Analysis of Test Result Data with non-Random Missingness

Master's Thesis

Mathematics, 3rd and 4th semester

Aalborg University

Date: 31.05.2022

AALBORG UNIVERSITET
STUDENTERRAPPORT

3rd and 4th semester at
Department of Mathematical Sciences
Skjernvej 4A
9220 Aalborg Ø
http://www.math.aau.dk

Titel:

Analysis of Test Result Data with non-Random Missingness

Themes:

Missing Data

Item Response Theory

Generalized Linear Mixed Models

Semester:

3rd and 4th Semester

Project Period :

01.09.2021 - 31.05.2022

ECTS:

50 credits

Supervisors:

Rasmus Waagepetersen

Group Members:

Mikkel Rúnason Simonsen

Printing Number: 3

Number of Pages: 149

Number of Attachments: 0

Abstract:

The main focus of the report is statistical inference based on test result data. In particular, the data generating model for the full data is assumed to be the Rasch model and missingness is modelled using the steps model for the dropout.
Within the framework of the Rasch model, several parameter estimators are considered, including the joint, conditional and marginal maximum likelihood estimators. The asymptotic properties of the estimators are derived and verified for practical use through simulation studies.
It is concluded that modelling the dropout is essential as there is found correlation between when a subject experiences dropout and the ability of said subject.

By approving submission in Digital Eksamen, each group member accepts that everyone has participated equally in the project work and that the group is collectively responsible for the contents of the report.

Preface

This report is developed in the fall of 2021 and spring of 2022 as a long master's thesis by a MSc student who studies mathematics at the Department of Mathematical Sciences at Aalborg university.

The main focus of the report is to conduct statistical inference and in particular parameter estimation for the Rasch model based on a dataset referred to in the report as the *test results data*. For this purpose theory regarding the Rasch model, generalized linear mixed models and missing data is presented.

To read the report it is suggested that the reader has a basic knowledge within the fields of probability theory and statistics.

In the report citations are denoted with [number] associated to a publication in the bibliography and when referring to a table, figure, definition or theorem the corresponding numbering is written in the text. For equations there are parentheses around the corresponding number.

I would like to thank my supervisor Rasmus Waagepetersen for good and appreciated supervision.

Contents

1 Introduction

The purpose of this report is to explore the *test results data* which is a dataset consisting of test responses of 663 students from the danish public school. The students range from 10 to 12 years old and originates from 19 different schools across the country. Each of the students where given the same test consisting of 36 questions regarding fractions, and they only had a limited time to respond to the questions, i.e. it was a *speeded test*. The data were kindly provided by Associate Professor Pernille Ladegaard Pedersen, Via University College, Aarhus.

Specifically, the test results data consists of the age, gender, class and school of each student, as well as a binary response pattern y_i indicating which questions were solved correctly. The data is structured such that $y_{ij} = 1$ implies that the ith subject correctly solved the jth item.

As the analysis of this dataset will be conducted using *item response theory* (IRT) models, in particular the Rasch model, the students will be referred to as *subjects* and the questions will be referred to as *items* for the rest of the report, in accordance with the IRT litterature.

Another important feature of the test results data is the fact that some of the data is missing. This is due to the subjects not responding to certain items for various reasons, for instance if a subject decided to skip a question deemed too difficult or if the subject ran out of time.

The number of missing responses for each item can be found in Figure 1.1.

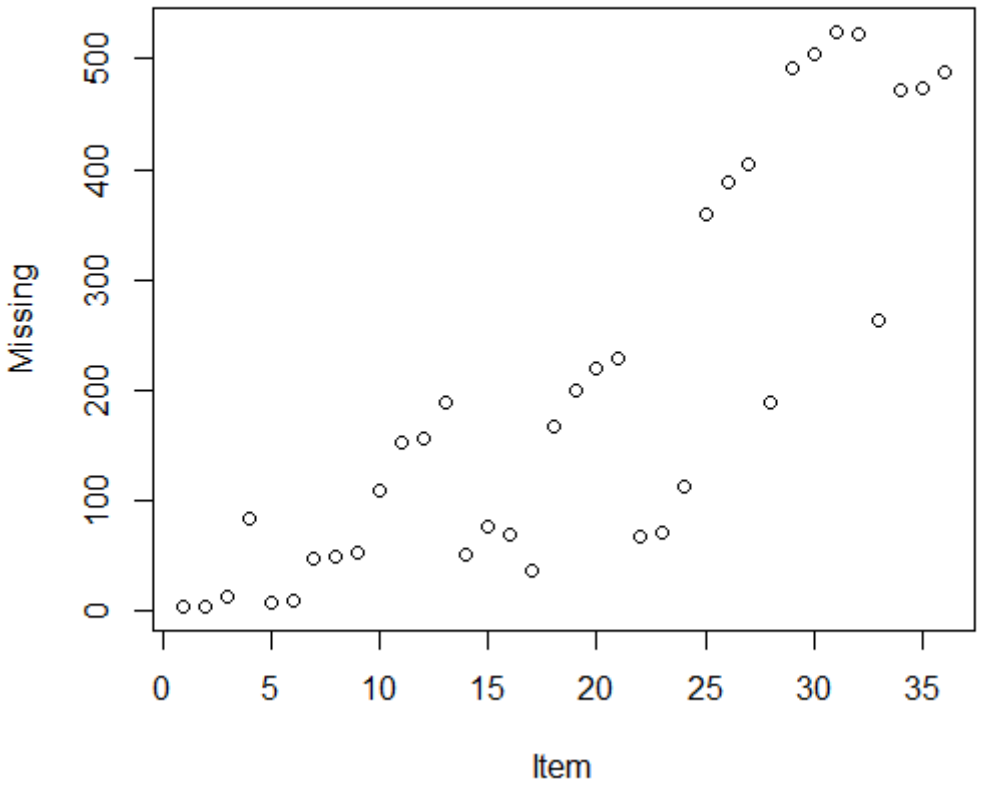

Figure 1.1: Figure showing the number of missing responses for each item.

It is clear from the figure that there is a tendency for more missingness to occur at higher item numbers.

It will be a working assumption throughout the report that the subjects solve the items in enumerated order and hence the large amount of missing responses among the last items is due to the time constraint, i.e. the last items where not answered because the subject ran out of time. This phenomenon is known as dropout and will be the central focus point of the modelling of the missing data mechanism in the report.

One could argue that, usually when a teacher grades a student, a missing response would simply be considered as an incorrect response, and hence the analysis could be conducted by changing all missing responses to incorrect responses and then conduct analysis on an augmented dataset with no missingness.

The problem with this is that if e.g. a subject runs out of time before answering the last items, then it is unknown whether or not the subject could answer them correctly. Simply assuming that the subject wouldn't have been able to answer them correctly would result in perceiving the subject as worse than he is.

On the other hand, ignoring the missing data and only consider analysis based on the observed responses is clearly also problematic, since one would throw away the information regarding the ability of the subject contained in the missing responses. It might for instance be the case that there is correlation between an early dropout and a low level of subject ability.

Hence it is clear that the test results data contains an interesting missing data problem which somehow needs to be modelled in order to conduct a data analysis.

In Chapter 2 item response theory will be presented with a particular focus on the Rasch model considered as a generalized linear model. Furthermore, theory regarding parameter estimation in the Rasch model when considered as a generalized linear model will be presented in Chapter 3. In Chapter 4, theory regarding generalized linear mixed models in general, and in particular methods for computing the likelihood of said models are presented. This framework is then used to reformulate the Rasch model as a generalized linear mixed model by introducing random effects. In Chapter 5, the general framework of missing data is presented and maximum likelihood estimation under the assumption of an ignorable missingness mechanism is considered with a focus on the EM algorithm. The chapter is concluded by considering several models for the missing data mechanism present in the test results dataset. Analysis of the test results data is then conducted in Chapter 6 first by ignoring the missingness mechanism using a complete cases approach and then by modelling the missing data mechanism using one of the models described in Chapter 5 .

2 | The Rasch Model

In this chapter the Rasch model is presented based on the general framework of item response theory. Futhermore, the Rasch model is also considered as a *generalized linear model* (GLM) as it is a special case of logstic regression. Therefore, for readers unfamiliar with GLM's it is advised to read Appendix A which is a supplement to the chapter presenting relevant definitions of the topic and introduces the logistic regression as a GLM.

2.1 Presentation of the Rasch Model

In this section the Rasch model will be presented and important properties of the model will be discussed. The section on based on [10][Chapter 1 & 2] and [13]. The *Rasch model* was first introduced in 1960 by the Danish mathematician Georg Rasch in his book "*Probabilistic Models for Some Intelligence and Attainment Tests*" and is an *item response theory* (IRT) model. In IRT subjects respond to a series of items all meant to measure one or more latent traits of said subject, where it is observed whether the subject solves the item or not. Usually *unidimensionality* is assumed, such that only one latent trait, often referred to as the *ability of the subject*, is measured. In the report this will be referred to as the *subject parameter*. Furthermore, *local stochastic independence* is assumed, such that the subjects answer the items independently of each other, often referred to as no cheating, and answer each item independently from the other items given the latent trait(s) of said subject, often referred to as no learning. The central idea of IRT is that the probability of a subject with certain latent trait(s) solves a given item can be modelled through a simple function depending on the latent trait(s) and one or more parameters characterizing the item. Therefore, associated to each item is an *item characteristic curve* (ICC) which gives the probability of solving the item as a function of the latent trait(s).

Let $n \in \mathbb{N}$ denote the number of subjects and $p \in \mathbb{N}$ denote the number of items.

The Rasch model is a unidimensional IRT model where the ICC of the jth item is given by

$$f_j(\theta) = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)}, \tag{2.1}$$

such that β_j is the only parameter characterizing the jth item and will be referred to as the *difficulty of the item* or the *item parameter* for $j = 1, \ldots, p$. Note in particular that the ICC is strictly increasing, which is intuitively clear, since a subject with greater abilities should have a higher probability of solving a given item.

Furthermore, it should be noted that the Rasch model is also a logistic regression model with linear predictor $\eta_{ij} = \theta_i - \beta_j$ since, by (2.1), it follows that

$$p_{ij} = P\left(Y_{ij} = 1\right) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{2.2}$$

where the response $Y_{ij} = 1$ is interpreted as the ith subject solving the jth item for $i = 1, \ldots, n, j = 1, \ldots, p$.

Therefore, for $y_{ij} \in \{0, 1\}$,

$$\begin{aligned} p(y_{ij}; \theta_i, \beta_j) &:= P(Y_{ij} = y_{ij}) \\ &= \left(\frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}\right)^{y_{ij}} \left(1 - \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}\right)^{1 - y_{ij}} \\ &= \left(\frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}\right)^{y_{ij}} \left(\frac{1}{1 + \exp(\theta_i - \beta_j)}\right)^{1 - y_{ij}} \\ &= \frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)}. \end{aligned} \tag{2.3}$$

Furthermore, as the Rasch model is a logistic regression it is natural to consider the odds. In particular, the odds that the ith subject solves the jth item is by Equation (A.3) given as

$$o_{ij} = \exp(\theta_i - \beta_j)$$

and the odds ratio between the ith and i'th subject to solve the jth item is by Equation (A.4) given as

$$\frac{o_{ij}}{o_{i'j}} = \exp(\theta_i - \theta_{i'})$$

for $i, i' = 1 \ldots, n, j = 1, \ldots, p$.

Let $Y_{i+} = \sum_{j=1}^{p} Y_{ij}$ and $Y_{+j} = \sum_{i=1}^{n} Y_{ij}$ for $i = 1, \ldots, n, j = 1, \ldots, p$. For the ith subject, Y_{i+} is the random number of solved items and will be referred to as the *score of the ith subject*. Similarly Y_{+j} is the random number of subjects that solved the jth item and is called the *score of the jth item*. These scores will be of great importance in what will follow, primarily because they are sufficient statistics of their respective subject and item parameters. In other words, all the information contained in a given dataset regarding e.g. the ability of a subject is contained in the score of that subject. It does not matter which items or the difficulty of said items the subject solved, only the number of items.

When the score of a subject is either 0, meaning that the subject did not solve any items, or p, meaning that the subject solved all items, the score will be called an *extreme score*. Similarly, item scores of 0 or n will be called extreme scores. These extreme scores will later be seen to be problematic which is intuitively clear: how could e.g. the ability of a subject be estimated if said subject has not solved any items and thus not demonstrated any abilities at all?

It will now be shown that the sufficiency of the subject score is somewhat unique to the Rasch model.

Theorem 2.1.1. Sufficient Statistics and the Rasch Model

Consider a unidimensional IRT model such that

- For the jth item the corresponding ICC $g_j : \mathbb{R} \to (0, 1)$ is continuous and strictly increasing and satisfies that $\lim_{\xi \to -\infty} g_j(\xi) = 0, \lim_{\xi \to \infty} g_j(\xi) = 1$ for $j = 1, \ldots, p$,
- Local stochastic independence is satisfied, that is, for the ith subject

$$P(Y_{i1} = y_{i1}, \cdots, Y_{ip} = y_{ip}) = \prod_{j=1}^{p} P(Y_{ij} = y_{ij}) = \prod_{j=1}^{p} g_j(\xi_i)^{y_{ij}}(1 - g_j(\xi_i)^{y_{ij}})^{1-y_{ij}}$$

for $i = 1, \ldots, n$.

If the subject score Y_{i+} is a sufficient statistic for the corresponding subject parameter ξ_i then the item response model is equivalent to the Rasch model.

Proof. It follows by the assumption of sufficiency that the conditional probability of obtaining response pattern $y \in \Gamma(p) = \{0,1\}^p$, where $\Gamma(p)$ is the set of possible response patterns, given subject score r does not depend on the subject parameter ξ, i.e.

$$p(y \mid r) = \frac{p(y;\xi)}{p(r;\xi)} = c_r(y)$$

where $c_r(y)$ denotes a constant given r and y and does in particular not depend on ξ . Here, it should be noted that the conditional density of the response pattern given the subject score, the density of the response pattern and the density of the subject score are all denoted by p without the use of subscripts. This is done in order to simplify notation and it should be clear from the input which density is in question.

Let y be a response pattern with corresponding subject score $r = \sum_{j=1}^{p} y_j$ such that $y_1 = 1$ and $y_j = 0$ for some $j \in \{2, \ldots, p\}$.

Furthermore, define y' such that $y'_1 = 0, y'_j = 1$ and $y'_{j'} = y_{j'}$ for $j' = 2, \ldots j-1, j+1, \ldots, p$ and hence $\sum_{j'=1}^{p} y'_{j'} = \sum_{j'=1}^{p} y_{j'} = r$.

Let $y^{-(1,j)} = (y_2, \ldots y_{j-1}, y_{j+1}, \ldots, y_p)$ such that $y^{-(1,j)} = y'^{-(1,j)}$.

Since local stochastic independence implies that

$$\frac{p(y^{-(1,j)};\xi)g_1(\xi)(1-g_j(\xi))}{p(r;\xi)} = p(y \mid r) = c_r(y), \quad \frac{p(y'^{-(1,j)};\xi)(1-g_1(\xi))g_j(\xi)}{p(r;\xi)} = p(y' \mid r) = c_r(y')$$

it follows that

$$\begin{aligned} \frac{c_r(y)}{c_r(y')} &= \frac{p(y^{-(1,j)};\xi)g_1(\xi)(1-g_j(\xi))}{p(y'^{-(1,j)};\xi)(1-g_1(\xi))g_j(\xi)} \\ &= \frac{g_1(\xi)(1-g_j(\xi))}{(1-g_1(\xi))g_j(\xi)}. \end{aligned} \tag{2.4}$$

Since $g_1(\xi)$ is assumed to be strictly increasing there exists a strictly monotone function $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(\xi) = \text{logit}(g_1(\xi))$ such that

$$g_1(\xi) = \frac{\exp(\phi(\xi))}{1+\exp(\phi(\xi))} =: f_1(\phi(\xi)).$$

Furthermore, consider $f_j : \mathbb{R} \to \mathbb{R}$ such that $f_j(\phi(\xi)) = g_j(\xi)$. Then, by insertion into Equation (2.4) it follows that

$$\begin{aligned}\frac{c_r(y)}{c_r(y')} &= \frac{f_1(\phi(\xi))(1 - f_j(\phi(\xi)))}{(1 - f_1(\phi(\xi)))f_j(\phi(\xi))} \\ &= \frac{\frac{\exp(\phi(\xi))}{1+\exp(\phi(\xi))}(1 - f_j(\phi(\xi)))}{(1 - \frac{\exp(\phi(\xi))}{1+\exp(\phi(\xi))})f_j(\phi(\xi))}\end{aligned}$$

which implies that

$$f_j(\phi(\xi))\frac{c_r(y)}{c_r(y')} = \frac{\frac{\exp(\phi(\xi))}{1+\exp(\phi(\xi))}(1 - f_j(\phi(\xi)))}{\frac{1}{1+\exp(\phi(\xi))}} = \exp(\phi(\xi)) - \exp(\phi(\xi))f_j(\phi(\xi)))$$

and hence

$$f_j(\phi(\xi)) = \frac{\exp(\phi(\xi))}{\frac{c_r(y)}{c_r(y')} + \exp(\phi(\xi))} = \frac{\exp\left(\phi(\xi) - \log\left(\frac{c_r(y)}{c_r(y')}\right)\right)}{1 + \exp\left(\phi(\xi) - \log\left(\frac{c_r(y)}{c_r(y')}\right)\right)}.$$

Since $\xi \in \mathbb{R}$ was entirely unspecified it follows that $\phi(\xi) \in \mathbb{R}$ is also unspecified and will now be denoted θ. Furthermore, from Equation (2.4) it is known that both $c_r(y)$ and $c_r(y')$ are positive but otherwise unspecified and hence $\log\left(\frac{c_r(y)}{c_r(y')}\right)$ is a well defined real number which will now be denoted by β_j. In conclusion,

$$f_1(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)}, \quad f_j(\theta) = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)}, \text{ for } j = 2, \ldots, p,$$

such that the IRT model is equivalent to the Rasch model with item parameters $\beta_1 = 0$ and β_j for $j = 2, \ldots, p$. □

Theorem 2.1.1 shows, under appropriate conditions, that sufficiency of the subject score implies that the IRT model is equivalent to the Rasch model. The reverse implication, that the subject score is a sufficient statistic for the subject parameter in the Rasch model, will be shown in the derivation of the conditional maximum likelihood in the following chapter where parameter estimation for the Rasch model will be considered.

3 | Parameter Estimation in the Rasch Model

In this chapter methods to conduct parameter estimation for the Rasch model presented in Chapter 2 will be explored. Specifically, the joint maximum likelihood (JML) and conditional maximum likelihood (CML) will be presented for the Rasch model in respectively Section 3.1 and Section 3.2. Furthermore, a goodness of fit (GOF) test based directly on asymptotic results of the conditional maximum likelihood is derived in Section 3.3. The Chapter is then concluded in Section 3.4 where a simulation study is conducted to illustrate some of the asymptotic properties of the estimators and test statistics presented in the Chapter.

Remark 3.0.1. Uniqueness of Parameter Estimates Consider a Rasch model with subject parameters $\theta = (\theta_1, \ldots, \theta_n)^\top$ and item parameters $\beta = (\beta_1, \ldots, \beta_p)^\top$. By Equation (2.2), an equivalent model would be obtained by translating the parameters by some $c \in \mathbb{R}$. To ensure uniqueness of the parameter estimates in the following, it will be assumed that $\beta_1 = 0$.

Consider in the following binary response data $y = \{y_{ij}\}_{\substack{i=1,\ldots,n \\ j=1,\ldots p}}$ which is assumed to be generated from a Rasch model with item parameter $\beta_0 \in \mathbb{R}^{p-1}$.

3.1 Joint maximum likelihood

This section regarding joint maximum likelihood estimation is based on [11].

It follows by local stochastic independence and Equation (2.3) that the *joint likelihood* is given by

$$\begin{aligned} L_J(\theta, \beta \mid y) &= \prod_{i=1}^{n} \prod_{j=1}^{p} p(y_{ij}; \theta_i, \beta_j) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{p} \left(\frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)} \right) \\ &= \frac{\exp\left(\sum_{i=1}^{n} \theta_i y_{i+}\right) \exp\left(-\sum_{j=1}^{p} \beta_j y_{+j}\right)}{\prod_{i=1}^{n} \prod_{j=1}^{p} (1 + \exp(\theta_i - \beta_j))}, \end{aligned} \tag{3.1}$$

and hence the *joint log-likelihood* is given by

$$l_J(\theta, \beta \mid y) = \sum_{i=1}^{n} \theta_i y_{i+} - \sum_{j=1}^{p} \beta_j y_{+j} - \sum_{i=1}^{n} \sum_{j=1}^{p} \log(1 + \exp(\theta_i - \beta_j)).$$

By taking the partial derivatives wrt. each of the $n + p - 1$ parameters, recalling that $\beta_1 = 0$, the *joint score* is obtained as

$$s_J(\theta, \beta \mid y) = \begin{pmatrix} y_{1+} - \sum_{j=1}^{p} \frac{\exp(\theta_1 - \beta_j)}{1 + \exp(\theta_1 - \beta_j)} \\ \vdots \\ y_{n+} - \sum_{j=1}^{p} \frac{\exp(\theta_n - \beta_j)}{1 + \exp(\theta_n - \beta_j)} \\ -y_{+2} + \sum_{i=1}^{n} \frac{\exp(\theta_i - \beta_2)}{1 + \exp(\theta_i - \beta_2)} \\ \vdots \\ -y_{+p} + \sum_{i=1}^{n} \frac{\exp(\theta_i - \beta_p)}{1 + \exp(\theta_i - \beta_p)} \end{pmatrix}. \tag{3.2}$$

By setting the score function equal zero the *joint solution equations* are obtained as

$$y_{i+} = \sum_{j=1}^{p} \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = \sum_{j=1}^{p} p_{ij}, \quad i = 1, \ldots, n, \tag{3.3}$$

$$y_{+j} = \sum_{i=1}^{n} \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = \sum_{i=1}^{n} p_{ij}, \quad j = 2, \ldots, p. \tag{3.4}$$

Solving the joint solution equations, assuming a solution exists, yields the unique *joint maximum likelihood estimates* $\hat{\theta}_J = (\hat{\theta}_{1,J}, \ldots, \hat{\theta}_{n,J})^\top$ and $\hat{\beta}_J = (\hat{\beta}_{2,J}, \ldots, \hat{\beta}_{p,J})^\top$.

It should be noted that the JML estimates does not necessarily exist, which is for instance the case if $y_{i+} = p, y_{i+} = 0, y_{+j} = n$ or $y_{+j} = 0$ for some $i = 1, \ldots, n$ or $j = 1, \ldots, p$, i.e. if there are any extreme scores in the data. This follows directly from the joint solution equations and fact that $0 < p_{ij} < 1$ for $i = 1, \ldots, n, j = 1, \ldots, p$.

The joint solution equations also illustrates the sufficiency of the subject and item scores since it is clear that e.g. $y_{i+} = y_{i'+}$ implies $\hat{\theta}_{i,J} = \hat{\theta}_{i',J}$ for $i, i' = 1, \ldots, n, i \neq i'$, and similarly for the item parameter estimates.

It will now be shown that the parameter estimates obtained by solving Equations (3.3) and (3.4) are unique and maximizes the joint likelihood function, which will be done by showing that the joint observed information matrix is positive definit. Let $k_{ij} = \frac{\exp(\theta_i - \beta_j)}{(1+\exp(\theta_i - \beta_j))^2}$, then it follows from Equation (3.2) that

$$
\begin{aligned}
\frac{\partial^2}{\partial \theta_i^2} l(\theta, \beta \mid y) &= -\sum_{j=1}^{p} \frac{\exp(\theta_i - \beta_j)}{(1 + \exp(\theta_i - \beta_j))^2} = -\sum_{j=1}^{p} k_{ij}, \quad i = 1, \ldots, n, \\
\frac{\partial^2}{\partial \beta_j^2} l(\theta, \beta \mid y) &= -\sum_{i=1}^{n} \frac{\exp(\theta_i - \beta_j)}{(1 + \exp(\theta_i - \beta_j))^2} = -\sum_{i=1}^{n} k_{ij}, \quad j = 2, \ldots, p, \\
\frac{\partial^2}{\partial \theta_i \partial \beta_j} l(\theta, \beta \mid y) &= \frac{\exp(\theta_i - \beta_j)}{(1 + \exp(\theta_i - \beta_j))^2} = k_{ij}, \quad i = 1, \ldots, n \; j = 2, \ldots, p, \\
\frac{\partial^2}{\partial \theta_i \partial \theta_{i'}} l(\theta, \beta \mid y) &= 0, \quad i, i' = 1, \ldots, n, i \neq i', \\
\frac{\partial^2}{\partial \beta_j \partial \beta_{j'}} l(\theta, \beta \mid y) &= 0, \quad j, j' = 2, \ldots, p, j \neq j'.
\end{aligned}
$$

From the above the *joint observed information* can be constructed as

$$J_J(\theta, \beta \mid y) = \begin{bmatrix} \sum_{j=1}^{p} k_{1j} & 0 & \cdots & 0 & -k_{12} & \cdots & \cdots & -k_{1p} \\ 0 & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \sum_{j=1}^{p} k_{nj} & -k_{n2} & \cdots & \cdots & -k_{np} \\ -k_{12} & \cdots & \cdots & -k_{n2} & \sum_{i=1}^{n} k_{i2} & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & 0 \\ -k_{1p} & \dots & \dots & -k_{np} & 0 & \cdots & 0 & \sum_{i=1}^{n} k_{ip} \end{bmatrix}.$$

The joint observed information matrix is positive definite, which follows from the fact that for any $z \in \mathbb{R}^{n+p-1} \setminus \{0\}$

$$\begin{aligned} z^\top J_j(\theta, \beta \mid y) z &= \sum_{i=1}^{n} z_i^2 \sum_{j=1}^{p} k_{ij} + \sum_{j=2}^{p} z_{n+j-1}^2 \sum_{i=1}^{n} k_{ij} - \sum_{i=1}^{n} z_i \sum_{j=2}^{p} z_{n+j-1} k_{ij} - \sum_{j=2}^{p} z_{n+j-1} \sum_{i=1}^{n} z_i k_{ij} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} (z_i^2 + z_{n+j-1}^2 - z_i z_{n+j-1} - z_{n+j-1} z_i) k_{ij} + z_i^2 k_{i1} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} (z_i - z_{n+j-1})^2 k_{ij} + z_i^2 k_{i1} > 0. \end{aligned}$$

Remark 3.1.1. If the assumption $\beta_1 = 0$ was removed and the joint score and joint observed information matrix also included partial derivatives wrt. β_1, then the joint observed information matrix would be positive semidefinite, i.e. the found maximum would not necessarily be unique. This is in agreement with our considerations in Remark 3.0.1 regarding uniqueness.

Theorem 3.1.2. Existence of JML estimates

Fix the number of items p. Then the probability that the joint maximum likelihood estimates exists approaches zero as the number of subjects n approaches infinite.

Proof. Let $P_{is} = P(y_{i+} = s)$ denote the probability that the ith subject has score s. Since P_{is} is strictly positive for all $i = 1, \ldots, n$ and $s = 0, \ldots, p$ the probability that the ith subject has an extreme score is strictly positive, i.e. $P_{i0} + P_{ip} > 0$. If atleast one subject has an extreme score then the JML estimates does not exist and hence the probability that JML estimates exists is bounded above by $\prod_{i=1}^{n}(1 - (P_{i0} + P_{ip}))$, the probability that none of the subjects have an extreme score. The theorem now follows from the fact that $\prod_{i=1}^{n}(1 - (P_{i0} + P_{ip})) \to 0$ as $n \to \infty$. □

The following Theorem regarding the bias of the JML estimator is stated in [6][Page 36] and the proof is beyond the scope of this report.

Theorem 3.1.3. Bias of JML Estimator

The joint maximum likelihood estimates for both the subject and item parameters are biased. Futhermore, the bias is of order $\frac{p}{1-p}$.

It should be noted that Theorem 3.1.3 is a special case of a more general result derived in [16] regarding bias when conducting joint estimation of parameters.

Theorem 3.1.3 shows that the bias of the parameter estimates does not converge towards zero asymptotically and therefore it is clear that the estimators would also be inconsistent as the following theorem based on [4][Page 66-69] states.

Theorem 3.1.4. Inconcistency of JML Estimates

Fix the number of items p and keep the item parameters $\beta_j, j = 1, \ldots, p$ constant. Then the joint maximum likelihood estimates $\hat{\theta}_{i,J}$ and $\hat{\beta}_{j,J}$ of respectively θ_i and β_j for $i = 1, \ldots, n, j = 1, \ldots, p$ are inconsistent as n approaches infinity.

Intuituvely, the asymptotic problems with the JML come from the fact that the number of parameters in the model increases as the number of subjects increases.

Another approach to parameter estimation which avoids this problem will now be presented.

3.2 Conditional maximum likelihood

This section regarding the conditional maximum likelihood is based on [6] and [2]. The basic idea of the CML is to condition on the subject score, which is a sufficient statistic for the subject parameter as will be shown in the following, such that the CML only depends on the item parameters. Consider the conditional probability that a subject with parameter $\tilde{\theta}$ obtains result pattern $\tilde{y} \in \Gamma(p)$ given the score of the subject $s = \tilde{y}_+$:

$$p(\tilde{y} \mid s; \tilde{\theta}, \beta) = \frac{p(\tilde{y}; \tilde{\theta}, \beta)}{p(s; \tilde{\theta}, \beta)}. \tag{3.5}$$

By local stochastic independence and Equation (2.3), the probability of a subject to obtain result pattern $\tilde{y}$ is given by

$$p(\tilde{y}; \tilde{\theta}, \beta) = \frac{\exp\left(\sum_{j=1}^{p} \tilde{y}_j \left(\tilde{\theta} - \beta_j\right)\right)}{\prod_{j=1}^{p} \left(1 + \exp(\tilde{\theta} - \beta_j)\right)} = \frac{\exp(s\tilde{\theta}) \exp\left(\sum_{j=1}^{p} -\tilde{y}_j \beta_j\right)}{\prod_{j=1}^{p} \left(1 + \exp(\tilde{\theta} - \beta_j)\right)}. \tag{3.6}$$

Furthermore, the probability that the subject would have obtained a result pattern with score s is given as the following sum over result patterns with score s:

$$p(s; \tilde{\theta}, \beta) = \sum_{\substack{y \in \Gamma(p): \\ y_+ = s}} p(y; \tilde{\theta}, \beta) = \sum_{\substack{y \in \Gamma(p): \\ y_+ = s}} \frac{\exp(s\tilde{\theta}) \exp\left(\sum_{j=1}^{p} -y_j \beta_j\right)}{\prod_{j=1}^{p} \left(1 + \exp(\tilde{\theta} - \beta_j)\right)} \tag{3.7}$$

where Equation (3.6) is used in the second equality.

Inserting Equations (3.6) and (3.7) in Equation (3.5) yields

$$p(\tilde{y} \mid s; \beta) = \frac{\exp\left(\sum_{j=1}^{p} -\tilde{y}_j \beta_j\right)}{\sum_{\substack{y \in \Gamma(p): \\ y_+ = s}} \exp\left(\sum_{j=1}^{p} -y_j \beta_j\right)}$$

which does not depend on the subject parameter $\tilde{\theta}$ thus showing that the subject score is a sufficient statistic for the subject parameter.

By defining the *symmetric functions*

$$\gamma_s(\beta) = \sum_{\substack{y \in \Gamma(p): \\ y_+ = s}} \exp\left(\sum_{j=1}^{p} -y_j \beta_j\right)$$

for $s = 1, \ldots, p$, it follows that

$$p(\tilde{y} \mid s; \beta) = \frac{\exp\left(\sum_{j=1}^{p} -\tilde{y}_j \beta_j\right)}{\gamma_s(\beta)}. \tag{3.8}$$

Remark 3.2.1. The symmetric functions inherit their names from the well known *elementary symmetric polynomials* given by

$$S_k(x) = \sum_{1 \leq j_1 < \cdots < j_k \leq n} \prod_{i=1}^{k} x_{j_i}$$

for $x \in \mathbb{R}^n, k, n \in \mathbb{N}$ since

$$\gamma_s(\beta) = S_s(\exp(-\beta_1), \ldots, \exp(-\beta_p)).$$

The following lemma regarding the symmetric functions is important for the derivations in the rest of the section.

