simulating a single data set and assessing parameter recovery on the basis of MSE between the estimated and simulated parameters. Azevedo, Fox, and Andrade (2015) uses a similar approach on the basis of 10 data sets, again a small number. Although repeating this procedure over many data sets (Monte Carlo) is preferred, the computational cost of the MCEM procedure outweighs its benefits given the fact $N$ is relatively large. Furthermore, the simulation study is used to verify the (stochastic) convergence of the procedure.

In this proof of concept by simulation, we choose a simple set-up of the latent level without covariates. By fixing item parameters (as discussed in 4.5) it becomes important to match variance parameters (including those in the kernel) to plausible combinations. Including covariates makes this process overly complex in the simulation set-up.

5.1 Simulation set-up

Complete data for $N = 1000$ subjects on 57 items spread over ten visits were simulated according to the scheme in table 4. This scheme is the study design of the real data. Visit times in years were sampled from a normal distribution with standard deviation 1/26, such that approximately 95% of appointments occurred within 4 weeks of the scheduled times of $(0.08, 0.15, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5)$ in years. Item parameters are fixed at those obtained by Van Buuren (2014).

The ability traits are generated with quadratic (i.e. $q = 2$) polynomial parameters $\zeta_i$ and $\beta = (5.0, 51.0, -11.5)$, excluding covariates $u$ and associated parameters $\gamma$. The $\beta$ non-varying parameters are endowed with non-informative priors (i.e. $\Sigma$ in step 5 taken to be zero, a similar choice was made in Fox and Glas (2001)).

Subject-specific variation around $\zeta_i$ is modelled as zero mean multivariate normal with an unstructured covariance matrix $\Sigma_b$ as specified below. The Gaussian process is parametrized as squared exponential using $(v_0, w_0) = (0.65, 0.03)$. At the scheduled times, this results in draws from the GP as displayed ten times in the middle of figure 3. The $v_0$ parameter determines the scale of the y-axis. For illustration, the figure on the left and right show draws using $v_0$ and $0.1 \cdot w_0$ respectively $10 \cdot w_0$.

$$\Sigma_b = \begin{bmatrix} \sigma_1 & \sigma_2 & \sigma_4 \\ \sigma_2 & \sigma_3 & \sigma_5 \\ \sigma_4 & \sigma_5 & \sigma_6 \end{bmatrix} = \begin{bmatrix} 31.5 & -19.3 & 4.2 \\ -19.3 & 11.9 & -2.6 \\ 4.2 & -2.6 & 0.60 \end{bmatrix}$$

Figure 4 includes profile plots of the simulated abilities and the (fixated) item parameters (Van Buuren, 2014). Starting from these abilities and difficulties, a probit link was used to arrive at a total of 114,000 scores $Y_{ijt} \in \{0, 1\}$. 

22
5.2 Initialization and perfect response profiles

Since the Gibbs sampler is circular by nature, initializing it in a low posterior likelihood region has a strong impact on convergence, particularly so in multilevel models. Unlike Azevedo, Fox, and Andrade (2015), these issues are mentioned in Fox and Glas (2001), raising concern regarding too small initialization of variance parameters. In our model, if \( v_0 \) or \( \Sigma_b \) become unrealistically small, it takes a very long time for the Gibbs sampler to alter this behaviour.

As a remedy, Fox and Glas (2001) suggest using item estimates from standard software as initialization. Furthermore, not being in the longitudinal setting, they use the procedure of Albert (1992) to generate plausible starting values for the ability estimates based on a standard normal prior (Bock and Aitkin, 1981; Bock and Mislevy, 1982). Upon convergence of this starting procedure, average estimates serve as input to start their overall Gibbs sampling procedure. Remark that this resembles the Empirical Gibbs procedure to a fair extent.

Since observations per visit in the longitudinal study design scheme are low, the ability variance is too high for the procedure of Albert (1992) to be used at initialization. Still, it is important that abilities (and the variance parameters they imply throughout several levels of the hierarchical model) are started sensibly. As
an alternative to the initialization procedure of Fox and Glas (2001), we suggest determining starting values by the following guidelines. Note that our start is crucially dependent on the difficulty estimates, but as already pointed out this is desirable for comparison to Van Buuren (2014). Furthermore, our procedure is equally dependent on item estimates as that of Fox and Glas (2001) is.

**Step 1**

Define the set of items administered to subject $i$ at time $t \in \tau_i$ as $J_{it}$. We use a weighted average of passed difficulties corrected by their relative positioning to determine an estimate $\hat{\theta}_{it}$.

\[
\hat{\theta}_{it} := \min(\delta_j : j \in J_{it}) + (\max(\delta_j : j \in J_{it}) - \min(\delta_j : j \in J_{it})) \cdot \frac{\sum_{j \in J_{it}} Y_{ijt} \cdot \delta_j}{\sum_{j \in J_{it}} \delta_j} \quad (17)
\]

Note that (17) interpolates abilities linearly across the $\delta$ scale and does not take the link function of model definition 2.1 into account. This is refined during the Empirical Gibbs procedure.

**Step 2**

In the next step, as a first approximation $\zeta_i$ is estimated using least squares $\forall i \in [1,n]$.

\[
\hat{\zeta}_i := (X_i'X_i)^{-1}X_i'\hat{\theta}_i \quad (18)
\]

We use $\hat{\zeta}_i$ to compute fitted values and residuals $\forall i \in [1,n]$.

\[
\hat{\theta}_i := X_i\hat{\zeta}_i \quad (19)
\]

\[
\hat{f}_i := \hat{\theta}_i - X_i\hat{\zeta}_i \quad (20)
\]

The quality of $\hat{\theta}$ can be assessed by plugging it into the binomial likelihood (which introduces the link although not taken into account yet) or by computing the relative amount of scores $Y$ fitted correctly.

**Definition 5.1 (Binomial likelihood and accuracy).** Let $g(\cdot)$ denote the link function (i.e. probit or logit). Then, conditional on $\hat{\theta}$ and $\delta$ we define the binomial likelihood and take its logarithm:

\[
ll := \sum_{i=1}^{n} \sum_{j \in J_{it}} \sum_{t \in \tau_i} \log \left( Y_{ijt} \cdot g^{-1}(\hat{\theta}_{it} - \delta_j) + (1 - Y_{ijt}) \cdot (1 - g^{-1}(\hat{\theta}_{it} - \delta_j)) \right)
\]

Secondly, define the fitted scores $\hat{Y}$ conditional on $\hat{\theta}$ and $\delta$:

\[
\hat{Y}_{ijt} := 1_{(\hat{\theta}_{it} - \delta_j) > 0.5}
\]

Let $|Y|$ denote the cardinality of $Y$ and $\hat{Y}$. The accuracy $Acc \in [0,1]$ is defined as:

\[
Acc(Y, \hat{Y}) := \frac{1}{|Y|} \sum_{i=1}^{n} \sum_{j \in J_{it}} \sum_{t \in \tau_i} 1_{Y_{ijt} = \hat{Y}_{ijt}}
\]

Remark that the $\hat{\zeta}_i$ estimation did not take $\Psi_{\theta,i}$ into account and is not guaranteed to maximize the binomial likelihood. One way to improve this is through a GLM estimation, in which $\delta$ is used as offset. However, perfect response profiles discussed next give a preference for the procedure outlined above.
Perfect response adjustment (step 2)

A perfect response at a visit is said to occur when all scores are either zero or one. Sometimes the term zero score is used to indicate the perfect response in case of zeroes, but we will adopt the same terminology for both extremes. In such a case, the administered items are unable to determine a subject’s ability at that time point, only that it is outside the respective range of difficulties. Being unbounded, a maximum likelihood procedure (e.g. a GLM) will place the ability in the tail of the link function, having an adverse effect on the \( \zeta \) estimation. However, forcing the perfect response ability on the \( \delta \) range as is done by equation (17) is arbitrary as well. One way to deal with perfect responses is to stipulate a minimum and maximum perfect response (ability) profile, which lower respectively upper bind a subject’s ability at perfect visits.

We choose a simple definition of the perfect response profile by introducing two parameters \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \).

**Definition 5.2 (Perfect response profile).** Let \( \lambda_{\text{min}}, \lambda_{\text{max}} \geq 0 \). Define the perfect response profiles at time \( t \) with respect to the administered difficulties \( J_{it} \):

\[
\begin{align*}
\theta_{\text{min}}^{it} & := \min(\delta_j : j \in J_{it}) - \lambda_{\text{min}} \\
\theta_{\text{max}}^{it} & := \max(\delta_j : j \in J_{it}) + \lambda_{\text{max}}
\end{align*}
\]

As mentioned before, optimizing the binomial likelihood is hindered by the unboundedness of the problem, but the accuracy (definition 5.1) can be used as surrogate to find \( \hat{\lambda}_{\text{min}} \) and \( \hat{\lambda}_{\text{max}} \).

\[
(\hat{\lambda}_{\text{min}}, \hat{\lambda}_{\text{max}}) := \arg \max_{\lambda} \text{Acc}(Y, \hat{Y}(\lambda)) \quad \lambda \in \mathbb{R}^2^+ \cup \{0\}
\]

In the simplest case, \( \lambda_{\text{min}} = \lambda_{\text{max}} = 0 \). The first generalization is allowing positive \( \lambda \), but restricting this to the symmetric case in which \( \lambda_{\text{min}} = \lambda_{\text{max}} \). Most general is the asymmetric scenario, in which \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are free in the space of non-negative real numbers. The optimal one-dimensional \( \lambda \) based on the simulated data is displayed at the left-hand side of figure 5. Contour and surface plots of the asymmetric optimization are included there as well.

Finally then, \( \zeta \) is refitted using the \( \lambda \)-adjusted \( \hat{\theta} \), resulting in improved binomial likelihood and accuracy as listed in table 1.

---

**Figure 5:** \( \lambda \) optimization based on accuracy.
Figure 6: Profile plots based on $\hat{\lambda}$.

Table 1: Binomial likelihood and accuracy before and after adjustment.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Raw $\hat{\theta}$ Symm.</th>
<th>Adjusted $\hat{\theta}$ Symm.</th>
<th>Adjusted $\hat{\theta}$ Asymm.</th>
<th>Raw $\hat{\theta}$ quadratic Symm.</th>
<th>Adjusted $\hat{\theta}$ quadratic Symm.</th>
<th>Adjusted $\hat{\theta}$ quadratic Asymm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{\text{min}}$</td>
<td>0</td>
<td>2.25</td>
<td>1.75</td>
<td>0</td>
<td>2.25</td>
<td>1.75</td>
</tr>
<tr>
<td>$\lambda_{\text{max}}$</td>
<td>0</td>
<td>2.25</td>
<td>3.50</td>
<td>0</td>
<td>2.25</td>
<td>3.50</td>
</tr>
<tr>
<td>ll probit ($\times 1e5$)</td>
<td>-0.545</td>
<td>-0.524</td>
<td>-0.525</td>
<td>-0.457</td>
<td>-0.455</td>
<td>-0.450</td>
</tr>
<tr>
<td>Acc</td>
<td>0.895</td>
<td>0.899</td>
<td>0.899</td>
<td>0.884</td>
<td>0.888</td>
<td>0.889</td>
</tr>
</tbody>
</table>

Step 3

Grouping information over subjects, the (adjusted) $\hat{\zeta}$ result in estimates of non-varying parameters $\hat{\beta}$ and $\hat{\gamma}$ by least squares. This step is model specific depending on the covariate relationships (e.g. additive effects). Profile plots of several choices for $\lambda$ are included in figure 5. Define the residuals that result from the least squares fit as $\hat{b}_i$.

Step 4

Finally, we imply variance parameters $\hat{v}_0, \hat{w}_0$ and $\hat{\Sigma}_b$ using the residuals $\hat{f}_i$ and $\hat{b}_i$. Since sensitivity of the overall procedure to $w_0$ is expected, a safe choice is to initiate at $w_0 = 0$ (see footnote §). Other than that, this procedure is analogous to that described by equations (15) and (16) with $M = 1$ and $B = 0$.