Lemma 3.2.2. Symmetric functions

For any $j = 2, \ldots, p$ and $s = 1, \ldots, p-1$ it follows that the symmetric functions can be expressed recursively as

$$\gamma_s(\beta) = \exp(-\beta_j)\gamma_{s-1}^{(j)}(\beta) + \gamma_s^{(j)}(\beta), \tag{3.9}$$

where $\gamma_s^{(j)}(\beta)$ denotes the symmetric function with the jth item omitted. Furthermore, the partial derivative of $\gamma_s(\beta)$ wrt. β_j can be expressed as

$$\frac{\partial}{\partial \beta_j}\gamma_s(\beta) = -\exp(-\beta_j)\gamma_{s-1}^{(j)}(\beta) \tag{3.10}$$

and the partial derivative of $\gamma_s^{(j)}(\beta)$ wrt. $\beta_{j'}$, where $j' = 2, \ldots, p, j' \neq j$, is given by

$$\frac{\partial}{\partial \beta_{j'}}\gamma_s^{(j)}(\beta) = -\exp(-\beta_{j'})\gamma_{s-1}^{(j,j')}(\beta) \tag{3.11}$$

where $\gamma_{s-1}^{(j,j')}(\beta)$ denotes the symmetric function with both the jth and the j'th item omitted.

Proof. Since the omission of the jth item is equivalent to letting $y_j = 0, y \in \Gamma(p)$, it follows that

$$\gamma_s^{(j)}(\beta) = \sum_{\substack{y \in \Gamma(p): \\ y_+ = s \\ y_j = 0}} \exp\left(\sum_{j'=1}^{p} -y_{j'}\beta_{j'} \right), \tag{3.12}$$

and hence

$$\begin{aligned} \exp(-\beta_j)\gamma_{s-1}^{(j)}(\beta) &= \sum_{\substack{y \in \Gamma(p): \\ y_+ = s-1 \\ y_j = 0}} \exp\left(\sum_{j'=1}^{p} -y_{j'}\beta_{j'} - \beta_j \right) \\ &= \sum_{\substack{y \in \Gamma(p): \\ y_+ = s \\ y_j = 1}} \exp\left(\sum_{j'=1}^{p} -y_{j'}\beta_{j'} \right). \end{aligned} \tag{3.13}$$

Combining Equations (3.12) and (3.13) immediately yields (3.9).

Equation (3.10) follows directly from Equation (3.9) by noting that $\gamma_s^{(j)}(\beta)$ does not depend on β_j.

Equation (3.11) follows since

$$\begin{aligned} \frac{\partial}{\partial \beta_{j'}} \gamma_s^{(j)}(\beta) &= \sum_{\substack{y \in \gamma(p) \\ y_+ = s \\ y_j = 0}} -y_{j'} \exp\left(-\sum_{k=1}^{p} y_k \beta_k \right) \\ &= -\exp(-\beta_{j'}) \sum_{\substack{y \in \gamma(p) \\ y_+ = s-1 \\ y_j = 0, y_{j'} = 0}} \exp\left(-\sum_{k=1}^{p} y_k \beta_k \right) \\ &= -\exp(-\beta_{j'}) \gamma_{s-1}^{j,j'}(\beta). \end{aligned}$$

□

Define the *conditional likelihood* as

$$L_C(\beta \mid y) = \prod_{i=1}^{n} p(y_i \mid y_{i+}; \beta) = \prod_{i=1}^{n} \frac{\exp\left(\sum_{j=1}^{p} -y_{ij}\beta_j \right)}{\gamma_{y_{i+}}(\beta)}$$

where the second equality follows from Equation (3.8). Let $y^{(1:p-1)}$ denote the vector composed of y_i such that $y_{i+} = 1, \ldots, p-1$, i.e. the dataset restricted to subjects with non-extreme scores.

Furthermore, let n_s denote the number of subjects with score s for $s = 0, \ldots, p$. Then the conditional likelihood can be written as

$$L_C(\beta \mid y) = \frac{\exp\left(-\sum_{j=1}^{p} y_{+j}\beta_j\right)}{\prod_{s=0}^{p} \gamma_s(\beta)^{n_s}} = \frac{\exp\left(-\sum_{j=1}^{p} y_{+j}^{(1:p-1)}\beta_j\right)}{\prod_{s=1}^{p-1} \gamma_s(\beta)^{n_s}} \tag{3.14}$$

and where in the last equality it is used that $\gamma_0(\beta) = 1$, $\gamma_p(\beta) = \exp\left(-\sum_{j=1}^{p} \beta_j\right)$ and that

$$\sum_{j=1}^{p} y_{+j}\beta_j = \sum_{j=1}^{p} y_{+j}^{(1:p-1)}\beta_j + n_p \sum_{j=1}^{p} \beta_j.$$

Hence the *conditional log-likelihood* is given as

$$l_C(\beta \mid y) = -\sum_{s=1}^{p-1} n_s \log(\gamma_s(\beta)) - \sum_{j=1}^{p} y_{+j}^{(1:p-1)}\beta_j.$$

It should be noted that subjects with extreme scores contain no information regarding the difficulty of the items, which intuitively explains why they are not part of the CML function.

The partial derivative wrt. the jth item parameter is given by

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} l_C(\beta \mid y) &= -\sum_{s=1}^{p-1} n_s \frac{\partial \log(\gamma_s(\beta))}{\partial \beta_j} - y_{+j}^{(1:p-1)} \\
&= \sum_{s=1}^{p-1} n_s \frac{\exp(-\beta_j)\gamma_{s-1}^{(j)}(\beta)}{\gamma_s(\beta)} - y_{+j}^{(1:p-1)} \\
&= \sum_{s=1}^{p-1} n_s \frac{\gamma_s(\beta) - \gamma_s^{(j)}(\beta)}{\gamma_s(\beta)} - y_{+j}^{(1:p-1)}
\end{aligned} \tag{3.15}$$

where the second and third equality follows from respectively Equation (3.10) and Equation (3.9). Therefore, the *conditional score* is given as

$$s_C(\beta \mid y) = \begin{pmatrix} \sum_{s=1}^{p-1} n_s \frac{\gamma_s(\beta) - \gamma_s^{(2)}(\beta)}{\gamma_s(\beta)} - y_{+2}^{(1:p-1)} \\ \vdots \\ \sum_{s=1}^{p-1} n_s \frac{\gamma_s(\beta) - \gamma_s^{(p)}(\beta)}{\gamma_s(\beta)} - y_{+p}^{(1:p-1)} \end{pmatrix}. \tag{3.16}$$

Setting the conditional score equal to zero yields the *conditional solution equations*

$$y_{+j}^{(1:p-1)} = \sum_{s=1}^{p-1} n_s \frac{\gamma_s(\beta) - \gamma_s^{(j)}(\beta)}{\gamma_s(\beta)}, \text{ for } j = 2, \ldots, p. \tag{3.17}$$

Solving the conditional solution equations wrt. β, assuming a solution exist, yields the *conditional maximimum likelihood estimates* $\hat{\beta}_C = (\hat{\beta}_{2,C}, \ldots, \hat{\beta}_{p,C})^\top$, which is a unique solution cf. [9][Theorem 3 and 4]. Note that a solution does not exist in the case of an extreme item score since $0 < \frac{\gamma_s(\beta) - \gamma_s^{(j)}(\beta)}{\gamma_s(\beta)} < 1$. However, the probability of an extreme item score goes to zero as the number of subjects increases.

From Equation (3.15) it follows that for $j, j' = 2, \ldots p$, the second order partial derivatives of the conditional log-likelihood are given by

$$\begin{aligned}
\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell_C(\beta \mid y) &= \sum_{s=1}^{p-1} n_s \frac{\partial}{\partial\beta_{j'}} \frac{\gamma_s(\beta) - \gamma_s^{(j)}(\beta)}{\gamma_s(\beta)} \\
&= \sum_{s=1}^{p-1} n_s \frac{\left(\frac{\partial}{\partial\beta_{j'}}\gamma_s(\beta) - \frac{\partial}{\partial\beta_{j'}}\gamma_s^{(j)}(\beta)\right)\gamma_s(\beta) - \left(\gamma_s(\beta) - \gamma_s^{(j)}(\beta)\right)\frac{\partial}{\partial\beta_{j'}}\gamma_s(\beta)}{\gamma_s(\beta)^2} \\
&= \sum_{s=1}^{p-1} n_s \frac{\left(\gamma_s(\beta) - \gamma_s^{(j)}(\beta)\right)\left(\gamma_s(\beta) - \gamma_s^{(j')}(\beta)\right) - (\gamma_s(\beta) - \gamma_s^{(j')}(\beta))\gamma_s(\beta)}{\gamma_s(\beta)^2} \\
&\quad - n_s \frac{\frac{\partial}{\partial\beta_{j'}}\gamma_s^{(j)}(\beta))\gamma_s(\beta)}{\gamma_s(\beta)^2} \\
&= \sum_{s=1}^{p-1} n_s \frac{-\gamma_s^{(j)}(\beta)(\gamma_s(\beta) - \gamma_s^{(j')}(\beta)) - \frac{\partial}{\partial\beta_{j'}}\gamma_s^{(j)}(\beta)\gamma_s(\beta)}{\gamma_s(\beta)^2}.
\end{aligned} \tag{3.18}$$

For $j = j'$ Equation (3.18) yields

$$\frac{\partial^2}{\partial\beta_j^2}\ell_C(\beta \mid y) = \sum_{s=1}^{p-1} n_s \frac{-\gamma_s^{(j)}(\beta)(\gamma_s(\beta) - \gamma_s^{(j)}(\beta))}{\gamma_s(\beta)^2} \tag{3.19}$$

since $\gamma_s^{(j)}(\beta)$ does not depend on β_j.

Furthermore, for $j \neq j'$ Equation (3.18) yields

$$\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell_C(\beta \mid y) = \sum_{s=1}^{p-1} n_s \frac{\exp(-\beta_{j'})\gamma_{s-1}^{(j,j')}(\beta)\gamma_s(\beta) - \gamma_s^{(j)}(\beta)(\gamma_s(\beta) - \gamma_s^{(j')}(\beta))}{\gamma_s(\beta)^2} \tag{3.20}$$

where Equation (3.11) have been applied.

Thus a closed form expression have been obtained of every entry in the *conditional observed information* given by

$$J_C(\beta|y) = -\frac{\partial^2}{\partial\beta^\top\partial\beta}\ell_C(\beta, y) \tag{3.21}$$

which in particular only depends on y through n_s for $s = 1, \ldots, p-1$. The conditional score and conditional observed information can be used to obtain the CML estimates using e.g. the Newtons-Raphson algorithm assuming that a solution exists.

Once the conditional maximum likelihood item parameter estimates have been obtained, these can be used to estimate the subject parameters by insertion into the joint solution equations for the JML estimates.

The following Theorem is stated in [12] and the proof is beyond the scope of this report.

Theorem 3.2.3. Bias of CML estimators

The CML estimators of the item parameters are unbiased, i.e.

$$\mathbb{E}\left[\hat{\beta}_C\right] = \beta_0.$$

Consider the *restricted conditional likelihoods* which are the conditional likelihood where the subjects with score y_{i+} different from s is omitted for $s = 1, \ldots, p-1, i = 1, \ldots, n$, i.e.

$$L_C^{(s)}(\beta \mid y) = \prod_{i=1}^{n} p(y_i \mid s, \beta)^{1[y_{i+}=s]} = \frac{\exp\left(-\sum_{j=1}^{p} y_{+j}^{(s)}\beta_j\right)}{\gamma_s(\beta)^{n_s}}$$

where $y^{(s)}$ denotes the vector composed of y_i such that $y_{i+} = s$ for $i = 1, \ldots, n$ i.e. the data restricted to subjects with score s.

By defining the *restricted conditional log-likelihood*, *restricted conditional score* and *restricted conditional observed information* in the obvious way, it follows by the above derivation of the quantities of the CML that

$$\begin{aligned}
\ell_C^{(s)}(\beta \mid y) &= -\sum_{j=1}^{p} y_{+j}^{(s)}\beta_j - n_s\gamma_s(\beta) \\
s_C^{(s)}(\beta \mid y) &= \begin{pmatrix} -n_s\frac{\gamma_s(\beta)-\gamma_s^{(2)}(\beta)}{\gamma_s(\beta)} - y_{+2}^{(s)} \\ \vdots \\ -n_s\frac{\gamma_s(\beta)-\gamma_s^{(p)}(\beta)}{\gamma_s(\beta)} - y_{+p}^{(s)} \end{pmatrix} \\
J_C^{(s)}(\beta \mid y) &= -\frac{\partial^2}{\partial\beta\partial\beta^\top}\ell_C^{(s)}(\beta \mid y)
\end{aligned}$$

with elements

$$\frac{\partial^2}{\partial\beta_j^2}\ell^{(s)}(\beta \mid y) = n_s\frac{-\gamma_s^{(j)}(\beta)(\gamma_s(\beta)-\gamma_s^{(j)}(\beta))}{\gamma_s(\beta)^2}, \quad \text{for } j = 2,\ldots,p$$

and

$$\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell^{(s)}(\beta \mid y) = n_s\frac{\exp(-\beta_{j'})\gamma_{s-1}^{(j,j')}(\beta)\gamma_s(\beta) - \gamma_s^{(j)}(\beta)(\gamma_s(\beta)-\gamma_s^{(j')}(\beta))}{\gamma_s(\beta)^2},$$

for $j, j' = 2,\ldots,p, j \neq j'$.

Define the *restricted conditional maximum likelihood estimate* $\hat{\beta}_C^{(s)} = \left(\hat{\beta}_{2,C}^{(s)},\ldots,\hat{\beta}_{p,C}^{(s)}\right)^\top$ as the solution to the *restricted conditional solution equations* given by

$$n_s\frac{\gamma_s(\beta)-\gamma_s^{(j)}(\beta)}{\gamma_s(\beta)} = y_{+j}^{(s)}, \quad \text{for } j = 2,\ldots,p, s = 1\ldots,p-1.$$

Furthermore, it follows immediately that the relation between the CML and restricted CML is given by

$$L_C(\beta \mid y) = \prod_{s=1}^{p-1} L_C^{(s)}(\beta \mid y) \tag{3.22}$$

and similarly it follows that

$$s_C(\beta \mid y) = \sum_{s=1}^{p-1} s_C^{(s)}(\beta \mid y) \tag{3.23}$$

and

$$J_C(\beta \mid y) = \sum_{s=1}^{p-1} J_C^{(s)}(\beta \mid y). \tag{3.24}$$

Theorem 3.2.4. Asymptotic Normality of Restricted CML Estimator
The restricted CML estimate is asymptotically normally distributed with mean β_0 and variance $J_C^{(s)}(\beta_0)^{-1}$ for $s = 1, \ldots, p-1$, i.e.

$$\sqrt{n_s}\left(\hat{\beta}_C^{(s)} - \beta_0\right) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0, I_C^{(s)}(\beta_0)^{-1}\right)$$

where $I_C^{(s)}(\beta_0) = \frac{1}{n_s} J_C^{(s)}(\beta_0 \mid y)$.

Proof. Recall that $Y_1, Y_2, \ldots, Y_{n_s}$ are i.i.d given that they share the same subject score, i.e. $Y_{i+} = s$ for $i = 1, \ldots, n_s$.

It follows immediately by standard results, see e.g. [8][Theorem 18], that

$$\sqrt{n_s}(\hat{\beta}_C^{(s)} - \beta_0) \mid (Y_{i+})_{i=1,\ldots,n} \xrightarrow[n_s\to\infty]{d} N_{p-1}\left(0_{p-1}, I_C^{(s)}(\beta_0)^{-1}\right)$$

since $I_C^{(s)}(\beta_0)$ is the observed information and hence also the Fisher information for a single observation as it does not depend on y.

In order to extend this result from the conditional distribution to the marginal distribution, consider a bounded and continuous function $f : \mathbb{R}^{p-1} \to \mathbb{R}$. Note that

$$\begin{aligned}\mathbb{E}\left[f\left(\sqrt{n_s}(\hat{\beta}_C^{(s)} - \beta_0)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[f\left(\sqrt{n_s}(\hat{\beta}_C^{(s)} - \beta_0)\right) \mid (Y_{i+})_{i=1,\ldots,n}\right]\right] \\ &\xrightarrow[n\to\infty]{} \mathbb{E}\left[\mathbb{E}\left[f(z)\right]\right], \quad \text{for } z \sim N_{p-1}\left(0_{p-1}, I_C^{(s)}(\beta_0)^{-1}\right)\end{aligned}$$

where it is used that $n \to \infty \implies n_s \to \infty$. Thus it follows that

$$\mathbb{E}\left[f\left(\sqrt{n_s}(\hat{\beta}_C^{(s)} - \beta_0)\right)\right] \to \mathbb{E}\left[f(z)\right]$$

and as f was an arbitrarily chosen bounded and continous function the result follows.

□

Theorem 3.2.4 will be used to show a similar result regarding the CML estimator.

In [2][Theorem 2] it is stated that the following result regarding consistency of the CML estimator is obtained from Theorem 3.2.4, Equation (3.23) and some elementary continuity arguments. However, as this is not clear to the author of this report, the proof has been omitted.

Theorem 3.2.5. Consistency of CML Estimator

The CML estimator is consistent, i.e.

$$\hat{\beta}_C \xrightarrow[n\to\infty]{p} \beta_0.$$

Theorems 3.2.4 and 3.2.5 can now be used to show asymptotic normality of the CML estimator.

Theorem 3.2.6. Asymptotic Normality of CML Estimator

The CML estimates is asymptotically normally distributed with mean β_0 and variance $J_C(\beta_0 \mid y)^{-1}$, i.e.

$$J_C(\beta_0 \mid y)^{1/2}\left(\hat{\beta}_C - \beta_0\right) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0_{p-1}, I_{p-1}\right)$$

Proof. A zero'th order multivariate Taylor expansion of $\frac{\partial}{\partial\beta_j}\ell_C(\beta \mid y)$ around $\hat{\beta}_C$ yields

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\ell_C(\beta_0 \mid y) &= \frac{\partial}{\partial\beta_j}\ell_C(\hat{\beta}_C \mid y) + \sum_{j'=2}^{p}(\hat{\beta}_{j',C} - \beta_{j',0})\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell_C(\beta^* \mid y) \\ &= \sum_{j'=2}^{p}(\hat{\beta}_{j',C} - \beta_{j',0})\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell_C(\beta^* \mid y)\end{aligned}$$

for some β^* such that $|\beta_j^* - \beta_{j,0}| \leq |\hat{\beta}_{j,C} - \beta_{j,0}|$ for $j = 2, \ldots, p$. Theorem 3.2.5 implies that $\hat{\beta}_{j',C} \xrightarrow[n\to\infty]{p} \beta_{j,0}$ and hence $\beta_{j'}^* \xrightarrow[n\to\infty]{p} \beta_{j,0}$.

Therefore, it follows that $\frac{\partial}{\partial\beta_j}\ell_C(\beta_0 \mid y)$ has the same limiting distribution as $\sum_{j'=2}^{p}(\hat{\beta}_{j',C} - \beta_{j',0})\frac{\partial^2}{\partial\beta_j\partial\beta_{j'}}\ell_C(\beta_0 \mid y)$.

Applying this for $j = 2, \ldots p$, it is obtained that $s_C(\beta_0 \mid y)$ has the same limiting distribution as $\left(\hat{\beta}_C - \beta_0\right) J_C(\beta_0 \mid y)$.

The standard results referenced in the proof of Theorem 3.2.4 also yields that

$$s_C^{(s)}(\beta_0 \mid y) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0_{p-1}, J_C^{(s)}(\beta_0 \mid y)\right).$$

Combining this with Equation (3.23) and Equation (3.24) implies that

$$s_C(\beta_0 \mid y) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0_{p-1}, J_C(\beta_0 \mid y)\right)$$

such that

$$J_C(\beta_0 \mid y)\left(\hat{\beta}_C - \beta_0\right) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0_{p-1}, J_C(\beta_0 \mid y)\right)$$

and hence

$$J_C(\beta_0 \mid y)^{1/2}\left(\hat{\beta}_C - \beta_0\right) \xrightarrow[n\to\infty]{d} N_{p-1}\left(0_{p-1}, I_{p-1}\right)$$

□

When comparing Theorems 3.2.3 and 3.2.6 to Theorems 3.1.3 and 3.1.4 it is evident why the CML estimator is usually preferred in application.

Informally, there are three main problems with the JML estimator which results in the non-existence of JML estimates, bias and inconsistency. The first problem is that the occurrence of extreme scores implies that the JML estimates do not exists. This is particularly problematic as the number of subjects increases since this increases the probability of an extreme subject score. The extreme subject scores do not occur in the CML function and are therefore not problematic. Furthermore, as the number of subjects increases, the probability of extreme item scores decreases since $0 < p_{ij} < 1$.

The second problem is the joint estimation of subject and item parameters which leads to bias. Since CML estimation conditions on the subject score, a sufficient statistic for the subject parameter, CML estimates the item parameters independently of the subject parameters, thus resulting in an unbiased estimate.

The third problem is that when the number of subjects increases then the number of parameters to be estimated in JML increases likewise. Furthermore, an increased number of subjects brings no additional information regarding the other subjects parameters, only further information regarding the item parameters. This is not an issue for the CML since only item parameters are to be estimated. Therefore, as the number of subjects increases, the number of parameters to be estimated remains the same, and additional information regarding these parameters are obtained, which results in consistent estimates.

Another approach to parameter estimation in the Rasch model is the maximum marginal likelihood which will be discussed in Section 4.3.

The asymptotic results regarding the CML will be utilized in the following section to formulate a goodness of fit test for the Rasch model.

3.3 Goodness of Fit Test

In this section a goodness of fit test for the Rasch model will be presented based on the framework of the CML and restricted CML from Section 3.2. The section is based on [2]. Consider the *conditional likelihood-ratio* between the Rasch model fitted to each score group and the Rasch model fitted to the full data given by

$$\lambda = \frac{L_C(\hat{\beta}_C \mid y)}{\prod_{s=1}^{p-1} L_C^{(s)}(\hat{\beta}_C^{(s)} \mid y)}. \tag{3.25}$$

Clearly $\lambda \leq 1$ with $\lambda = 1$ if $\hat{\beta}_{C,s} = \hat{\beta}_C$ for all $s = 1, \ldots, p-1$ by Equation (3.22).

The null hypothesis for this test is that the data is generated by the Rasch model and this is rejected for small values of λ, i.e. small values of the test statistic are critical. This is intuitively clear because no matter which score group is considered, all the subjects took the same test and hence answered the same items with fixed item parameters. Therefore, no matter which score group is considered, the item estimates ought to be approximately the same, such the test statistic λ would be close to one. This phenomenon where the difficulty of the items can be measured based on a sample not drawn at random from the total subject population, i.e. that the estimates of the item parameters is somehow independent of the ability of the students constituting the sample, is often referred to as *specific objectivity*.

As per usual when working with likelihood-ratio tests, the test statistic

$$Z = -2\log(\lambda) = 2\sum_{s=1}^{p-1} \ell_C^{(s)}(\hat{\beta}_{C,s} \mid y) - 2\ell(\hat{\beta}_C \mid y) \tag{3.26}$$

is considered where large values of Z are critical.

Asymptotic properties of Z will be described shortly and for this purpose the following result is needed. The result is a corollary of Fisher-Cochran's theorem presented in [21][Page 188] and the proof is beyond the scope of this report.

Corollary 3.3.1. Let $Q = Q_1 + Q_2$ where $Q \sim \chi^2(a)$, $Q_1 \sim \chi^2(b), Q_2 \geq 0$ and $a > b$. Then $Q_2 \sim \chi^2(a-b)$.

Using Corollary 3.3.1 is can now be shown that Z is asymtotically chi-squared distributed.

Theorem 3.3.2. Asymptotic Distribution of Test Statistic
The test statistic Z given by Equation 3.26 is asymptotically chi-squared distributed with $(p-1)(p-2)$ degrees of freedom, i.e.

$$Z \xrightarrow[n\to\infty]{d} \chi^2((p-1)(p-2)).$$

Proof. Equation (3.22) implies that

$$L_C(\beta \mid y) - \sum_{s=1}^{p-1} L_C^{(s)}(\beta \mid y) = 0$$

for any $\beta \in \mathbb{R}^p$ and hence it follows that Z can be written as

$$Z = -2\ell_C(\hat{\beta}_C \mid y) + 2\ell_C(\beta_0 \mid y) + 2\sum_{s=1}^{p-1} \ell_C^{(s)}(\hat{\beta}_C^{(s)} \mid y) - 2\sum_{s=1}^{p-1} \ell_C^{(s)}(\beta_0 \mid y). \tag{3.27}$$

A first order multivariate Taylor expansion of $\ell_C(\beta_0 \mid y)$ around $\hat{\beta}_C$ yields

$$\begin{aligned}
\ell_C(\beta_0 \mid y) &= \ell_C(\hat{\beta}_C \mid y) + \sum_{j=2}^{p} \frac{\partial}{\partial \beta_j} \ell(\hat{\beta}_C \mid y)(\beta_{j,0} - \hat{\beta}_{j,C}) \\
&\quad + \sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial \beta_j \partial \beta_j'} \frac{\ell_C(\beta^* \mid y)}{2} (\beta_{j,0} - \hat{\beta}_{j,C})(\beta_{j',0} - \hat{\beta}_{j'C}) \\
&= \ell_C(\hat{\beta}_C \mid y) + \sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial \beta_j \partial \beta_j'} \frac{\ell_C(\beta^* \mid y)}{2} (\beta_{j,0} - \hat{\beta}_{j,C})(\beta_{j',0} - \hat{\beta}_{j',C})
\end{aligned}$$

for some β^* such that $|\beta_j^* - \beta_{j,0}| \leq |\hat{\beta}_{j,C} - \beta_{j,0}|$ for $j = 2, \ldots, p$, where it in the second equality is used that $\frac{\partial}{\partial \beta_j} \ell_C(\hat{\beta}_C \mid y)$ is zero for $j = 2, \ldots, p$ by the definition of $\hat{\beta}_C$.

Hence it follows that

$$-2\ell(\hat{\beta}_C \mid y) + 2\ell(\beta_0 \mid y) = \sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial\beta_j\partial\beta_{j'}}\ell(\beta^* \mid y)(\beta_{j,0} - \hat{\beta}_{j,C})(\beta_{j',0} - \hat{\beta}_{j',C}). \quad (3.28)$$

Similarly, for $s = 1, \ldots, p$ a first order multivariate Taylor expansion of $\ell_C^{(s)}(\beta_0 \mid y)$ around $\hat{\beta}_C^{(s)}$ yields

$$\ell_C^{(s)}(\beta_0 \mid y) = \ell_C^{(s)}(\hat{\beta}_C^{(s)} \mid y) + \sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial\beta_j\partial\beta_j'}\frac{\ell_C^{(s)}(\beta^{*(s)} \mid y)}{2}(\beta_{j,0} - \hat{\beta}_{j,C}^{(s)})(\beta_{j',0} - \hat{\beta}_{j',C}^{(s)})$$

for some $\beta^{*(s)}$ such that $|\beta_j^{*(s)} - \beta_{j,0}| \leq |\hat{\beta}_{j,C}^{(s)} - \beta_{j,0}|$ for $j = 2, \ldots, p$ and hence

$$2\sum_{s=1}^{p-1}\ell_C^{(s)}(\hat{\beta}_C^{(s)} \mid y) - 2\sum_{s=1}^{p-1}\ell_C^{(s)}(\beta_0 \mid y) = -\sum_{s=1}^{p-1}\sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial\beta_j\partial\beta_j'}\frac{\ell_C^{(s)}(\beta^{*(s)} \mid y)}{2}(\beta_{j,0} - \hat{\beta}_{j,C}^{(s)})(\beta_{j',0} - \hat{\beta}_{j',C}^{(s)}). \quad (3.29)$$

Insertion of Equations (3.28) and (3.29) in Equation (3.27) yields

$$Z = \sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial\beta_j\partial\beta_j'}\ell_C(\beta^* \mid y)(\beta_{j,0} - \hat{\beta}_j)(\beta_{j',0} - \hat{\beta}_{j'}) - \sum_{s=1}^{p-1}\sum_{j=2}^{p}\sum_{j'=2}^{p} \frac{\partial}{\partial\beta_j\partial\beta_j'}\ell_C^{(s)}(\beta^{*(s)} \mid y)(\beta_{j,0} - \hat{\beta}_{j,C}^{(s)})(\beta_{j',0} - \hat{\beta}_{j',C}^{(s)})$$

Theorems 3.2.6 and 3.2.4 imply that $\hat{\beta}_C$ and $\hat{\beta}_C^{(s)}$ for $s = 1, \ldots, p-1$ converges in probability to β_0 for $n \to \infty$ and therefore so does β^* and $\beta^{*(s)}$.

It follows that Z has the same limiting distribution as

$$-(\beta_0 - \hat{\beta}_C)^\top J_C(\beta_0 \mid y)(\beta_0 - \hat{\beta}_C) + \sum_{s=1}^{p-1}\left(\beta_0 - \hat{\beta}_C^{(s)}\right)^\top J_C^{(s)}(\beta_0 \mid y)\left(\beta_0 - \hat{\beta}_C^{(s)}\right). \quad (3.30)$$

By Theorem 3.2.6 it follows that

$$(\beta_0 - \hat{\beta}_C)^\top J_C(\beta_0 \mid y)(\beta_0 - \hat{\beta}_C) = \left(\sqrt{J_C(\beta_0 \mid y)}(\beta_0 - \hat{\beta}_C)\right)^\top \sqrt{J_C(\beta_0 \mid y)}(\beta_0 - \hat{\beta}_C) \sim \chi^2(p-1)$$

since $\sqrt{J_C(\beta_0 \mid y)}(\beta_0 - \hat{\beta}_C) \sim N_{p-1}(0, I_{p-1})$. Similarly, by Theorem 3.2.4 it follows that

$$\sum_{s=1}^{p-1}\left(\beta_0 - \hat{\beta}_C^{(s)}\right)^\top J_C^{(s)}(\beta_0 \mid y)\left(\beta_0 - \hat{\beta}_C^{(s)}\right) \sim \chi^2((p-1)(p-1)).$$

Insertion into Equation (3.30) yields that Z has limiting distribution as the difference between two chi-squared distributions with degrees of freedom $(p-1)(p-1)$ and $(p-1)$ respectively.

Therefore, it follows by Corollary 3.3.1, since Z is strictly positive, that Z is asymptotically chi-squared distributed with $(p-1)(p-1) - (p-1) = (p-1)(p-2)$ degrees of freedom.

□

3.4 Simulation Study

In this section a simulation study will be conducted in order to illustrate the practical implications of Theorems 3.1.2, 3.2.6 and 3.3.2. The implementations and code discussed in this section can be found in Appendix B.

In this simulation study, 1.000 datasets are generated from a Rasch model with $n = 500$, $p = 10$, $\beta_0 = (\beta_{2,0}, \ldots, \beta_{p,0})^\top$ where $\beta_{j,0} = 0.2(j-1)$ for $j = 2, \ldots, p$ and $\theta_0 = (\theta_{1,0}, \ldots, \theta_{n,0})^\top$ where each $\theta_{i,0}$ is a realization of a normal distribution with mean and standard deviation equal to one and are fixed across the 1.000 simulated datasets.

Note that n is chosen as 500 as it is similar in size to the 663 subjects in the test result data. This choice of n also have the implication that JML estimation cannot be conducted on any of the 1.000 simulated datasets, since each of them contains extreme subject scores. Considering the size of n this is also to be expected based on Theorem 3.1.2. The Monte Carlo estimate, and corresponding Monte Carlo error, of the expected number of subjects in each score group can be found in Table 3.1.