Starting values

In summary, the asymmetric initialization procedure gives the following starting values. $\beta = (-0.35, 58.3, -13.7)$, $v_0 = 2.45$, $w_0 = 0$ and

$$
\Sigma_b = \begin{bmatrix}
30.6 & -23.7 & 6.4 \\
-23.7 & 21.7 & -6.4 \\
6.4 & -6.4 & 2.0
\end{bmatrix}
$$

Since the squared exponential kernel in definition 2.1 is defined such that the reciprocal of $w_0^2$ is used, $w_0 = 0$ is indefinite. Instead, a proxy of $w_0 = 1e^{-6}$ is used throughout.
Inner and outer loop

We use the following specification of inner and outer loop, where \( \text{rep}(10, 500) \) means repeat inner loop of length 10 a number of 500 times.

\[
M_2 = (\text{rep}(10, 500), \text{rep}(100, 50), \text{rep}(1000, 5), 5000, 10000, 20000)
\]

In total, 558 outer loops generate \( M = 50,000 \) iterations. A burn-in of \( B_p = 20\% \) is used in each inner loop before updating the outer loop parameters.

5.3 Parameter recovery and convergence

Inference on inner loop parameters is made based on all iterations of the final outer loop (no burn-in). Table 2 includes summary information on posterior non-varying parameters. Credible intervals are displayed under \( \alpha = 0.05 \).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Truth</th>
<th>Mean</th>
<th>Median</th>
<th>CI low</th>
<th>CI high</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>5.0</td>
<td>4.67</td>
<td>4.67</td>
<td>4.37</td>
<td>4.96</td>
<td>0.14</td>
<td>0.38</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>51.0</td>
<td>51.24</td>
<td>51.24</td>
<td>51.04</td>
<td>51.44</td>
<td>0.07</td>
<td>0.27</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-11.5</td>
<td>-11.56</td>
<td>-11.56</td>
<td>-11.61</td>
<td>-11.51</td>
<td>0.00</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 3: Parameter recovery variance parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Truth</th>
<th>Est.</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_1 )</td>
<td>31.50</td>
<td>31.75</td>
<td>0.064</td>
<td>0.253</td>
</tr>
<tr>
<td>( \sigma_2 )</td>
<td>-19.30</td>
<td>-19.32</td>
<td>0.000</td>
<td>0.017</td>
</tr>
<tr>
<td>( \sigma_3 )</td>
<td>11.90</td>
<td>11.88</td>
<td>0.000</td>
<td>0.017</td>
</tr>
<tr>
<td>( \sigma_4 )</td>
<td>4.20</td>
<td>4.08</td>
<td>0.015</td>
<td>0.124</td>
</tr>
<tr>
<td>( \sigma_5 )</td>
<td>-2.60</td>
<td>-2.53</td>
<td>0.004</td>
<td>0.066</td>
</tr>
<tr>
<td>( \sigma_6 )</td>
<td>0.60</td>
<td>0.580</td>
<td>0.000</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Since it was already noted that the parameter recovery may be somewhat suboptimal due to the simulation set up with fixed difficulties (and exclusion of Monte Carlo loop), we are satisfied with the results of tables 2 and 3. Another important proof of concept here is the convergence of parameters.

Figure 7 displays the entire trace on which table 2 is based. The plots seem to suggest stochastic convergence on all \( \beta \) parameters. Figure 8 includes trace plots on the \( \Sigma_\beta \) outer loop variance parameters, which seem to suggest convergence of the Empirical Gibbs procedure. The same holds for the \( v_0 \) parameter in figure 9. The trace of \( w_0 \) seems more noisy, but this may be due to the differently scaled vertical axes of the plots. A sensitivity analysis on this kernel parameter will be performed in Part III.
Figure 7: Trace plots $\beta$ final inner loop, $M = 20,000$, [simulation].

Figure 8: Figure panel outer loop trace plots $\Sigma_b$ [simulation].
6 Sequential Bayesian estimation and inference

In the second phase of the estimation procedure, the reference parameters are used to shrink subject specific parameters in a fully Bayesian framework. Here, the mean posterior estimates on \((\beta, \gamma, \Sigma_b, v_0, w_0)\) from the reference procedure fulfill the role of the prior. In case \(\delta\) is fixed, it again carries over directly here, otherwise its mean posterior estimate enters.

As was already discussed at the end of section 2, the most important contribution of subject specific estimation using the reference prior is shrinkage in case of low sample sizes. As data becomes increasingly available sequentially, the ability profile is allowed to be more tailored to the subject.

6.1 Gibbs sampler using reference parameters

The subject specific estimation is made by a four step Gibbs sampler outlined in Algorithm 2, which may be compared against Algorithm 2.

\begin{algorithm}
\caption{Gibbs sampler for \((Z_i, \theta_i, \xi_i)\) posterior, given reference parameters \((\delta, \beta, \gamma, \Sigma_b, v_0, w_0)\).}
\begin{algorithmic}
\State \textbf{initialization} \((\theta^0_i, \xi^0_i)\);
\For \(m \in [1 : M] \)
\State \textbf{Step 1} Sample \(Z^m_i\) from \(\pi((Z_i | \theta^{m-1}_i) | \delta)\);
\State \textbf{Step 2} Sample \(\theta^m_i\) from \(\pi((\theta_i | Z^m_i, \xi^{m-1}_i) | \delta, v_0, w_0)\);
\State \textbf{Step 3} Sample \(\xi^m_i\) from \(\pi((\xi_i | \theta^m_i) | \beta, \gamma, \Sigma_b, v_0, w_0)\);
\EndFor
\end{algorithmic}
\end{algorithm}

The full conditionals were specified in subsection 4.4. Remark that when information \(Y_i\) becomes available sequentially, the resulting estimates become less shrunk to the reference parameters (most notably, \(\beta\) and \(\gamma\)).

We make a clear distinction between sequential updates versus full information Bayesian estimation. In the case of full information, we receive (all) longitudinal data on a subject and update the posterior ability based on it. In practice however, most information becomes available sequentially and earlier made estimates become suboptimal. Whether or not to change these historical values is best decided by the practitioner. Throughout, we assume that earlier made ability estimates are fixed, but all available information is used to determine the new estimate at the time in question.
6.2 Inference

The Bayesian estimation of algorithm 2 leads to \( \tilde{M} \) samples of the posterior longitudinal ability at each visit after burn in \( B \). Summary measures on the matrix of posterior abilities are readily made. Most important for practical purposes is the median posterior ability. Furthermore, we can take the spread in the posterior distribution into consideration by looking at the credible interval.

6.2.1 Subject specific inference

The definition of a posterior credible interval is included in appendix A.1.2. Let \( \tau_i \) denote the observation times of (new) subject \( i \). For every visit, a posterior credible interval on ability can be computed. Although the GP allows for a more refined interpolation (Rasmussen and Williams, 2006), simply connecting the intervals piecewise linearly gives an approximation to a region of posterior subject specific ability.

6.2.2 Reference lines

Van Buuren (2014) uses a diagram of reference ability curves (“the population”) for inference (see section 7). As many ability profiles are observed, summaries (over time) can be made to compare profiles of new subjects against. When a credible interval on posterior ability (in our case of the reference estimation) must be obtained, this approximation is intricate as times vary. Using a bandwidth parameter and dividing the time horizon into \( H \) points, (median) posterior abilities of a group of subjects can be summarized by computing the \( \alpha \) quantile in the range of one bandwidth left and right of each point \( h \in H \).

**Definition 6.1** (Smooth quantiles using bandwidth). Let the time horizon be divided into \( H \) equally spaced points \( \{ t_h : h \in [1, H] \} \). Define the bandwidth parameter \( bw > 0 \) and let \( \alpha \in [0, 1] \). Let \( \theta^{[h]} := \{ \theta_{j,t} : t \in (t_h - bw, t_h + bw) \} \) be a vector of all \( N_h \) ability estimates falling into time section \( h \). Define the following section quantile \( \forall h \in [1, H] \):

\[
Q_{\alpha}^{[h]} := \inf \{ x \in \theta^{[h]} : \alpha \leq \frac{1}{N_h} \sum_{k=1}^{N_h} 1_{\{ \theta_{k}^{[h]} \leq x \}} \} \tag{21}
\]

Where the expression involving the sum is known as the empirical distribution function of the values in \( \theta^{[h]} \). A smooth line (e.g. polynomial degree 3) then connects the \( H \) section quantiles to form the smooth quantile curve.

For future reference we move slightly ahead towards notation used in definition 7.4 and label this smooth curve as \( C_{S100}^{[100]}(t) \), where the \( S \) indicates the method we propose in definition 6.1. We found \( H = 100 \) equally spaced points between the minimum and maximum observed time and a bandwidth of 0.05 (years) to work well for the SMOCC data (see section 9). Remark that other summary procedures, e.g. based on LMS curves (Cole and Green, 1992), can serve as alternative (see further subsection 7.5).

6.2.3 Covariate corrected reference diagram

Notice that we have more information for subject specific inference than the population ability estimate. Based on the covariate information, \( \beta, \gamma \) and variance parameters \( \Sigma_{\theta}, \nu_0, w_0 \), we create empirical reference quantiles by drawing from multivariate normals throughout the hierarchy model definition 2.1. Combining levels, we have that:

\[
\theta_{i}^{REF} | \beta, \gamma, \Sigma_{\theta}, \nu_0, w_0 \sim N_{|\tau_i|}(X_iW_i \left[ \begin{array}{c} \beta \\ \gamma \end{array} \right], X_i\Sigma_{\theta}X_i^t + \Psi_{\theta,j}) \tag{22}
\]

Smooth quantile lines (definition 6.1) of draws from this distribution plotted over time are defined as covariate corrected reference diagram. These concepts are illustrated on the SMOCC data in section 9.
Part III:
Applications to Developmental score data TNO
7 Introduction to the D-score and SMOCC data

In part III, the proposed method is applied to the SMOCC data and compared to earlier work in which the Developmental score (D-score) was described using the same data set (Van Buuren, 2014; Jacobusse, Van Buuren, and Verkerk, 2006).

7.1 Description and background

For this application, we use data from the "Social Medical survey Of Children attending child health Clinics" (SMOCC) project (Herngreen et al., 1992). On targeted ages between one and 30 months, \( n = 2038 \) subjects were tested on 57 items (fail/pass) spread across ten visits according to the Van Wiechen scheme. A novelty in the SMOCC data was that, in addition to the Van Wiechen scheme, all subjects were tested on the scheduled items of the next visit as well, hence doubling observations on each item. This links the study design in such a way that the Rasch model becomes identifiable (Jacobusse, Van Buuren, and Verkerk, 2006).

<table>
<thead>
<tr>
<th>Visit</th>
<th>Items</th>
<th>( n_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1 2 3 4 5 6 7</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
<td>12</td>
</tr>
<tr>
<td>III</td>
<td>8 9 10 11 12 13 14 15 16 17 18</td>
<td>11</td>
</tr>
<tr>
<td>IV</td>
<td>13 14 15 16 17 18 19 20 21 22 23 24 25</td>
<td>13</td>
</tr>
<tr>
<td>V</td>
<td>19 20 21 22 23 24 25 26 27 28 29 30 31</td>
<td>13</td>
</tr>
<tr>
<td>VI</td>
<td>26 27 28 29 30 31 32 33 34 35 36 37</td>
<td>12</td>
</tr>
<tr>
<td>VII</td>
<td>32 33 34 35 36 37 38 39 40 41 42 43</td>
<td>12</td>
</tr>
<tr>
<td>VIII</td>
<td>38 39 40 41 42 43 44 45 46 47 48 49 50</td>
<td>13</td>
</tr>
<tr>
<td>IX</td>
<td>44 45 46 47 48 49 50 51 52 53 54 55 56 57</td>
<td>14</td>
</tr>
<tr>
<td>X</td>
<td>51 52 53 54 55 56 57</td>
<td>7</td>
</tr>
</tbody>
</table>

7.2 Descriptive statistics SMOCC

Table 5 includes descriptive statistics on the SMOCC data extended with covariates gender (categorical) and duration of pregnancy (continuous, displayed as histogram). With respect to Van Buuren (2014), the data is slightly filtered. Starting from \( n = 2038 \) with a total number of 164,885 observations, one subject was removed for not having an entry on the duration of pregnancy variable. Furthermore, observations of 74 subjects with less than 4 visits were removed to accommodate fitting polynomials up to order \( q = 3 \) in the new method. This filtered data is used as consistent basis for all analyses in the coming sections, descriptive statistics to which are presented in table 5.

Remark that the analysis of Jacobusse, Van Buuren, and Verkerk (2006) did not take any covariates into account and uses the Rasch model instead. Furthermore, the longitudinal character of the data was not exploited to full extent, treating each ability estimate \( \theta_i \) as independent \( \theta_i \). To avoid confusion in notation, we denote this larger set of hypothetical individuals by index \( \kappa \in [1, N_\kappa] \) with abilities \( \theta_\kappa := \theta_i \). In the case of the SMOCC data, \( N_\kappa \) is roughly \( 9 \cdot n \), since most subjects are observed over 9 visits.
### Table 5: Descriptive statistics SMOCC data.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Aggregated</th>
<th>Per subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
</tr>
<tr>
<td>Observations</td>
<td>163,652</td>
<td>83.4</td>
</tr>
<tr>
<td>Pass</td>
<td>127,025</td>
<td>0.776</td>
</tr>
<tr>
<td>Fail</td>
<td>36,627</td>
<td>0.224</td>
</tr>
<tr>
<td>Subjects n</td>
<td>1963</td>
<td>-</td>
</tr>
<tr>
<td>Visits</td>
<td>16,372</td>
<td>8.3</td>
</tr>
<tr>
<td>Covariates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girl Observations</td>
<td>83,417</td>
<td>83.8</td>
</tr>
<tr>
<td>Girl n and Pass</td>
<td>996</td>
<td>0.788</td>
</tr>
<tr>
<td>Boy Observations</td>
<td>80,235</td>
<td>83.0</td>
</tr>
<tr>
<td>Boy n and Pass</td>
<td>967</td>
<td>0.763</td>
</tr>
<tr>
<td>Preg. duration (wks)</td>
<td>1963</td>
<td>39.6</td>
</tr>
</tbody>
</table>

### 7.3 Rasch Model

**Definition 7.1 (Rasch Model).** Let $Y_{k,j} \in \{0,1\}$ be a binary random variable denoting the score of subject $k \in [1, N_k]$ on item $j \in [1, J]$. The Rasch model states:

$$Y_{k,j} | p_{k,j} \sim \text{Bin}(1, p_{k,j})$$

$$\logit(p_{k,j}) = \theta_k - \delta_j$$

$$\Rightarrow p_{k,j} := \mathbb{P}(Y_{k,j} = 1 | \theta_k, \delta_j) = \frac{\exp(\theta_k - \delta_j)}{1 + \exp(\theta_k - \delta_j)}$$

Where $\theta_k$ denotes a subject specific continuous latent trait ("ability") and $\delta_j$ an item specific ("difficulty") parameter taking values in $\Theta$.

Details of Rasch estimation can for instance be found in Fisher and Molenaar (1995). Typically, difficulty estimation is more robust than ability estimation, since there are more subjects than items in most scenarios. Therefore, a possible reason to move from the frequentist setting to a Bayesian approach is to shrink the ability estimate under small sample sizes. Assuming the density $\pi$ of the posterior and prior both exist (see further appendix A.1.2), we have the following notion of posterior distribution.

**Proposition 7.1 (Posterior ability distribution).** Let $y_k \in \{0,1\}^{|J|}$ be a vector of observations of $Y_{k,j}$ with $j \in J$, a subset of items surveyed at the visit of interest. Assume observed item scores are conditionally i.i.d. and item difficulties are known, such that the likelihood (i.e. joint conditional distribution $\mathbb{P}^{(Y)}$ as in definition A.2 and proposition A.2) is of the following product form:

$$\mathbb{P}^{(Y)}(Y_{k,J} = y_k | \theta_k, \delta_j) := L_J(\theta_k | Y_{k,J} = y_k, \delta_j) = \prod_{j \in J} \mathbb{P}(Y_{k,j} = 1 | \theta_k, \delta_j)^y_{kj} \cdot (1 - \mathbb{P}(Y_{k,j} = 1 | \theta_k, \delta_j))^{1-y_{kj}}$$

By Bayes Rule (proposition A.1 and A.2), we have $\forall \theta \in \Theta$:

$$\pi(\theta_k = \theta | Y_{k,J} = y_k, \delta_j) = \frac{\mathbb{P}^{(Y)}(Y_{k,J} = y_k | \theta_k = \theta, \delta_j) \cdot \pi(\theta_k = \theta)}{\int_\Theta \mathbb{P}^{(Y)}(Y_{k,J} = y_k | \theta_k = u, \delta_j) \cdot \pi(\theta_k = u) \, du}$$

Where the integral in the denominator follows from the Law of Total Probability.
7.4 D-score estimation

Van Buuren (2014) uses standard Rasch estimation software (RUMM2020) to arrive at ability and difficulty estimates in which all observations are treated independently. He then suggests the EAP method (cf. subsection 7.4) to summarize information of item subsets into ability estimates at the respective time points. The point estimate of the expectation of posterior 7.1 is known as the Expected A Posteriory (EAP) estimator (Bock and Aitkin, 1981; Bock and Mislevy, 1982).

**Definition 7.2 (EAP).** The point estimate of the Expected A Posteriori (EAP) ability for subject \( \kappa \) after seeing the set of \( \mathcal{J} \) items is defined as the expectation with respect to the posterior in proposition 7.1:

\[
EAP(\theta_\kappa)_{\mathcal{J}} = \int_\Theta u \cdot \pi(\theta_k | Y_{k,\mathcal{J}} = y_k, \delta_j) \, du = \frac{\int_\Theta u \cdot \Phi(y_{k,\mathcal{J}} | \theta_k = y_k, \delta_j) \cdot \pi(\theta_k = u) \, du}{\int_\Theta \Phi(y_{k,\mathcal{J}} | \theta_k = y_k, \delta_j) \cdot \pi(\theta_k = u) \, du}
\]

**Computation**

Bock and Aitkin (1981) and Bock and Mislevy (1982) assume a standard normal prior on \( \theta_k \) and suggest the probit link instead of the logit link function in definition 7.1. In that case, both integrals in definition 7.2 can be approximated by Gauss-Hermite quadrature. More generally, numerical integration methods can be used to approximate the EAP solution, also in case of the logit link and any parametric prior.

Before applying the EAP method, Van Buuren (2014) transforms the abilities into D-scores using definition 7.3. The scaling is such that two difficulties are anchored at practically relevant values. Remark that the remaining difficulties are transformed analogously. These transformations are made without loss of generality since only the difference between ability and difficulty is identified (see further subsection 4.5).

**Definition 7.3 (D-score (Van Buuren, 2014)).** \( \forall \kappa \in [1, N_\kappa] \)

\[
D_\kappa := 38.906 + 2.1044 \cdot \theta_k
\]

Then, Van Buuren (2014) chooses the prior such that at every time point \( t \), the prior is set to \( N(\mu_t, 5) \), where \( \mu_t := \frac{1}{N_\kappa} \sum_{\kappa=1}^{N_\kappa} D_\kappa(t) \), the mean of observed D-scores at time \( t \). Remark that past information of the individual is not taken into account. Finally, this prior is used to imply \( D_{it} \) estimates \( \forall i \in [1, n] \), in a sense summarizing the \( D_\kappa \) information belonging to that specific subject.

7.5 Inference in the reference diagram

In Van Buuren (2014), inference of D-scores is made by displaying them on a reference diagram (figure 10), in which the LMS method (Cole and Green, 1992) is used to construct quantile lines based on D-scores of all \( N \) subjects. For every (scheduled) observation time \( t \), the LMS method assumes that a standardization of \( D_{.,t} \) followed by a power transformation is standard normally distributed. Across time, these distributions are assumed independent. A penalized maximum likelihood procedure finds optimal values for \( L, M \) and \( S \) at all input times, in which three penalties regulate smoothness. Using optimal values for \( L(t), M(t) \) and \( S(t) \), the \( \alpha \) quantiles (or 100\( \alpha \) "centiles" in Cole and Green (1992)) can be found as follows:

**Definition 7.4 (Centiles (Cole and Green, 1992)).** Let \( \Phi(\cdot) \) denote the standard normal cumulative distribution function and \( \alpha \in [0, 1] \)

- For \( L(t) \neq 0 \): \( C_{100\alpha}(t) = M(t) \cdot (1 + L(t)S(t)\Phi^{-1}(\alpha))^{1/L(t)} \)
- For \( L(t) = 0 \): \( C_{100\alpha}(t) = M(t) \exp(S(t)\Phi^{-1}(\alpha)) \)

Worked out details of the penalized ML procedure of Cole and Green (1992) can be found in Appendix B.4. Table 4 in Van Buuren (2014) lists the optimized \( L(t), M(t) \) and \( S(t) \) values.

Figure 10 includes the centile lines positioned at \( +0 \) SD (the median), \( +1 \) SD (\( \approx \alpha \approx 0.84 \)), \( +2 \) SD (\( \approx \alpha \approx 0.977 \)), \( -1 \) SD (\( \approx \alpha \approx 0.16 \)), \( +2 \) SD (\( \approx \alpha \approx 0.023 \)), since SD=1. Furthermore, two exemplary D-score curves are drawn in.
8 Reference estimation on the SMOCC data

In this section the reference estimation using Empirical Bayes is made on the SMOCC data. After initializing values have been found, several set-ups are tested with respect to the degree of the polynomial, the covariates, the kernel parameters and the link.

8.1 Initialization

Using the initialization procedure described in subsection 5.2, we generated figure 11 in which the $\text{Acc}$ (definition 5.1) optimization is summarized. All information displayed is based on $q = 3$. The symmetric adjustment parameter was found based on a grid search from 0 to 20 by 0.5 step length. The asymmetric $\lambda$ are found in the same way, but now with step sizes of 1.0 for computational efficiency. The profile plots based on three different choices of $\lambda$ (no adjustment, symmetric adjustment respectively asymmetric adjustment) are displayed in figure 12. Similar to table 1 based on definition 5.1, we summarize the adjustments in table 6. The asymmetric results show an interesting trait of the data: perfect (1) response profiles are far more common than perfect (0). This is also expressed in the low $\hat{\lambda}_{\text{min}}$ parameter in table 6. There, results of the fitted (double hat) $\hat{\theta}$ polynomials of equation (19) are included as well. Observe that the raw quadratic polynomial fits rather well and only leads to a minor loss in accuracy and binomial likelihood with respect to the $\hat{\theta}$ fit.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Raw</th>
<th>Adjusted</th>
<th>$\hat{\theta}$ quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{\text{min}}$</td>
<td>0</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>$\lambda_{\text{max}}$</td>
<td>0</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>$ll \times 1e5$</td>
<td>-1.50</td>
<td>-1.45</td>
<td>-1.45</td>
</tr>
<tr>
<td>Acc</td>
<td>0.781</td>
<td>0.804</td>
<td>0.804</td>
</tr>
</tbody>
</table>

Table 6: Binomial likelihood and accuracy before and after adjustment - SMOCC.
In case of \( q = 2 \), an optimal asymmetric \( \hat{\lambda} \) of (8,14) is found, which is comparable to the \( q = 3 \) case looking at the contour plot in figure 11.