	$\mathbf{n_0}$	$\mathbf{n_1}$	$\mathbf{n_2}$	$\mathbf{n_3}$	$\mathbf{n_4}$	$\mathbf{n_5}$	$\mathbf{n_6}$	$\mathbf{n_7}$	$\mathbf{n_8}$	$\mathbf{n_9}$	$\mathbf{n_{10}}$
$\boldsymbol{\mu_{MC}}$	8	25	41	56	66	72	71	64	51	33	13
$\boldsymbol{\sigma_{MC}}$	0.12	0.19	0.26	0.31	0.34	0.35	0.34	0.32	0.28	0.22	0.15

Table 3.1: Monte Carlo estimates μ_{MC} of the expected number of subjects in each score group rounded to integer values, and Monte Carlo error σ_{MC} rounded to two decimal places, based on the subject scores of the 1.000 simulated datasets.

The CML estimates are obtained in R for each dataset using the clogistic function from the Epi package, which is a function for maximizing conditional likelihoods in logistic regression models.

From Theorem 3.2.6 it follows that if the CML estimates $\hat{\beta}_C = (\hat{\beta}_{2,C}, \ldots, \hat{\beta}_{p,C})^\top$ are normalized as $J_C(\beta_0|y)^{1/2}(\hat{\beta}_C - \beta_0)$ then the normalized CML estimates should follow a multivariate normal distribution with mean 0_{p-1} and covariance I_{p-1}.

Therefore, in order to obtain the normalized CML estimates the conditional observed information $J_C(\beta_0 \mid y)$ is implemented in R. Recalling Section 3.2, the entries $J_C(\beta_0 \mid y)$ are given by Equations (3.19) and (3.20) and hence the symmemtric functions must be implemented in R. This is done based on Remark 3.2.1 which states that

$$\gamma_s(\beta) = S_s(\exp(-\beta_1), \ldots, \exp(-\beta_p))$$

for $s = 1, \ldots, p-1$ since elementary symmetric polynomials are easily computed in R using the sum, combn and prod functions.

Furthermore, this approach also immediately yields all other terms in Equations (3.19) and (3.20) as e.g.

$$\gamma_s^{(j)}(\beta) = S_s(\exp(-\beta_1), \ldots, \exp(-\beta_{j-1}), \exp(-\beta_{j+1}), \ldots, \exp(-\beta_p))$$

for $s = 1, \ldots, p-1, j = 2, \ldots, p$.

Based on the above, the conditional joint observation and hence the normalized CML estimates are computed for each dataset.

In Figure 3.1 it is clearly seen that the marginal distribution of the normalized CML estimates seems to follow a standard normal distribution for each item.

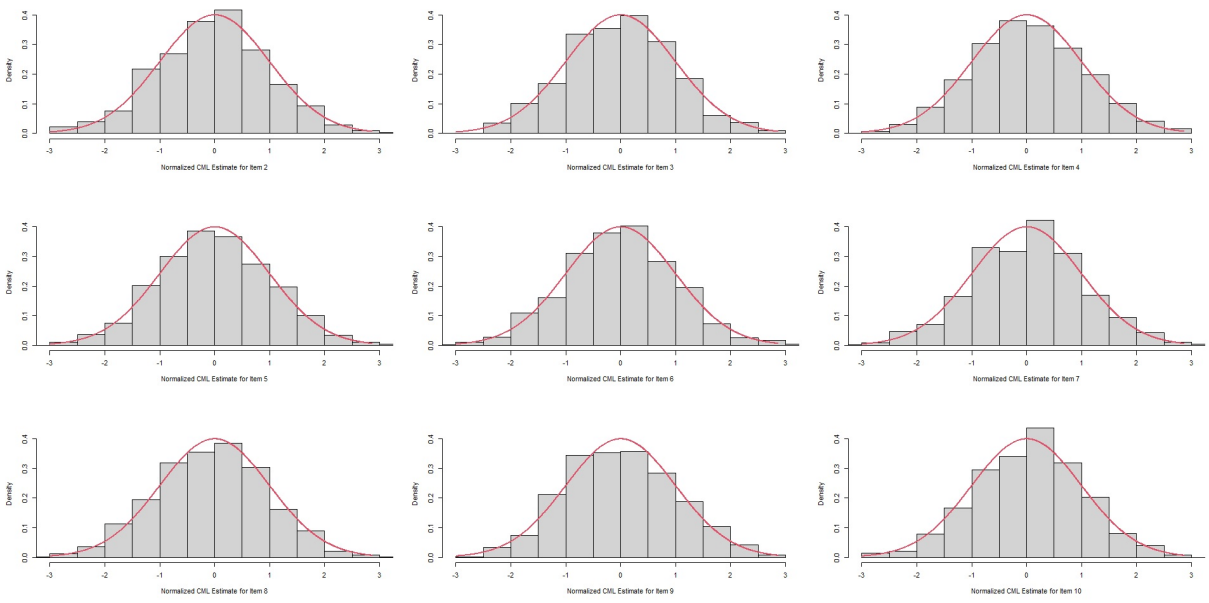

Figure 3.1: Histograms of the normalized CML estimates for each item parameter plotted with the standard normal density for reference and comparison.

This is also supported by applying the Andersen-Darling test for univariate normality using the mvn from the package MVN in R. Here it is found that the test statistic for each item yields a p-value above the corrected significance level when using an overall significance level of 5% and when using the Bonferroni correction since this a multiple hypothesis testing setup.

Furthermore, the mvn function also conducts the Henze-Zirkler test for multivariate normality which is accepted with a p-value of 0.64.

Regarding correlation between the items, the covarariance matrix based on the sample of normalized CML estimates is computed in R and the diagonal values are contained in $(0.95, 1.06)$ while the off-diagonal values belongs to $(-0.08, 0.05)$.

The covariance matrix suggests that there is little to no correlation between the normalized CML item estimates.

Thus all of the above illustrates the asymptotic normality of the normalized CML estimator given by Theorem 3.2.6.

Furthermore, in order to consider the asymptotic distribution of the test statistic of the goodness of fit test described in Section 3.3, the test statistic Z is computed for each simulated dataset.

In Figure 3.2 the GOF test statistics are shown in a histogram together with the density of the χ^2-distribution with $(p-1)(p-2)$ degrees of freedom for comparison as per Theorem 3.3.2.

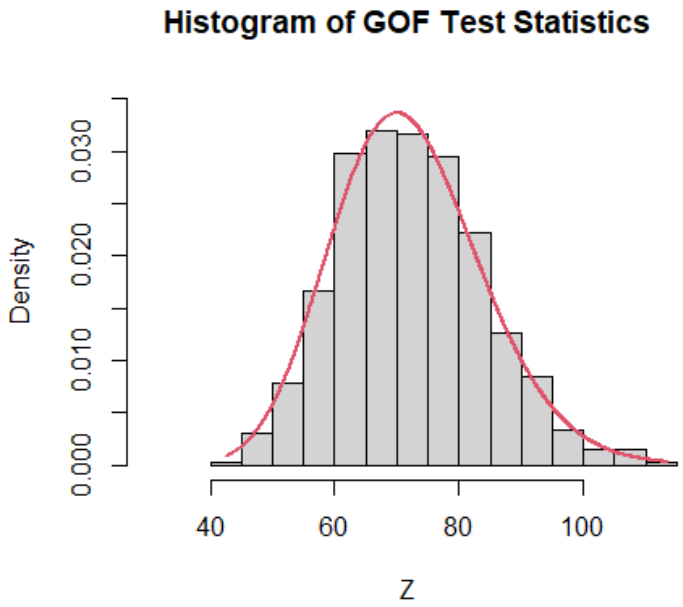

Figure 3.2: Histogram of the GOF test statistic for each simulated dataset compared to the density of the χ^2-distribution with $(p-1)(p-2)$ degrees of freedom.

Furthermore, Figure 3.3 shows a QQ-plot of the GOF test statistics against the χ^2-distribution.

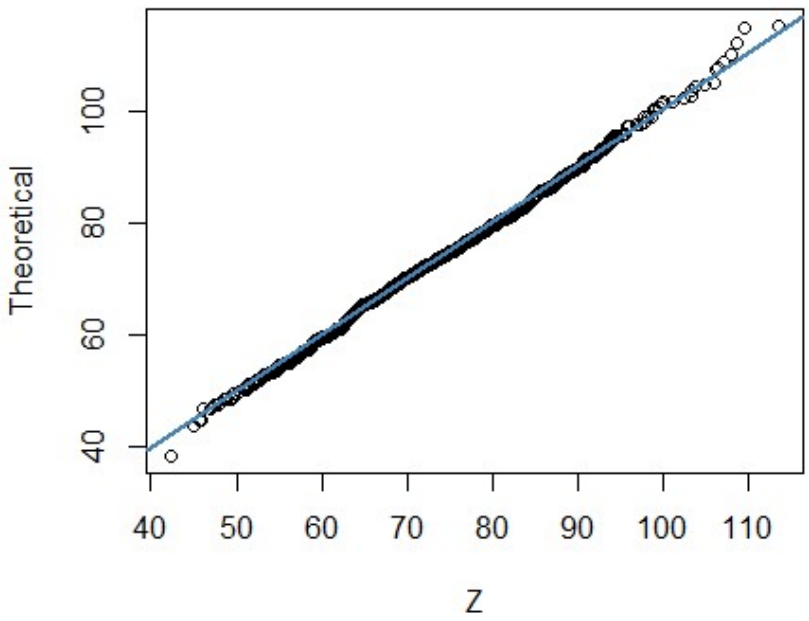

Figure 3.3: QQ-plot of the GOF test statistics against the χ^2-distribution with $(p-1)(p-2)$ degrees of freedom.

While both Figures 3.2 and 3.3 clearly illustrates Theorem 3.3.2, it should be noted that in an application, the most important aspect is the probability of rejecting the null hypothesis.

Therefore, for each dataset it is controlled whether the GOF test statistic is larger than the critical value, or equivalently, whether the p-value is smaller than the significance level. A significance level of 5% yields a critical value of approximately 93, corresponding to the rejection of 54 out of 1000 null hypotheses, i.e. a rejection rate of approximately 5%.

In conclusion, the above analysis supports Theorem 3.3.2 and illustrates that the estimator already exhibits its asymptotic behavior for $n = 500$.

4 | The Rasch Model as a Generalized Linear Mixed Model

The Rasch model was presented in Chapter 2 as a GLM, and within this framework parameter estimation was considered in Chapter 3. Here two estimators were considered, namely the JML estimator which is biased and inconsistent cf. Theorems 3.1.3 and 3.1.4, and the CML estimator which is unbiased and consistent cf. Theorems 3.2.3 and 3.2.5 yet only provides estimates for the item parameters. Hence it is clear that if one is interested in not only the items but also the subjects, another approach is needed.

Furthermore, in many situations the interest is not in individual subjects but rather in the entire subject population. This could for instance be the case when considering the test results data, where a researcher might not be interested in the particular ability level of a specific subject, but rather wish to make conclusions regarding all students of the danish public school similar to those represented in the data.

This is done by imposing a distribution over the subject parameters and hence considering them as random effects rather than fixed effects. Then the parameters of interest are the parameters of the distribution rather than the individual subject ability.

In Section 4.1 generalized linear mixed models (GLMM) are presented in general and the Rasch model is presented within this framework. The Laplace approximation and Gauss-Hermite quadrature is then presented as methods for computing the likelihood function for GLMMs in Section 4.2. Appendix D is a supplement to Section 4.2 regarding Monte Carlo methods for computing the likelihood of a GLMM and simulation from the conditional distribution of random effects given the data using rejection sampling. Then marginal maximum likelihood for the Rasch model is considered in Section 4.3 which is how parameter estimation is conducted when the Rasch model is considered as a GLMM. Then finally in Section 4.4 a simulation study is conducted to illustrate the asymptotic properties of the MML estimator.

4.1 Generalized Linear Mixed Models

This section is based on [25]. In the following a definition of a GLMM is presented which is a direct modification of the definition of a GLM, see A.0.2, to include random effects in the linear predictor.

Definition 4.1.1. Generalized Linear Mixed Model

Let $Y = (Y_1, \ldots, Y_k)$ and $U = (U_1, \ldots, U_m)$ for $k, m \in \mathbb{N}$ be random vectors, and let $X \in \mathbb{R}^{k \times p}, Z \in \mathbb{R}^{k \times m}$ and $\beta \in \mathbb{R}^p$. Furthermore, let $g : M \to \mathbb{R}$ be an intervertible function for $M \subseteq \mathbb{R}$.

Then Y is said to follow a *generalized linear mixed model* with *fixed effects design matrix* X, *fixed effects parameter* β, *random effects design matrix* Z, *random effects* U and *link function* g if the following holds:

- $U \sim N_m(0, \Sigma)$ for some covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$.
- Conditional on $U = u$, the Y_i's for $i = 1, \ldots, k$ are independent, their distributions belongs to the exponential dispersion familiy and $\mathbb{E}[Y_i \mid u] = g^{-1}(\eta_i) \in M$ where $\eta = X\beta + Zu$ is the linear predictor.

Essentially Definition 4.1.1 states that Y follows a GLMM if conditioned on the random effects U, Y follows a GLM.

Example 4.1.2. The Rasch Model as a GLMM Consider the Rasch model and assume that the ability parameter of a random subject follows a normal distribution with mean zero and variance σ^2, i.e.

$$\theta \sim N(0, \sigma^2).$$

This assumption appears rather reasonable as the closely related attribute of IQ is known to follow a normal distribution see e.g. [14]. However, it is not easy to test for this assumption as the random effects are only observed indirectly through binary response patterns.

Conditioned on a realization from the ability distribution, it follows that the probability of solving the jth item is given by

$$\mathbb{P}(Y_j = 1 \mid \theta; \beta_j) = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)}, \quad \text{for } j = 1, \ldots, p.$$

Assuming independence between the ability of the subjects, it follows that the ability parameter vector for all subjects follows a multivariate normal distribution with mean giving by the n-dimensional zero vector and variance $\sigma^2 I_n$, i.e. $\theta = (\theta_1, \ldots, \theta_n) \sim N(0_n, \sigma^2 I_n)$ such that θ constitutes the random effects of the model. Furthermore, the vector consisting of the item difficulties $\beta = (\beta_1, \ldots, \beta_p)$ is the fixed effects parameter. The fixed effects design matrix X is given such that the row corresponding to η_{ij}, say x_{ij}, is given by zeroes in all entries except for -1 in the jth entry. Similarly the random effects design matrix Z is given such that z_{ij}, the row corresponding to the ith subject and jth item, is given by zeroes except for the ith entry which is 1.

The marginal density of Y can be determined as

$$f_Y(y; \beta, \Sigma) = \int_{\mathbb{R}^m} f_{Y,U}(y, u; \beta, \Sigma) \mathrm{d}u = \int_{\mathbb{R}^m} f_{Y|U}(y \mid u; \beta) f_U(u; \Sigma) \mathrm{d}u. \tag{4.1}$$

In the following the densities will not be written with subscript to indicate which distribution the density belongs to. Although this is abuse of notation it will be clear from the context which densities are meant.

Since $f(y; u, \beta)$ is given as a product of densities on the form of Equation (A.1) since $Y_i|(U = u)$ has distribution belonging to the exponential dispersion family and $f(u; \Sigma)$ is the density of a multivariate normal distribution, it follows that $f(y \mid u; \beta) f(u; \Sigma)$ is a fairly complicated function and hence the integral in Equation (4.1) does not have a closed form solution in general.

In Section 4.2 different methods for computing the marginal likelihood will be discussed.

In some situations interest could be at predicting the value of the random effects which led to the observation. This could e.g. be to predict the ability of a subject given the observed response pattern. The following proposition regarding the minimum mean square error predictor is presented in [25] and will not be proven.

Proposition 4.1.3. Conditional Mean as Predictor

The conditional mean

$$\mathbb{E}[U \mid Y = y]$$

is the minimum mean square error predictor of the random effects U.

For applications of 4.1.3 it should be noted that the conditional denisty of the random effects U given the data $Y = y$ is not explicitly given in the model specification of the GLMMs and it follows that the conditional mean would generally be determined as

$$\begin{aligned}\mathbb{E}[U \mid Y = y] &= \int_{\mathbb{R}^m} u f(u \mid y; \beta, \Sigma) \mathrm{d}u \\ &= \int_{\mathbb{R}^m} u \frac{f(y \mid u; \beta) f(u; \Sigma)}{f(y; \beta, \Sigma)} \mathrm{d}u \end{aligned} \tag{4.2}$$

which, like the density of Y, is problematic to evaluate.

Proposition 4.1.4. Score and Observed Information

Let ψ denote the covariance parameter vector, $\theta = (\beta, \psi)^\top$ and let $\tilde{s}(\theta \mid y, u) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y, u; \beta, \Sigma))$ denote the score of the joint likelihood Y and U. Then under regularity conditions it follows that the marginal score is given as

$$s(\theta \mid y) = \mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right] \tag{4.3}$$

and the marginal observed information as

$$\begin{aligned} j(\theta \mid y) = &- \mathbb{E}_\theta \left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \tilde{s}(\theta \mid y, U) \right) \mid Y = y \right] \\ &- \mathbb{V}\mathrm{ar}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right]. \end{aligned} \tag{4.4}$$

Proof. Equation (4.3) follows since

$$\begin{aligned}
s(\theta \mid y) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y; \beta, \Sigma)) \\
&= \frac{1}{f(y; \beta, \Sigma)} \frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathbb{R}^m} f(y, u; \beta, \Sigma) \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(y, u; \beta, \Sigma)}{f(y; \beta, \Sigma)} \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(y, u; \beta, \Sigma)}{f(y, u; \beta, \Sigma)} f(u \mid y; \beta, \Sigma) \mathrm{d}u \\
&= \mathbb{E}_\theta \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(y, U; \beta, \Sigma)) \mid Y = y \right] \\
&= \mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right]
\end{aligned}$$

where the regularity conditions are used in the third equality to interchange the differentiation and the integration.

In order to prove Equation (4.4), first note that

$$\begin{aligned}
j(\theta \mid y) &= -\frac{\mathrm{d}^2}{\mathrm{d}\theta^\top \mathrm{d}\theta} \log(f(y; \beta, \Sigma)) \\
&= -\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right] \\
&= -\int_{\mathbb{R}^m} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} \tilde{s}(\theta \mid y, u) f(u \mid y; \beta, \Sigma) \mathrm{d}u \\
&= -\int_{\mathbb{R}^m} \left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \tilde{s}(\theta \mid y, u) \right) f(u \mid y; \beta, \Sigma) \mathrm{d}u \\
&\quad - \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} f(u \mid y; \beta, \Sigma) \right) \mathrm{d}u \\
&= -\mathbb{E}_\theta \left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \tilde{s}(\theta \mid y, U) \right) \mid Y = y \right] \\
&\quad - \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \frac{f(u, y; \beta, \Sigma)}{f(y; \beta, \Sigma)} \right) \mathrm{d}u
\end{aligned} \tag{4.5}$$

where Equation (4.3) is used in the second equality and the regularity conditions are used in the third equality to interchange the differentiation and the integration.

Since

$$\begin{aligned}
&\int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \frac{f(u, y; \beta, \Sigma}{f(y; \beta, \Sigma)} \right) \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \frac{1}{f(y; \beta, \Sigma)} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} f(u, y; \beta, \Sigma) \mathrm{d}u \\
&\quad + \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) f(u, y; \beta, \Sigma) \frac{\mathrm{d}}{\mathrm{d}\theta^\top} \frac{1}{f(y; \beta, \Sigma)} \mathrm{d}u,
\end{aligned}$$

where the first term can be written as

$$\begin{aligned}
&\int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \frac{1}{f(y; \beta, \Sigma)} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} f(u, y; \beta, \Sigma) \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \frac{f(u \mid y; \beta, \Sigma)}{f(y, u; \beta, \Sigma)} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} f(u, y; \beta, \Sigma) \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \tilde{s}(\theta \mid y, U)^\top f(u \mid y; \beta, \Sigma) \mathrm{d}u \\
&= \mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \tilde{s}(\theta \mid y, U)^\top \mid Y = y \right]
\end{aligned}$$

and the second term as

$$\begin{aligned}
&\int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) f(u, y; \beta, \Sigma) \frac{\mathrm{d}}{\mathrm{d}\theta^\top} \frac{1}{f(y; \beta, \Sigma)} \mathrm{d}u \\
&= \int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) f(u, y; \beta, \Sigma) \left(-\frac{1}{f(y; \beta, \Sigma)^2} \right) \frac{\mathrm{d}}{\mathrm{d}\theta^\top} f(y; \beta, \Sigma) \mathrm{d}u \\
&= -\int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) f(u \mid y; \beta, \Sigma) s(\theta \mid y)^\top \mathrm{d}u \\
&= -\mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right] \mathbb{E}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right]^\top
\end{aligned}$$

it follows that

$$\int_{\mathbb{R}^m} \tilde{s}(\theta \mid y, u) \left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top} \frac{f(u, y; \beta, \Sigma}{f(y; \beta, \Sigma)} \right) \mathrm{d}u = \mathbb{V}\mathrm{ar}_\theta \left[\tilde{s}(\theta \mid y, U) \mid Y = y \right]. \tag{4.6}$$

Inserting Equation (4.6) in Equation (4.5) yields (4.4). □

This section will now be concluded by noting that in order to evaluate the likelihood in Equation (4.1), the conditional mean in Equation (4.2) as well as the expectations and variances in Proposition 4.1.4 some methods are needed.

The following section, which in particular focuses on computation of the likelihood, will present methods that can be utilized for these purposes.

4.2 Computation of Likelihood for GLMMs

This section is based on [25]. In some situtations Equation (4.1) can be simplified by factorizing such that the m-dimensional integral can be replaced by a product of one-dimensional integrals. Consider for instance the case where the random effects are mutually independent and the data can be written on the form $Y = (Y_{ij})_{ij}$ for $i = 1, \ldots, m, j = 1, \ldots, p_i$ where $p_i \in \mathbb{N}$ such that $Y_i = (Y_{ij})_j$ only depends on U through U_i. This is exactly the case for the Rasch model where the random effects $\theta = (\theta_1, \ldots, \theta_n)$ are mutually independent and the response pattern for the ith subject $y_i \in \Gamma(p)$ only depends on θ through θ_i, the ability parameter for that subject, see Example 4.1.2.

Then the marginal likelihood of Y can be written on the form

$$f(y; \beta, \Sigma) = \prod_{i=1}^{m} \int_{\mathbb{R}} f(y_i \mid u_i; \beta) f(u_i; \sigma_i^2) \mathrm{d}u_i \tag{4.7}$$

where σ_i^2 denotes the ith diagonal entry of Σ, i.e. σ_i^2 is the variance of U_i.

In conclusion, methods are needed for computing integral of the form

$$\int_{\mathbb{R}} f(y \mid u; \beta) f(u; \sigma^2) \mathrm{d}u. \tag{4.8}$$

In the following the Laplace approximation and Gauss-Hermite quadrature is presented as methods to approximate the integral. Futhermore, in Appendix D Monte Carlo methods for computing the likelihood of a GLMM is discussed along with a technique to simulate from the conditional distribution $U|(Y = y)$ using rejection sampling.

Laplace Approximation

This subsection regarding the Laplace approximation is based on [27] and [26].

Let $g(u) = \log\left(f(y \mid u;\beta)f(u;\sigma^2)\right)$ and assume there exists $\hat{u} = \arg\max_{u\in\mathbb{R}} g(u)$ such that $g'(\hat{u}) = 0$.

A second order taylor expansion yields

$$g(u) \approx g(\hat{u}) + (u-\hat{u})g'(\hat{u}) + \frac{1}{2}(u-\hat{u})^2 g''(\hat{u}) = g(\hat{u}) - \frac{1}{2}(u-\hat{u})^2(-g''(\hat{u})) \tag{4.9}$$

where it is seen that $\exp(g(u))$ is approximately proportional to a normal density with mean $\mu_{\text{LP}} = \hat{u}$ and variance $\sigma^2_{\text{LP}} = \frac{-1}{g''(\hat{u})}$. Thus the integral in Equation (4.8) can be approximated as

$$\begin{aligned}\int_{\mathbb{R}} f(y \mid u;\beta)f(u;\sigma^2)\mathrm{d}u &= \int_{\mathbb{R}} \exp(g(u))\mathrm{d}u \\ &\approx \exp(g(\hat{u})) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(u-\hat{u})^2(-g''(\hat{u}))\right)\mathrm{d}u \\ &= \exp(g(\mu_{\text{LP}}))\sqrt{2\pi\sigma^2_{\text{LP}}}. \end{aligned} \tag{4.10}$$

Equation (4.10) is called the *Laplace approximation* of the integral given in Equation (4.8).

Furthermore, since

$$f(u \mid y;\beta,\sigma^2) = \frac{f(y \mid u;\beta)f(u;\sigma^2)}{f(y;\beta,\sigma^2)} \propto \exp(g(u))$$

it follows that

$$U|(Y=y) \overset{d}{\approx} N(\mu_{\text{LP}}, \sigma^2_{\text{LP}})$$

yielding that μ_{LP} is an estimate of the conditional mean given in Proposition 4.1.3.

> **Remark 4.2.1.** Even though the attention of the report has been directed towards the one-dimensional integrals thus far, the Laplace approximation can easily be modified to the higher dimensional case simply by utilizing a multivariate Taylor expansion in Equation (4.9).

From the motivation of Equation (4.7) it is clear that

$$f(y \mid u; \beta) = \prod_{j=1}^{p} f(y_j \mid u; \beta)$$

such that the integrand of Equation (4.8) can be written as

$$\exp\left(\sum_{j=1}^{p} \log(f(y_j \mid u; \beta))\right) f(u; \sigma^2). \tag{4.11}$$

In the following, results regarding asymptotic convergence of the Laplace approximation and the order of the convergence will be discussed under certain conditions. The proofs of the results are rather technical and can be found in Appendix C.

Ideally, the results should be presented and proven for integrands on the form $\exp(ph_p(u))g(u)$, because this would include the integrands in Equation (4.11), which is seen by choosing

$$h_p(u) = \frac{1}{p}\sum_{j=1}^{p} \log(f(y_j \mid u)), \quad g(u) = f(u; \sigma^2).$$

However, if the h depends on p this would imply that the associated maximizer u_{LP} also depends on p, making the proof even more complicated and technical.

For simplicity, the results will be presented and proven for integrands of the form $\exp(ph(u))g(u)$ in order to illustrate the asymptotic properties of the Laplace approximation.

With this integrand, clearly the Laplace parameter is given as

$$\mu_{\text{LP},p} = \hat{u}_p = \arg\max_{u \in \mathbb{R}} \left(ph(u) + \log(g(u))\right).$$

In particular it should be noted that $\mu_{\text{LP},p}$ depends on p. However, it is clear that the term $ph(x)$ is asymptotically dominant in the sense that

$$\mu_{\text{LP},p} \underset{p\to\infty}{\longrightarrow} \arg\max_{u \in \mathbb{R}} h(u) =: \hat{u}.$$

For simplicity $\mu_{\text{LP}} = \hat{u}$ is therefore chosen, which is asymptotically equivalent.

Similarly

$$\sigma^2_{\text{LP},p} = \frac{-1}{g''(u)} = \frac{-1}{ph''(u) + \frac{g''(u)g(u) - g'(u)^2}{g(u)^2}}$$

is replaced by the asymptotically equivalent

$$\sigma^2_{\text{LP},p} = \frac{-1}{ph''(u)}.$$

Theorem 4.2.2. Convergence of Laplace Approximation

Let $g, h : \mathbb{R} \to \mathbb{R}$ be functions, assume that h is three times differentiable and that there exists some $\hat{u} \in \mathbb{R}$ such that the following conditions are satisfied:

1. $\hat{u}$ is a local maximum for h, i.e. $H = -h''(\hat{u}) > 0$ and $h'(\hat{u}) = 0$,
2. $\hat{u}$ is a global maximum of h in the sense that

$$\forall \Delta > 0 \exists \epsilon > 0 : |u - \hat{u}| \leq \Delta \implies h(\hat{u}) - h(u) \geq \epsilon,$$

3. $h^{(3)}$ and g is bounded near $\hat{u}$, i.e.

$$\exists \delta > 0 : |u - \hat{u}| \leq \delta \implies |h^{(3)}(u)| \leq K, |g(u)| \leq C \text{ for some } K, C > 0,$$

4. either $\int_{\mathbb{R}} |g(u)| \mathrm{d}u \leq K_a$ or $\int_{\mathbb{R}} \exp(h(u))|g(u)| \mathrm{d}u \leq K_b$ for some $K_a, K_b > 0$.

Then the Laplace approximation of

$$I_p = \int_{\mathbb{R}} \exp(ph(u))g(u) \mathrm{d}x$$

converges to I_p, i.e.

$$\frac{I_p}{\exp(ph(\hat{u}))g(\hat{u})\sqrt{2\pi p^{-1} H^{-1}}} \xrightarrow{p \to \infty} 1.$$

Proof. See Appendix C. □

Theorem 4.2.2 shows not only asymptotic convergence of the Laplace approximation but also that the relative error goes to zero asymptotically.

The following Theorem regarding the order of the relative error expands on this result.

Theorem 4.2.3. Order of the Relative Error of Laplace Approximation
Consider the assumptions of Theorem 4.2.2 and assume further that h is four times differentiable, and extend condition 3 by assuming that $|\hat{u}-u| \leq \delta \implies |h^{(4)}(u)| < C'$ for some $C' > 0$.

Then the relative error is of order (p^{-1}), i.e.

$$\frac{I_p - \exp(ph(\hat{u}))g(\hat{u})\sqrt{2\pi p^{-1}H^{-1}}}{I_p} = O(p^{-1}).$$

Proof. See Appendix C. □

Assume that Theorems 4.2.2 and 4.2.3 can be extended to integrands of the form $\exp(ph_p(u))g(u)$. Then it is clear that when the Laplace approximation is applied in the context of the Rasch model later in the report, see e.g. Section 4.4, then the number of items p will be the determining factor regarding the accuracy of the approximation.

Gauss-Hermite Quadrature

This subsection regarding Guass-Hermite quadrature is based on [25].

The idea of Gauss-Hermite quadrature is to approximate the integral of $f(x)\phi(x)$ as

$$\int_{\mathbb{R}} f(x)\phi(x)\mathrm{d}x \approx \sum_{i=1}^{M} w_i f(x_i). \tag{4.12}$$

where $\phi(\cdot)$ denotes the density of the standard normal distribution, $f : \mathbb{R} \to \mathbb{R}$ is some real function and $w_i, x_i \in \mathbb{R}$ for $i = 1, \ldots, M$, where $M \in \mathbb{N} \setminus \{0\}$ is the number of quadrature points.

For any $M \in \mathbb{N} \setminus \{0\}$, w_i and x_i can be found in an appropiate table, and they are determined as follows. Equation (4.12) should be exact for polynomials of degree less than $2M$. Thus by the linearity of the integral it is enough to solve the following system of equations

$$\int_{\mathbb{R}} x^k \phi(x)\mathrm{d}x = \sum_{i=1}^{M} w_i x_i^k, \quad k = 0, \ldots, 2M-1. \tag{4.13}$$

Since the left-hand side of Equation (4.13) is simply the kth moment of the standard normal distribution, which is known to equal zero for odd k and equal $(k-1)!!$ for even k, it follows that the system of equations given by (4.13) can be written as

$$1[k \text{ even}](k-1)!! = \sum_{i=1}^{M} w_i x_i^k, \quad k = 0, \ldots, 2M-1 \tag{4.14}$$

which has a unique solution cf. [25][Page 10].

Recall that the density of the random effect $f(u;\sigma^2)$ is the density of a normal distribution with mean zero and variance σ^2 such that $\frac{U}{\sigma} \sim N(0,1)$ and hence Equation (4.8) can be written as

$$\int_{\mathbb{R}} f(y \mid u;\beta) f(u;\sigma^2) \mathrm{d}u \overset{x=\frac{u}{\sigma}}{=} \int_{\mathbb{R}} f(y \mid \sigma x;\beta)\phi(x)\mathrm{d}x. \tag{4.15}$$

Directly applying Guass-Hermite quadrature on Equation (4.15) yields

$$\int_{\mathbb{R}} f(y \mid u;\beta) f(u;\sigma^2) \mathrm{d}u \approx \sum_{i=1}^{M} w_i f(y \mid \sigma x_i;\beta) \tag{4.16}$$

and is called *naive Gauss-Hermite quadrature*. The problem with this is that $f(y \mid \sigma x;\beta)$ might be very different from a polynomial and there are therefore no guarantee regarding the quality of the approximation.

Therefore, another approach to Gauss-Hermite quadrature is needed. Consider the Laplace approximation of Equation (4.8) and let ϕ_{LP} denote the density of the normal distribution with mean μ_{LP} and variance σ^2_{LP}. Then

$$\begin{aligned}\int_{\mathbb{R}} f(y \mid u;\beta) f(u;\sigma^2)\mathrm{d}u &= \int_{\mathbb{R}} \frac{f(y \mid u;\beta) f(u;\sigma^2)}{\phi_{\text{LP}}(u)} \phi_{\text{LP}}(u)\mathrm{d}u \\ &\overset{x=\frac{u-\mu_{\text{LP}}}{\sigma_{\text{LP}}}}{=} \int_{\mathbb{R}} \frac{f(y \mid \sigma_{\text{LP}}x+\mu_{\text{LP}};\beta) f(\sigma_{\text{LP}}x+\mu_{\text{LP}};\sigma^2)}{\phi(x)} \sigma_{\text{LP}}\phi(x)\mathrm{d}x\end{aligned} \tag{4.17}$$

and the application of Gauss-Hermite quadrature to Equation (4.17) yields

$$\int_{\mathbb{R}} f(y \mid u;\beta) f(u;\sigma^2)\mathrm{d}u \approx \sum_{i=1}^{M} w_i \frac{f(y \mid \sigma_{\text{LP}}x_i+\mu_{\text{LP}};\beta) f(\sigma_{\text{LP}}x_i+\mu_{\text{LP}};\sigma^2)}{\phi(x_i)} \sigma_{\text{LP}} \tag{4.18}$$

which is called *adaptive Guass-Hermite quadrature*.

Adaptive Guass-Hermite quadrature is generally considered to be more precice than naive Gauss-Hermite. This results from Equation (4.9) which showed that $f(y \mid u; \beta) f(u; \sigma^2)$ was approximately proportional to a normal density with mean μ_{LP} and variance σ^2_{LP}, such that $\frac{f(y \mid u; \beta) f(u; \sigma^2)}{\phi_{\text{LP}}(x)}$ is approximately equal the normalizing constant $f(y; \beta, \sigma^2)$.

Example 4.2.4. Adaptive Gauss-Hermite for Conditional Mean Recall that in order to assume that the likelihood can be factorized as have been done in this section, it must hold that the random effects are mutually independent.

This would however also imply that the condition mean giving by Equation (4.2) can be factorized into one-dimensional integrals on the form

$$\int_{\mathbb{R}} u \frac{f(y \mid u; \beta) f(u; \Sigma)}{f(y; \beta, \Sigma)} \mathrm{d}u.$$

Adaptiv Gauss-Hermite quadrature is for particularly useful when computing the conditional mean. This is seen by considering

$$\int_{\mathbb{R}} u \frac{f(y \mid u; \beta) f(u; \Sigma)}{f(y; \beta, \Sigma)} \mathrm{d}u = \frac{1}{f(y)} \int_{\mathbb{R}} (\sigma_{\text{LP}} x + \mu_{\text{LP}}) \frac{f(y \mid \sigma_{\text{LP}} x + \mu_{\text{LP}}; \beta) f(\sigma_{\text{LP}} x + \mu_{\text{LP}}; \sigma^2)}{\phi(x)} \sigma_{\text{LP}} \phi(x) \mathrm{d}x$$

where

$$(\sigma_{\text{LP}} x + \mu_{\text{LP}}) \frac{f(y \mid \sigma_{\text{LP}} x + \mu_{\text{LP}}) f(\sigma_{\text{LP}} x + \mu_{\text{LP}})}{\phi(x)} \sigma_{\text{LP}}$$

is approximately a first order polynomial by the same arguments as above, and hence is in particular a polynomial of degree below $2M$ for any $M \in \mathbb{N} \setminus \{0\}$.

The following proposition shows that the Laplace approximation is essentially just a special case of adaptive Gauss-Hermite quadrature.

Proposition 4.2.5. Equivalence Between Adaptive Gauss-Hermite and Laplace Approximation

The Laplace approximation given by Equation (4.10) and adaptive Gauss-Hermite quadrature given by Equation (4.18) of the integral

$$\int_{\mathbb{R}} f(y \mid u; \beta) f(u; \sigma^2) \mathrm{d}u$$

coincides when there is one quadrature point.

Proof. For $M = 1$ it follows from Equation (4.13) that x_1 and w_1 is determined as the solution to

$$1 = \int_{\mathbb{R}} x^0 \phi(x) \mathrm{d}x = w_i x_i^0 \tag{4.19}$$

$$0 = \int_{\mathbb{R}} x^1 \phi(x) \mathrm{d}x = w_i x_i^1 \tag{4.20}$$

such that $x_1 = 0$ and $w_1 = 1$.

Adaptive Gauss-Hermite quadrature with $M = 1$ of the integral $\int_{\mathbb{R}} f(y \mid u; \beta) f(u; \sigma^2) \mathrm{d}u$ is given as

$$\begin{aligned}
\int_{\mathbb{R}} f(y \mid u; \beta) f(u; \sigma^2) \mathrm{d}u &\approx w_1 \frac{f(y \mid \sigma_{\text{LP}} x_1 + \mu_{\text{LP}}; \beta) f(\sigma_{\text{LP}} x_1 + \mu_{\text{LP}}; \sigma^2)}{\phi(x_1)} \sigma_{\text{LP}} \\
&= \frac{f(y \mid \mu_{\text{LP}}; \beta) f(\mu_{\text{LP}}; \sigma^2)}{\frac{1}{\sqrt{2\pi}} \exp(0)} \sigma_{\text{LP}} \\
&= \exp\left(\log(f(y \mid \mu_{\text{LP}}; \beta) f(\mu_{\text{LP}}; \sigma^2))\right) \sqrt{2\pi\sigma_{\text{LP}}^2}
\end{aligned}$$

which is exactly the Laplace approximation as given in Equation (4.10). □

Remark 4.2.6. Gauss-Hermite quadrature suffers heavily from the curse of dimensionality since evaluations in a grid in a high dimensional space can be computationally heavy. Guass-Hermite quadrature is therefore most reasonably used for small values of M in high dimensional spaces.

However a large number of quadrature points might be needed in order to obtain a certain precision, and in a high dimensional space this is simply not feasible.

In this situation other methods are needed such as the Monte Carlo methods described in Appendix D.

4.3 Marginal Maximum Likelihood for the Rasch Model

This section is based on [22].

In this section response data on the form $y = (y_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,p}}$ is considered and the Rasch model will be interpreted as an GLMM as in Example 4.1.2 such that

$$\theta = (\theta_1, \dots, \theta_n) \sim N_n(0_n, \sigma^2 I_n) \tag{4.21}$$

and

$$P\left(y_{ij} = 1 \mid \theta_i; \beta_j\right) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \quad \text{for } i = 1, \dots, n, j = 1, \dots, p. \tag{4.22}$$

The purpose of the section is to present a commonly used method for conducting parameter estimation in the Rasch model using marginal maximum likelihood.

Remark 4.3.1. In Remark 3.0.1 it was argued that in order to ensure uniqueness of the parameter estimates it was needed to somehow fix a parameter to the real line. This was done by assuming that $\beta_1 = 0$. However, if the parameter estimates is fixated to the real line in such a way that $\beta_1 = 0$, it would not necessarily be the case that $\mathbb{E}[\theta] = 0_n$. Therefore, the parameter estimates will be shifted relative to Chapter 3 by assuming that $\mathbb{E}[\theta] = 0_n$ instead of $\beta_1 = 0$.

It follows by local stochastic independence and Equation (2.3) that the conditional probability of result pattern y_i given ability parameter θ_i is

$$\begin{aligned} p(y_i \mid \theta_i; \beta) &= \prod_{j=1}^{p} p(y_{ij} \mid \theta_i; \beta_j) \\ &= \prod_{j=1}^{p} \left(\frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)} \right) \\ &= \frac{\exp\left(\theta_i y_{i+}\right) \exp\left(-\sum_{j=1}^{p} \beta_j y_{ij}\right)}{\prod_{j=1}^{p} \left(1 + \exp(\theta_i - \beta_j)\right)}. \end{aligned} \tag{4.23}$$

Since the abilities of the subjects are independent it follows by Equation 4.7 that the *marginal likelihood* is given by

$$\begin{aligned} L_M(\beta, \sigma^2 \mid y) &:= \int_{\mathbb{R}^n} f(y, \theta; \beta, \sigma^2) \mathrm{d}\theta \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} p(y_i \mid \theta_i; \beta) f(\theta_i; \sigma^2) \mathrm{d}\theta_i \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} \frac{\exp\left(\theta_i y_{i+}\right) \exp\left(-\sum_{j=1}^{p} \beta_j y_{ij}\right)}{\prod_{j=1}^{p} \left(1 + \exp(\theta_i - \beta_j)\right)} f(\theta_i; \sigma^2) \mathrm{d}\theta_i. \end{aligned}$$

The choice $(\hat{\beta}_M, \hat{\sigma}^2_M)$ of (β, σ^2) which maximizes the marginal likelihood will be referred to as the *marginal maximum likelihood estimate* (MML estimate) of (β, σ^2).

As discussed in Section 4.2, there are numerous approaches to do parameter estimation for GLMMs in general. Later in the report, when conducting data analysis in Chapter 6, the Laplace approximation and Gauss-Hermite is going to be used as they are standard methods implemented in R. However, for IRT models in particular, the EM-algorithm approach suggested in [22] has seen great use in applications and will now be presented.

Consider the posterior density of θ, that is, the conditional density of θ given the data

$$\begin{aligned} f(\theta \mid y, \beta, \sigma^2) &= \frac{p(y \mid \theta; \beta) f(\theta; \sigma^2)}{p(y; \beta, \sigma^2)} \\ &= \frac{\prod_{i=1}^{n} f(\theta_i; \sigma^2) \prod_{j=1}^{p} p(y_{ij} \mid \theta_i; \beta_j)}{p(y; \beta, \sigma^2)} \end{aligned} \tag{4.24}$$

where the second equality follows from local stochastic independence.

Choose an initial estimate $(\beta^{(0)}, \sigma^{2(0)})$ for (β, σ^2). Often $\beta^{(0)}$ is chosen as the negative of the item scores, which makes sense as e.g. the more difficult items would tend to have a lower item score and hence a greater initial difficulty estimate.

Then iteratively the parameter estimates are updated by maximizing the conditional expectation of $\log(f(y, \theta; \beta, \sigma^2))$ given y and the current parameter estimates. Hence at the kth iteration for some $k \in \mathbb{N}$, $(\beta^{(k+1)}, \sigma^{2(k+1)})$ is obtained as

$$(\beta^{(k+1)}, \sigma^{2(k+1)}) = \max_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+} \mathbb{E}_{(\beta^{(k)},\sigma^{2(k)})}\left[\log\left(f(y, \theta; \beta, \sigma^2)\right) \mid y\right]. \tag{4.25}$$

The procedure will then be terminated once some convergence criteria has been reached, e.g. at the dth iteration for some $d \in \mathbb{N}$, and the MML estimate $(\hat{\beta}_M, \hat{\sigma}^2_M) = (\beta^{(d)}, \sigma^{2(d)})$ is obtained.

Remark 4.3.2. The method defined by Equation (4.25) is known as the EM-algorithm and is a standard method for maximization for missing data problems. The EM-algorithm will be presented and discussed in more detail in Chapter 5 regarding missing data. Intuitively it makes sense to apply the EM-algorithm as an maximization technique for models with random effects since the realizations of the random effects θ_i for $i = 1, \ldots, n$ can be viewed as missing data.

Insertion of Equation (4.24) in the right-hand side of Equation (4.25) yields that

$$\begin{aligned}
&\mathbb{E}_{(\beta^{(k)},\sigma^{2(k)})}\left[\log\left(f(y, \theta; \beta, \sigma^2)\right) \mid y\right] \\
&= \mathbb{E}_{(\beta^{(k)},\sigma^{2(k)})}\left[\log\left(f(\theta \mid y; \beta, \sigma^2)p(y; \beta, \sigma^2)\right) \mid y\right] \\
&= \mathbb{E}_{(\beta^{(k)},\sigma^{2(k)})}\left[\log\left(\prod_{i=1}^{n} f(\theta_i; \sigma^2) \prod_{j=1}^{p} p(y_{ij} \mid \theta_i; \beta_j)\right) \mid y\right] \\
&= \sum_{i=1}^{n} \int_{\mathbb{R}} \log\left(f(\theta_i; \sigma^2)\right) f(\theta_i \mid y; \beta^{(k)}, \sigma^{(2(k))}) \mathrm{d}\theta_i \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{p} \int_{\mathbb{R}} \log\left(p(y_{ij} \mid \theta_i; \beta_j)\right) f(\theta_i \mid y; \beta^{(k)}, \sigma^{(2(k))}) \mathrm{d}\theta_i
\end{aligned} \tag{4.26}$$

should be maximized. It is clear from Equation 4.26 that maximation wrt. σ^2 only involves the first sum and maximization wrt. β_j only involves the summation over i in the double sum with j fixed.

Insertion of the known density of the random effects and Equation (2.3) into Equation 4.26 yields

$$\sum_{i=1}^{n} \int_{\mathbb{R}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\theta_i^2}{2\sigma^2}\right)\right) f(\theta_i \mid y; \beta^{(k)}, \sigma^{(2(k))}) \mathrm{d}\theta_i \tag{4.27}$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{p} \int_{\mathbb{R}} \log\left(\frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)}\right) f(\theta_i \mid y; \beta^{(k)}, \sigma^{(2(k))}) \mathrm{d}\theta_i$$

Therefore, each iteration consists of computing

$$\sigma^{2(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \theta_i^2 f\left(\theta_i \mid y; \beta^{(k)}, \sigma^{2(k)}\right) \mathrm{d}\theta_i \tag{4.28}$$

and solving for β_j in

$$y_{+j} = \sum_{i=1}^{n} \int_{\mathbb{R}} \frac{1}{1 + \exp(\beta_j - \theta_i)} f\left(\theta_i \mid y; \beta^{(k)}, \sigma^{2(k)}\right) \mathrm{d}\theta_i. \tag{4.29}$$

Notice that $\int_{\mathbb{R}} \frac{1}{1+\exp(\beta_j - \theta_i)} f\left(\theta_i \mid y; \beta^{(k)}, \sigma^{2(k)}\right) \mathrm{d}\theta_i = \mathbb{E}_{(\beta^{(k)}, \sigma^{2(k)})}\left[\frac{1}{1+\exp(\beta_j - \theta_i)} \mid y\right]$ which is strictly between zero and one. Therefore, it follows that a solution to Equation (4.29) for each j only exists if the jth item does not have an extreme score. In other words, in order to obtain the MML estimate it is needed that $0 < y_{+j} < n$ for $j = 1, \ldots, p$. In this case it is suggested in [22] to approxmiate the solution to Equation (4.29) by use of the secant method.

Remark 4.3.3. The MML method described in this section can be used in a more general setting with a broader class of models, e.g. IRT models with two (2PL) and three (3PL) item parameters. This is because unlike CML, MML does not utilize the conditional framework presented Section 3.2 which can only be used for the Rasch model cf. Theorem 2.1.1.

Alternatively, other approaches have been presented specifically for the Rasch model where it is still considered as a mixed model, but where the item parameters and the parameters of the ability distribution are estimated separately. In particular, CML is used to estimate the item parameters and then with the estimates inserted into the likelihood function either the EM algorithm, see [3], or the Newton-Raphson procedure applied to the so called population likelihood, see [1], is utilized to estimate the distribution parameters.

4.4 Simulation Study

In this section a simulation study will be conducted in order to investigate the properties of the MML estimator when parameter estimation is conducted in R using the glmer function from the lme4 package, which is a function for fitting GLMM's. The code used in this section has been omitted as it is very similiar to the code in Appendix B and is explained in detail throughout the section.

The glmer function approximates the marginal likelihood using Gauss-Hermite approximation with the default number of quadrature point as one, yielding the Laplace-approximation. Furthermore, the approximated marginal likelihood is then maximized using Nelder-Mead.

If the marginal likelihood could be evaluated exactly then maximizing it would simply yield a maximum likelihood estimator which is known to be unbiased and asymptotically normally distributed under regularity conditions. However, as the marginal likelihood has to be estimated, the goal of this section is to study whether the approximation of the marginal likelihood has influence on the bias and asymptotic distribution of the estimator.

For this purpose, 1.000 datasets are generated from a Rasch model with $n = 500$, $p = 10$, $\beta_0 = (\beta_{1,0}, \ldots, \beta_{p,0})^\top$ where $\beta_{j,0} = 0.2(j-1) - 1$ for $j = 1, \ldots, p$ and $\sigma_0 = 1$. That is, for each dataset the abilities of the students $\theta = (\theta_1, \ldots, \theta_n)^\top$ are first generated, and then the response patterns are generated based on the item parameters β and the subject parameters θ.

Table 4.1 shows that the MML estimator is approximately unbiased, as the Monte Carlo estimates of the expected value of the MML estimator almost agrees perfectly with the actual parameter values.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	σ
(β_0, σ_0)	-1	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1
μ_{MC}	-1.00	-0.80	-0.61	-0.40	-0.20	0.00	0.20	0.41	0.61	0.80	0.98
σ_{MC}	0.004	0.004	0.003	0.004	0.003	0.004	0.004	0.003	0.003	0.004	0.002

Table 4.1: Table containing the true parameter values (β_0, σ_0), the Monte Carlo estimates μ_{MC} of the expected values of the MML estimators rounded off to two decimal places, and Monte Carlo errors σ_{MC} rounded to three decimal places, based on the MML estimates of the 1.000 simulated datasets.

Furthermore, Figure 4.1 suggests that the marginal distribution of the MML estimator agrees with a normal distribution for each parameter.

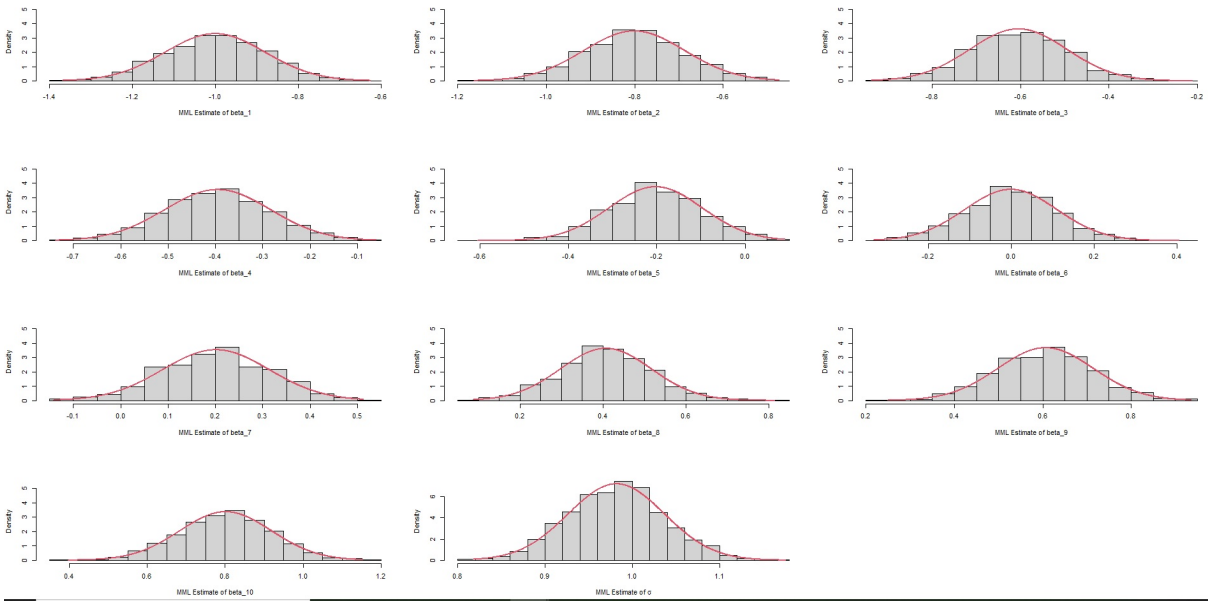

Figure 4.1: Histograms of the MML estimates for each parameter plottet with the normal density with mean and standard deviation estimated from the sample.

This is also supported by applying the Andersen-Darling test for univariate normality using the mvn from the package MVN in R. Here it is found that the test statistic for each item yields a p-value above the corrected significance level when using an overall significance level of 5% and when using the Bonferroni correction since this a multiple hypothesis testing setup.

Furthermore, the mvn function also conducts the Henze-Zirkler test for multivariate normality which is accepted with a p-value of 0.25.

To summarize the results obtained in this section, the fact that the marginal likelihood cannot be evaluated exactly and hence is approximated using the Laplace approximation does not seem to interfere with the unbiasedness nor the asymptotic properties of the ML estimator. This is despite the fact that $p = 10$ was chosen relatively small, which as discussed in Section 4.2 is important regarding the accuracy of the Laplace approximation. Therefore, reasonable approximations can be expected when the Laplace approximation is used on the test result data with $p = 36$ in Chapter 6.

This chapter will be concluded by comparing the accuracy of the conditional and marginal ML estimators using the *root mean squared error* (RMSE).

For $n = 10, 100, 1.000,\ 1000$ datasets is simulated from the Rasch model using the parameters already specified in this section. Then, for each dataset the conditional and marginal ML estimate of each item parameter is obtained using the functions clogistic and glmer in R.

The RMSE for the conditional and marginal ML estimator is given as respectively

$$\text{RMSE}_{\text{cml}} = \left(\sqrt{\mathbb{E}\left[\left(\hat{\beta}_{2,C,n} - \beta_{2,0} \right)^2 \right]}, \ldots, \sqrt{\mathbb{E}\left[\left(\hat{\beta}_{p,C,n} - \beta_{p,0} \right)^2 \right]} \right)^\top$$

and

$$\text{RMSE}_{\text{mml}} = \left(\sqrt{\mathbb{E}\left[\left(\hat{\beta}_{2,M,n} - \beta_{2,0} \right)^2 \right]}, \ldots, \sqrt{\mathbb{E}\left[\left(\hat{\beta}_{p,M,n} - \beta_{p,0} \right)^2 \right]} \right)^\top .$$

The RMSEs are estimated using a Monte Carlo estimate based on the 1.000 simulated datasets for each $n = 10, 100, 1.000$ and can be found in Table 4.2.

n=10	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$\mathbf{RMSE_{cml}}$	-	3.240	2.574	2.861	2.794	2.643	3.107	2.992	3.266	3.702
$\mathbf{RMSE_{mml}}$	3.123	2.726	1.331	1.801	1.042	0.918	1.472	1.944	2.093	2.864
n=100	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$\mathbf{RMSE_{cml}}$	-	0.336	0.334	0.324	0.326	0.311	0.324	0.324	0.334	0.340
$\mathbf{RMSE_{mml}}$	0.263	0.264	0.254	0.252	0.249	0.239	0.254	0.247	0.253	0.245
$n = 1000$	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
$\mathbf{RMSE_{cml}}$	-	0.103	0.106	0.104	0.105	0.102	0.101	0.105	0.110	0.105
$\mathbf{RMSE_{mml}}$	0.082	0.079	0.080	0.077	0.078	0.075	0.077	0.074	0.081	0.078

Table 4.2: Table containing the Monte Carlo estimates of RMSE of respectively the conditional and marginal ML estimator for each item. The estimates are rounded to three decimal places and are based on 1.000 simulated datasets for each $n \in \{10, 100, 1000\}$

From Table 4.2 it is clear that the RMSE reduces as the sample size n increases, which is to be expected as both estimators are consistent and hence the RMSE should converge to zero for each item and estimator.

Table 4.2 also reveals that the RMSE of the marginal ML estimator for each item is smaller than the corrosponding RMSE of the conditional ML estimator for all sample sizes. Hence it is concluded that not only does the marginal ML estimator provide a measure of homogeneity of the subject population, but it also yields more accurate item parameter estimates compared to the conditional ML estimator. This is, of course, under the assumption that the model is correctly specified. For instance, if subject ability was not normal distributed, then this could potentially have great impact on the marginal ML estimator in constrast to the conditional ML estimator which would be unaffected.

5 | Missing Data

Throughout the report thus far focus has been on the Rasch model, i.e. the proposed data generating model for the test result data. However, as explained in Chapter 1, the test result data contains a lot of missing information which cannot reasonably be ignored without careful assessment.

The purpose of this chapter is to present general theory regarding missing data and propose a model for the missingness mechanism present in the test result data.

In Section 5.1 the general framework of missing data is presented including the different types of missingness. Then in Section 5.2 maximum likelihood inference under the assumption that the missingness mechanism is ignorable will be considered. The EM algorithm is introduced in Section 5.3 and some of its properties are discussed. Furtheremore, Appendix E is a supplement to Section 5.3 containing the proof of Theorem 5.3.2 and a result regarding convergence of the EM algorithm.

Finally the Chapter is concluded by Section 5.4 where a model for missingness mechanism for the test result data is proposed.

5.1 Framework

The section is based on [5][Chapter 1 and 2]. In this chapter it is assumed that a study is conducted where information is supposed to be collected on n subjects. The *full data*

$$Y = (Y_1, \ldots, Y_p)$$

is the random real-valued p-dimensional vector that is intended to be collected for each subject.

Furthermore, let

$$R = (R_1, \ldots, R_p) \tag{5.1}$$

denote the random *missingness pattern* for Y, where R_j indicates whether or not the jth component of Y is observed for $j = 1, \ldots, p$, i.e.

$$R_j = \begin{cases} 1 & \text{if } Y_j \text{ is observed,} \\ 0 & \text{otherwise} \end{cases}$$

and let $\bar{R} = (1 - R_1, \ldots, 1 - R_p)$.

For a given missingness pattern $r \in \Gamma(p)$ let $Y_{(r)}$ and $Y_{(\bar{r})}$ denote respectively the observed and the missing subset of Y such that the *observed data* is given as

$$(R, Y_{(R)}).$$

It should be noted that the observed data for a subject depends on two random vectors, namely the full data Y and the random missingness pattern R. The following definition describes different plausible situations regarding the dependence between R and Y.

Definition 5.1.1. Missingness Mechanisms

Let Y denote the full data and R the missingness pattern. Then the missingness mechanism is said to be

- *missing completely at random* (MCAR) if

$$P(R = r \mid Y) = P(R = r),$$

- *missing at random* (MAR) if

$$P(R = r \mid Y) = P(R = r \mid Y_{(r)}), \tag{5.2}$$

- and *missing not at random* (MNAR) if

$$P(R = r \mid Y) \neq P(R = r \mid Y_{(r)}).$$

Remark 5.1.2. Unlike MCAR, which has the simple interpretation that Y and R are indepedent, the meaning of MAR is relatively unclear. Often MAR is interpreted as conditional independence between R and $Y_{(\bar{R})}$ given $Y_{(R)}$ which is incorrect. Note that r appears on both sides of the conditioning symbol in the right-hand side of Equation (5.2) making the statement rather unclear. The correct intepreration of MAR is that the function $y \mapsto P(R = r \mid Y = y)$ takes on the same value for all $y \in \Gamma(p)$ which agrees on all coordinates i where $r_i = 1$ for $i = 1, \ldots, p$. It is however the case that conditional independence implies MAR, and that is exactly the special case of MAR which is usually considered.

While the plausible dependence structures between Y and R described in Definition 5.1.1 are important in the following theoretical derivations, it should be noted that there are obvious practical limitations. For instance, as only the observed data is available, it is not possible to determine whether the data is MAR or MNAR. Assuming MAR, it is however possible to test for MCAR cf. [5][Page 20].

Therefore, when handling missing data, the data analyst must make assumptions regarding the missingness mechanism without use of the data. The analyst must do this based on other information available regarding the study itself such as the methodology of data collection.

Before considering systematic approaches to handle missing data using maximum likelihood estimation and the EM algorithm, the section will be concluded by considering some naive methods commonly used when dealing with missing data. In particular, it will be illustrated why these methods are not advisable to use, dependent on the nature of the missingness mechanism, in particular whether the data is missing at random or not.

Naive Methods

There are a great number of naive methods for handling missing data. In this subsection two such methods will be considered, namely the *complete cases approach* and *last observation carried forward* (LOCF).

In the complete cases approach one simply ignores all the subjects for which some of the data is missing, or equivalenty only considers the *complete cases*, i.e. the subjects for which the full data has been observed. Intuitively, there is an underlying assumption that the complete cases are somehow representative for the whole dataset and hence it is sufficient to only consider the complete cases, i.e. it is assumed that $Y|(R = 1_p)$ has the same distribution as Y.

This need not be the case and the following example demonstrates that the effectiveness of this method greatly depends on this assumption.

Example 5.1.3. Complete Cases for Estimation of Mean

Suppose that Y_i for $i = 1, \dots, n$ are i.i.d real-valued random variables with expected value μ such that R_i indicates whether Y_i has been observed or not.

Then the expected value μ can be estimated by taking the sample mean of the complete cases, i.e.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}. \tag{5.3}$$

In the case of MCAR it is easily seen that the estimator is unbiased since

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[R_i Y_i]}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[R_i]} \\
&= \frac{\mathbb{E}[R_1 Y_1]}{\mathbb{E}[R_1]} \\
&= \frac{\mathbb{E}[R_1]\mathbb{E}[Y_1]}{\mathbb{E}[R_1]} \\
&= \mu
\end{aligned}$$

where the third equality follows from independence between Y_1 and R_1.

By similar arguments it follows from the weak law of large numbers that

$$\hat{\mu} \xrightarrow[n\to\infty]{p} \frac{\mathbb{E}[R_1 Y_1]}{\mathbb{E}[R_1]} = \mu$$

showing that the complete cases sample mean is consistent. It is readily seen that the complete data sample mean under the assumption of MCAR have the same properties as the usual sample mean for the full data. The main difference is essentially just that the complete data sample mean utilizes a smaller sample of size $\sum_{i=1}^{n} R_i$ but from the same distribution.

However, in the case of MNAR it follows that $\mathbb{E}[R_1 Y_1] \neq \mathbb{E}[R_1]\mathbb{E}[Y_1]$ in general, and thus $\hat{\mu}$ is not necessarily unbiased nor consistent. If e.g. it is more likely to observe larger values of Y, i.e. $P(R_i = 1 \mid Y = y)$ is increasing as a function of y then clearly the complete cases sample mean would be positively biased.

LOCF is an approach used in a more specific framework. Consider a *longitudinal* study where the full data $Y = (Y_1, \ldots, Y_p)$ are to be collected at prespecified times $t_1, \ldots, t_p$. Now suppose that there is missing data due to *dropout*, i.e. a subjects drops out of the study at time t_j for some $j \in \{2, \ldots, p\}$ such that $Y_1, \ldots, Y_{j-1}$ is observed and $Y_j, \ldots, Y_p$ is missing. Then the LOCF approach is to replace the missing values in the dataset with the last observed value for each subject. This would imply that a new dataset on the form $Y_{\text{LOCF}} = (Y_1, \ldots, Y_{j-1}, \ldots, Y_{j-1})$ for a subject with dropoup at time t_j will be constructed, which will then be treated as if it is the full data.

Again, as in the complete case approach, it is evident that the succes of the method depends greatly on the missingniss mechanism. Say for instance that the subject is a patient participating in a study regarding his condition, and then the patients condition takes a drastic turn for the worse causing the patient to drop out of the study. Using LOCF the missing data is simply replaced by the last observation such that the new dataset treated as the full data will consist solely of observation before the worsening of the patients condition, which clearly is problematic.

Remark 5.1.4. It should be noted that while dropout is assumed to take place in the test result data as the subjects run out of time, LOCF would clearly not be reasonable approach no matter the type of missingness mechanism due to the Rasch model property of local stochastic independence.

It is clear from the above that other approaches are needed, which will be presented in the following section.

5.2 Maximum Likelihood Estimation under MAR

This section regarding maximum likelihood inference on datasets with missing data is based on [5][Chapter 3].