### 8.2 Model set-ups: polynomial degree and covariates

We use the following specification of inner and outer loop, where \( \text{rep}(10,500) \) means repeat inner loop of length 10 a number of 500 times.

\[
M2 = (\text{rep}(10,500), \text{rep}(100,50), \text{rep}(1000,5), 5000, 10000, 20000)
\]

In total, 558 outer loops generate \( M = 50,000 \) iterations. A burn-in of \( Bp = 20\% \) is used in each inner loop before updating the outer loop parameters. Inference on inner loop parameters is made based on all iterations of the final outer loop (no burn-in).

Reference procedures of four different set-ups of model definition 2.1 are computed, where the polynomial degree \( q \) is varied between 2 and 3, with and without additionally including covariates. Using the average posterior \( \theta \) of the final outer loop, table 7 summarizes model fit based on definition 5.1. Furthermore,
the Deviance Information Criterion ((DIC) Spiegelhalter, Best, and Carlin, 2002) is used as model fit criterion penalized for the number of effective parameters.

**Definition 8.1 (DIC (Spiegelhalter, Best, and Carlin, 2002)).** Let $\theta$ denote $M$ posterior samples $\theta^{(m)}$ of the parameter of interest. Furthermore, let $Y$, $\delta$ and link function $g(\cdot)$ be defined as before.

$$DIC(\theta) := \bar{D}(\theta) + 2 \cdot \left( \bar{D}(\theta) - D(\bar{\theta}) \right)$$

$$\bar{D}(\theta) := \frac{1}{M} \sum_{m=1}^{M} \left( -2 \cdot ll(Y, \theta^{(m)}, \delta, g) \right)$$

where we recognize $-2 \cdot ll(Y, \cdot, \delta, g)$ as the deviance and $\bar{\theta}$ as the posterior mean.

Remark that when written in this way, the DIC resembles the form of penalization in AIC and BIC. A notable difference is that the DIC uses variation in posterior distribution (treating the parameter as random variable) to compute the penalty, making it more appropriate here than the other criteria. On a standard

<table>
<thead>
<tr>
<th>Covariates</th>
<th>With $q=3$</th>
<th>Without $q=3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree $q$</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>ll probit</td>
<td>-1.606</td>
<td>-1.805</td>
</tr>
<tr>
<td>Acc</td>
<td>0.7958</td>
<td>0.7882</td>
</tr>
<tr>
<td>DIC probit</td>
<td>1.767</td>
<td>1.368</td>
</tr>
</tbody>
</table>

Table 7: Accuracy and log likelihood model set-ups - SMOCC.

Dell Inspiron laptop with 8 GB RAM and i7 Intel processor, fitting the reference procedure (given the inner and outer $M$ specifications) takes about 4 hours. Details are included in table 14 (Appendix C.1).

<table>
<thead>
<tr>
<th>Par.</th>
<th>$q=3$ CI low</th>
<th>Mean</th>
<th>Median</th>
<th>CI high</th>
<th>$q=2$ CI low</th>
<th>Mean</th>
<th>Median</th>
<th>CI high</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>7.314</td>
<td>7.530</td>
<td>7.530</td>
<td>7.747</td>
<td>10.68</td>
<td>10.85</td>
<td>10.85</td>
<td>11.03</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>67.85</td>
<td>68.62</td>
<td>68.62</td>
<td>69.40</td>
<td>48.69</td>
<td>49.02</td>
<td>49.02</td>
<td>49.35</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-35.78</td>
<td>-34.98</td>
<td>-34.98</td>
<td>-34.20</td>
<td>-11.89</td>
<td>-11.75</td>
<td>-11.75</td>
<td>-11.61</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>6.905</td>
<td>7.138</td>
<td>7.139</td>
<td>7.375</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.189</td>
<td>0.333</td>
<td>0.334</td>
<td>0.476</td>
<td>0.182</td>
<td>0.329</td>
<td>0.329</td>
<td>0.477</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.289</td>
<td>0.329</td>
<td>0.330</td>
<td>0.369</td>
<td>0.293</td>
<td>0.334</td>
<td>0.334</td>
<td>0.375</td>
</tr>
</tbody>
</table>

Table 8: Parameter estimates and credible sets ($\alpha = 0.05$) for $q=3$ and $q=2$ with covariates.

From table 8 we note that the posterior distribution of the covariate effects remains roughly unchanged between the two polynomial degrees. On average, a girl has a 0.33 higher ability than a boy, and each week of pregnancy duration increases the ability by 0.33 ceteris paribus. Depending on the administered item and link function, this increase in ability carries over (non-linearly) to an increase in probability of answering the item correctly. We stress that the additive gender effect (no interaction with time) and linear pregnancy duration are simplifications.

Figure 13 illustrates the ability profiles the mean posterior polynomial and covariate effects parametrize. Despite coefficients being different, both polynomials show similar trends, particularly in the 0-24 month range. Beyond that range, as the mean posterior ability (boy at 40 weeks implies no covariate effect addition) clearly indicates, the cubic trend places abilities relatively far in excess of difficulties (as seen in the left plot). Inference in this region is only supported by low sample sizes, since relatively few children follow up past the age of 2 years. We conclude that $q=3$ overfits the data, as was already indicated by the DIC in table 7.
Figure 13: Random profiles (based on mean posterior estimates) including covariates for $q = 3$ (left) and $q = 2$ (right). In the middle, the mean posterior ability of a boy born at 40 weeks is drawn for $q = 3$ respectively $q = 2$ (dotted). All item difficulties are drawn in as well ($\times$).

Figure 14: Figure panel trace plots $q = 2$ with covariates.
Figure 15: Figure panel outer loop trace plots $\Sigma_b$ for $q = 2$ with covariates.

Figure 16: Figure panel outer loop trace plots $(v_0, w_0)$ for $q = 2$ with covariates.
8.3 Sensitivity analysis \( w_0 \)

We select the cubic set-up with covariates (saturated model) and analyse the sensitivity of parameter estimates against changes in \( w_0 \). Remark that the scenario \( w_0 = 0 \) is equivalent to a model with a diagonal homoskedastic covariance matrix.

<table>
<thead>
<tr>
<th>Polyn. degree</th>
<th>Model fit</th>
<th>( w_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>( q = 3 )</td>
<td>ll probit (( \times 1e5 ))</td>
<td>-1.606</td>
</tr>
<tr>
<td></td>
<td>Acc</td>
<td>0.796</td>
</tr>
<tr>
<td>( q = 2 )</td>
<td>ll probit (( \times 1e5 ))</td>
<td>-1.805</td>
</tr>
<tr>
<td></td>
<td>Acc</td>
<td>0.788</td>
</tr>
</tbody>
</table>

From the sensitivity analysis we conclude that both for \( q = 3 \) as well as \( q = 2 \), there is little difference in model fit between small values of \( w_0 \). Although we must be careful since all trace plots are scaled differently on the vertical axis, this conclusion seems to be in line with the relatively noisy pattern of \( w_0 \) in the outer trace (figure 16). It seems that the procedure is indifferent to a range of low \( w_0 \) values. However, for larger values of \( w_0 \) the results become sensitive as this covariance specification aligns poorly with the detailed polynomial modelling.

8.4 Probit vs logit and the link scale

Forcing \( w_0 \) to be a constant significantly speeds up the procedure, from which the study in this subsection benefits. Given the comparison made by table 9, simplifying to \( w_0 = 0 \) seems sensible. Throughout, we assume that (small) changes to the data do not alter the preference for fixing \( w_0 = 0 \).

The model set-up under study is \( q = 2 \) with covariates and \( w_0 = 0 \). The second column of table 10 presents \( ll \) and \( Acc \) for both the probit and logit (normal mixture \( k = 5 \); section 3) link. Here we use \( M_1 \) and \( M_2 \) loops as defined before. The logit procedure takes little about 6.2 hours, which is 1.7 times longer than the respective probit procedure in this set-up (see appendix C.1).

Under a scale of 100, the difficulties of Van Buuren (2014) are transported directly and fixed for parameter identification 4.5. As was mentioned in subsection 3.4, varying this scale factor influences the power to identify differences between difficulty and ability when transformed over the link. To investigate the resulting model fits, a sensitivity analysis is performed. To limit the computational expense, a shorter simulation set-up \( M_2^S \) is used. This shorter set-up is to some extent justified by the trace plots in figures 14 and 15, but should be seen as an approximation in any case.

\[
M_2^S = (rep(10, 400), 1000)
\]

Table 10 displays summary statistics on a selection of scales included in the study. Firstly, verify that the values in the second column (scale=100), which are taken from table 9 with \( w_0 = 0 \), are close to those in the sensitivity scale=100 column under the short simulation \( M_2^S \). This gives some reassurance that the shorter set-up is not too short.

Secondly, remark that the overall procedure is sensitive to the scale. Figure 17 contains summary statistics on all scales in the sensitivity analysis in a graphical illustration, which we use to lead the discussion. Most informative is plot (a), from which we conclude that scales under approximately 100 lead to suboptimal \( v_0 \) estimates. An explanation of this phenomenon was given in subsection 3.4. Plot (b) as well as (c) show that the fit criteria have local optima as function of scale, but we argue against there use since they occur at scale values for which the rescaled \( v_0 \) is unstable. The fact the model fits well under low scales is caused by the extremity of the link, which coincidentally matches most scoring patterns (many subjects score \( Y = 1 \) by
Table 10: Sensitivity analysis scale of difficulties.

<table>
<thead>
<tr>
<th>Scale Type</th>
<th>Scale</th>
<th>M2 100</th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
</tr>
</thead>
<tbody>
<tr>
<td>probit</td>
<td>ll (×1e5)</td>
<td>-1.805</td>
<td>-0.782</td>
<td>-0.670</td>
<td>-0.665</td>
<td>-0.738</td>
<td>-1.018</td>
<td>-1.803</td>
<td>-2.115</td>
<td>-2.367</td>
<td>-2.570</td>
</tr>
<tr>
<td></td>
<td>Acc</td>
<td>0.788</td>
<td>0.780</td>
<td>0.812</td>
<td>0.812</td>
<td>0.800</td>
<td>0.788</td>
<td>0.786</td>
<td>0.785</td>
<td>0.784</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v0 rescaled</td>
<td>2.959</td>
<td>29.735</td>
<td>2.487</td>
<td>2.307</td>
<td>2.490</td>
<td>2.755</td>
<td>2.961</td>
<td>2.992</td>
<td>3.007</td>
<td>3.012</td>
</tr>
<tr>
<td>logit</td>
<td>ll (×1e5)</td>
<td>-1.078</td>
<td>-1.077</td>
<td>-0.870</td>
<td>-0.792</td>
<td>-0.757</td>
<td>-0.766</td>
<td>-1.078</td>
<td>-1.274</td>
<td>-1.478</td>
<td>-1.687</td>
</tr>
<tr>
<td></td>
<td>Acc</td>
<td>0.771</td>
<td>0.783</td>
<td>0.795</td>
<td>0.786</td>
<td>0.782</td>
<td>0.778</td>
<td>0.771</td>
<td>0.771</td>
<td>0.771</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v0 rescaled</td>
<td>2.148</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>2.148</td>
<td>2.314</td>
<td>2.394</td>
<td>2.440</td>
</tr>
</tbody>
</table>

nature of the SMOCC data, see table 5). Furthermore, plot (c) suggests that the accuracy stabilizes along with rescaled $v_0$, which is another signal we reach desired invariance at higher scales. The log likelihoods of plot (b) do not share this property, but this is to be expected since the link distribution is (by design) not invariant.

Comparing probit and logit (dotted) results, we see that the accuracy is generally higher under the probit link for the displayed scales. Furthermore, observe that it stabilizes more quickly in terms of rescaled $v_0$, which is particularly evident from table 10. Remark that the (much) fatter tails of the truncated logit distribution (see figure 2) have such an extreme effect on the procedure under low scales that $v_0$ collapses to zero, which suggests the summarized ability over visits is in excess of all difficulties.

The higher accuracy, slightly robuster results and faster computation time dictate a preference for the probit link unless practice (e.g. interpretation) demands otherwise. A choice of scale=100 seems well-informed, certainly for the probit link, and practically appealing as well since this is exactly aligned with earlier work by Van Buuren (2014).

8.5 Summary of reference procedure

With the scale of the analysis reassured, we briefly summarize the model set-ups and sensitivities. Table 9 suggests that a simple covariance specification $w_0 = 0$ can be used, but for the moment we prefer a flexible $w_0$. In this covariance specification, table 11 points towards $q = 2$ with covariates as being the best model in terms of DIC. The resulting estimates were summarized in table 8 and figures 14, 15 and 16.

Figure 17: Figure panel link-scale sensitivities for the probit link and logit link (dotted). Plot (a) includes the rescaled $v_0$ estimate on the respective final outer loop, which should be constant. Plot (b) shows the log likelihood and plot (c) the accuracy as introduced in definition 5.1.
mean posterior estimates of inner loop parameters respectively the final ($M^{th}$) estimate of the outer loop parameters are used as reference (i.e. prior) for the subject specific sequential updating section that follows.

$$\beta = \begin{bmatrix} 10.85 \\ 49.02 \\ -11.75 \end{bmatrix} \quad \gamma = \begin{bmatrix} 0.329 \\ 0.334 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 5.666 & -8.022 & 2.792 \\ -8.022 & 16.13 & -5.861 \\ 2.792 & -5.861 & 2.143 \end{bmatrix} \quad \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} 2.959 \\ 0.011 \end{bmatrix}$$

The reference parameters displayed here are rounded. Finally, for completeness we mention that the difficulties $\delta$ are imported without transformation (i.e. scale 100) from Van Buuren (2014).

9 Subject specific updating on the SMOCC data

In this section, the methodology discussed in section 6 is illustrated on the SMOCC data using the reference parameters of subsection 8.5. For all of the following analysis, we use Algorithm 2 with $M = 25,000$ and $B = 0.2 \cdot M$, such that $\tilde{M} = 20,000$. For a subject with 10 visits (i.e. the full SMOCC scheme), results are obtained within three seconds. Throughout we work with the same two subjects as illustration that were also used by Van Buuren (2014, Fig.1).

9.1 Sequential estimation and shrinkage

We start by investigating sequential updating versus a Bayesian update after all longitudinal information is available at once. Figure 18 shows the difference between the two for the focus subjects not to be very profound. Two points to mention are the first visit estimate of the blue profile and the 9 month point of the red profile. There, a somewhat stronger adjustment towards subject specific information seems to occur when considered after all outcomes have become available.

![Figure 18: Median posterior ability of two subjects with sequential updates (a) and all information available at once (b).](image)

*Readability of most figures is not colour dependent, but here we stay consistent in layout with Van Buuren (2014). The “red” profile is the bottom one at the second half of the time horizon.*
Although discussed in detail in section 10, already at this stage it is relevant to compare figure 18 against the results of Van Buuren (2014). Notice that the sequential update is most relevant in practice, but the Bayesian update based on full information is most natural to compare the method of Van Buuren (2014) against on the SMOCC data. Generally speaking, profiles generated by both methods resemble each other, although the red subject profile seems to be positioned somewhat less extreme in the new method. Interestingly, there is a clear difference in the positioning of the 9 month point for the red subject profile between the methods. This holds true both for the sequential as well as the Bayesian update on full information. The first visit estimate of the blue subject profile is positioned differently only when figure 10 is compared against figure 18 (b).

### 9.2 Covariate corrected reference diagrams

Using the procedure described in section 6, covariate corrected reference diagrams are generated for both subjects under investigation (figure 19). Remark that in case of this reference model, the diagram is shifted by approximately $(38 - 36) \cdot 0.334$, since both subjects are boys born at 36 (blue) respectively 38 (red) weeks. This implies that the under performance of the red subject profile becomes somewhat more striking. The quantile lines plotted are those of $\alpha = (0.0023, 0.16, 0.5, 0.84, 0.977)$ which is equivalent to the median and $\pm 1SD$, $\pm 2SD$ on the standard normal. Again, briefly comparing figures 10 and 19, the reference diagram of Van Buuren (2014) seems to increase more steeply and end higher than that of the new method does. However, the $q = 3$ versus $q = 2$ comparison in subsection 8.2 (in particular figure 13) already concluded that estimates are based on a low amount of subjects in that region and the risk of overfitting increases.

The next section contains an elaborative comparison between the method of Van Buuren (2014) and the new method. This includes a more detailed look at subject specific updates beyond the two subjects that were brought to attention here.

---

**Figure 19:** Reproduction of Van Buuren (2014) figure 1 using the new method on covariate corrected reference diagram - including posterior 95% credible set per visit.
9.3 Subject specific inference

We end this section with a potentially interesting idea for future research. Notice that from the subject specific update procedure, the entire ability distribution at the respective visit becomes available. The question to be addressed (outside the scope of this thesis) is whether the credible set can be used for inference when compared against the covariate corrected reference diagram.

The most basic inference would be to make a statement regarding the extent to which the credible set and centile lines in the diagram overlap, but caution is at place e.g. with respect to multiple testing. The 95% credible set per visit is indicated by the error bars in figure 19, again based on sequentially obtained information. Interestingly, the credible set of the 9 month point is particularly wide for the red subject profile and that of the first visit in case of the blue subject profile.

10 Comparison of methods

In this section, we compare our results to those obtained by Van Buuren (2014). Since the models and procedures underlying both estimations are inherently different, a comparison of likelihood using traditional methods like AIC, BIC or DIC (Spiegelhalter, Best, and Carlin, 2002) is by no means straightforward. Both methods are only comparable up to the level that posterior ability implies observed scores. Furthermore, since Van Buuren (2014) uses the EAP (definition 7.2), comparisons are made most naturally against the mean of posterior ability estimates stemming from the reference procedure. We leave the application of multilevel (Bayesian) information criteria to model (or rather, method) selection as open extension and focus on measures of likelihood at the observed level instead (definition 5.1).

10.1 Reference parameters

As the method of Van Buuren (2014) does not include covariates, comparisons are best made based on a reference procedure excluding covariates. Based on the analyses in section 8, we select a polynomial degree $q = 2$ and fix $w_0 = 0$ for computational efficiency. This leads to the following reference parameters, which are taken as mean posterior in case of the inner loop and final estimates in case of the outer loop.

$$\beta = \begin{bmatrix} 10.90 \\ 48.96 \\ -11.73 \end{bmatrix}, \quad \Sigma_b = \begin{bmatrix} 6.930 & -8.848 & 3.028 \\ -8.848 & 16.44 & -5.965 \\ 3.028 & -5.965 & 2.182 \end{bmatrix}, \quad \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} 2.957 \\ 0 \end{bmatrix}$$

A summary table including credible sets on the $\beta$ parameters, trace plots of the $\beta$ parameters and trace plots on all outer loop parameters are included in appendix C.3. The figures show no reason to question (stochastic) convergence of the reference procedure.

10.2 Cross-validation

In this subsection, a K-fold cross-validation (CV) is used to assess out-of-sample properties of both the EAP method of Van Buuren (2014) as well as the proposed EB reference procedure combined with Bayesian subject-specific updates. In order to limit computational expense, $K = 5$ is used. The probit link is used to illustrate the new method, for this is faster than the procedure involving (a normal mixture approximation to) the logit link.

In K-fold CV, (K-1) folds are used as training data and the remaining fold as test data, which is then repeated until all folds are tested. The folds consist of all observations on a subset of randomly selected individuals. For the SMOCC data, 4 folds consist of 393 subjects and one fold of 391.

In case of the EAP method, the mean curve of the $N(\mu_t, 5)$ prior is recalibrated to the training data, details are included in appendix C.2. This procedure was also used in computation of the Dscore estimates.
for \( n = 1963 \). In case of the new method, the EB reference procedure of section 4 is fitted on the training data, i.e. the parameters shown in the previous subsection are recalibrated. The test estimates are found by the subject specific sequential updates of section 9. Figure 20 plots the CV estimated profiles on all subjects by the respective methods.

![Figure 20: Test set estimates Van Buuren (2014) Dscore (left) and median posterior ability new method (right).](image)

Table 11 includes details on model fit based on the full SMOCC data as well as expected model fit based on 5-fold CV. Remark that by nature of the reference procedure (new method), only the mean posterior ability is available and comparisons must be made on this account. For the subject specific updates, the median is available as well, for which we have a general preference given the possibility of an asymmetric posterior distribution. However, the results in table 11 suggest there is little difference here (compare the final two columns). A cautionary note should be placed when comparing the new method results on the full data and CV. The full data results are not based on a Bayesian update and hence use full information, whereas the CV test set estimation is made by the sequential updating scheme, as this is practically most relevant. Using the updating procedure on the full data would make this table consistent but also statistically invalid, since the data would then be used twice. In any case, observe that the CV model fit shows a remarkable improvement against the full data procedure, which suggests the subject specific sequential updates are successful in reducing noise.

Comparing the model fits of Van Buuren (2014), note that there is little difference between the apparent and expected prediction error. This is caused by the robustness of the EAP prior against (small) changes of the data (details in appendix C.2).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Selection Full data</th>
<th>Full data CV</th>
<th>New Method Full data</th>
<th>CV Mean post.</th>
<th>Median post.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ll probit ((\times 1e5))</td>
<td>-1.201 -1.202</td>
<td>-1.799</td>
<td>-1.192</td>
<td>-1.193</td>
<td></td>
</tr>
<tr>
<td>Acc</td>
<td>0.843 0.843</td>
<td>0.788</td>
<td>0.815</td>
<td>0.815</td>
<td></td>
</tr>
</tbody>
</table>
Finally, comparing the two methods, we see that the new method outperforms in terms of log likelihood, but the method of Van Buuren (2014) does better in terms of expected (CV) accuracy. In summary, the CV results seem to be indifferent between methods when it comes to model fit. That said, upon being forced to choose on the sole basis of table 11, we would favour the better accuracy of Van Buuren (2014) over the higher log likelihood of the new method. In the next subsection, we broaden this discussion.

10.3 Summary of comparison

As was already pointed out, it is not straightforward to compare model fits on both methods, since they are inherently different. The method of Van Buuren (2014) treats all time points separately, effectively modelling a spline of degree $|\tau_i|$ for each subject $i$, whereas the new method parametrizes this trend more stringently through the polynomial. That this comes at cost of model fit is no surprise and the question remains to what extent the procedure benefits from the new approach. The introduction mentioned four issues, which we briefly address here.

The first issue concerns the smoothness of individual profiles (to the eye of the practitioner). Section 9 showed that the profiles of two focus subjects were somewhat more smooth in the new method than in that of Van Buuren (2014), but it is difficult to make a formal conclusion on this basis. If we define a "smooth" line to be the model fit of a $d$-th degree polynomial (one linear model for each subject), we can compare the two methods based on the average (over subjects) MSE with regard to the smooth ability curves. Table 12 reads that the median posterior is relatively smooth, with the EAP showing 1.3 times higher MSE on average.

Table 12: Average MSE as measure of smoothness. The relative column compares EAP against median posterior rounded to one digit.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 3$</td>
<td>5.86</td>
<td>4.66</td>
<td>4.50</td>
<td>1.3</td>
</tr>
<tr>
<td>$d = 4$</td>
<td>3.41</td>
<td>2.78</td>
<td>2.66</td>
<td>1.3</td>
</tr>
<tr>
<td>$d = 5$</td>
<td>2.25</td>
<td>1.77</td>
<td>1.68</td>
<td>1.3</td>
</tr>
</tbody>
</table>

The second point of concern is the longitudinal character of the data. In the EAP procedure of Van Buuren (2014), grouping of measurements within subjects is not taken into account. Clearly, the new method has a theoretical advantage here.