Consider the *ideal full data* (R, Y) consisting of both the missingness pattern and the full data. While the ideal full data is unattainable in practice, it still relevant to consider its density $p_{R,Y}$. In particular, consider the *selection model factorization* given by

$$p_{R,Y}(r, y) = p_{R|Y}(r \mid y)p_Y(y).$$

Assuming parametric models for both the full data with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ and the missingness mechanism with parameter $\psi \in \Psi \subseteq \mathbb{R}^k$ yields the *selection model framework*, with models on the form

$$p_{R,Y}(r, y; \theta, \psi) = p_{R|Y}(r \mid y; \psi)p_Y(y; \theta).$$

Consider i.i.d observed data $(r_i, y_{i(r_i)})$ for $i = 1, \ldots, n$, let $\mathbf{r} = (r_1, \ldots, r_n), \mathbf{y}_{(\mathbf{r})} = (y_{1(r_1)}, \ldots, y_{n(r_n)})$ and define the *observed data likelihood* as

$$L_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \prod_{i=1}^{n} p_{R,Y_{(R)}}(r_i, y_{i(r_i)}; \theta, \psi).$$

The following definition of the separability condition is a technical prerequisite for most of the theory in the section.

Definition 5.2.1. Separability Condition

Consider the parameters $\theta \in \Theta$ for the full data Y and $\psi \in \Psi$ for the missingness mechanism. Then the *separability condition* is satisfied if the parameter space of $(\theta^\top, \psi^\top)^\top$ is $\Theta \times \Psi$.

The separability condition implies that the range of θ does not depend on ψ and vice verca, and thus intuitively information regarding one of the parameters contains no information regarding the other.

The separability condition together with the MAR assumption implies *ignorability* of the missingness mechanism, meaning that the missingness mechanism can be ignored when considering the observed data likelihood wrt. θ, as the following theorem shows.

Theorem 5.2.2. Ignorability for the Observed Data Likelihood

Assume MAR and the separability condition. Then the *observed data likelihood* wrt. θ is given by

$$L_{obs}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \prod_{i=1}^{n} p_{Y_{(r_i)}}(y_{i(r_i)}; \theta)$$

and does in particular not depend on the missingness mechanism.

Proof. The joint density of $(R, Y_{(R)})$ evaluated at r and $y_{(r)}$ is given as

$$\begin{aligned} p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) &= \int p_{R|Y}(r \mid y; \psi) p_Y(y; \theta) d\nu(y_{(\bar{r})}) \\ &= p_{R|Y_{(r)}}(r \mid y_{(r)}; \psi) \int p_Y(y; \theta) d\nu(y_{(\bar{r})}) \\ &= p_{R|Y_{(r)}}(r \mid y_{(r)}; \psi) p_{Y_{(r)}}(y_{(r)}; \theta) \end{aligned}$$

where the second equality follows from the MAR assumption and ν denotes the Lebesgue measure if Y is continous and the counting measure if Y is discrete.

Thus clearly the observed data likelihood is given as

$$L_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \prod_{i=1}^{n} p_{R|Y_{(r_i)}}(r_i \mid y_{i(r_i)}; \psi) p_{Y_{(r_i)}}(y_{i(r_i)}; \theta)$$

where the first factor regarding the missingness mechanism only depends on ψ and can therefore be ignored by the separability condition, thus yielding the result.

□

It follows immediately from Theorem 5.2.2, assumming MAR and the separability condition, that the *observed data log-likelihood* and the *observed data observed information* is given by respectively

$$\ell_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \sum_{i=1}^{n} \log\left(p_{Y_{(r_i)}}(y_{i(r_i)}; \theta)\right) \tag{5.4}$$

and

$$J_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log(p_{Y_{(r_i)}}(y_{i(r_i)}; \theta)). \tag{5.5}$$

Define the *observed data score* for the ith subject as

$$s_{\text{obs}}(\theta \mid r_i, y_{i(r_i)}) = \frac{\partial}{\partial\theta} \log\left(p_{Y_{(r_i)}}(y_{i(r_i)}; \theta)\right) \tag{5.6}$$

and define the *observed data maximum likelihood estimate* $\hat{\theta}_{\text{obs}}$ as the maximizer of the observed data likelihood, which under regularity conditions can be found as the solution to

$$\sum_{i=1}^{n} s_{\text{obs}}(\theta \mid r_i, y_{i(r_i)}) = 0.$$

Furthermore, define the *observed data Fisher information* as

$$i_{\text{obs}}(\theta) = \mathbb{E}_\theta\left[s_{\text{obs}}(\theta \mid R, Y_{(R)}) s_{\text{obs}}(\theta \mid R, Y_{(R)})^\top\right]. \tag{5.7}$$

In the following it will be shown that observed data maximum likelihood inference is essentially equivalent to usual maximum likelihood inference, where the random variable of interest is simply $(R, Y_{(R)})$ rather than Y. Firstly it will be shown that the observed data likelihood satisfies the two Bartlett identities. Next a result regarding the asymptotic normality of the observed data ML estimate will be presented.

Proposition 5.2.3. Barlett identities

Assume the separability condition and MAR. Then under regularity conditions the observed data maximum likelihood function satisfies the Bartlett identities, i.e.

$$\mathbb{E}[s_{\text{obs}}(\theta \mid R, Y_{(R)})] = 0_d$$

and

$$i_{\text{obs}}(\theta) = \mathbb{E}_\theta[J_{\text{obs}}(\theta \mid R, Y_{(R)})].$$

Proof. Recall from Theorem 5.2.2 that $p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \propto p_{Y_{(r)}}(y_{(r)}; \theta)$ such that

$$\begin{aligned}
\mathbb{E}_\theta[s_{\text{obs}}(\theta \mid R, Y_{(R)})] &= \sum_{r \in \Gamma(p)} \int \frac{\partial}{\partial \theta} \left(\ell_{\text{obs}}(\theta \mid r, y_{(r)}) \right) p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \sum_{r \in \Gamma(p)} \int \frac{\partial}{\partial \theta} p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \frac{\partial}{\partial \theta} \sum_{r \in \Gamma(p)} \int p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \frac{\partial}{\partial \theta} 1 \\
&= 0_d
\end{aligned}$$

where the third equality follows by interchanging the differentiation and the integral, proving the first Bartlett identity.

The second Bartlett identity follows by

$$\begin{aligned}
0_{d \times d} &= \frac{\partial^2}{\partial \theta \partial \theta^\top} \sum_{r \in \Gamma(p)} \int p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \sum_{r \in \Gamma(p)} \int \frac{\partial^2}{\partial \theta \partial \theta^\top} p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \sum_{r \in \Gamma(p)} \int \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta^\top} \log(p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi)) p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \right) \mathrm{d}\nu(y_{(r)}) \\
&= \sum_{r \in \Gamma(p)} \int \frac{\partial^2}{\partial \theta \partial \theta^\top} \log(p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi)) p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&\quad + \sum_{r \in \Gamma(p)} \int \frac{\partial}{\partial \theta^\top} \log(p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi)) \frac{\partial}{\partial \theta} \log(p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi)) \\
&\quad \times p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \psi) \mathrm{d}\nu(y_{(r)}) \\
&= \mathbb{E}_\theta \left[s_{\text{obs}}(\theta \mid R, Y_{(R)}) s_{\text{obs}}(\theta \mid R, Y_{(R)})^\top \right] - \mathbb{E}_\theta \left[J_{\text{obs}}(\theta \mid R, Y_{(R)}) \right]
\end{aligned}$$

where the third equality follows by interchanging the differentiation and the integral. □

Remark 5.2.4. The observed data score and Fisher information is defined based on the likelihood for a single subject, unlike the observed data likelihood, log-likelihood and observed information which is defined wrt. the joint density for all n subjects. This has the implication that e.g.

$$\mathbb{E}_\theta \left[J_{\text{obs}}(\theta \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}) \right] = n i_{\text{obs}}(\theta)$$

by Proposition 5.2.3.

While this may seem confusing, it will simplify notation in the following.

The following Theorem regarding the asymptotic normal distribution of the observed data MLE is based on [5][Page 60] and will not be proven.

Theorem 5.2.5. Asymptotic Normality of Observed Data MLE
Assume the separability condition and MAR. Furthermore, assume that the full data model is correctly specified with true parameter θ_0. Then under regularity conditions

$$\sqrt{n}(\hat{\theta}_{\text{obs}} - \theta_0) \xrightarrow[n\to\infty]{d} N_d(0_d, i_{\text{obs}}(\theta_0)^{-1}).$$

The following lemma will be used in the proof of Proposition 5.2.7 and Theorem 5.3.1.

Lemma 5.2.6. Assume MAR, then for a realization $(r, y_{(r)})$ of $(R, Y_{(R)})$ it holds that

$$p_{Y|R,Y_{(R)}}(y \mid r, y_{(r)}; \theta, \psi)) = p_{Y|Y_{(r)}}(y \mid y_{(r)}; \theta).$$

Proof. The result follows since

$$\begin{aligned} p_{Y|R,Y_{(R)}}(y \mid r, y_{(r)}; \theta, \psi) &= \frac{p_{R,Y}(r, y; \theta, \psi)}{\int p_{R,Y}(r, y; \theta, \psi) \mathrm{d}\nu(y_{(\bar{r})})} \\ &= \frac{p_{R|Y_{(r)}}(r \mid y_{(r)}; \psi) p_Y(y; \theta)}{p_{R|Y_{(r)}}(r \mid y_{(r)}; \psi) \int p_Y(y; \theta) \mathrm{d}\nu(y_{(\bar{r})})} \\ &= p_{Y|Y_{(r)}}(y \mid y_{(r)}; \theta) \end{aligned} \tag{5.8}$$

where MAR is used in the second equality. □

Using Lemma 5.2.6 an important relation between the full data and observed data score can now be derived.

Proposition 5.2.7. Assume the separability condition and MAR. Then under regularity conditions the observed data score is the conditional expection of the full data score given the observed data, i.e.

$$s_{\text{obs}}(\theta \mid r, y_{(r)}) = \mathbb{E}_\theta[s_{\text{full}}(\theta \mid Y) \mid R = r, Y_{(R)} = y_{(r)}]$$

where

$$s_{\text{full}}(\theta \mid y) = \frac{\partial}{\partial \theta} \log\left(p_Y(y; \theta)\right)$$

denotes the full data score.

Proof. Because of the separability condition and MAR, the missingness mechanism is ignorable and hence the observed data score is given as

$$\begin{aligned}
s_{\text{obs}}(\theta \mid r, y_{(r)}) &= \frac{\partial}{\partial \theta} \log(p_{Y_{(r)}}(y_{(r)}; \theta)) \\
&= \frac{\frac{\partial}{\partial \theta} p_{Y_{(r)}}(y_{(r)}; \theta)}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= \frac{\int \frac{\partial}{\partial \theta} p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= \frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= \int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_{Y \mid Y_{(r)}}(y_{(r)}, y_{(\bar{r})} \mid y_{(r)}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \qquad (5.9)
\end{aligned}$$

where the third equality follows from interchanging the order of integration and differentiation.

Applying Lemma 5.2.6 on (5.9) yields

$$\begin{aligned}
s_{\text{obs}}(\theta \mid r, y_{(r)}) &= \int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_{Y \mid R, Y_{(R)}}(y_{(r)}, y_{(\bar{r})} \mid r, y_{(r)}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \\
&= \mathbb{E}_\theta \left[s_{\text{full}}(\theta \mid Y) \mid R = r, Y_{(R)} = y_{(r)} \right].
\end{aligned}$$

□

It should be noted that Proposition 5.2.7 is essentially a missing data result equivalent to Equation (4.3) for GLMMs.

Furthermore, the Proposition can now be used to show the famous missing information principle.

Theorem 5.2.8. Missing Information Principle

Assume the separability condition and MAR. Then under regularity conditions it follows that the full data Fisher information equals the observed data Fisher information plus a term representing the lost information, i.e.

$$i_{\text{full}}(\theta) = i_{\text{obs}}(\theta) + i_{\text{miss}}(\theta)$$

where

$$i_{\text{miss}}(\theta) = \mathbb{E}_\theta \left[\mathbb{V}\text{ar}_\theta[s_{\text{full}}(\theta \mid Y) \mid R, Y_{(R)}] \right].$$

Proof. It is a well known result following from the first Bartlett identity that the full data Fisher information equals the variance of the full data score under regularity conditions. Thus it follows by the law of total variation that

$$\begin{aligned} i_{\text{full}}(\theta) &= \mathbb{V}\text{ar}_\theta[s_{\text{full}}(\theta \mid Y)] \\ &= \mathbb{V}\text{ar}_\theta \left[\mathbb{E}_\theta[s_{\text{full}}(\theta \mid Y) \mid R, Y_{(R)}] \right] + \mathbb{E}_\theta \left[\mathbb{V}\text{ar}_\theta[s_{\text{full}}(\theta \mid Y) \mid R, Y_{(R)}] \right] \\ &= i_{\text{obs}}(\theta) + i_{\text{miss}}(\theta) \end{aligned}$$

where Proposition 5.2.7 and Proposition 5.2.3 are applied in the third equality. □

It is clear that the observed data ML estimate $\hat{\theta}_{\text{obs}}$ can be obtained using the usual numerical optimization algorithms such as the Newton-Raphson algorithm. However, in practice this have proven to be computationally challenging as $p_{Y_{(r)}}(y_{(r)}; \theta)$ for any realization $(r, y_{(r)})$ of $(R, Y_{(R)})$ is not known directly but can be obtained by integrating out the missing subset of the full data density which is assumed to be known, i.e.

$$p_{Y_{(r)}}(y_{(r)}; \theta) = \int p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})}).$$

Therefore other techniques have been developed such as the *Expectation-Maximization* (EM) algorithm presented in the following section.

5.3 Expectation-Maximization Algorithm

This section regarding the EM algorithm is based on [5][Chapter 3].

The section will be started by rewriting the observed data log-likelihood in a way which is convenient for the rest of the section.

Theorem 5.3.1. Observed Data Log Likelihood for EM Algorithm
Assume the separability condition and MAR. Then

$$\ell_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = Q(\theta; \theta') - H(\theta; \theta')$$

for any fixed value $\theta' \in \Theta$ where

$$\begin{aligned} Q(\theta; \theta') &= \mathbb{E}_{\theta'}\left[\ell_{\text{full}}(\theta \mid \mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right], \\ H(\theta; \theta') &= \mathbb{E}_{\theta'}\left[\log\left(p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta)\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \end{aligned}$$

and $\ell_{\text{full}}(\theta \mid \mathbf{Y})$ denotes the full data log-likelihood.

Proof. Since

$$p_{\mathbf{Y}}(\mathbf{y}; \theta) = p_{\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y}_{(\mathbf{r})}; \theta) p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta)$$

it follows that

$$\ell_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \ell_{\text{full}}(\theta \mid \mathbf{y}) - \log(p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta))$$

and hence

$$\begin{aligned} \ell_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) &= \int \ell_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta') \mathrm{d}\nu(\mathbf{y}) \\ &= \int \ell_{\text{full}}(\theta \mid \mathbf{y}) p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta') \mathrm{d}\nu(\mathbf{y}) \\ &\quad - \int \log\left(P_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta')\right) p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y} \mid \mathbf{y}_{(\mathbf{r})}; \theta') \mathrm{d}\nu(\mathbf{y}) \\ &= \mathbb{E}_{\theta'}\left[\ell_{\text{full}}(\theta \mid \mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &\quad - \mathbb{E}_{\theta'}\left[\log\left(p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta)\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &= Q(\theta; \theta') - H(\theta; \theta') \end{aligned}$$

where Lemma 5.2.6 is applied in the third equality, proving the result. □

Theorem 5.3.1 motivates the EM algorithm presented in the following, because it is clear that a potential strategy to maximize the observed data likelihood is simply to maximize $Q(\theta \mid \theta')$ wrt. θ iteratively, assuming that the $H(\theta \mid \theta')$ term can be controlled.

Expectation-Maximization Algorithm

Given initial value $\theta^{(0)} \in \Theta$ the kth iteration of the EM algorithm for $k \in \mathbb{N}_0$ is performed by first evaluating $Q(\theta \mid \theta^{(k)})$ and then maximazing it wrt. θ, referred to as respectively the expectation-step (E-step) and the maximization-step (M-step), i.e.

- *E-Step*:

$$Q(\theta; \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}} \left[\ell_{\text{full}}(\theta \mid \mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})} \right]$$

- *M-step*:

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta; \theta^{(k)}).$$

The algorithm is then terminated when some convergence criteria have been satisfied e.g. if $\|\theta^{(l+1)} - \theta^{(l)}\|$ for some $l \in \mathbb{N}$ gets below a prespecified tolerance for some norm $\|\cdot\|$.

Evidently the goal is that the output of the EM algorithm $\theta^{(l+1)}$ is a reasonable estimate of the observed data MLE $\hat{\theta}_{obs}$, which will be shown later. Before that, however, standard errors will be considered. Since the EM algorithm only returns an approximation of the observed data MLE, standard errors for the estimates must be computed separately. Thus a method for determining the observed data observed information is needed, which is given by the following theorem.

Theorem 5.3.2. Observed Data Observed Information

Assume the separability condition and MAR. Then the observed data observed information is given as

$$\begin{aligned} J_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = &\sum_{i=1}^{n} \mathbb{E}_\theta \left[J_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)} \right] \\ &- \mathbb{E}_\theta \left[s_{\text{full}}(\theta \mid Y_i) s_{\text{full}}(\theta \mid Y_i)^\top \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)} \right] \\ &+ \mathbb{E}_\theta \left[s_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)} \right] \mathbb{E}_\theta \left[s_{\text{full}}(\theta \mid Y_i)^\top \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)} \right] \end{aligned}$$

where $J_{\text{full}}(\theta \mid Y_i)$ denotes the full data observed information for the ith subject.

Proof. See Appendix E. □

Remark 5.3.3. The expression for the observed data observed information in Theorem 5.3.2 demonstrates a conditional version of the missing information principle described in Theorem 5.2.8. In particular, the first term is given by

$$\sum_{i=1}^{n} \mathbb{E}\left[J_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right]$$

and the second and third term together can be written as

$$-\sum_{i=1}^{n} \mathbb{V}\text{ar}\left[s_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right]$$

which when combined yields

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[J_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right] =& J_{\text{obs}}(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) \\ &+ \sum_{i=1}^{n} \mathbb{V}\text{ar}\left[s_{\text{full}}(\theta \mid Y_i) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right].\end{aligned}$$

Taking expectations on both sides of the equality yields Theorem 5.2.8.

Convergence of the EM algorithm will now be considered in the following subsection.

Convergence of the EM Algorithm

This subsection regarding the convergence of the EM algorithm is based on [5][Chapter 3] and [7].

While the EM algorithm is very simple in its formulation, it is not immediately clear why it would result in a reasonable approximation of the observed data MLE since neither the E-step nor the M-step controls the $H(\theta \mid \theta')$ term in Theorem 5.3.1. However, the following theorem shows that each iteration increases the observed data log-likelihood.

Theorem 5.3.4. EM Iteration Increases Log-Likelihood
Assume the separability condition and MAR. Then the observed data log-likelihood increases for each iteration of the EM algorithm, i.e. for $k \in \mathbb{N}_0$ then

$$\ell_{\text{obs}}(\theta^{(k+1)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) \geq \ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}).$$

Proof. It follows by Theorem 5.3.1 by choosing $\theta' = \theta^{(k)}$ and $\theta' = \theta^{(k+1)}$ respectively, that

$$\begin{aligned}&\ell_{\text{obs}}(\theta^{(k+1)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) - \ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) \\ &= Q\left(\theta^{(k+1)}; \theta^{(k)}\right) - Q\left(\theta^{(k)}; \theta^{(k)}\right) - \left(H\left(\theta^{(k+1)}; \theta^{(k)}\right) - H\left(\theta^{k}; \theta^{(k)}\right)\right).\end{aligned}$$

Clearly

$$Q(\theta^{(k+1}; \theta^{(k)}) - Q(\theta^{(k)}; \theta^{(k)}) \geq 0$$

since $\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta; \theta^{(k)})$ by the M-step.

Thus to prove the result it is sufficient to show that

$$H(\theta^{(k+1)}; \theta^{(k)}) - H(\theta^{k}; \theta^{(k)}) \leq 0.$$

By the definition of $H(\cdot \mid \cdot)$ it follows that

$$\begin{aligned}H(\theta; \theta^{(k+1)}) - H(\theta; \theta^{(k)}) &= \mathbb{E}_{\theta^{(k)}}\left[\log\left(p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k+1)})\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &\quad - \mathbb{E}_{\theta^{(k)}}\left[\log\left(p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k)})\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &= \mathbb{E}_{\theta^{(k)}}\left[\log\left(\frac{p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k+1)})}{p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k)})}\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &\leq \log\left(\mathbb{E}_{\theta^{(k)}}\left[\frac{p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k+1)})}{p_{\mathbf{Y}|\mathbf{R},\mathbf{Y}_{(\mathbf{R})}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{Y}_{(\mathbf{R})}; \theta^{(k)})} \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right]\right) \\ &= \log\left(\int p_{\mathbf{Y}|\mathbf{Y}_{(\mathbf{r})}}(\mathbf{y}_{(\mathbf{r})}, \mathbf{y}_{(\bar{\mathbf{r}})} \mid \mathbf{y}_{(\mathbf{r})}; \theta^{(k+1)}) \mathrm{d}\nu(\mathbf{y}_{(\bar{\mathbf{r}})})\right) \\ &= 0\end{aligned}$$

where the inequality follows from Jensen's inequality for conditional expectations since $\log(\cdot)$ is a concave function, proving the result. □

As Theorem 5.3.4 shows that each iteration of the EM algorithm increases the log-likelihood, conditions can now be presented under which the EM algorithm converges which is proven in Appendix E.

In [7] it is suggested to consider the mapping $M : \Theta \to \Theta$ implicitly defined by the EM algorithm as

$$M(\theta^{(k)}) = \theta^{(k+1)}, \quad \text{for } k = 0, 1, \ldots.$$

In particular, Theorem 5.3.4 implies that if $\theta^* \in \Theta$ satisfies that $\ell(\theta^* \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) \geq \ell(\theta \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})$ for all $\theta \in \Theta$ then

$$\ell(M(\theta^*) \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) = \ell(\theta^* \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}).$$

Therefore, if the observed data MLE $\hat{\theta}_{\text{obs}}$ exists, which is an unique maximizer wrt. θ, then

$$M(\hat{\theta}_{\text{obs}}) = \hat{\theta}_{\text{obs}},$$

i.e. the maximum likelihood estimate is a fixpoint for the EM algorithm.

Furthermore, given a fixed point $\theta^* \in \Theta$, it follows by a first order Taylor expansion that

$$M(\theta^{(k)}) \approx M(\theta^*) + \left(\theta^{(k)} - \theta^*\right) \frac{\partial}{\partial \theta} M(\theta^*)$$

which yields

$$\theta^{(k+1)} - \theta^* \approx \left(\theta^{(k)} - \theta^*\right) \frac{\partial}{\partial \theta} M(\theta^*) \tag{5.10}$$

suggesting that near the fixed point θ^* the EM algorithm have approximately linear convergence. In particular, it is shown in [7][Page 10] that the rate of convergence is directly related to the proportion of missing information.

It has been shown that an EM step increases the likelihood function, see Theorem 5.3.4, that under appropriate conditions the EM algorithm converges, see Theorem E.0.1, that the MLE is a fixed point for the EM algorithm and that the EM algorithm converges approximately linearly near its fixed points. However, results regarding convergence of the EM algorithm towards the MLE can be found in [7] and is beyond the scope of this report.

A special familiy of models for which the EM algorithm is particularly convenient will now be considered.

EM Algorithm for Exponential Familiy Models

A major challenge in writing this subsection was modifying derivations and results from [17][Page 2-4] and [15][Page 4,7] in order to generalize the results to natural parameters. This was done in order to prove the results regarding the EM algorithm in the exponential familily case presented in [5][Page 68-69] given by Equations (5.11) and (5.12).

and then modify these results to the case of missing data.

It is evident that the EM algorithm is more convient in situtation where the full data log-likelihood has a simple form and in particular where the E-step and the M-step are easily performed.

As it will be shown in this subsection, the *exponential familiy* is a class of models where to EM algorithm in particularily simple and convenient.

Consider the case where the distribution of Y belongs to the exponential familiy, that is, the density is given by

$$p_Y(y;\theta) = c(y)\exp\left(\sum_{l=1}^{d}\eta_l(\theta)t_l(y) - b(\theta)\right)$$

for functions $c : \mathbb{R}^p \to \mathbb{R}, b : \Theta \to \mathbb{R}$, *natural parameters* $\eta_l(\theta)$ for $l = 1,\dots,d$, where $\eta_l : \Theta \to \mathbb{R}$, and *sufficient statistics* $t_l(y)$ where $t_l : \mathbb{R}^p \to \mathbb{R}$. Note that the natural exponential familiy defined in Definition A.0.1 is a subfamily of the exponential family.

It follows that the joint density for all n i.i.d observations is given by

$$p_{\mathbf{Y}}(\mathbf{y};\theta) := \prod_{i=1}^{n} p_Y(y_i;\theta) = \tilde{c}(\mathbf{y})\exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y}) - \tilde{b}(\theta)\right)$$

where

$$\tilde{c}(\mathbf{y}) = \prod_{i=1}^{n} c(y_i) \quad \tilde{b}(\theta) = nb(\theta), \quad \tilde{t}(\mathbf{y}) = (\tilde{t}_1(\mathbf{y}),\dots,\tilde{t}_p(\mathbf{y}))^\top = \left(\sum_{i=1}^{n} t_1(y_i),\dots,\sum_{i=1}^{n} t_p(y_i)\right)^\top$$

and $\eta(\theta) = (\eta_1(\theta),\dots,\eta_d(\theta))^\top$.

The following result is needed for further study of how the EM algorithm can be simplified when applying to a model beloning the exponential family.

Lemma 5.3.5. Suppose that the distribution of Y belongs to the exponential family such that

$$p_{\mathbf{Y}}(\mathbf{y};\theta) = \tilde{c}(\mathbf{y})\exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y}) - \tilde{b}(\theta)\right)$$

Under regularity conditions it follows that

$$\frac{\partial}{\partial\theta}\tilde{b}(\theta) = \frac{\partial}{\partial\theta}\left(\eta(\theta)^\top\right)\mathbb{E}_\theta\left[\tilde{t}(\mathbf{Y})\right].$$

Proof. Since $p_{\mathbf{Y}}(\mathbf{y};\theta)$ is a density, it follows that

$$1 = \int \tilde{c}(\mathbf{y}) \exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y}) - \tilde{b}(\theta)\right) \mathrm{d}\nu(\mathbf{y}) \implies \tilde{b}(\theta) = \log\left(\int \tilde{c}(\mathbf{y}) \exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y})\right) \mathrm{d}\nu(\mathbf{y})\right)$$

and hence

$$\begin{aligned}
\frac{\partial}{\partial\theta}\tilde{b}(\theta) &= \frac{\partial}{\partial\theta} \log\left(\int \tilde{c}(\mathbf{y}) \exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y})\right) \mathrm{d}\nu(\mathbf{y})\right) \\
&= \frac{\frac{\partial}{\partial\theta}\int \tilde{c}(\mathbf{y}) \exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y})\right) \mathrm{d}\nu(\mathbf{y})}{\exp(\tilde{b}(\theta))} \\
&= \int \frac{\partial}{\partial\theta}\left(\eta(\theta)^\top\right) \tilde{t}(\mathbf{y}) \tilde{c}(\mathbf{y}) \exp\left(\eta(\theta)^\top \tilde{t}(\mathbf{y}) - \tilde{b}(\theta)\right) \mathrm{d}\nu(\mathbf{y}) \\
&= \int \frac{\partial}{\partial\theta}\left(\eta(\theta)^\top\right) \tilde{t}(\mathbf{y}) p_{\mathbf{y}}(\mathbf{y};\theta) \mathrm{d}\nu(\mathbf{y}) \\
&= \frac{\partial}{\partial\theta}\left(\eta(\theta)^\top\right) \mathbb{E}_\theta\left[\tilde{t}(\mathbf{Y})\right]
\end{aligned}$$

where the differentiation and the integral was interchanged in the third equality. □

The following proposition regarding maximum likelihood estimation for exponential family models is a well known result for exponential families.

Proposition 5.3.6. ML Solution Equation for Exponential Family Models
Suppose that the distribution of Y belongs to the exponential family. Then under regularity conditions it follows that the maximum likelihood solution equation is given by

$$\tilde{t}(\mathbf{y}) = \mathbb{E}_\theta\left[\tilde{t}(\mathbf{Y})\right]$$

Proof. As the full data log-likelihood is given by

$$\ell_{\text{full}}(\theta \mid \mathbf{y}) = \log(\tilde{c}(\mathbf{y})) + \eta(\theta)^\top \tilde{t}(\mathbf{y}) - \tilde{b}(\theta)$$

it follows differentiation wrt. θ and setting equal zero yields

$$\frac{\partial}{\partial\theta}\eta(\theta)^\top \tilde{t}(\mathbf{y}) = \frac{\partial}{\partial\theta}\tilde{b}(\theta).$$

By applying Lemma 5.3.5 it follows that the maximum likelihood solution equation is given by

$$\frac{\partial}{\partial\theta}\eta(\theta)^\top\tilde{t}(\mathbf{y}) = \frac{\partial}{\partial\theta}\left(\eta(\theta)^\top\right)\mathbb{E}_\theta\left[\tilde{t}(\mathbf{Y})\right] \implies \tilde{t}(\mathbf{y}) = \mathbb{E}_\theta\left[\tilde{t}(\mathbf{Y})\right]$$

where it is assumed that $\frac{\partial}{\partial\theta}\eta(\theta)^\top$ is invertible.

□

Utilizing the above, it is clear that the EM algorithm becomes particularly simple when considering exponential family models.

Consider the proof of Proposition 5.3.6 but replace the full data log-likelihood with the expected full data log-likelihood given the observed data, i.e.

$$\begin{aligned}Q(\theta;\theta^{(k)}) &= \mathbb{E}_{\theta^{(k)}}\left[\log(\tilde{c}(\mathbf{Y})) + \eta(\theta)^\top\tilde{t}(\mathbf{Y}) - \tilde{b}(\theta) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &= \eta(\theta)^\top\mathbb{E}_{\theta^{(k)}}\left[\tilde{t}(\mathbf{Y}) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] - \tilde{b}(\theta) + \mathbb{E}_{\theta^{(k)}}\left[\log(\tilde{c}(\mathbf{Y})) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right]\end{aligned}$$

which, by differentiating wrt. θ and setting equal zero yields

$$\begin{aligned}&\frac{\partial}{\partial\theta}\eta(\theta)^\top\mathbb{E}_{\theta^{(k)}}[\tilde{t}(\mathbf{Y}) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}] = \frac{\partial}{\partial\theta}\eta(\theta)^\top\mathbb{E}_\theta[\tilde{t}(\mathbf{Y})] \\ &\implies \mathbb{E}_{\theta^{(k)}}[\tilde{t}(\mathbf{Y}) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}] = \mathbb{E}_\theta[\tilde{t}(\mathbf{Y})]\end{aligned}$$

where it is assumed that $\frac{\partial}{\partial\theta}\eta(\theta)^\top$ is invertible and Lemma 5.3.5 is applied. Intuitively this result is the natural extension of Proposition 5.3.6 in the case of missing data, where the sufficient statistic of the full data is replaced by the conditional expectation of the sufficient statistic given the observed data.

In conclusion, if the distribution of Y belongs to the exponential family then the EM algorithm can be simplified to

- Given $\theta^{(k)}$, compute
$$\mathbb{E}_{\theta^{(k)}}[\tilde{t}(\mathbf{Y}) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}] \tag{5.11}$$
- Determine $\theta^{(k+1)}$ as the solution to the system of d equations given by
$$\mathbb{E}_{\theta^{(k)}}[\tilde{t}(\mathbf{Y}) \mid \mathbf{R}=\mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}] = \mathbb{E}_\theta[\tilde{t}(\mathbf{Y})]. \tag{5.12}$$

The following example regarding the use of the EM algorithm in the case of a univariate normal sample demonstrates the simplified version of the EM algorithm for exponential family models.

Example 5.3.7. EM Algorithm for Univariate Normal Sample

Suppose that $Y_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$, $\mathbf{Y} = (Y_1, \dots, Y_n)$ and that $Y_1, \dots, Y_m$ are observed and that $Y_{m+1}, \dots, Y_n$ are missing, i.e. $R_1 = \cdots = R_m = 1, R_{m+1} = \cdots = R_n = 0$ and let $\mathbf{R} = (R_1, \dots, R_n)$ such that $\mathbf{Y}_{(\mathbf{R})} = (Y_1, \dots, Y_m)$. Furthermore, let $(\mathbf{r}, \mathbf{y}_{(\mathbf{r})})$ be a realization of $(\mathbf{R}, \mathbf{Y}_{\mathbf{R}})$. The normal density is given by

$$\begin{aligned} f_Y(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{n\mu^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right) \\ &= c(y)\exp\left(\eta(\mu, \sigma^2)^\top t(Y) - b(\mu, \sigma^2)\right) \end{aligned}$$

where

$$\eta(\theta) = \left(\frac{\mu}{\sigma^2}, \quad -\frac{1}{2\sigma^2}\right), \quad t(y) = (y, y^2), \quad b(\mu, \sigma^2) = \frac{n\mu^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})$$

and $c(y) = 1$, showing that the univariate normal distribution belongs to the exponential family. Therefore, by Equation (5.11) when considering the EM algorithm, taking the conditional expectation of the full data log-likelihood can be replaced by simply taking the conditional expectation of the sufficient statistics, i.e. the computation of

$$Q(\mu, \sigma^2; \mu^{(k)}, \sigma^{2^{(k)}}) = \mathbb{E}_{\mu^{(k)}, \sigma^{2^{(k)}}}\left[\ell_{\text{obs}}(\theta \mid \mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right]$$

is replaced by computing

$$\begin{aligned} \mathbb{E}_{\mu^{(k)}, \sigma^{2^{(k)}}}\left[\tilde{t}(\mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] &= \mathbb{E}_{\mu^{(k)}, \sigma^{2^{(k)}}}\left[\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2\right) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}\right] \\ &= \left(\sum_{i=1}^m y_i + \sum_{i=m+1}^n \mathbb{E}_{\mu^{(k)}, \sigma^{2^{(k)}}}[Y_i], \sum_{i=1}^m y_i^2 + \sum_{i=m+1}^n \mathbb{E}_{\mu^{(k)}, \sigma^{2^{(k)}}}[Y_i^2]\right) \\ &= \left(\sum_{i=1}^m y_i + (n-m)\mu^{(k)}, \sum_{i=1}^m y_i^2 + (n-m)(\sigma^{2^{(k)}} + \mu^{2(k)})\right) \\ &=: (t_1^{(k)}, t_2^{(k)}). \end{aligned}$$

Furthermore, by Equation (5.12) the $(k+1)$th M-step simply consists of solving

$$\left(t_1^{(k)}, t_2^{(k)}\right) = \mathbb{E}_{\mu^{(k)},\sigma^{2(k)}}[\tilde{t}(\mathbf{Y}) \mid \mathbf{R} = \mathbf{r}, \mathbf{Y}_{(\mathbf{R})} = \mathbf{y}_{(\mathbf{r})}] = \mathbb{E}_{\mu,\sigma^2}[\tilde{t}(\mathbf{Y})] = \left(n\mu, n(\sigma^2 - \mu^2)\right)$$

wrt. μ and σ^2, with solution

$$\mu^{(k+1)} = \frac{t_1^{(k)}}{n}, \quad \sigma^{2^{(k+1)}} = \frac{t_2(k)}{n} - \mu^{(k+1)^2}.$$

Note that this example is mainly for illustrative purposes as the observed data maximum likelihood estimates for (μ, σ^2) is simply given by

$$\hat{\mu}_{\text{obs}} = \sum_{i=1}^{m} \frac{y_i}{m}, \quad \hat{\sigma}^2_{\text{obs}} = \sum_{i=1}^{m} \frac{y_i^2}{m} - \hat{\mu}^2.$$

Remark 5.3.8. Monte Carlo EM Algorithm While this subsection focuses on a special case where to EM algorithm is particularly simple and hence more attractive to use, it should be noted that the EM algorithm can be used on a far wider range of models.

In particular, for models where there is a complex model for the full data the E-step might not even be carried out on closed form.

In such situtations it natural to use a Monte Carlo approach as in Section 4.2.

For the ith individual use Monte Carlo approximation to obtain

$$\mathbb{E}_{\theta^{(k)}}\left[\log(p_Y(Y_i;\theta)) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right] \approx \frac{1}{M}\sum_{l=1}^{M} \log\left(p_Y((y_{i(r_i)}, y^l_{i(\bar{r}_i)}));\theta\right)$$

where $y^l_{i(\bar{r})}$ are i.i.d with density $p_{Y_{(\bar{r})}|Y_{(r)}}(y_{(\bar{r})} \mid y_{i(r_i)}; \theta^{(k)})$ for $l = 1, \ldots, M \in \mathbb{N}$. At the $(k+1)$th E-step the *Monte Carlo EM algorithm* consists of computing the numerical approximation given by

$$\begin{aligned} Q(\theta;\theta^{(k)}) &= \sum_{i=1}^{n} \mathbb{E}_{\theta^{(k)}}\left[\log(p_Y(Y_i;\theta)) \mid R_i = r_i, Y_{i(R_i)} = y_{i(r_i)}\right] \\ &\approx \sum_{i=1}^{n} \frac{1}{M}\sum_{l=1}^{M} \log\left(p_Y((y_{i(r_i)}, y^l_{i(\bar{r}_i)}));\theta\right) \end{aligned}$$

and the $(k+1)$th M-step is done by maximizing this numerical approximation.

> It should be noted that the result given in Theorem 5.3.4 regarding each iteration of the EM algorithm increases the log-likelihood is not necessarily true for the Monte Carlo EM algorithm and therefore there are no garentees regarding convergence.
>
> Furthermore, if the Monte Carlo EM algorithm does convergence, it does so approximately linearly by Equation (5.10). This combined with the computationally heavy E-step yields a method which might be rather unwieldy and thus care must taken when using this approach.

5.4 Modelling the Dropout Effect

In this section inspired by [20][Chapter 3] the dropout effect for the test result data will be modelled. Recall Chapter 1 where the idea of dropout in the test result data was first introduced. Here it was vaguely argued that the dropout could potentially contain information relevant to the ability of the subjects and would have to be modelled. With the terminology presented in this chapter these considerations can be reformulated as the dropout might not be ignorable.

In this section several situations are considered for which the missingness is either MNAR or the seperability condition is not satisfied, such that the dropout is not ignorable.

For simplicity it will be assumed in this section that all other missingness than the dropout effect due to the time constraint is ignorable.

Furthermore, the (sequential) steps model as presented in [23] will be used to model the missing data mechanism. Modelling the missing data mechanism in IRT speeded test by the steps model has been investigated in [20][Chapter 3] where it was found using extensive simulation studies that this method could in fact negate the bias which would otherwise occour by ignoring the missing data mechanism.

The steps model specifies the probability of the ith subject to dropout at some item d_i, where $d_i \in \{1, \dots, p+1\}$, recalling that dropout at item d_i implies that

$$r_{i1} = 1, \dots, r_{i(d_i-1)} = 1, r_{id_i} = 0, \dots, r_{ip} = 0,$$

where $d_i = p + 1$ is interpreted as no dropout, i.e. $r_{i1} = 1, \dots, r_{ip} = 1$.

The steps model is given by

$$P(r_{i1} = 1, \ldots, r_{i(d_i-1)} = 1, r_{id_i} = 0, \ldots, r_{ip} = 0) = p(d_i; \xi_i, \delta) = \left(\prod_{j=1}^{d_i-1} \tilde{p}_{ij} \right) (1 - \tilde{p}_{id_i})$$

where

$$\tilde{p}_{ij} = \frac{\exp(\xi_i - \delta_j)}{1 + \exp(\xi_i - \delta_j)}, \quad \tilde{p}_{i(p+1)} = 0, \quad \text{for } i = 1, \ldots, n, \ j = 1, \ldots, p \tag{5.13}$$

and $\delta = (\delta_1, \ldots, \delta_n)^\top$.

It should be noted that while Equation (5.13) agrees with Equation (2.2) in the specification of the Rasch model, of course with different parameters, there is still a fundamental distinction between the steps model and the Rasch model.

Under the Rasch model, local stochastic independence is assumed such that given the parameters, a given subject answers each item independently. However for the steps model $r_{ij} = 1$ implies that $r_{i\tilde{j}} = 1$ for $\tilde{j} = 1, \ldots, j$ and similarly $r_{ij} = 0$ implies $r_{i\tilde{j}} = 0$ for $\tilde{j} = j, \ldots, p$.

Furthermore it should be noted that the parameters of the steps model have a different interpretation as well.

The subject parameter ξ_i is interpreted as the speed of the ith subject, such that a large ξ_i would imply a high probability that the ith subject can respond to each item without running out of time.

Similarly, the item parameter δ_j represents the total workload needed to respond to all items before and including the jth item.

The joint model for the ideal full data will now be specified in the case where all the parameters are considered as fixed effects. Note that the model specification when considering random effects are equivalent to the fixed effects case when conditioning on the random effects.

Let $r_{i(-d_i)}$ denote all missingness other than the dropout s.t. r_i and $(r_{i(-d_i)}, d_i)$ are equivalent for $i = 1, \ldots, n$, and let $\psi = (\xi^\top, \delta^\top, \tilde{\psi}^\top)^\top$ denote the missingness parameters, were $\tilde{\psi}$ denotes the missingness parameters not associated to the dropout.

It is assumed that the dropout occurs as specified by the steps model independently of the full data generated from the Rasch model, and that given the dropout and full data ignorable missingness occurs, i.e.

$$\begin{aligned} p(r_i, y_i; \theta, \beta, \psi) &= p(r_{i(-d_i))}, d_i, y_i; \theta_i, \beta, \psi) \\ &= p(r_{i(-d_i)} \mid d_i, y_i; \tilde{\psi}) p(d_i; \xi_i, \delta) p(y_i; \theta_i, \beta). \end{aligned}$$

The definition of ignorability is slightly different when it is only assumed for part of the missingness mechanism compared to when ignorability is assumed for the whole missingness mechanism.

Specifically, in this context the assumption of MAR is replaced with

$$p(r_{i(-d_i)} \mid d_i, y_i; \tilde{\psi}) = p(r_{i(-d_i)} \mid d_i, y_{i(r_i)}; \tilde{\psi}).$$

Furthermore, the assumption of separability between the missingness parameters ψ and parameters for the full data model θ, β is replaced with separability between the parameter for the ignorable part of the missingness mechanism $\tilde{\psi}$ and the parameters for the dropout as well as the parameters for the full data model $\theta, \xi, \beta, \delta$. It is clear that these assumptions implies that the missingness not due to dropout is ignorable, since

$$\begin{aligned} p(r_i, y_{i(r_i)}; \theta_i, \beta, \psi) &= \int p(r_i, y_i; \theta_i, \beta, \psi) \mathrm{d}\nu(y_{i(\bar{r}_i)}) \\ &= \int p(r_{i(-d_i)} \mid d_i, y_i; \tilde{\psi}) p(d_i; \xi_i, \delta) p(y_i; \theta_i, \beta) \mathrm{d}\nu(y_{i(\bar{r}_i)}) \\ &= p(r_{i(-d_i)} \mid d_i, y_{i(r_i)}; \tilde{\psi}) \int p(d_i; \xi_i, \delta) p(y_i; \theta_i, \beta) \mathrm{d}\nu(y_{i(\bar{r}_i)}) \\ &= p(r_{i(-d_i)} \mid d_i, y_{i(r_i)}; \tilde{\psi}) p(d_i, y_{i(r_i)}; \theta_i, \xi_i, \beta, \delta) \\ &\propto p(d_i, y_{i(r_i)}; \theta_i, \xi_i, \beta, \delta), \end{aligned} \tag{5.14}$$

where the proportionality is wrt. the parameters of interest $\theta, \xi, \beta, \delta$. However, considerations has to made regarding $p(d_i, y_{i(r_i)}; \theta_i, \xi_i, \beta, \delta)$.

Clearly D_i and $Y_{i(R_i)}$ are not independent, but since D_i and Y_i are, it follows that

$$\begin{aligned} p(d_i, y_{i(r_i)}; \theta_i, \xi_i, \beta, \delta) &= \int p(d_i, y_i; \theta_i, \xi_i, \beta, \delta) \mathrm{d}\nu(y_{i\bar{r}_i}) \\ &= p(d_i; \xi_i, \delta) \int p(y_i; \theta_i, \beta) \mathrm{d}\nu(y_{i\bar{r}_i}) \\ &= p(d_i; \xi_i, \delta) p(y_{i(r_i)}; \theta_i, \beta) \end{aligned} \tag{5.15}$$

where it should be noted that the above argument is a MCAR version of the proof of Theorem 5.2.2.

Equations (5.14) and (5.15) yields that

$$p(r_i, y_{i(r_i)}; \theta_i, \beta, \psi) \propto p(y_{i(r_i)}; \theta_i, \beta) p(d_i; \xi_i, \delta), \quad \text{for } i = 1, \ldots, n \tag{5.16}$$

which is of great importance in the rest of the section.

In the following possible forms of δ and ξ will be proposed.

It is clear that δ_j should be monotonically increasing as a function of j. Since the subjects does not run out of time at the first few items, it is not be possible to estimate δ_j for small values of j unless some restrictions are made on the form of the item parameters. For simplicity it will be assumed that

$$\delta_j = \tau + j\eta \tag{5.17}$$

where τ is a baseline level and η represents the monotone increasing log odds of dropout with increasing item number.

Remark 5.4.1. The assumption of a monotonically increasing δ_j as a function of j as seen in Equation (5.17) is an explicit usage of the assumption that all subjects try to solve the items in order of the item enumeration.

If the subject speed parameters $\xi = (\xi_1, \ldots, \xi_n)^\top$ along with τ and η are simply assumed to be fixed parameters unrelated to the subject speed θ or item difficulty β, then clearly the separability condition is satisfied. This, in combination with Equation (5.16) implies ignorability of the entire missingness mechanism, yielding the case considered in Section 6.1. While this would be particularly simple, the modelling of the dropout using the steps model would contribute with no new information regarding the data.

Furthermore, one issue is that intuitively there ought to be correlation between the ability and speed of a subject.

Another option is to model the speed deterministically given the ability as

$$\xi_i = \alpha\theta_i, \quad \text{for } \alpha \in \mathbb{R}, i = 1, \ldots, n. \tag{5.18}$$

In this case the distribution of the dropout depends explicitly on the ability of the subject, implying that the separabililty conditions is not satisfied.

Let $\psi = (\theta^\top, \alpha, \tau, \eta, \tilde{\psi}^\top)^\top$ denote the parameter vector of all missingness where $\tilde{\psi}$ is the parameter vector for the missingness not due to dropout, which is assumed to be ignorable.

The joint likelihood is then on the form

$$\begin{aligned} L_{J,\text{obs}}(\theta, \alpha, \beta, \tau, \eta \mid r, y_{(r)}) &= p(r, y_{(r)}; \theta, \beta, \psi) \\ &= \prod_{i=1}^{n} p(r_i, y_{i(r_i)}; \theta_i, \beta, \psi) \\ &\propto \prod_{i=1}^{n} p(y_{i(r_i)}; \theta_i, \beta) p(d_i; \theta_i, \alpha, \tau, \eta), \end{aligned}$$

where the proportionality follows by Equation (5.16).

By similar arguments it follows that

$$\begin{aligned} p(r, y_{(r)}; \theta, \beta, \psi) &\propto \prod_{i=1}^{n} p(y_{i(r_i)}; \theta_i, \beta) p(d_i; \theta_i, \alpha, \tau, \eta) \\ &= \prod_{i=1}^{n} p(y_{i(r_i)} \mid y_{i(r_i)+}; \beta) p(y_{i(r_i)+}; \theta_i, \beta) p(d_i; \theta_i, \alpha, \tau, \eta) \end{aligned}$$

and hence the conditional likelihood is given as

$$L_{C,\text{obs}}(\beta \mid r, y_{(r)}) = \prod_{i=1}^{n} p(y_{i(r_i)} \mid y_{i(r_i)+}; \beta)$$

which is simply a complete cases version of the conditional likelihood, which is also considered in Section 6.1.

Furthermore, by considering the GLMM framework, and letting $\psi = (\alpha, \tau, \eta, \tilde{\psi}^\top)^\top$, the marginal likelihood is given by

$$\begin{aligned} L_{M,\text{obs}}(\alpha, \beta, \tau, \eta, \sigma^2 \mid r, y_{(r)}) &:= \int_{\mathbb{R}^n} f(r, y_{(r)}, \theta; \beta, \sigma^2, \psi) \mathrm{d}\theta \\ &= \int_{\mathbb{R}^n} p(r, y_{(r)} \mid \theta; \beta, \psi) f(\theta; \sigma^2) \mathrm{d}\theta \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} p(r_i, y_{i(r_i)} \mid \theta_i; \beta, \psi) f(\theta_i; \sigma^2) \mathrm{d}\theta_i \\ &\propto \prod_{i=1}^{n} \int_{\mathbb{R}} p(y_{i(r_i)} \mid \theta_i; \beta) p(d_i \mid \theta_i; \alpha, \tau, \eta) f(\theta_i; \sigma^2) \mathrm{d}\theta_i. \end{aligned}$$

Parameter estimation can be conducted from either the observed data joint, conditional or marginal likelihood presented above.

However, this has been omitted since the assumption of the deterministic relationship between the ability and speed of subjects given by Equation (5.18) is too restrictive.

Intuitively, it is quite clear that there is a distinction between the ability of a subject and the speed of said subject. One could easily imagine a diligent subject who has a high θ, yet is slow and rigorous in his work and hence has a low ξ. On the other hand, it is also intuitively clear that there ought to be a dependence structure between the two since a high level of ability in general would imply that the subject can solve the items quicker. Here it is also worth noting that ξ measures the speed of a subject in regards to responding to items but not necessarily solving them correctly, and hence a student with low θ could potentially also have a high ξ.

Therefore, by considering the GLMM framework, it will be assumed that

$$(\theta, \xi) \sim N_2(0_2, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_\theta^2 & \rho\sigma_\theta\sigma_\xi \\ \rho\sigma_\theta\sigma_\xi & \sigma_\xi^2 \end{bmatrix}. \tag{5.19}$$

Remark 5.4.2. Consider the interpretation of MAR presented in Remark 5.1.2 and suppose $\rho \neq 0$. Then clearly, if $y_{(\bar{r})}$ is a vector of ones then the conditional distribution of θ given $Y = y$ is skewed such that higher values of θ are more likely than if e.g. $y_{(\bar{r})}$ is a vector of zeroes. But as $\rho \neq 0$, this implies that the value of $y_{(\bar{r})}$ influences the conditional distribution of ξ given $Y = y$, which again influences the conditional distribution of R given $Y = y$. In conclusion it is evident that the mapping $y \mapsto P(R = r \mid Y = y)$ does not coincide for all $y \in \Gamma(p)$ with the same $y_{(r)}$ but potentially different $y_{(\bar{r})}$.

On the other hand if $\rho = 0$ then clearly the missingness mechanism is MCAR.

Let $\lambda = (\sigma_\theta, \sigma_\xi, \rho)^\top$ denote the covariance parameter vector for Σ, $\psi = (\tau, \eta, \tilde{\psi}^\top)^\top$ denote the missingness parameters and $v = (\beta^\top, \tau, \eta, \lambda^\top)^\top$ denote the parameters of interest. Since the abilities and the speeds between the subjects are independent it follows by Equation (4.7) that the marginal likelihood is given by

$$\begin{aligned} L_M(v \mid r, y_{(r)}) &= \int_{\mathbb{R}^{2n}} f(r, y_{(r)}, \theta, \xi; \beta, \psi, \lambda) \mathrm{d}\xi \mathrm{d}\theta \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}^2} p(r_i, y_{i(r_i)} \mid \theta_i, \xi_i; \beta, \psi) f(\theta_i, \xi_i; \lambda) \mathrm{d}\xi_i \mathrm{d}\theta_i \\ &\propto \prod_{i=1}^{n} \int_{\mathbb{R}^2} p(y_{i(r_i)} \mid \theta_i; \beta) p(d_i \mid \xi_i; \tau, \eta) f(\theta_i, \xi_i; \lambda) \mathrm{d}\xi_i \mathrm{d}\theta_i. \end{aligned} \tag{5.20}$$

In Section 6.2 focus will be on the implementation and maximization of this marginal likelihood.

Furthermore, the marginal score and observed information can be obtained using Proposition 4.1.4. The proposition yields that the score is given by

$$s_M(v \mid y_{(r)}, r) = \sum_{i=1}^{n} \mathbb{E}_v \left[\tilde{s}_{iM}(v \mid y_{i(r_i)}, r_i, \theta_i, \xi_i) \mid r_i, y_{i(r_i)} \right] \tag{5.21}$$

where

$$\begin{aligned} \tilde{s}_{iM}(v \mid y_{i(r_i)}, r_i, \theta_i, \xi_i) &= \frac{\partial}{\partial v} \log(f(y_i, r_i, \theta_i, \xi_i; \beta, \psi, \lambda)) \\ &= \frac{\partial}{\partial v} \log \left(p(y_{i(r_i)}; \theta_i \mid \beta) p(d_i; \xi_i \mid \tau, \eta) f(\theta_i, \xi_i \mid \lambda) \right) \\ &=: \frac{\partial}{\partial v} g_i(\theta_i, \xi_i). \end{aligned} \tag{5.22}$$

The components of $\tilde{s}_{iM}(v \mid y_{i(r_i)}, r_i, \theta_i, \xi_i)$ will now be derived.

Since

$$\begin{aligned}\frac{\partial}{\partial \beta_j} p(y_{ij} \mid \theta_i; \beta_j) &= \frac{\partial}{\partial \beta_j} \frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)} \\ &= \frac{y - {}_{ij} \exp(y_{ij}(\theta_i - \beta_j))(1 + \exp(\theta_i - \beta_j)) + \exp(\theta_i - \beta_j) \exp(y_{ij}(\theta_i - \beta_j))}{(1 + \exp(\theta_i - \beta_j))^2}\end{aligned}$$

such that

$$\frac{\partial}{\partial \beta_j} \log(p(y_{ij} \mid \theta_i; \beta_j)) = p_{ij} - y_{ij}$$

and hence

$$\begin{aligned}\frac{\partial}{\partial \beta_j} g_i(\theta_i, \xi_i) &= \frac{\partial}{\partial \beta_j} \log\left(p(y_{i(r_i)} \mid \theta_i; \beta)\right) \\ &= r_{ij}(p_{ij} - y_{ij}), \quad \text{for } j = 1, \ldots, p. \end{aligned} \tag{5.23}$$

Similarly, since

$$\begin{aligned}\frac{\partial}{\partial \tau} \tilde{p}_{ij} &= \frac{\partial}{\partial \tau} \frac{\exp(\xi_i - (\tau + j\eta))}{1 + \exp(\xi_i - (\tau + j\eta))} \\ &= \frac{-\exp(\xi_i - (\tau + j\eta))(1 + \exp(\xi_i - (\tau + j\eta))) + \exp(\xi_i - (\tau + j\eta)) \exp(\xi_i - (\tau + j\eta))}{(1 + \exp(\xi_i - (\tau + j\eta)))^2}\end{aligned}$$

it follows that

$$\frac{\frac{\partial}{\partial \tau} \tilde{p}_{ij}}{\tilde{p}_{ij}} = \tilde{p}_{ij} - 1$$

and

$$\frac{\frac{\partial}{\partial \tau} \tilde{p}_{id_i}}{1 - \tilde{p}_{id_i}} = -\tilde{p}_{id_i}$$

such that

$$\begin{aligned}\frac{\partial}{\partial \tau} g_i(\theta_i, \xi_i) &= \frac{\partial}{\partial \tau} \log\left(p(d_i \mid \xi_i; \tau, \eta)\right) \\ &= \left(\sum_{j=1}^{d_i - 1} \frac{\partial}{\partial \tau} \log(\tilde{p}_{ij})\right) + \frac{\partial}{\partial \tau} \log(1 - \tilde{p}_{id_i}) \\ &= \left(\sum_{j=1}^{d_i - 1} \tilde{p}_{ij} - 1\right) + \tilde{p}_{id_i} \\ &= \left(\sum_{j=1}^{d_i} \tilde{p}_{ij}\right) - (d_i - 1) \end{aligned} \tag{5.24}$$

and similarly

$$\frac{\partial}{\partial \eta} g_i(\theta_i, \xi_i) = \left(\sum_{j=1}^{d_i} j\tilde{p}_{ij})\right) - \left(\sum_{j=1}^{d_i - 1} j\right). \tag{5.25}$$

Furthermore,

$$\begin{aligned}\frac{\partial}{\partial\sigma_\theta}g_i(\theta_i,\xi_i) &= \frac{\partial}{\partial\sigma_\theta}\log\left(f(\theta_i,\xi_i;\lambda)\right)\\ &= -\frac{\partial}{\partial\sigma_\theta}\log(2\pi\sigma_\theta\sigma_\xi(1-\rho^2)) + \frac{\partial}{\partial\sigma_\theta}\left(-\frac{1}{2(1-\rho^2)}\left(\frac{\theta_i^2}{\sigma_\theta^2}+\frac{\xi_i^2}{\sigma_\xi^2}-2\rho\frac{\theta\xi}{\sigma_\theta\sigma_\xi}\right)\right)\\ &= -\frac{1}{\sigma_\theta}+\frac{\theta_i^2}{(1-\rho^2)\sigma_\theta^3}-\frac{\rho\theta\xi}{(1-\rho^2)\sigma_\theta^2\sigma_\xi}\end{aligned} \tag{5.26}$$

which by symmetry also implies that

$$\begin{aligned}\frac{\partial}{\partial\xi}g_i(\theta_i,\xi_i) &= \frac{\partial}{\partial\sigma_\xi}\log\left(f(\theta_i,\xi_i;\lambda)\right)\\ &= -\frac{1}{\sigma_\xi}+\frac{\xi_i^2}{(1-\rho^2)\sigma_\xi^3}-\frac{\rho\theta\xi}{(1-\rho^2)\sigma_\xi^2\sigma_\theta}.\end{aligned} \tag{5.27}$$

Lastly, since

$$\begin{aligned}\frac{\partial}{\partial\rho}g_i(\theta_i,\xi_i) &= \frac{\partial}{\partial\rho}\log\left(f(\theta_i,\xi_i;\lambda)\right)\\ &= \frac{\rho}{1-\rho^2}-\frac{\rho}{(1-\rho^2)^2}\left(\frac{\theta_i^2}{\sigma_\theta^2}+\frac{\xi_i^2}{\sigma_\xi^2}\right)+\frac{1+\rho^2}{(1-\rho^2)^2}\frac{\theta\xi}{\sigma_\theta\sigma_\xi}\end{aligned} \tag{5.28}$$

it follows that the components of $\tilde{s}_{iM}(v \mid y_{i(r_i)}, r_i, \theta_i, \xi_i)$ are given by Equations (5.23), (5.24), (5.25), (5.26), (5.27) and (5.28).

Thus the marginal score $s_M(v \mid y_{(r)}, r)$ can be determined by summing the conditional expectation of $\tilde{s}_{iM}(v \mid y_{i(r_i)}, r_i, \theta_i, \xi_i)$ over all n subjects.

Furthermore, Proposition 4.1.4 also yields the observed information

$$\begin{aligned}j_M(v \mid r, y_{(r)}) = -\sum_{i=1}^{n}&\mathbb{E}_v\left[\left(\frac{\mathrm{d}}{\mathrm{d}v}\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i)\right) \mid r_i, y_{i(r_i)}\right]\\ &+ \mathbb{V}\mathrm{ar}_v\left[\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i) \mid r_i, y_{i(r_i)}\right].\end{aligned}$$

Here $\left(\frac{\mathrm{d}}{\mathrm{d}v}\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i)\right)$ is a sparse matrix with nonzero components given by

$$
\begin{aligned}
\frac{\partial^2}{\partial \beta_j^2} g_i(\theta_i, \xi_i) &= r_{ij}(p_{ij}^2 - p_{ij}) \\
\frac{\partial^2}{\partial \tau^2} g_i(\theta_i, \xi_i) &= \sum_{j=1}^{d_i} \tilde{p}_{ij}^2 - \tilde{p}_{ij} \\
\frac{\partial^2}{\partial \tau \partial \eta} g_i(\theta_i, \xi_i) &= \sum_{j=1}^{d_i} j\tilde{p}_{ij}^2 - \tilde{p}_{ij} \\
\frac{\partial^2}{\partial \eta^2} g_i(\theta_i, \xi_i) &= \sum_{j=1}^{d_i} j^2\tilde{p}_{ij}^2 - \tilde{p}_{ij} \\
\frac{\partial^2}{\partial \sigma_\theta^2} g_i(\theta_i, \xi_i) &= \frac{1}{\sigma_\theta^2} - \frac{3\theta^2}{(1-\rho^2)\sigma_\theta^4} + \frac{2\rho\theta\xi}{(1-\rho^2)\sigma_\theta^3\sigma_\xi} \\
\frac{\partial^2}{\partial \sigma_\theta \partial \sigma_\xi} g_i(\theta_i, \xi_i) &= \frac{\rho\theta\xi}{(1-\rho^2)\sigma_\theta^2\sigma_\xi^2} \\
\frac{\partial^2}{\partial \sigma_\theta \partial \rho} g_i(\theta_i, \xi_i) &= \frac{2\rho\theta_i^2}{(1-\rho^2)^2\sigma_\theta^3} - \frac{(\rho^2+1)\theta\xi}{(1-\rho^2)^2\sigma_\theta^2\sigma_\xi} \\
\frac{\partial^2}{\partial \sigma_\xi^2} g_i(\theta_i, \xi_i) &= \frac{1}{\sigma_\xi^2} - \frac{3\xi^2}{(1-\rho^2)\sigma_\xi^4} + \frac{2\rho\theta\xi}{(1-\rho^2)\sigma_\xi^3\sigma_\theta} \\
\frac{\partial^2}{\partial \sigma_\xi \partial \rho} g_i(\theta_i, \xi_i) &= \frac{2\rho\xi^2}{(1-\rho^2)^2\sigma_\xi^3} - \frac{(\rho^2+1)\theta\xi}{(1-\rho^2)^2\sigma_\xi^2\sigma_\theta} \\
\frac{\partial^2}{\partial \rho^2} g_i(\theta_i, \xi_i) &= \frac{\rho^2+1}{(1-\rho^2)^2} - \frac{3\rho^2+1}{(1-\rho^2)^3}\left(\frac{\theta_i^2}{\sigma_\theta^2} + \frac{\xi_i^2}{\sigma_\xi}\right) + \frac{2\rho(\rho^2+3)\theta\xi}{(1-\rho^2)^3\sigma_\theta\sigma_\xi}.
\end{aligned}
$$

The score and observed information cannot be specified any further before the data analysis in the following chapter because the expected values and variance is taken wrt. the conditional distribution of the random effects given the data.

6 | Data Analysis

The purpose of this chapter is to conduct a data analysis and make statistical inference, in particular parameter estimation, based on the test result data presented in Chapter 1 and the theory presented in Chapters 2, 3, 4 and 5. Throughtout the report thus far the Rasch model has been proposed as the data generating model for the full data and the steps model for the dropout.

For simplicity the parameter estimation will first be conducted under the assumption of ignorability of the missingness mechanism in Section 6.1. Then in Section 6.2 parameter estimation will be conducted where dropout mechanism is modelled by the steps model.

For the rest of the chapter let $y = \{y_{ij}\}_{\substack{i=1,\dots,n \\ j=1,\dots p}}$ denote the test result data and let $r = \{r_{ij}\}_{\substack{i=1,\dots,n \\ j=1,\dots p}}$ denote the missing data indicators of y, such that $r_{ij} = 1$ implies that y_{ij} has been observed and $r_{ij} = 0$ implies that y_{ij} is missing.

6.1 Assuming Ignorable Missingness Mechanism

In this section parameter estimation in the Rasch model will be conducted under the assumption that the missingness mechanism is ignorable. The R code described in this section can be found in Appendix F. In particular, MAR and the separability condition is assumed such that the proof of Theorem 5.2.2 implies that

$$p_{R,Y_{(R)}}(r, y_{(r)}; \theta, \beta, \psi) \propto p_{Y_{(r)}}(y_{(r)}; \theta, \beta) \tag{6.1}$$

where ψ denotes the parameter vector for the missingness mechanism. Densities will in the following be denoted without the use of subscripts to ease notation.