The third issue that was mentioned relates to the sensitivity of the EAP procedure to the $N(\mu_1, 5)$ prior. The new method replaces this dependence with the reference procedure prior, which is better aligned with the longitudinal design. Interestingly, the comparison of $q = 3$ versus $q = 2$ for the polynomial degree seems to suggest that the EAP prior is based on an very low sample size towards the end of the time horizon. This implies that the reference diagram of Van Buuren (2014) may be too high towards the end, although this is difficult to substantiate without further research on out-of-sample coverage of the credible sets.

Fourthly, the new method is better positioned to include covariate corrections, both in the estimation as well as in the reference diagram. This was illustrated in section 9.
Part IV:
Discussion and Conclusion
Discussion

The new method we propose leaves several future directions and points of discussion, of which we mention a few. We do so along the lines of varying (and variance) parameters, non-varying parameters, identification, scaling, modelling the SMOCC data and generalizations to other fields.

With respect to the variance modelling, we found the distinctive $w_0$ parameter to the GP not to add much to the model fit. We must conclude that despite its added value in non-equidistant measurements over time, it has not made a contribution here. However, generalizations to multiple $w$ parameters (time-dependent) may change this. Furthermore, our method could benefit from other structures not tested (e.g. continuous AR1). Along similar lines, one may wish to reconsider the use of an unstructured $\Sigma_b$ matrix, which was not discussed so far.

Regarding the non-varying parameters, transformations at the latent level (either at the vertical scale or the horizontal scale) could possibly reduce the amount of parameters needed in the polynomial. Care must be taken that the model remains linear as this is crucial for the conjugacy of the Gibbs procedure. Furthermore, no perfect response profile adjustment was made in the reference procedure, although the polynomial fit was shown to improve by this at initialization.

Besides contributing to interpretation, identifying the latent scale through difficulty parameters, as opposed to fixing a variance parameter, decreases the amount of steps in the Gibbs cycle and hence speeds up computation. However, for non-trivial longitudinal study designs, determining which item parameters to fix becomes an intricate point. Furthermore, it makes the estimation inherently consist of two stages and hence potentially suboptimal. Finally, it was shown that the entire procedure is sensitive to the scale at which these difficulties are introduced.

This scaling not only has an effect on the variance parameters, but also on the extent to which the probit and logit link give different results. In addition to this remark, the quality of the normal mixture approximation (order 5) after truncation deserves some further research, where attention should be paid to the computational expense of the resulting reference procedure.

The settings of the Empirical Gibbs procedure (most notably, the loop sizes $M_1$ and $M_2$) were determined ad hoc and convergence was assessed informally. A straightforward improvement can be made by including a stopping criterion on the outer loop. Furthermore, more elaborate techniques are available for monitoring both loops (Levine and Casella, 2001).

The SMOCC data was primarily used as illustration of the new method and further attempts to model it (e.g. by covariate interactions) are likely to improve its performance. Anyway, a model without covariates was the best basis of comparison against earlier work of Van Buuren (2014). We mention this further in the conclusion. Nevertheless, the covariate corrected reference diagram seems particularly promising to meet practical demand. A possible line of future research focuses on subject specific inference with respect to such diagrams. Besides inference, subject specific prediction is highly relevant as well in practice, for which our methods can be extended. There, the GP can again play a role.

Finally, we stress that applicability of the methods we propose is not limited to studies of human development, but general to situations of longitudinal 0/1 data.
Conclusion

This thesis concerns (Empirical) Bayesian estimation of longitudinal 0/1 data, with particular attention to smoothing and inference on the latent (link) level. We have extended the Gibbs sampling framework of Fox and Glas (2001) and Azevedo, Fox, and Andrade (2015) to allow indirect variance specification in the spirit of the GLMM, with inclusion of the GP (as suggested by Shi and Wang (2014)) to handle non-equidistant measurement times flexibly. Furthermore, given the detailed account of mean curve modelling including covariates and freedom to change the variance structure by using item parameters for identification, the studies of Fox and Glas (2001) and Azevedo, Fox, and Andrade (2015) are essentially united. Moreover, a point of future research mentioned by Fox and Glas (2001) concerning the extension towards the logit link is addressed by a normal mixture approximation. With respect to hyper prior specification, we prefer to be pragmatic rather than subjective and therefore adapt the framework towards Empirical Gibbs (Casella, 2001).

A novelty is the use of the mean posterior of this estimation as prior in a (fully) Bayesian subject specific sequential update procedure. The resulting (individual) posterior ability distribution can be used for inference, e.g. when displayed on a reference diagram as suggested by Van Buuren (2014). We suggest a covariate corrected reference diagram as improvement, which has great potential for practical use in human development monitoring.

All code is hand-written in R and relies heavily on C++ through the Rcpp package (Eddelbuettel and François, 2011; Bates and Eddelbuettel, 2013) to increase speed. On a standard Dell Inspiron laptop with 8 GB RAM and i7 Intel processor, the reference procedure takes a few hours, but a subject specific update is made within three seconds.

The proposed methods were illustrated on the SMOCC human development data from TNO, and a comparison was made between the new method and that of Van Buuren (2014). In terms of model fit both procedures are close, with perhaps a slight advantage for the method of Van Buuren (2014). On the other hand, the new method does better in terms of smoothing the latent ability curves. Furthermore, the new method models the longitudinal nature of the data more appropriately and is able to correct for covariate information.
References


REFERENCES


Appendices
Appendix Part I

A.1 Formal definition Frequentist and Bayesian Statistics

A.1.1 Frequentist Framework

In the frequentist framework, any statistical model is parametrized by a "true" population parameter (vector) $\theta_0$ in the sense that such a truth generally exists. This concept is often operationalized in practice by determining a maximum likelihood estimator (MLE) $\hat{\theta}$ after observing a small proportion of the population as sample in an experiment.

**Definition A.1 (Likelihood).** Let real-valued random variable $Y$ with observations $y := \{ y_i : i \in [1, n] \}$ be i.i.d. distributed with density $f$, parametrized by $\theta \in \Theta \subseteq \mathbb{R}^k$. Then considering $y$ fixed, the likelihood $\Theta \mapsto [0, 1]$ is defined as:

$$L(\theta | y) : \theta \mapsto \prod_{i=1}^{n} f_\theta(y_i)$$

Upon taking the natural logarithm, $L(Y|\theta)$ is known as the log-likelihood $l(Y|\theta)$:

$$l(\theta | y) : \theta \mapsto \sum_{i=1}^{n} \log(f_\theta(y_i))$$

A.1.2 Bayesian Statistics

In Bayesian statistics, rather than postulating that a general true parameter $\theta_0$ exists, it is considered to be a random variable itself.

**Definition A.2 (Bayesian Model and distributions).** Let $Y$ and $\vartheta$ be random variables on a joint measurable sample space $(\mathbb{R} \times \Theta, \sigma(B \times D))$ with (non-product) distribution $\Pi$.

The marginal distribution of $\vartheta$, denoted $\Pi^\vartheta : D \mapsto [0, 1]$ is called the prior distribution.

The conditional distribution of $Y$ given $\vartheta$, $\Pi_{Y|\vartheta} := (B \times \Theta) \mapsto [0, 1]$, is referred to as the model (upon specification of the prior):

$$\Pi_{Y|\vartheta}(B|\vartheta = \theta) \quad \forall B \in B, \theta \in \Theta$$

Furthermore, the conditional distribution of $\vartheta$ given $Y$ is known as the posterior distribution:

$$\Pi(D|Y = y) \quad \forall D \in D, y \in \mathbb{R}$$

The joint probability measure is not of the product type, since practically speaking we are only interested in models for which parameter and data are dependent. Although practically apparent, the existence of the conditional distributions is not shown yet. The following proposition states Bayes’ Rule including sufficient conditions for existence.

**Proposition A.1 (Bayes’ Rule (extended)).** Let $Y$ and $\vartheta$ be random variables taking values in a joint measurable sample space $(\mathbb{R} \times \Theta, B \times D)$ with (non-product) distribution $\Pi$. Assume $\Theta$ is a Polish space and $D$ is the Borel-algebra on $\Theta$. Let $\vartheta$ have prior distribution $\Pi^\vartheta$ and let the model $\Pi_{Y|\vartheta}$ be dominated by a $\sigma$-finite measure $\mu$ on $\mathbb{R}$, such that it has conditional density $f_\vartheta(y)$. Then, we have $\forall D \in D$ and $y \in \mathbb{R}$

$$\Pi(D|y) = \frac{\int_{D} f_\vartheta(y) \ d\Pi^\vartheta(\vartheta)}{\int_{\Theta} f_\vartheta(y) \ d\Pi^\vartheta(\vartheta)}$$

Progress towards practical application is made when $Y$ is considered as random vector, a collection of identically distributed random variables $Y_i$ with $i \in [1, n]$ that are conditionally independent such that the model can be factorized (in density form): $f_\vartheta(Y) = \prod_{i=1}^{n} f_\vartheta(Y_i)$. From hereon, such random vectors $Y$ will be referred to as conditional i.i.d. data.
Proposition A.2 (Bayes’ rule with conditional i.i.d. data). Let \( y \in \mathbb{R}^n \) be a sampled vector of \( Y \) and \( \vartheta \) a random variable with sample probability space \((\Theta, \mathcal{D}, \Pi^\vartheta)\) and density \( \pi^\vartheta \) with respect to the Lebesgue measure. Assume the model is dominated. Then \( \forall D \subseteq \mathcal{D} \)

\[
\Pi(D|y) = \frac{\int_D \prod_{i=1}^n f_\vartheta(y_i)) \; d\Pi^\vartheta(\vartheta)}{\int_\Theta \prod_{i=1}^n f_\vartheta(y_i)) \; d\Pi^\vartheta(\vartheta)} = \frac{\int_D \prod_{i=1}^n f_\vartheta(y_i)) \pi^\vartheta(\vartheta) \; d\vartheta}{\int_\Theta \prod_{i=1}^n f_\vartheta(y_i)) \pi^\vartheta(\vartheta) \; d\vartheta}
\]

Remark that the existence of the density (providing the second equality) is practically relevant (for computation of the integrals), but theoretically unnecessary for the existence of the posterior (the first equality).

Furthermore, note that \( \vartheta \) was assumed to be a random vector (i.e. dimension one) so far, but generalizations to random vector \( \vartheta \in \Theta \subseteq \mathbb{R}^k \) are readily made.

From proposition A.2 it becomes clear that for dominated models of conditional i.i.d. data, the likelihood plays a central role in Bayesian statistics, since by definition A.1, we have \( \prod_{i=1}^n f_\vartheta(y_i) = L(\vartheta|y) \).

Definition A.3 (Credible Set). Assume we have a Bayesian procedure with posterior distribution \( \Pi(\vartheta|y) \) after observing \( y \in \mathbb{R}^n \) and \( k \)-dimensional random parameter vector \( \vartheta \). Let \( \alpha \in (0,1) \). The level-\( \alpha \) credible set \( \mathcal{G} \subseteq \mathbb{R}^n \) is defined such that:

\[
\Pi(G|y) \geq 1 - \alpha
\]

Besides Bayesian inference with respect to the random vector of parameters, one may also be interested in prediction of the outcome variable \( Y \).