Joint Likelihood Estimation

In order to conduct joint likelihood estimation, the *observed data joint likelihood*, from now on simply referred to as the joint likelihood, given by

$$L_{J,\text{obs}}(\theta, \beta \mid r, y_{(r)}) = p(r, y_{(r)}; \theta, \beta, \psi) \propto \prod_{i=1}^{n} p(y_{i(r_i)}; \theta_i, \beta)$$

will be considered. By noting that

$$\begin{aligned} p(y_{i(r_i)}; \theta_i, \beta) &= \int p(y_{i(r_i)}, y_{i(\bar{r}_i)}; \theta_i, \beta) \mathrm{d}\nu(y_{i(\bar{r}_i)}) \\ &= \int \prod_{j=1}^{p} p(y_{ij}; \theta_i, \beta_j) \mathrm{d}\nu(y_{i(\bar{r}_i)}) \\ &= \prod_{j=1}^{p} p(y_{ij}; \theta_i, \beta_j)^{r_{ij}} \end{aligned}$$

it is immediately clear that when the missing data mechanism is ignorable, the joint likelihood is simply the full data joint likelihood where each factor containing missing values is removed, i.e. a complete cases version of the full data joint likelihood. It is also clear that the observed data joint likelihood inherits the problems of the full data joint likelihood. For instance, joint maximum likelihood estimation cannot be conducted if there are extreme scores present in the data, i.e. if $y_{i(r_i)+} = r_{i+}$ or $y_{i(r_i)+} = 0$ for any $i = 1, \ldots n$.

Furthermore, since an increased number of subjects implies an equivalent increment in the number of parameters in the model, inconsistency is still to be expected.

The JML estimates are computed in R using the glm function which is a function used to fit generalized linear models.

The item parameter estimates $\hat{\beta}_J$ can be found in Table 6.1.

While the 663 subject parameter estimates have not been included in the report, the sample variance of the subjet parameter estimates is $\hat{\sigma}_J = 1.98$.

Conditional Likelihood Estimation

Recall how the full data conditional likelihood was obtained in Chapter 3 by factorizing the density as

$$p(y;\theta,\beta)=\prod_{i=1}^{n}p(y_i;\theta_i,\beta)=\prod_{i=1}^{n}p(y_i \mid y_{i+};\beta)p(y_{i+};\theta_i,\beta)$$

and then simply disregarding the second factor such that

$$L_C(\beta \mid y)=\prod_{i=1}^{n}p(y_i \mid y_{i+};\beta).$$

Similarly, since

$$\begin{aligned}p(r,y_{(r)};\psi,\theta,\beta) &\propto p(y_{(r)};\theta,\beta)\\ &=\prod_{i=1}^{n}p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)p(y_{i(r_i)+};\theta,\beta)\end{aligned}$$

the *observed data conditional likelihood*, from now on simply referred to as the conditional likelihood, is given as

$$L_{C,\text{obs}}(\beta \mid r,y_{(r)})=\prod_{i=1}^{n}p(y_{i(r_i)} \mid y_{i(r_i)+};\beta).$$

While a rigorous study of the observed data conditional likelihoods properties will be omitted in this report, it should be noted the *observed data conditional score*

$$s_{C,\text{obs}}(\beta|r,y_{(r)})=\frac{\partial}{\partial\beta}\log(L_{C,\text{obs}}(\beta \mid r,y_{(r)}))$$

satisfies the first Bartlett identity, thus giving it merit as a reasonable estimating function.

This follows since

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\beta}\log(p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)) \mid y_{i(r_i)+}\right] &= \int\frac{\frac{\partial}{\partial\beta}p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)}{p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)}p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)\mathrm{d}y_{i(r_i)}\\ &=\frac{\partial}{\partial\beta}\int p(y_{i(r_i)} \mid y_{i(r_i)+};\beta)\mathrm{d}y_{i(r_i)}\\ &=0_p\end{aligned}$$

and hence it follows by the law of total expectation that

$$\mathbb{E}\left[s_{C,\text{obs}}(\beta \mid r,y_{(r)})\right]=0.$$

The CML estimates are obtained in R using the clogistic function from the Epi package, which is a function for maximizing conditional likelihoods in logistic regression models.

Parameter estimates $\hat{\beta}_J$ can be found in Table 6.1.

Furthermore, a goodness of fit test as described in Section 3.3 is also conducted in yielding a test statistic of $Z = 1267.00$. Comparing this test statistic to a χ^2-distribution with $(p-1)(p-2) = 1190$ degrees of freedom as per Theorem 3.3.2 yields a p-vaue of 0.059. That is, with a significance level of 5% the Rasch model is accepted as the data-generating model.

Marginal Maximum Likelihood

The *observed data marginal likelihood*, from now on simply referred to as the marginal likelihood, is given as

$$\begin{aligned} L_{M,\text{obs}}(\beta, \sigma^2 \mid r, y_{(r)}) &:= \int_{\mathbb{R}^n} f(r, y_{(r)}, \theta; \psi, \beta, \sigma^2) \mathrm{d}\theta \\ &= \int_{\mathbb{R}^n} p(r, y_{(r)} \mid \theta; \psi, \beta) f(\theta; \sigma^2) \mathrm{d}\theta \\ &\propto \int_{\mathbb{R}^n} p(y_{(r)} \mid \theta; \beta) f(\theta; \sigma^2) \mathrm{d}\theta \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} p(y_{i(r_i)} \mid \theta_i; \beta) f(\theta_i; \sigma^2) \mathrm{d}\theta_i \end{aligned}$$

Asymptotic results regarding parameter estimates using the marginal likelihood is obtained from Theorem 5.2.5.

The MML estimates are obtained in R using the glmer function from the lme4 package, which is a function for fitting GLMMS.

The glmer function approximates the marginal likelihood using Gauss-Hermite approximation with the default number of quadrature point as one, yielding the Laplace-approximation. Furthermore, the marginal likelihood is then maximized using Nelder-Mead and convergence is verified by controlling that $max|grad|$ is below some prespecified tolerance, where $max|grad|$ denotes the numerically largest entry of the score approximated using finite-difference methods.

When using the Laplace-approximation, the tolerance was not met as the greatest numeric value of entries in the score was

$$max|grad| = 0.0938797.$$

Therefore, the number of quadrature point is increased to 5, yielding

$$max|grad| = 0.0117677$$

which is a considerable improvement from the Laplace-approximation case. Using Gauss-Hermite quadrature with 5 quadrature points, the estimated variance of subject ability is given by $\hat{\sigma}_M = 1.36$ and the item difficulty parameter estimates $\hat{\beta}_M$ as well as the associated standard errors SE_M can be found in Table 6.1. It should be noted that parameter estimation has been made assuming $\mathbb{E}[\theta] = 0$ such that $\hat{\beta}_1 \neq 0$. Therefore, $\hat{\beta}_M$ has been translated such that $\hat{\beta}_1 = 0$ in order to make comparisons between the item difficulty parameter estimates simpler.

Discussion and Comparison

When deriving the observed data version of the joint, conditional and marginal likelihood under the assumption of ignorable missingness mechanism, it was seen that these were simply the complete cases equivalents of the full data versions.

The default handling of NA values in glm, clogistic and glmer is na.omit which directly removes all records with NA values from the dataset, thus yielding exactly the complete cases analysis.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
$\hat{\beta}_{\mathbf{J}}$	0	-1.27	-0.84	2.13	-0.83	-1.20	1.08	2.22	1.74	-0.41	3.07	1.68
$\hat{\beta}_{\mathbf{C}}$	0	-1.20	-0.79	2.01	-0.78	-1.13	1.02	2.10	1.65	-0.39	2.90	1.60
$\hat{\beta}_{\mathbf{M}}$	0	-1.21	-0.81	2.01	-0.79	-1.15	1.02	2.09	1.64	-0.42	2.87	1.56
$\mathbf{SE_M}$	0.11	0.14	0.13	0.11	0.13	0.14	0.11	0.11	0.11	0.13	0.13	0.12
	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}	β_{22}	β_{23}	β_{24}
$\hat{\beta}_{\mathbf{J}}$	3.33	0.82	1.54	0.55	-0.69	2.18	5.87	5.24	7.49	-0.53	-1.26	0.86
$\hat{\beta}_{\mathbf{C}}$	3.15	0.77	1.46	0.52	-0.66	2.07	5.54	4.95	7.03	-0.51	-1.19	0.81
$\hat{\beta}_{\mathbf{M}}$	3.10	0.77	1.45	0.52	-0.67	2.04	5.40	4.91	6.93	-0.53	-1.22	0.81
$\mathbf{SE_M}$	0.14	0.11	0.11	0.11	0.13	0.12	0.27	0.23	0.52	0.13	0.15	0.11
	β_{25}	β_{26}	β_{27}	β_{28}	β_{29}	β_{30}	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
$\hat{\beta}_{\mathbf{J}}$	0.24	0.80	2.36	0.26	3.09	4.08	3.70	5.05	1.89	1.25	1.07	2.46
$\hat{\beta}_{\mathbf{C}}$	0.22	0.76	2.24	0.24	2.94	3.89	3.53	4.79	1.79	1.18	1.01	2.34
$\hat{\beta}_{\mathbf{M}}$	0.15	0.68	2.14	0.22	2.87	3.78	3.42	4.62	1.79	1.10	0.92	2.23
$\mathbf{SE_M}$	0.17	0.16	0.16	0.13	0.21	0.26	0.26	0.34	0.13	0.19	0.19	0.19

Table 6.1: Table containing item parameter estimates rounded to two decimal places obtained from the test result data using respectively joint, conditional and marginal maximum likelihood assuming ignorability of the missing mechanism. The table also includes the standard error associated to each marginal ML parameter estimate.

Considering Table 6.1 it is seen that there are no statistically significant difference in the parameter estimates obtained from the different estimators, as all three estimates associated to each item parameter is contained within two standard errors of the marginal ML estimate.

In particular, the estimated orderings of item difficulty are in agreement. Furthermore, using the L_2 norm it follows that

$$\|\hat{\beta}_J - \hat{\beta}_C\| = 0.87 \quad \|\hat{\beta}_J - \hat{\beta}_M\| = 1.19 \quad \|\hat{\beta}_C - \hat{\beta}_M\| = 0.38$$

showing that the estimates are similar but that the conditional and marginal maximum likelihood estimates are closer to each other than to the JML estimates.

However, although the item parameter estimates are somewhat in agreement, there is quite big difference between the estimated standard deviation in subject ability when considering the sample standard deviation of the JML subject ability estimates $\hat{\sigma}_J = 1.98$ and the MML estimate $\hat{\sigma}_M = 1.36$. Intuitively, it is reasonable that the MML variance estimate is smaller since the GLMM framework imposes a distribution on the subject abilities whereas the subject ability parameters can be chosen "freely" in the joint maximum likelihood estimation.

6.2 Assumming the Steps Model for Dropout

In this section parameter estimation will be conducted in the Rasch model when modelling the dropout effect using the steps model within the GLMM framework as presented in Section 5.4. In particular, the marginal likelihood given by Equation (5.20) will be maximized. The R code described in this section can be found in Appendix G.

In order to maximize the marginal likelihood, it is first implemented in R and then maximized using R's default optimization function optim.

Implementation of the Marginal Likelihood in R

In order to evaluate the marginal likelihood given by Equation (5.20) for a given set of parameters $v = (\beta^\top, \tau, \eta, \lambda^\top)^\top$, the subject specific integral must be computed for each subject.

This is done using the two-dimensional Laplace approximation, which recalling Section 4.2 is given by

$$\int_{\mathbb{R}^2} p(y_{i(r_i)} \mid \theta_i; \beta) p(d_i \mid \xi_i; \tau, \eta) f(\theta_i, \xi_i; \lambda) \mathrm{d}\xi_i \mathrm{d}\theta_i \approx \exp(g_i(\theta_{\mathrm{LP}}, \xi_{\mathrm{LP}})) \sqrt{2\pi |\Sigma_{\mathrm{LP}}|}$$

where

$$g_i(\theta, \xi) = \log \Big(p(y_{i(r_i)} \mid \theta_i; \beta) p(d_i \mid \xi_i; \tau, \eta) f(\theta_i, \xi_i; \lambda) \Big), \quad \text{for } i = 1, \ldots, n$$

is the logarithm of the ith integrand, $(\theta_{i\mathrm{LP}}, \xi_{i\mathrm{LP}})$ is a maximizer of g_i, $|\Sigma_{i\mathrm{LP}}|$ denotes the determinant of $\Sigma_{i\mathrm{LP}}$ and $\Sigma_{i\mathrm{LP}}$ is the negative of the inverse Hessian of g evaluated at $(\theta_{i\mathrm{LP}}, \xi_{i\mathrm{LP}})$.

In order to obtain $(\theta_{i\mathrm{LP}}, \xi_{i\mathrm{LP}})$, g_i must be maximized for $i = 1, \ldots, n$, which is done using R's default optimization function optim. In particular, the BFGS algorithm is chosen for optimization as it is known for quick convergence and efficiency.

In the BFGS algorithm the gradient of the objective function is used, which if not supplied, will be replaced by a finite difference method to approximate the gradient. In order to obtain fast convergence and avoid unnecessary computations, the exact gradient of g_i is supplied, which is derived in the following.

First $g_i(\theta_i, \xi_i)$ is written out as

$$\begin{aligned}
g_i(\theta_i, \xi_i) &= \log \Big(p(y_{i(r_i)} \mid \theta_i; \beta) p(d_i \mid \xi_i; \tau, \eta) f(\theta_i, \xi_i; \lambda) \Big) \\
&= \log \left(\prod_{j=1}^{d_i - 1} \left(p(y_{ij} \mid \theta_i; \beta_j)^{r_{ij}} \tilde{p}_{ij} \right) (1 - \tilde{p}_{id_i}) \right) f(\theta_i, \xi_i; \lambda) \Big) \\
&= \left(\sum_{j=1}^{d_i - 1} r_{ij} \log(p(y_{ij} \mid \theta_i; \beta_j)) + \log(\tilde{p}_{ij}) \right) + \log(1 - \tilde{p}_{id_i}) + \log(f(\theta_i, \xi_i; \lambda)).
\end{aligned}$$

By Equation (2.3) it follows for $j = 1, \ldots, p$ that

$$\begin{aligned}\frac{\partial}{\partial \theta_i} p(y_{ij} \mid \theta_i; \beta_j) &= \frac{\partial}{\partial \theta_i} \frac{\exp(y_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)} \\ &= \frac{y_{ij} \exp(y_{ij}(\theta_i - \beta_j))(1 + \exp(\theta_i - \beta_j)) - \exp(\theta_i - \beta_j) \exp(y_{ij}(\theta_i - \beta_j))}{(1 + \exp(\theta_i - \beta_j))^2}\end{aligned}$$

such that

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \log(p(y_{ij} \mid \theta_i; \beta_j)) &= \frac{\frac{\partial}{\partial \theta_i} p(y_{ij} \mid \theta_i; \beta_j)}{p(y_{ij} \mid \theta_i; \beta_j)} \\ &= \frac{y_{ij} \exp(y_{ij}(\theta_i - \beta_j))(1 + \exp(\theta_i - \beta_j)) - \exp(\theta_i - \beta_j) \exp(y_{ij}(\theta_i - \beta_j))}{(1 + \exp(\theta_i - \beta_j)) \exp(y_{ij}(\theta_i - \beta_j))} \\ &= y_{ij} - p_{ij}. \end{aligned} \tag{6.2}$$

Furthermore,

$$\frac{\partial}{\partial \theta_i} \log(\tilde{p}_{ij}) = 0 \tag{6.3}$$

and

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \log(f(\theta_i, \xi_i; \lambda)) &= \frac{\partial}{\partial \theta_i} \left(-\frac{1}{2(1-\rho^2)} \left(\left(\frac{\theta_i}{\sigma_\theta} \right)^2 + \left(\frac{\xi_i}{\sigma_\xi} \right)^2 - 2\rho \frac{\theta}{\sigma_\theta} \frac{\xi}{\sigma_\xi} \right) \right) \\ &= -\frac{1}{2(1-\rho^2)} \left(2\frac{\theta_i}{\sigma_\theta^2} - 2\frac{\rho\xi}{\sigma_\theta \sigma_\xi} \right) \\ &= \frac{\rho\xi}{(1-\rho^2)\sigma_\theta \sigma_\xi} - \frac{\theta_i}{(1-\rho^2)\sigma_\theta^2}. \end{aligned} \tag{6.4}$$

Combining equations (6.2), (6.3) and (6.4) yields

$$\frac{\partial}{\partial \theta_i} g_i(\theta_i, \xi_i) = \sum_{j=1}^{d_i - 1} r_{ij} \left(y_{ij} - p_{ij} \right) + \frac{\rho \xi_i}{(1-\rho^2)\sigma_\theta \sigma_\xi} - \frac{\theta_i}{(1-\rho^2)\sigma_\theta^2}. \tag{6.5}$$

Through similar arguments it follows that

$$\frac{\partial}{\partial \xi_i} g_i(\theta_i, \xi_i) = d_i - 1 - \left(\sum_{j=1}^{d_i} \tilde{p}_{ij} \right) + \frac{\rho \theta_i}{(1-\rho^2)\sigma_\theta \sigma_\xi} - \frac{\xi_i}{(1-\rho^2)\sigma_\xi^2}. \tag{6.6}$$

Equations (6.5) and (6.6) specify the exact gradient utilized in the BFGS algorithm.

Considering

$$|\Sigma_{i\text{LP}}| = | - \frac{\partial^2}{\partial (\theta_i, \xi_i)^2} g_i(\theta_{i\text{LP}}, \xi_{i\text{LP}})^{-1}|,$$

it should be noted that the negative sign can be omitted since it is a 2×2 matrix.

Furthermore, since the determinant of the inverse is the inverse of the determinant and the determinant of a 2×2 matrix can be efficiently computed exactly, it follows that the only challenge remaining in determining $|\Sigma_{i\text{LP}}|$ is to determine an expression for the Hessian of g as

$$|\Sigma_{i\text{LP}}| = \frac{1}{|\frac{\partial^2}{\partial(\theta_i,\xi_i)^2} g_i(\theta_{i\text{LP}}, \xi_{i\text{LP}})|}.$$

While the optim function provides the option to return a numeric approximation of the Hessian evaluated at the optimum, this would result in less precise Laplace approximation and would increase the computational demands of the R implementation drastically.

Therefore, the exact Hessian is used which is derived in the following.

It follows immediately from Equation (6.5) that

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i^2} g_i(\theta_i, \xi_i) &= -\sum_{j=1}^{d_i-1} r_{ij} \frac{\partial}{\partial\theta_i} p_{ij} - \frac{1}{(1-\rho^2)\sigma_\theta^2} \\
&= -\sum_{j=1}^{d_i-1} r_{ij} \frac{\exp(\theta_i - \beta_j)(1 + \exp(\theta_i - \beta_j)) - \exp(\theta_i - \beta_j)^2}{(1 + \exp(\theta_i - \beta_j))^2} - \frac{1}{(1-\rho^2)\sigma_\theta^2} \\
&= -\sum_{j=1}^{d_i-1} r_{ij} \frac{p_{ij}}{1 + \exp(\theta_i - \beta_j)} - \frac{1}{(1-\rho^2)\sigma_\theta^2}.
\end{aligned}$$

Similarly it follows by Equation (6.6) that

$$\frac{\partial^2}{\partial\xi_i^2} g_i(\theta_i, \xi_i) = -\sum_{j=1}^{d_i-1} \frac{\tilde{p}_{ij}}{1 + \exp(\xi_i - (\tau + j\eta))} - \frac{1}{(1-\rho^2)\sigma_\theta^2}$$

and

$$\frac{\partial^2}{\partial\theta_i \partial\xi_i} g_i(\theta_i, \xi_i) = \frac{\rho}{(1-\rho^2)\sigma_\theta \sigma_\xi}.$$

In conclusion, once $(\theta_{i\text{LP}}, \xi_{i\text{LP}})$ has been determined using optim they are then inserted in the expressions above yielding the Hessian and hence also the determinant of $\Sigma_{i\text{LP}}$. This in turn yields the Laplace approximation of the two-dimensional integral corresponding to the ith subject for $i = 1, \ldots, n$.

Once all n integrals have been approximated, the marginal likelihood can simply be estimated as the product of the Laplace approximations.

Remark 6.2.1. Since each of the integrals usually is a very small number, this product of integrals would be indistinguishable from zero due to the limited precision of floating point numbers. Therefore, the marginal log likelihood is estimated instead.

A major challenge in this implementation is computational efficiency and numeric instability.

In particular, efficient vectorized computations in R have been utilized whenever possible and the function parLapply from the parallel package has been used in order to utilize multiple cores when computing the Laplace approximations. The code was run on a computer with i7-7700HQ processor and 8GB RAM, where all 8 virtual cores where utilized using parallelized computations, resulting in an average time of approximately 10 seconds per likelihood evaluation, of which thousand are needed for maximization.

Regarding numerical instability, functions such as exp are used multiple times throughout the implementation, which due to the limited precision of floating point numbers yields Inf whenever the input becomes somewhat large (above 710 when using IEEE 64-Bit floating point numbers). Therefore, one have to be careful when choosing parameters in order to ensure that e.g. the exponential function does not return Inf.

Another example is when considering the density for the dropout, where

$$1 - \frac{\exp(\xi_i - (\tau + d_i\eta))}{1 + \exp(\xi_i - (\tau + d_i\eta))}$$

is replaced with

$$\frac{1}{1 + \exp(\xi_i - (\tau + d_i\eta))}$$

as the former simply returns 0 and the latter returns small positive values for reasonably large $\xi_i - (\tau + d_i\eta)$.

Maximization of the Marginal Likelihood

With the marginal likelihood implemented in R, such that the marginal likelihood can be evaluated given parameters $v = (\beta^\top, \tau, \eta, \lambda^\top)^\top$, the marginal likelihood can also be maximized.

This is also done using optim, but unlike the implementation of the marginal likelihood the BFGS method could not be used. This is due to the numerical instability discussed earlier, since the BFGS method with optim had a tendency to choose parameter values which are numerically much larger than the current iterate, resulting in e.g. the exponential of large numbers.

In contrast, the Nelder-Mead method with optim generally v parameters not too dissimilar from the current iterate, avoiding these numerical problems.

Furthermore, some parameters have a parameter space which is not the whole real line, in particular σ_θ and σ_ξ which are strictly positive and ρ which is between -1 and 1.

These bounds have to be modelled implicitly as optim does not support the use of bounds when using the Nelder-Mead method.

Therefore, instead of maximizing wrt. $\sigma_\theta, \sigma_\xi$ and ρ, the maximization is done wrt. $\log(\sigma_\theta), \log(\sigma_\xi)$ and $\tan(\frac{\pi}{2}\rho)$ as $\exp : \mathbb{R} \to \mathbb{R}_+$ and $\frac{2}{\pi}\arctan : \mathbb{R} \to (-1, 1)$.

The MML estimate $\hat{\beta}_M$ obtained in Section 6.1 under the assumption of ignorability is used as the initial value for the item difficulty parameters $\beta = (\beta_1, \ldots, \beta_p)^\top$. In particular, these are the item parameter estimates when $\mathbb{E}[\theta] = 0$, i.e. before the parameter estimates where translated by $-\hat{\beta}_{1M}$ to make $\hat{\beta}_{1M} = 0$.

Furthermore, $\tau = -3, \eta = 0.1, \sigma_\theta = \sigma_\xi = 1$ and $\rho = 0.1$ were chosen as initial values. Evaluating the marginal log likelihood in the initial values yields -29076.24.

After approximately 63.500 iterations and one week of computations the optim function converged with parameter estimates $\hat{v}_D$ given in Table 6.2. The table also includes standard errors of the parameter estimates which where obtained using Proposition 4.1.4 which states that

$$\begin{aligned} j_M(v \mid r, y_{(r)}) = -\sum_{i=1}^{n} \mathbb{E}_v \left[\left(\frac{\mathrm{d}}{\mathrm{d}v} \tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i) \right) \mid r_i, y_{i(r_i)} \right] \\ + \mathbb{V}\mathrm{ar}_v \left[\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i) \mid r_i, y_{i(r_i)} \right] \end{aligned}$$

where $\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i)$ and $\frac{\mathrm{d}}{\mathrm{d}v}\tilde{s}_{iM}(v \mid r_i, y_{i(r_i)}, \theta_i, \xi_i)$ were derived in Section 5.4 and the conditional expectations were obtained using Monte Carlo estimation by simulating random effects θ_i, ξ_i from their conditional distribution given $r_i, y_{i(r_i)}$. In particular, the result

$$(\theta_i, \xi_i) | (r_i, y_{i(r_i)} \overset{d}{\approx} N(\mu_{\mathrm{i,LP}}, \Sigma_{\mathrm{i,LP}})$$

from Section 4.2 where utilized as the Laplace parameters were already obtained for each subject due to the Laplace approximations in the implementation of the likelihood. Another possible approach would be to simulate from the exact conditional distribution using rejection sampling as discussed in Proposition D.0.1.

Note that the item difficulty parameter estimates presented in Table 6.2 have been translated by -0.9 such that $\hat{\beta}_{1,D} = 0$.

	Estimate	**Standard Error**		**Estimate**	**Standard Error**
$\hat{\beta}_{1,D}$	0	0.01	$\hat{\beta}_{22,D}$	-0.40	0.01
$\hat{\beta}_{2,D}$	-0.93	0.01	$\hat{\beta}_{23,D}$	-0.89	0.01
$\hat{\beta}_{3,D}$	-0.65	0.01	$\hat{\beta}_{24,D}$	0.62	0.01
$\hat{\beta}_{4,D}$	1.55	0.01	$\hat{\beta}_{25,D}$	-0.06	0.02
$\hat{\beta}_{5,D}$	-0.62	0.01	$\hat{\beta}_{26,D}$	0.37	0.02
$\hat{\beta}_{6,D}$	-0.88	0.01	$\hat{\beta}_{27,D}$	1.49	0.02
$\hat{\beta}_{7,D}$	0.75	0.01	$\hat{\beta}_{28,D}$	0.16	0.01
$\hat{\beta}_{8,D}$	1.61	0.01	$\hat{\beta}_{29,D}$	2.16	0.03
$\hat{\beta}_{9,D}$	1.26	0.01	$\hat{\beta}_{30,D}$	2.94	0.06
$\hat{\beta}_{10,D}$	-0.33	0.01	$\hat{\beta}_{31,D}$	2.60	0.05
$\hat{\beta}_{11,D}$	2.18	0.01	$\hat{\beta}_{32,D}$	3.55	0.10
$\hat{\beta}_{12,D}$	1.16	0.01	$\hat{\beta}_{32,D}$	1.39	0.01
$\hat{\beta}_{13,D}$	2.31	0.01	$\hat{\beta}_{34,D}$	0.74	0.02
$\hat{\beta}_{14,D}$	0.60	0.01	$\hat{\beta}_{35,D}$	0.60	0.02
$\hat{\beta}_{15,D}$	1.14	0.01	$\hat{\beta}_{36,D}$	1.57	0.03
$\hat{\beta}_{16,D}$	0.40	0.01	$\hat{\tau}_D$	-7.84	0.02
$\hat{\beta}_{17,D}$	-0.50	0.01	$\hat{\eta}_D$	0.19	0.00
$\hat{\beta}_{18,D}$	1.52	0.01	$\hat{\sigma}_{\theta,D}$	0.31	0.00
$\hat{\beta}_{19D}$	4.41	0.07	$\hat{\sigma}_{\xi,D}$	1.38	0.00
$\hat{\beta}_{20D}$	3.99	0.05	$\hat{\rho}_D$	0.54	0.01
$\hat{\beta}_{21D}$	6.03	0.34			

Table 6.2: Parameter estimates and standard errors rounded to two decimal places.

Evaluating the marginal log likelihood in $\hat{v}_D$ yields -10672.05.

Code Validation

In order to validate the R implementation and maximization of Equation (5.20) a simulation study is conducted. Here 100 datasets are simulated from a Rasch model with dropout based on the steps model using $n = 100, p = 5$ and the parameters $\beta_0 = (\beta_{1,0}, \ldots, \beta_{p,0})$ with $\beta_j = 0.5j - 1.5$, $\tau_0 = -6, \eta_0 = 1, \sigma_{\theta,0} = 0.5, \sigma_{\xi,0} = 1.5, \rho_0 = 0.5$ and $v_0 = (\beta_0^\top, \tau_0, \eta_0, \sigma_{\theta,0}, \sigma_{\xi,0}, \rho_0)^\top$.

It should be noted that the number of datasets, n and p have been chosen relatively small in order to make the simulation study computationally feasible.

For each dataset maximization of the marginal likelihood is conducted using the method described previously.

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$	$\boldsymbol{\tau}$	$\boldsymbol{\eta}$	$\boldsymbol{\sigma_\theta}$	$\boldsymbol{\sigma_\xi}$	$\boldsymbol{\rho}$
$\mathbf{v_0}$	-1.0	-0.5	0.0	0.5	1.0	-6.0	1.0	0.5	1.5	0.5
$\boldsymbol{\mu}_{\text{MC}}$	-0.94	-0.49	0.03	0.47	1.04	-6.01	0.99	0.48	1.60	0.51
$\boldsymbol{\sigma}_{\text{MC}}$	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.03	0.01

Table 6.3: Table containing the true parameter values v_0, the Monte Carlo estimates μ_{MC} of the expected values of the MML estimators rounded off to two decimal places, and Monte Carlo errors σ_{MC} rounded to three decimal places, based on the MML estimates of the 100 simulated datasets.

Table 6.3 shows that the R implementation obtains reasonable parameter estimates in regards to the bias of the estimator, since the Monte Carlo estimate for each parameter deviates less than 2 times the Monte Carlo error from the true parameter value, except for th Monte Carlo estimate for σ_ξ which is still reasonably close to the true parameter value.

In conclusion, this simulation study shows that the R implementation and maximization of the likelihood works as intended.

Discussion

A rather important parameter to consider is the correlation between ability and speed, ρ. Clearly, if $\rho = 0$ then the ability and speed of a subject would be uncorrelated and hence independent, such that the missingness mechanism would be ignorable. This would be the same situation as studied in section 6.1 and hence $\hat{\beta}_D = \hat{\beta}_M, \hat{\sigma}_{\theta D} = \hat{\sigma}_{\theta M}$. That is, modelling the dropout effect would have contributed with no additional information regarding the distribution of subject ability parameters or item difficulty parameters.

However, as $\hat{\rho}_D = 0.54$ with an associated standard error 0.01, this is not the case. In particular, the ability and speed of a subject is positively correlated, such that a high speed of a subject increases the probability of the subject to also have high ability and vice versa. This also means that an early dropout indicates low ability.

Consider e.g. the case of a subject who drops out early but has a high correct rate before dropout. In the case of ignorable missingness this subject would have a large ability estimate, as an early dropout is assumed to be unrelated to the ability of the subject. However, in the case of a positive correlation between the ability and speed of a subject, the student would have a lower ability estimate as the early dropout indicates lower ability.

It is seen that $\hat{\sigma}_D$ is different from both $\hat{\sigma}_M$ and $\hat{\sigma}_J$, and in particular significantly lower.

Furthermore, when comparing $\hat{\beta}_D$ with $\hat{\beta}_J, \hat{\beta}_C$ and $\hat{\beta}_M$ in Table 6.1 it is seen that numerical values of the $\hat{\beta}_D$ estimates are significantly smaller than $\hat{\beta}_J, \hat{\beta}_C$ and $\hat{\beta}_M$ for every item. However, the ordering of the items are almost unchanged, such that e.g. item 21 is the most difficult item, item 19 is the second most difficult item and so on no matter which of the estimates are considered.

Thus the dropout mechanism is MNAR by Remark 5.4.2 and therefore in particular non-ignorable, and choosing to ignore it results in numerically larger parameter estimates for both β and σ_θ compared to when the dropout is modelled.

7 | Conclusion

The test result data was presented in Chapter 1 where it was apparent that both a model for the full data and the missingenss mechanism would be needed, as the missingness was potentially MNAR. Specifically, as motivated by Figure 1.1, it was decided to model the dropout.