Definition A.4 (Posterior Predictive Distribution). Similar to proposition A.2, let \( y \in \mathbb{R}^n \) be a vector of observations and let \( \Pi(\vartheta|y) \) denote the posterior distribution of random variable \( \vartheta \). Assume \( Y \) has a conditionally i.i.d. model. For \( m \geq 1 \), the posterior predictive distribution of the next \( m \) points of \( Y \) is defined as:

\[
\hat{\Pi}^Y_m(B_1, ..., B_m | y) := \int_{\Theta} (\prod_{i=1}^m \Pi(B_i|\vartheta)) \; d\Pi(\vartheta|y) \quad \forall B \in \mathcal{B}(\mathbb{R}^n)
\]

Definition A.5 (Prediction Set). Similar to Credible Set, the prediction set or prediction interval (in case \( m = 1 \)) is defined with respect to the posterior predictive distribution. Assume we have a Bayesian procedure with conditional i.i.d. data. Let \( \alpha \in (0,1) \). A set \( G \subseteq \mathcal{B}(\mathbb{R}^n) \) is called a level-\( \alpha \) (Bayesian) predictive set for \( \hat{Y} = (Y_{n+1}, ..., Y_{n+m}) \) on \((\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}^Y)\) if:

\[
\hat{\Pi}^Y_m(G|y) \geq 1 - \alpha
\]

A.2 Basic definitions

Definition A.6 (Stochastic Process). Let \( X := \{ X_t : t \in \mathcal{T} \} \) be a real-valued random vector with each \( X_t \) from the same probability space \((\Omega, \mathcal{F}, \mathbb{P})\) onto measure space \((\mathbb{R}, \mathcal{B})\). \( X \) is called a stochastic process with state space \((\mathbb{R}, \mathcal{B})\).

Throughout this thesis, \( \mathcal{T} \) is considered to be of finite cardinality, for instance a discrete time process.

Definition A.7 (Taylor polynomial of order \( p \) in dimension \( d \)). The following is taken from definition 1 in Raudenbush, Yang, and Yosef (2000) (stemming from Magnus and Neudecker (1988), slightly extended and adapted in notation. Let \( h(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R} \) have continuous partial derivatives of order \( p \) in a neighbourhood \( N_X \) of \( \tilde{v} \). Then, \( \forall \tilde{v} \in N_X \) we define the Taylor polynomial of order \( p \) as follows:

\[
\hat{h}_p(\tilde{v}) := h(\tilde{v}) + h^{(1)}(\tilde{v})(\tilde{v} - \tilde{v}) + \sum_{k=2}^{p} \frac{1}{k!} \left[ \otimes (\tilde{v} - \tilde{v})^T \right] h^{(k)}(\tilde{v})(\tilde{v} - \tilde{v})
\]

Here, \( h^{(k)}(\tilde{v}) := \frac{\delta h^{(k-1)}(v)}{\delta v^j} \bigg|_{v=\tilde{v}} \forall k \in [1, p] \), with vec \( h^{(k-1)}(v) \) the half-vectorization to obtain the lower-triangular matrix of partial derivatives (e.g. all unique elements of the Jacobian for \( k = 1 \) and the Hessian for \( k = 2 \)) in column vector form. Furthermore, \( \otimes u := u \prod_{i=1}^{k-1}(\otimes u) \), i.e. vector \( u \) being multiplied \( k \) times using Kronecker products.
Raudenbush, Yang, and Yosef (2000, Definition 1.) state that \( h(v) = \hat{h}(v) \) for \( v \in N_g \), although it is not true that the continuity of (in this case all) partial derivatives in the neighbourhood \( N_g \) is a sufficient condition for this to hold for general \( h(\cdot) \). Assuming the \( h(\cdot) \) under study to be locally \( (N_g) \) defined by a convergent power series (i.e. analytic) would be sufficient, but verifying this for general dimension \( d \) is out of scope.

**Definition A.8 (Kullback-Leibler divergence).** Kullback and Leibler (1951) states the following divergence (it is not a formal distance since it is asymmetric) between approximation density \( q \) and truth \( p \)

\[
KL(p||q) = \int p(x) \log_e \left( \frac{p(x)}{q(x)} \right) \, dx
\]

Where it must be provided that \( q(x) = 0 \) coincides with \( p(x) = 0 \), in which case the contribution is defined to be zero.
Appendix II: Methodology and Computational Statistics

B.1 Posterior distribution multivariate normal conjugate update

This appendix verifies the mean and variance expression of the posterior distributions in step 2 and 3 of the Gibbs sampling scheme (subsection 4.3.3). Only a proof of the result in step 3 is given, step 2 is similar.

\[
\pi(\beta|\theta, X, \mu_\beta, \Sigma_\beta, \Psi_\theta) \propto \pi(\theta|\beta, X, \gamma, \nu, \Psi_\theta) \times \pi(\beta|\mu_\beta, \Sigma_\beta)
\]

\[
\propto \exp((\theta - X\beta)'\Psi_\theta^{-1}((\theta - X\beta)) \times \exp((\beta - \mu_\beta)'\Sigma_\beta^{-1}(\beta - \mu_\beta))
\]

Then, focusing on the matrix algebra working towards a proportional quadratic form for \(\beta\) and dropping subscripts:

\[
(\theta - X\beta)'\Psi_\theta^{-1}(\theta - X\beta) + (\beta - \mu_\beta)'\Sigma_\beta^{-1}(\beta - \mu_\beta)
\]

\[
= \theta'\Psi^{-1}\theta + \beta'X'\Psi^{-1}X\beta - \theta'\Psi^{-1}X\beta + \beta'\Sigma_\beta^{-1}\beta - \mu_\beta'\Sigma_\beta^{-1}\mu + \mu_\beta'\Sigma_\beta^{-1}\mu
\]

\[
\propto \beta'X'\Psi^{-1}X\beta - \theta'\Psi^{-1}X\beta + \beta'\Sigma_\beta^{-1}\beta - \mu_\beta'\Sigma_\beta^{-1}\mu
\]

\[
:= (\beta - \bar{\mu})'\Sigma^{-1}(\beta - \bar{\mu}) + \bar{\mu}'\Sigma^{-1}\bar{\mu}
\]

with \(\Sigma = (X'\Psi^{-1}X + \Sigma^{-1})^{-1}\)

\[
\propto (\beta - \bar{\mu})'\Sigma^{-1}(\beta - \bar{\mu})
\]

Here, the \(\Sigma^{-1}\) substitution is clear from the expansion by looking at the terms that involve \(\beta\) twice. Then \(\bar{\mu}\) immediately follows from an implied equivalence:

\[
\Rightarrow \beta'\Sigma^{-1}\bar{\mu} = \beta'X'\Psi^{-1}\theta + \beta'\Sigma^{-1}\mu
\]

\[
\Leftrightarrow \bar{\mu} = \Sigma(X'\Psi^{-1}\theta + \Sigma^{-1}\mu)
\]

\[
= (X'\Psi^{-1}X + \Sigma^{-1})^{-1}(X'\Psi^{-1}\theta + \Sigma^{-1}\mu)
\]

Summarizing, we find:

\[
\pi(\beta|\theta, X, \mu_\beta, \Sigma_\beta, \Psi_\theta) \propto \exp((\beta - \bar{\mu})'\Sigma^{-1}(\beta - \bar{\mu}))
\]

Which specifies the multivariate normal distribution \(N_q(\bar{\mu}, \Sigma)\)

B.2 Mixture normal approximation to logistic density

On a grid of \(1e^5\) equally spaced \(x\) values on \([-10, 10]\), the logistic density is evaluated using R function \(dlogis()\) in the Stats package and compared to several approximation candidates. The distance between the logistic and the approximative density is measured as the L2 norm and referred to as "Loss". Some researchers (e.g. Savalei, 2006, with specific attention to IRT vs Normal Ogive) may prefer other "measures", e.g. the Kullback-Leibler divergence (Kullback and Leibler, 1951, cf. appendix A.2). The results that follow were found to be indistinguishable between these loss functions.\(^1\)

The approximations are based on \(k\) normal mixtures with weight parameters \(w\) and standard deviations \(sd\). Constrained Nelder-Mead optimization (Nelder and Mead, 1965) was used to find parameters that minimize the loss and are listed in table 13.\(^2\)

\(^1\)For high-dimensions, the KL estimates and L2-loss estimates are very different, which is caused by the high degree of freedom in combining weight and scale parameters. When evaluated in their mixture however, the resulting densities are indistinguishable to the eye (we tested this for \(k = 5\)).

\(^2\)Constraints were such that weights are positive and sum to one. Standard deviations were optimized on the log scale (presented after exponential transformation) to guarantee non-negativity.
Table 13: Optimal mixing weights and sd parameters for Gaussian approximations.

<table>
<thead>
<tr>
<th>Mixing Weights</th>
<th>St. deviations</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w1</td>
<td>w2</td>
<td>w3</td>
<td>w4</td>
<td>w5</td>
<td>sd1</td>
<td>sd2</td>
<td>sd3</td>
<td>sd4</td>
<td>sd5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=1</td>
<td>1.000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.644</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.2689</td>
</tr>
<tr>
<td>k=2</td>
<td>0.458</td>
<td>0.539</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.230</td>
<td>2.133</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1124</td>
</tr>
<tr>
<td>k=3</td>
<td>0.189</td>
<td>0.235</td>
<td>0.577</td>
<td>-</td>
<td>-</td>
<td>1.052</td>
<td>2.563</td>
<td>1.621</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0127</td>
</tr>
<tr>
<td>k=4</td>
<td>0.088</td>
<td>0.086</td>
<td>0.423</td>
<td>0.403</td>
<td>-</td>
<td>0.955</td>
<td>2.971</td>
<td>1.383</td>
<td>2.014</td>
<td>-</td>
<td>-</td>
<td>0.0018</td>
</tr>
<tr>
<td>k=5</td>
<td>0.436</td>
<td>0.377</td>
<td>0.017</td>
<td>0.109</td>
<td>0.061</td>
<td>1.935</td>
<td>1.337</td>
<td>1.107</td>
<td>2.879</td>
<td>0.920</td>
<td>-</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Based on these optimal parameters, the following comparisons are made as illustrated in figure 21.

As noted by Albert and Chib (1993), a t-distribution with 8 degrees of freedom, labelled “t8”, is (almost) linear in the qq-plot (figure 21 L) and after scaling an accurate approximation. Holmes and Held (2006) remark that by extending the probability range to 0.0001-0.9999, the t8 accuracy is lower than it appears in Albert and Chib (1993). Without scaling it is rather distant from both the density as well as the distribution function (figure 21 R). Furthermore, the t-distribution does not have the conjugacy advantage, for which we prefer to use a normal mixture approximation.

Remark, $k = 4$ is rather close in loss with respect to $k = 5$. Depending on the probability range of interest (we have taken 0.0001-0.9999), the computational cost of $k = 5$ over $k = 4$ is met by an increase in accuracy of the tail. For comparison, note that $k = 3$ seems to do well in the qq-plot domain [-5,5], which corresponds to logistic cumulative probabilities of 0.007 respectively 0.993. We choose to be on the safe side and conclude that $k = 5$ is an approximation with a decent accuracy to cost trade-off.

B.3 Sampling from a truncated normal

By conditional independence the covariance matrix of $e_i$ is $s \cdot I_{|\tau_i|}$ and we can sample from it by sampling the elements in $Z_{\tau_i}$ one-by-one from univariate truncated normals. A straightforward approach would be to sample from a normal distribution with the appropriate mean and variance until a positive respectively negative value is obtained (the truncation point is always 0). However, this may take a considerable amount of time when the truncation point is several standard deviations to the “wrong” side of the mean. Therefore, we use an algorithm by Robert (1995) based on acceptance-rejection, which increases the probability of acceptance at the first draw significantly.
In this appendix the acceptance-rejection algorithm of Robert (1995) based on his optimality parameter $\alpha^*$ is restated and briefly discussed for the case of left truncation at 0 with mean $m = \theta_{it} - \delta_j$ and standard deviation $s$.

**Algorithm 3**: Algorithm to sample from N($m, s$) left truncated at 0.

Set $\mu = 0$ and $\mu^- = -m/s$ ;

if $\mu^- < 0$ then

<table>
<thead>
<tr>
<th>Repetitive sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1A Sample $z$ from N(0,1) ;</td>
</tr>
<tr>
<td>Step 2A if $z &gt; \mu^-$ then</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

end

else

<table>
<thead>
<tr>
<th>Robert (1995)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1B Compute $\alpha^* = \frac{\mu^- + \sqrt{\mu^-^2 + 4}}{2}$ ;</td>
</tr>
<tr>
<td>Step 2B Sample $z$ from Exp($\alpha^*, \mu^-$) (cf. proposition B.1) ;</td>
</tr>
<tr>
<td>Step 3B Compute $\rho = \exp(-\frac{(z - \alpha^*)^2}{2})$ ;</td>
</tr>
<tr>
<td>Step 4B Sample $u \sim U(0, 1)$ ;</td>
</tr>
<tr>
<td>Step 5B if $u \leq \rho$ then</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

end

return $z \cdot s + m$

Since this is not included in Robert (1995), we provide some detail on step 2B.

**Proposition B.1** (Sampling from Exp($\alpha^*, \mu^-$)). The density of a random variable $X \sim \text{Exp}(\alpha^*, \mu^-)$ is defined as:

$$g(x; \alpha^*, \mu^-) = \alpha^* \exp(-\alpha^*(x - \mu^-)) I_{x \geq \mu^-}$$

We use the Inverse Transform Method to sample a realization of $X$, which we call $z$.

Sample $u \sim U(0, 1)$. Then, by Probability Integral Transform it holds:

$$u = \int_{\mu^-}^{z} g(x; \alpha^*, \mu^-) dx = \int_{\mu^-}^{z} g(x; \alpha^*, \mu^-) dx$$

$$\Leftrightarrow z = \mu^- - \frac{\ln(1-u)}{\alpha^*}$$

The rationale of Algorithm 3 is that as the left truncation point tends further to the right of the mean, the density tends to that of a translated exponential. The algorithm is efficient in the sense that when simulation by repetitive normal sampling becomes more difficult, the acceptance probability of Robert’s algorithm increases. Robert’s part has an acceptance probability at the first draw of at least 76%. The repetitive part accepts it’s first draw in at least 50% of the cases (Robert, 1995).

Remark that this procedure also works for the case $Y = 0$ by sampling from the left truncated distribution and multiplying the result by minus 1 to arrive at a sample from the right truncated distribution.
B.4 LMS Likelihood details

This appendix contains worked out details for the LMS likelihood equation in Cole and Green (1992).

We adopt similar notation as in Cole and Green (1992). Consider an observed outcome vector \( y \) on \( t \) in which the dependence on \( t \) is suppressed for ease of notation. Instead, the subindex \( i \) points to the observations \( z \) is standard normal distributed. Most notably, \( \forall t \) it is supposed that the power transformation \( y^{L(t)} \) or, if \( L(t) = 0 \) then \( \log_e(y) \), is normally distributed; the standardization that follows is rather trivial.

Provided \( Z \) is standard normal and given some \( t \), empirical quantiles (or "centiles" in Cole and Green (1992)) on \( y(t) \) can be computed. Let \( \alpha \in [0,1] \) be such that 100\( \alpha \)% is the quantile of interest, defined \( C_{100\alpha}(t) := \inf\{v : P[y(t) \leq v] \leq \alpha \} \). Similarly, the theoretical quantile of the standard normal as function of \( \alpha \) is defined as: \( \Phi^{-1}(\alpha) := \inf\{z : \Phi(z) \leq \alpha\} \)

For \( L(t) \neq 0\):

\[
\alpha = P[Z \leq \Phi^{-1}(\alpha)] = P[\left(\frac{(\frac{y(t)}{M(t)})^{L(t)}}{S(t) \cdot L(t)}\cdot L(t)\right) \leq \Phi^{-1}(\alpha)] \\
= P[\left(\frac{y(t)}{M(t)}\right)^{L(t)} \leq 1 + L(t)S(t)\Phi^{-1}(\alpha)] \\
= P[y(t) \leq M(t)(1 + L(t)S(t)\Phi^{-1}(\alpha))^{1/L(t)}] \\
\Rightarrow \\
C_{100\alpha}(t) = M(t) \cdot (1 + L(t)S(t)\Phi^{-1}(\alpha))^{1/L(t)}
\]

For \( L(t) = 0 \):

\[
\alpha = P[Z \leq \Phi^{-1}(\alpha)] = P[\log_e(\frac{y(t)}{M(t)}) \leq \Phi^{-1}(\alpha)] \\
= P[y(t) \leq M(t) \exp(S(t)\Phi^{-1}(\alpha))] \\
\Rightarrow \\
C_{100\alpha}(t) = M(t) \exp(S(t)\Phi^{-1}(\alpha))
\]

Remark that given \( t \), the \( z(t) \) are assumed independent (and identically distributed). The LMS method further assumes that this independence holds across \( t \) as well, arriving at the following likelihood expression, in which the dependence on \( t \) is suppressed for ease of notation. Instead, the subindex \( i \) points to the
specific \((t\text{-dependent})\) value for \(L, M, \) and \(S\).

\[
L(L, M, S; y) = \prod_{i=1}^{n} P[Y = y_i] = \prod_{i=1}^{n} \frac{dP[Y \leq y_i]}{dy} = \prod_{\{i: (i \in [1, n] \cap L \neq 0)\}} \frac{dP[Z \leq \frac{(\frac{y_i}{M_i})^{L_i} - 1}{L_i S_i}]}{dy} \prod_{\{i: (i \in [1, n] \cap L = 0)\}} \frac{1}{y_i S_i} \cdot \phi[Z \leq \frac{\log_e(\frac{y_i}{M_i})}{S_i}]
\]

Taking the natural logarithm and excluding constants to express dependence on \(L, M, \) and \(S\) only, we retrieve Cole and Green (1992) equation (5):

\[
I(L, M, S; y) \propto \sum_{\{i: (i \in [1, n] \cap L \neq 0)\}} \left[ -\frac{1}{2} z_i^2 + L_i \log_e(\frac{y_i}{M_i}) - \log_e(S_i) \right] + \sum_{\{i: (i \in [1, n] \cap L = 0)\}} \left[ -\frac{1}{2} z_i^2 - \log_e(S_i) \right]
\]

To find \(L, M, \) and \(S\), Cole and Green (1992) optimize a penalized likelihood using smoothness parameters \(\alpha_{\lambda}, \alpha_{\mu}, \) and \(\alpha_{\sigma}\). This expression (Cole and Green (1992) eq. (6)) is:

\[
I(L, M, S; y) - \frac{1}{2} \alpha_{\lambda} \int L''(t)^2 \, dt - \frac{1}{2} \alpha_{\mu} \int M''(t)^2 \, dt - \frac{1}{2} \alpha_{\sigma} \int S''(t)^2 \, dt
\]
C Appendix Part III: Application D-score TNO

C.1 Elapsed time

Table of elapsed time in hours per set-up. In all set-ups, $M_2$ of subsection 8.2 was used. Most labels are self-explanatory, but we mention a few. $w_0 = N$ implies the $w_0$ optimization is switched off (i.e. $w_0$ fixed at initial value). Link $P$ is short for probit, $L$ for logit. Output concerns trace plot updates of every inner loop in the R console, which may make the overall procedure slightly slower.

<table>
<thead>
<tr>
<th>data</th>
<th>q</th>
<th>covariates</th>
<th>w0</th>
<th>Link</th>
<th>Output</th>
<th>Time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim</td>
<td>2</td>
<td>N</td>
<td>Y</td>
<td>P</td>
<td>Y</td>
<td>2.88</td>
</tr>
<tr>
<td>SMOCC</td>
<td>3</td>
<td>Y</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>3.73</td>
</tr>
<tr>
<td>SMOCC</td>
<td>3</td>
<td>N</td>
<td>Y</td>
<td>P</td>
<td>Y</td>
<td>3.74</td>
</tr>
<tr>
<td>SMOCC</td>
<td>3</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>2.05</td>
</tr>
<tr>
<td>SMOCC</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>3.64</td>
</tr>
<tr>
<td>SMOCC</td>
<td>2</td>
<td>N</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>3.69</td>
</tr>
<tr>
<td>SMOCC</td>
<td>2</td>
<td>N</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>2.01</td>
</tr>
<tr>
<td>SMOCC</td>
<td>2</td>
<td>N</td>
<td>Y</td>
<td>L</td>
<td>N</td>
<td>6.24</td>
</tr>
</tbody>
</table>

All computation times are strongly reduced by use of C++ in R through the Rcpp package (Eddelbuettel and François, 2011; Bates and Eddelbuettel, 2013).

C.2 Reproduction prior mcurve Van Buuren (2014)

The EAP Dscore calculation (definition 7.2) uses a prior $N(\mu_t, 5)$ distribution on abilities, where $\mu_t$ is a point on a curve defined as the mcurve. Van Buuren (2014) uses an iterative procedure to determine the mcurve, which is taken to be the median from the LMS technique (appendix B.4).

In the cross-validation of subsection 10.2, the mcurve must be calibrated to the training data. Since the resulting Dscores on the test set only depend on the training set through this mcurve, a faster and easier method circumvents the iterative Dscore and LMS estimation. For completeness, we mention that in this cross-validation it is assumed that the stage one abilities and difficulties ($\theta_\kappa : \kappa \in [1, N_\kappa]$ and $\delta$) remain unchanged after excluding the test set.

**Definition C.1** (New mcurve). Let $\theta_{TR}$ denote all $\theta_\kappa$ for which index $\kappa$ belongs to the subjects included in the training set, where the notation $\theta_\kappa$ was introduced in section 7. Let $\tau$ denote the vector of respective times (i.e. ages in weeks) to which the elements in $\theta_{TR}$ belong. We define the new prior mean as follows:

\[
\mu_t := \begin{bmatrix} 1 & \log(t) & \log(t)^2 & \log(t)^3 \end{bmatrix} \hat{\alpha} \quad t \in [2, 145]
\]

Where $\hat{\alpha} \in \mathbb{R}^4$ is the solution to the following linear regression model with logarithmically transformed time variables:

\[
\theta_{TR} = a_0 1 + a_1 \log(\tau) + a_2 \log(\tau)^2 + a_3 \log(\tau)^3 + \epsilon \quad \epsilon \sim N_{|TR|}(0, \sigma^2 I)
\]

Using the full dataset of $n = 2038$ subjects (Van Buuren, 2014), the original and reproduced mcurves can be compared as shown in figure 22. For consistency with earlier work on mcurve estimation, weeks are used as input time unit, but months are used for plotting. A correlation of 0.99996 between the two curves is found at weekly $\mu_t$ values selected between week 2 and 145 (analogous to earlier analyses Van Buuren (2014)).
C.3 Summary of reference procedure \( q = 2, \) without covariates, \( w_0 = 0 \)

Table 15: Parameter estimates and credible sets (\( \alpha = 0.05 \)) for \( q = 2 \) without covariates and \( w_0 = 0 \).

<table>
<thead>
<tr>
<th>Par.</th>
<th>CI low</th>
<th>Mean</th>
<th>Median</th>
<th>CI high</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>10.73</td>
<td>10.90</td>
<td>10.91</td>
<td>11.08</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>48.62</td>
<td>48.96</td>
<td>48.96</td>
<td>49.30</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-11.88</td>
<td>-11.73</td>
<td>-11.73</td>
<td>-11.58</td>
</tr>
</tbody>
</table>

Figure 23: Figure panel trace plots \( q = 2 \) without covariates and \( w_0 = 0 \).
Figure 24: Figure panel outer loop trace plots $\Sigma_k$ for $q = 2$ without covariates and $w_0 = 0$.

Figure 25: Figure panel outer loop trace plots $(v_0, w_0)$ for $q = 2$ without covariates and $w_0 = 0$. 