Regarding the full data model, IRT was presented in Chapter 2 in a general setting, and the Rasch model was proposed as the full data model. Therefore, estimators for the Rasch model was presented and asymptotic properties derived. Furthermore, the derivation of the asymptotic properties of the CML estimator was also used to derive a GOF test based on the Rasch property of specific objectivity.

The Rasch model as a data generating model was accepted by the GOF test when assuming ignorability of the missingness mechanism in 6.1.

In Section 6.1 the joint, conditional and marginal ML estimates where obtained from the test result data under the assumption of ignorable missingness. Here it was found that there where no statistically significant differences between the different types of item difficulty estimates for each item. A measure of subject homogeneity was obtained from the MML with $\hat{\sigma}_M = 1.37$.

Regarding the modelling of the dropout, the steps model was introduced in Section 5.4 where the subject speed ξ and the correlation between subject ability and speed ρ was introduced.

A marginal likelihood was then derived including models for the full data, the dropout and the ability and speed of subject modelled as bivariate normally distributed random effects.

This marginal likelihood was then implemented in R and maximized in Section 6.2.

Here, a 95% confidence interval of ρ was obtained as $(0.52, 0.56)$, confirming that the dropout does in fact yield MNAR missingness cf. Remark 5.4.2 under the proposed model.

Furthermore, the standard deviation in subject ability was estimated to $\hat{\sigma}_D = 0.31$, considerably smaller than the estimate $\hat{\sigma}_M = 1.37$ obtained in Section 6.1.

Similarly it was found that estimates based on ignorable missingness yielded the same ordering of item difficulty as when modelling the dropout, but were however significantly larger in numeric values. That is, the difficulty of the items was seen to be significantly more similar when modelling the dropout.

In conlusion, under the stated assuptions the dropout effect was found to be MNAR and hence non-ignorable. Choosing to treat the missingness mechanism as ignorable regardless would result in severly skewed result in significant changes to the obtained parameter estimates, implying that modelling the missingness mechanism is essential to conduct statistical inference on the test result data.

8 | Final Remarks

This chapter contains the final remarks of the report including reflection of how the content of the report have been prioritized, and how the analysis could be continued if time was not a limitation.

The EM Aglorithm

The EM algorithm was introduced in Chapter 5 as an approach to conduct optimization in situations where e.g. $p_{Y_{(r)}}(y_{(r)};\theta)$ is difficult to compute, and hence the usual methods for maximizing the observed data likelihood are computationally infeasible.

However, this is clearly not the case for the Rasch model making the EM algorithm a less appealing choice of optimization approach. Furthermore, as seen in Section 4.3 the E-step is not explicitly given but instead has to be numerically approximated by e.g. a Laplace approximation. Furthermore, the EM algorithm only has linear convergence near to optimum and does not return standard errors as opposed to e.g. a Newton-Raphson approach with quadratic convergence which also computes the observed information at each iteration.

Based on the above, the EM algorithm have not been used to maximize the marginal likelihood in this report.

Nevertheless, this begs the question as to why the EM algorithm was considered in Sections 4.3 and 5.3 if it was not going to be used in the data analysis.

One of the reasons that the EM algorithm was been considered in this report is due to its prominent role in the missing data litterature, see e.g. [7], [5] and in the IRT litteature see e.g. [22], [20] and [3].

Furtheremore, as mentioned in Remark 4.3.2, the EM algorithm is applicable for GLMM's since the random effects can be considered as missing data which was the original purpose of the EM algorithm. Hence the EM algorithm somehow overlaps the topics of GLMM's and missing data making it especially relevant in this report.

Other IRT models

Within IRT there are numerous models which could have beeng used, for example the 2PL model where each item has two associated parameters, namely difficulty and discrimination, i.e.

$$P(Y_{ij} = 1) = \frac{\exp(\alpha_j \theta_i - \beta_j)}{1 + \exp(\alpha_j \theta_i - \beta_j)}.$$

Intuitively this is a natural extension of the Rasch model since not all questions requires the same utilization of ability. For instance, one could design a "coinflip"-type question where the subject has a 50% chance of answering correctly regardless of ability. This situation could not be explained within the Rasch model but in the 2PL model this is easily found by setting the item difficulty and discrimination to zero, hence yielding

$$P(Y_{ij} = 1) = \frac{\exp(0 \cdot \theta_i - 0)}{1 + \exp(0 \cdot \theta_i - 0)} = \frac{1}{2}, \quad \text{for } i = 1, \dots, n.$$

Clearly the Rasch model is a special case of the 2PL model where all discrimination parameters equals 1.

The reason this has not been prioritized in the report is because it would fundamentally not add anything new. The joint likelihood estimation in the 2PL model would obviously suffer from the same issues as in the Rasch model and since sufficiency of the subject score is unique to the Rasch model c.f. Theorem 2.1.1 it follows that the conditional approach is not viable. Hence the parameter estimation would simply be done using a marginal maximum likelihood approach by considering the subject ability as a random effect and where the likelihood would be estimated using the methods described in Section 4.2, see Remark 4.3.3

To summarize, applying other IRT models than the Rasch model to the test result data might add new insight to understand the data, but not new methods for the report and has therefore not been prioritized.

Utilizing all the Test Result Data

Recall Chapter 1 where the test result data is presented. Here it is mentioned that the dataset actually contains more information regarding the subjects than just the response pattern, includig age, gender and school. Although this additional information has not been utilized in the report, it is interresting to consider if it could have been.

Obviously as the Rasch model is a logistic regression additional covariates could have easily been added to the linear predictor yielding a modified Rasch model.

Another approach as to how this additional information could have been used is the following.

Recall Section 3.3 where the Rasch model property specific objectivity is presented, which intuitively means that the estimators of the item parameters is consistent no matter the sample of subjects. As explained then, the GOF test presented in Theorem 3.3.2 is a special case of specific objectivity where the test is based on consistent estimators for each score group.

However, one could also have constructed a similar GOF test by e.g. grouping the subjects by age, gender or school rather than subject score.

This has however been omitted from the report as one GOF test ought to be sufficient.

Modelling the Missingness Mechanism

When modelling the missingness mechanism in Section 5.4 it is assumed that all missingness other than dropout is ignorable such that

$$p(r_i, y_{i(r_i)} \mid \theta; \psi, \beta) \propto p(d_i, y_{i(r_i)} \mid \theta, \xi; \beta, \tau, \eta).$$

However this assumption it most likely wrong. Intuitively, especially since it's a speeded test, it is quite possible that subjects simply skip items deemed too difficult. Hence missing responses before dropout might indicate lower ability just like early dropout.

Therefore, it might contribute to the understanding of the dataset to model more of the missing data mechanism than just the dropout, and this could for instance be done using a Rasch model.

Expanding the modelling of the missingness mechanism would be my focus if more time was available.

Improving the Maximization Approach

The implementation and maximization of the marginal likelihood in Equation (5.20) as discussed in Section 6.2 was quite inefficient as the Nelder-mead algorithm converged 63.500 iterations. This implies that the implemented approach will most likely not be utilized in other applications or by other reasearchers.

As the marginal score and observed information have already been implemented it seems reasonable to maximize the marginal likelihood using the Newton-Raphson algorithm rather than the Nelder-Mead algorithm. However, attempts at this have so far not been successful as problems occur similar to those described in Section 6.2 regarding the use of the BFGS algorithm.

In conclusion, if more time were available, improving the maximization approach would be a priority along with expanding the modelling of the missingness mechanism.

Bibliography

[1] Erling Andersen and Mette Madsen. “Estimating the Parameters of the Latent Population Distribution”. In: *Psychometrika* 42.3 (1973), pp. 357–374.

[2] Erling B. Andersen. “A Goodness of Fit Test for the Rasch Model”. In: *Psychometrika* 38.1 (1973), pp. 123–140.

[3] Erling B. Andersen. “Asymptotic Properties of Conditional Maximum-Likelihood Estimators”. In: *Journal of the Royal Statistical Society* 32 (1970), pp. 283–301.

[4] Erling B. Andersen. *Conditional inference and models for measuring*. Kbh. : Mentalhygiejnisk Forlag, 1973.

[5] Marie Davidian and Anastasios A. Tsiatis. “Statistical Methods for Analysis With Missing Data”. 2015.

[6] Matthias Von Davier. *Rasch Model from: Handbook of Item Response Theory. CRC Press*. 2016. URL: https://www.routledgehandbooks.com/doi/10.1201/9781315374512-5 (visited on 09/23/2021).

[7] A.P Dempster, N.M. Laird, and D.B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society* 39.1 (1977), pp. 1–38.

[8] Thomas S. Ferguson. *A Course in Large Sample Theory*. Chapman Hall, 1996. ISBN: 0-412-04371-8.

[9] Gerhard H. Fischer. “On the Existence and Uniqueness of Maximum-Likelihood Estimates in the Rasch Model ”. In: *Psychometrika* 46.1 (1981), pp. 59–77.

[10] Gerhard H. Fischer and Ivo W. Molenaar. *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, 1995. ISBN: 978-1-4612-4230-7.

[11] Shelby J. Haberman. *Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses*. ETS, 2004.

[12] Reinhold Hatzinger. *Parameter Estimation in the Rasch Model*. 2010. URL: https://statmath.wu.ac.at/~hatz/psychometrics/10w/RM_handouts_3.pdf (visited on 09/24/2021).

[13] Reinhold Hatzinger. *The Rasch Model*. 2010. URL: https://statmath.wu.ac.at/~hatz/psychometrics/10w/RM_handouts_2.pdf (visited on 09/23/2021).

[14] Richard J. Herrnstein and Charles Murray. *The Bell Curve*. Free Press, 1994. ISBN: 0-02-914673-9.

[15] Micheal I. Jordan. "An Introduction to Probabilistic Graphical Models". 2003. URL: https://www.stat.berkeley.edu/~mjwain/Fall2012_Stat241a/reader_ch8.pdf (visited on 01/29/2022).

[16] J. Kiefer and J. Wolfowitz. "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: *Annals of Mathematical Statistics* 27 (1956), pp. 887–906.

[17] Steffen Lauritzen. "Maximum Likelihood in Exponential Families". 2004. URL: http://www.stats.ox.ac.uk/~steffen/teaching/bs2siMT04/si6bw.pdf (visited on 01/29/2022).

[18] Henrik Madsen and Poul Thyregod. *Introduction to General and Generalized Linear Models*. CRC Press, 2011. ISBN: 978-1-4200-9155-7.

[19] P. McGullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2009. ISBN: 978-0412317606.

[20] Jonald L. Pimentel. *Item Response Theory Modeling With NonIgnorable Missing Data*. PrintPartners Ipskamp B.V., 2005. ISBN: 90-365-2295-1.

[21] C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. John Wiley & Sons. Inc., 1973. ISBN: 0-471-21875-8.

[22] Steven E. Rigdon and Robert K. Tsutakawa. "Parameter Estimation in Latent Trait Models". In: *Psychometrika* 48.4 (1983), pp. 567–574.

[23] N.D. Verhelst C.A.W. Glas H.H. de Vries. "A steps model to analyze partial credit." In: *Handbook of modern item response theory.* (1997), pp. 123–138.

[24] Rasmus Waagepetersen. “Computation of the Likelihood Function for GLMMs - Monte Carlo Methods”. 2021. URL: https://people.math.aau.dk/~rw/Undervisning/TostaII/Slides/montecarlo.pdf (visited on 12/08/2021).

[25] Rasmus Waagepetersen. “Generalized Linear Mixed Models - Computation of the Likelihood Function”. 2021. URL: https://people.math.aau.dk/~rw/Undervisning/TostaII/Slides/2.pdf (visited on 12/06/2021).

[26] Rasmus Waagepetersen. “Generalized Linear Mixed Models - Computation of the Likelihood Function”. URL: https://people.math.aau.dk/~rw/Undervisning/TostaII/Slides/2.pdf (visited on 10/20/2021).

[27] Rasmus Waagepetersen. “Laplace approximation”. URL: https://people.math.aau.dk/~rw/Undervisning/TostaII/Slides/laplace.pdf (visited on 10/20/2021).

A | Generalized Linear Models

This appendix is based on [19][Chapter 2] and [18][Chapter 4], and is meant as a supplement to Chapter 2 with the purpose of introducing the reader to generalized linear models and logistic regression.

First, the exponential dispersion family is defined as follows.

Definition A.0.1. Exponential Dispersion Family

A distribution is said to be in the *exponential dispersion family* if its density can be written as

$$f(y;\theta,\lambda) = \exp(\lambda(\theta^\top y - b(\theta)) + c(y,\lambda)) \tag{A.1}$$

for *canonical parameter* $\theta \in \Omega \subseteq \mathbb{R}^k, k \in \mathbb{N}$, *dispersion parameter* $\lambda > 0$ and functions $b : \Omega \to \mathbb{R}$, $c : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$.

The subfamily where $\lambda = 1$ is called *the natural exponential family.*

Definition A.0.1 can now be used to define GLMs.

Definition A.0.2. Generalized Linear Models

Let $Y = (Y_1, \dots Y_n)$ be a random vector with expected values $\mathbb{E}[Y_i] = \mu_i$ for $i = 1, \dots, n$ and let $X \in \mathbb{R}^{n \times p}$ for $n, p \in \mathbb{N}$. Suppose that $\mu_i \in M \subseteq \mathbb{R}$ and let $\eta_i = x_i^\top \beta$ for $i = 1, \dots, n$ and some $\beta \in \mathbb{R}^p$. Furthermore, let $g : \mathrm{M} \to \mathbb{R}$ be an invertible function.

Then Y is said to follow a *generalized linear model* with *design matrix* X, *link function* g and *linear predictors* $\eta_1, \dots, \eta_n$ if Y_i follows a distribution from the exponential dispersion family, $Y_i \perp\!\!\!\perp Y_j$ and $\eta_i = g(\mu_i)$ for $i, j = 1, \dots, n, i \neq j$.

Logistic regression will now be presented and in particular it will be shown that is a GLM.

Example A.0.3. Logistic Regression as a GLM

Let $X_i \overset{\perp\!\!\!\perp}{\sim} \text{Bin}(n_i, p_i)$ and η_i denote the number of successes and the linear predictor for the ith subject for $i = 1, \ldots, n$.

For a logistic regression, p_i is modelled using the relation

$$\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) \tag{A.2}$$

which is well defined since $\text{logit} :]0, 1[\to \mathbb{R}$.

By Equation (A.2) it follows that

$$p_i = \frac{p_i}{1-p_i}(1-p_i) = \frac{p_i}{1-p_i}\frac{1}{\frac{p_i}{1-p_i}+1} = \frac{\exp(\eta_i)}{\exp(\eta_i)+1},$$

that is, the probability for succes for the ith subject is modelled as the standard logistic function of the linear predictor for that subject.

While odds are not a concept usually utilized in most of statistics, they are often considered when dealing with logistic regression because of their simple form. The odds for succes for the ith subject is given by

$$o_i = \frac{p_i}{1-p_i} = \exp(\eta_i) \tag{A.3}$$

such that the odds ratio between the ith and jth subject is given by

$$\frac{o_i}{o_j} = \exp(\eta_i - \eta_j). \tag{A.4}$$

It will be now shown that the logistic regression is a GLM.

Let $Y_i = \frac{X_i}{n_i}$ such that

$$\mathbb{E}[Y_i] = p_i,\ Y_i \perp\!\!\!\perp Y_j \text{ for } i \neq j$$

and

$$\begin{aligned} f_{Y_i}(y) &= \frac{d}{dy} F_{Y_i}(y) = \frac{d}{dy} F_{X_i}(n_i y) = n_i f_{X_i}(n_i y) = n_i \binom{n_i}{n_i y} p_i^{n_i y} (1-p_i)^{n_i - n_i y} \\ &= n_i \binom{n_i}{n_i y} \exp\left(n_i y \log(p_i) + (n_i - n_i y)\log(1-p_i)\right) \\ &= n_i \binom{n_i}{n_i y} \exp\left(n_i \left(y \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right)\right) \end{aligned}$$

showing that the distribution of Y_i belongs to the exponential dispersion family with $\theta_i = \text{logit}(p_i)$, $b(\theta) = \log(1+\exp(\theta))$ and $\lambda = n_i$.

B | R Code for Simulation Study of CML Estimator

This appendix contains the R code for the simulation study regarding the conditional likelihood discussed in Section 3.4. In particular, the R code was converted to latex code using RMarkdown and the knit function.

Data generation

```
K=1000
n= 500
p = 10
beta = (1:p-1)*0.2

theta_sim = rnorm(n,mean = 1)
#save(theta_sim, file = "theta_sim.RData")
load("theta_sim.RData")

response = rep(NA,0,n*p)

item = factor(rep(1:p,n))

subject = factor(rep(1:n, each=p))

dat = data.frame(subject,item,response)

simdata = vector(mode = "list", length = K)

prob = vector(mode = "list", length = p)
```

```
for(j in 1:p){
  prob[[j]] = exp(theta_sim-beta[j])/(1+exp(theta_sim-beta[j]))
}



for(k in 1:K){
  for (j in 1:p){
    dat$response[dat$item == j] =rbinom(n,1,prob[[j]])
  }
 simdata[[k]]=dat
}

#save(simdata, file = "simdata.RData")
load("simdata.RData")
```

Subject scores including MC estimates and error

```
#Computing subject scores
n_s = data.frame(matrix(nrow = K, ncol = p+1))
for(k in 1:K){
  data = simdata[[k]]

  data_wide =data.frame(matrix(nrow=n, ncol = p))
  for(i in 1:n){
    data_wide[i,]=data$response[data$subject==i]
  }

  subjectscore = rowSums(data_wide)
 n_s[k,]=table(subjectscore)
}
#save(n_s, file = "simstudy_n_s.RData")
load("simstudy_n_s.RData")
```

```
#No datasets without extreme scores, hence JML estimation not possible
min(n_s[1])
min(n_s[11])
#Monte Carlo estimates
colMeans(n_s)
#Monte Carlo error
sapply(n_s, sd)/sqrt(1000)
```

Conditional maximum likelihood estimation

```
library(Epi)
beta_cml = data.frame(matrix(nrow = K, ncol = p-1))

for(k in 1:K){
 cml <- clogistic(response ~ item, strata = subject, data = simdata[[k]])
 beta_cml[k,] =-coef(cml)
}
names(beta_cml) = c("item2","item3","item4","item5","item6", "item7",
                    "item8","item9", "item10")

#save(beta_cml, file = "simstudy_beta_cml.RData")
load("simstudy_beta_cml.RData")



#Evaluation of conditional observed information
sympoly <- function(s,X) sum(combn(X, s, prod))

gamma = sapply(1:(p-1),sympoly,exp(-beta))
gamma_j = data.frame(matrix(nrow = p-1, ncol = p-1))
gamma_ji = data.frame(matrix(nrow = (p-1)^2, ncol = p-1))

for(j in 2:(p)){
```

```
  gamma_j[j-1,] = sapply(1:(p-1),sympoly,exp(-beta[-j]))
  for(i in 2:(p)){
    gamma_ji[(j-2)*(p-1)+(i-1),]=sapply(0:(p-2),sympoly,exp(-beta[-c(j,i)]))
  }
}

J = function(n_s_k){
  J_C = matrix(nrow = p-1, ncol = p-1)
  for(j in 2:(p)){
    for(i in 2:(p)){
      if(i==j){
        factor1 = (-gamma_j[j-1,]*(gamma-gamma_j[j-1,]))/(gamma^2)
        J_C[j-1,j-1]=sum(n_s_k*factor1)
      } else {
          factor2 = (exp(-beta[i])*gamma_ji[(j-2)*(p-1)+(i-1),]*gamma
                      -gamma_j[j-1,]*(gamma-gamma_j[i-1,]))/gamma^2
          J_C[j-1,i-1]=sum(n_s_k*factor2)
      }
    }
  }
  return(-J_C)
}

#Computing normalized CML estimates

beta_cml_norm = data.frame(matrix(nrow=K, ncol=p-1))
for(k in 1:K){
  beta_cml_norm[k,] = sqrtm(J(n_s[k,][-c(1,11)]))
                          %*%as.vector(t(beta_cml[k,]-beta[-1]))
}
names(beta_cml_norm) = c("item2","item3","item4","item5","item6", "item7",
                         "item8","item9", "item10")
```

```
#save(beta_cml_norm, file = "simstudy_beta_cml_norm.RData")
load("simstudy_beta_cml_norm.RData")

#Histograms for normalized estimates associated to each item

x <- seq(min(beta_cml_norm$item3), max(beta_cml_norm$item3), length = 1000)
fun = dnorm(x)
par(mfrow=c(3,3))
for(j in 1:(p-1)){
  hist(prob=TRUE,beta_cml_norm[,j], breaks=15, main = NULL,ylim = c(0,0.43),
           xlim = c(-3,3), xlab =  paste("Normalized CML  for Item", j+1))
  lines(x, fun, col = 2, lwd = 2)
}
library(MVN)
mvn(beta_cml_norm)
cov(beta_cml_norm)
```

Goodness of fit tests

```
Z = rep(0,1000)
for(k in 1:K){
  vec = rep(0,p-1)
  data = simdata[[k]]
  cml = clogistic(response ~ item, strata = subject, data)
  data_wide =data.frame(matrix(nrow=n, ncol = p))
  for(i in 1:n){
    #data_wide[i,]=data$response[data$subject==i]
  }
  subjectscore = rowSums(data_wide)
  table(subjectscore)

  for(s in c(1:(p-1))){
   cml_s<- clogistic(response ~ item, strata = subject,
```

```
                    data[data$subject %in% which(subjectscore == s), ])
    vec[s] = cml_s$loglik[2]
  }

Z[[k]] =2*(sum(vec) - cml$loglik[2])
}
#save(Z, file = "simstudy_Z.RData")
load("simstudy_Z.RData")
df = (p-1)*(p-2)
x <- seq(min(Z), max(Z), length = 1000)
chisq <- dchisq(x, df)
hist(Z, prob = TRUE,ylim = c(0,0.035), xlim = c(30, max(Z)),
                  main = "Histogram of GOF Test Statistics")
lines(x, chisq, col = 2, lwd = 2)
qqplot(Z,rchisq(1000,df), xlab = "Z", ylab = "Theoretical",
                  main = "QQ-Plot of GOF test statistics against ")
qqline(Z,distribution = function(p) qchisq(p,df=df), col = "steelblue", lwd = 2)
sum(1-pchisq(Z, df)<0.05)
```

C | Proof of Asymptotic Results of the Laplace Approximation

This Appendix is meant as a supplement to Section 4.2 containing proofs of the asymptotic results of the Laplace Approximation.

Proof of Theorem 4.2.2

Proof. Let $\delta > 0$ be the constant that exists according to condition 3 and define $A_\delta = [\hat{x} - \delta, \hat{x} + \delta]$ and $A_\delta^c = \mathbb{R} \setminus A_\delta$. Since

$$\begin{aligned}\frac{I_n}{\exp(nh(\hat{x}))g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} &= \frac{I_n \exp(-nh(\hat{x}))}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} \\ &= \frac{\int_{A_\delta} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} + \frac{\int_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} \end{aligned} \tag{C.1}$$

it is sufficient to show that the first term of Equation (C.1) converges to 1 and the second term converges to 0.

To show that the second term in Equation (C.1) converges to 0 in the case $\int_{\mathbb{R}} |g(x)|\mathrm{d}x \leq K_a$ (condition 4a), note that

$$\begin{aligned}\int_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x &\leq \exp(-n\epsilon) \int_{A_\delta^c} g(x)\mathrm{d}x \\ &\leq \exp(-n\epsilon) \int_{\mathbb{R}} |g(x)|\mathrm{d}x \\ &\leq \exp(-n\epsilon)K_a\end{aligned}$$

where condition 2 is applied in the first inequality choosing $\Delta = \delta$ and condition 4a is applied in the final inequality. This implies that

$$\left| \frac{\int_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} \right| \leq \frac{\exp(-n\epsilon)K_a}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} \xrightarrow{n\to\infty} 0.$$

Similarly, in the case $\int_{\mathbb{R}} \exp(h(x))|g(x)|\mathrm{d}x \leq K_b$ (condition 4b),

$$\begin{aligned}\int_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x &= \int_{A_\delta^c} \exp((n-1)(h(x) - h(\hat{x})) + h(x) - h(\hat{x}))g(x)\mathrm{d}x \\ &\leq \exp(-(n-1)\epsilon)\exp(-h(\hat{x}))\int_{A_\delta^c} \exp(h(x))|g(x)|\mathrm{d}x \\ &\leq \exp(-(n-1)\epsilon)\exp(-h(\hat{x}))K_b\end{aligned}$$

where condition 2 is applied in the first inequality choosing $\Delta = \delta$ and condition 4b is applied in the final inequality. Thus

$$\left|\frac{\int_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}}\right| \leq \frac{\exp(-(n-1)\epsilon)\exp(-h(\hat{x}))K_b}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} \overset{n\to\infty}{\longrightarrow} 0.$$

In order to see that the first term in Equation (C.1) converges to 1, define

$$\begin{aligned}J_n &= \int_{A_\delta} \exp(nh(x) - nh(\hat{x}))g(x)\mathrm{d}x \\ &\overset{u=\sqrt{n}(x-\hat{x})}{=} n^{-1/2}\int_{B_{n,\delta}} \exp(nh(n^{-1/2}u + \hat{x}) - nh(\hat{x}))g(n^{-1/2}u + \hat{x})\mathrm{d}u\end{aligned}$$

where $B_{n,\delta} = [-\sqrt{n}\delta, \sqrt{n}\delta]$, so that it has to be shown that

$$\frac{J_n}{g(\hat{x})\sqrt{2\pi n^{-1}H^{-1}}} = \frac{\sqrt{n}J_n}{g(\hat{x})\sqrt{2\pi H^{-1}}} \to 1.$$

Note that

$$g(\hat{x})\sqrt{2\pi H^{-1}} = \int_{\mathbb{R}} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x$$

and let

$$f_n(u) = \exp(nh(n^{-1/2}u + \hat{x}) - nh(\hat{x}))g(n^{-1/2}u + \hat{x}) - \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})$$

such that

$$\begin{aligned}
&\left|\sqrt{n}J_n - \int_{\mathbb{R}} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x\right| \\
&= |\int_{B_{n,\delta}} \exp(nh(n^{-1/2}u + \hat{x}) - nh(\hat{x}))g(n^{-1/2}u + \hat{x})\mathrm{d}u - \int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x \\
&\qquad - \int_{B_{n,\delta}^c} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x| \\
&\leq \left|\int_{B_{n,\delta}} f_n(u)\mathrm{d}u\right| + \left|\int_{B_{n,\delta}^c} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x\right|.
\end{aligned}$$

Since

$$\lim_{n\to\infty} \left|\int_{B_{n,\delta}^c} \exp\left(-\frac{H}{2}u^2\right) g(\hat{x})\mathrm{d}x\right| = 0$$

it is sufficient to show that

$$\left|\int_{B_{n,\delta}} f_n(u)\mathrm{d}u\right| \overset{n\to\infty}{\longrightarrow} 0.$$

Consider the function $p(u) = nh(n^{-1/2}u + \hat{x})$ such that

$$\begin{aligned}
p(0) &= nh(\hat{x}), \\
p'(0) &= \sqrt{n}h'(\hat{x}) = 0, \\
p''(0) &= h''(\hat{x}) = -H.
\end{aligned}$$

A second order taylor expansion of $p(u)$ around zero, utilizing that h is three times differentiable and that linear combinations of differentiable functions are differentiable, yields

$$nh(n^{-1/2}u+\hat{x})-nh(\hat{x})=-\frac{H}{2}u^2+R_n(u)$$

where

$$R_n(u)=nn^{-3/2}\frac{h^{(3)}(n^{-1/2}c'+\hat{x})}{6}u^3=n^{-1/2}\frac{h^{(3)}(c)}{6}u^3$$

where c' is between zero and u and $c=n^{-1/2}c'+\hat{x}$ is between $\hat{x}$ and $n^{-1/2}u+\hat{x}$.

Since $u\in B_{n,\delta}$ implies that $|\hat{x}-c|<\delta$ such that condition 3 implies that $|h^{(3)}(c)|\leq K$ and hence $R_n(u)\to 0$ for $n\to\infty$, it follows for any $u\in\mathbb{R}$ that

$$\mathbb{1}[u\in B_{n,\delta}]\left(\exp\left(nh(n^{-1/2}u+\hat{x})-nh(\hat{x})\right)g(n^{-1/2}u+\hat{x})-\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})\right)\overset{n\to\infty}{\longrightarrow}0.$$

Since

$$\mathbb{1}[u\in B_{n,\delta}]\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})\overset{n\to\infty}{\longrightarrow}\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})$$

it follows that

$$\mathbb{1}[u\in B_{n,\delta}]\exp\left(nh(n^{-1/2}u+\hat{x})-nh(\hat{x})\right)g(n^{-1/2}u+\hat{x})\overset{n\to\infty}{\longrightarrow}\exp\left(-\frac{H}{2}u^2\right)g(\hat{x}).$$

Furthermore, for $u\in B_{n,\delta}$ then $|u|\leq\sqrt{n}\delta$ such that

$$|R_n(u)|\leq\frac{K}{6}\delta u^2,\quad|(n^{-1/2}u+\hat{x})-\hat{x}|\leq\delta$$

and hence, by applying condition 3,

$$\mathbb{1}[u\in B_{n,\delta}]\exp\left(nh(n^{-1/2}u+\hat{x})-nh(\hat{x})\right)|g(n^{-1/2}u+\hat{x})|\leq\exp\left(-\frac{H}{2}u^2+\frac{K}{6}\delta u^2\right)C$$

is obtained, where the righthand side is integrable for sufficiently small δ.

It follows by two applications of Lebesgue's dominated convergence theorem that

$$\begin{aligned}\int\limits_{B_{n,\delta}}f_n(u)\mathrm{d}u&=\int\limits_{B_{n,\delta}}\exp\left(nh(n^{-1/2}u+\hat{x})-nh(\hat{x})\right)g(n^{-1/2}u+\hat{x})\mathrm{d}u-\int\limits_{B_{n,\delta}}\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})\mathrm{d}u\\&\overset{n\to\infty}{\longrightarrow}\int\limits_{\mathbb{R}}\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})\mathrm{d}u-\int\limits_{\mathbb{R}}\exp\left(-\frac{H}{2}u^2\right)g(\hat{x})\mathrm{d}u=0\end{aligned}$$

and hence

$$\left|\int\limits_{B_{n,\delta}}f_n(u)\mathrm{d}u\right|\overset{n\to\infty}{\longrightarrow}0.$$

□

Proof of Theorem 4.2.3

Proof. Assume for simplicity that $g(x) = 1$. Note that this implies that condition 4b is satisfied since 4a is not.

From Equation (C.1) in the proof of Theorem 4.2.2 it is known that

$$\frac{I_n}{\exp(nh(\hat{x}))\sqrt{2\pi n^{-1}H^{-1}}} = \frac{\sqrt{n}J_n}{\sqrt{2\pi H^{-1}}} + \frac{\int\limits_{A_\delta^c} \exp(nh(x) - nh(\hat{x}))\mathrm{d}x}{\sqrt{2\pi n^{-1}H^{-1}}}.$$

Therefore, as it was shown that the second term has an upper bound of order $O(\exp(-n))$, it is sufficient to show that

$$\frac{\sqrt{n}J_n}{\sqrt{2\pi H^{-1}}} = 1 + O(n^{-1}).$$

As in the proof of Theorem 4.2.2, consider at taylor expansion of $nh(n^{-1/2}u + \hat{x})$, this time of order three, around zero:

$$nh(n^{-1/2}u + \hat{x}) - nh(\hat{x}) = -\frac{H}{2}u^2 + n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u) \tag{C.2}$$

where $R_n(u) = n^{-1}\frac{h^{(4)}(c)}{24}u^4$ for c between $\hat{x}$ and $n^{-1/2}u + \hat{x}$. Furthermore, a first order taylor expansion of $\exp(\cdot)$ around zero yields

$$\exp\left(n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)\right) = 1 + n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u) + \frac{\exp(c_n')}{24}\left(n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)\right)^2 \tag{C.3}$$

for c_n' between zero and $n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)$.

Inserting Equation (C.2) and then Equation (C.3) into $\sqrt{n}J_n$ yields

$$\begin{aligned}
\sqrt{n}J_n &= \int\limits_{B_{n,\delta}} \exp\left(nh(n^{-1/2}u + \hat{x}) - nh(\hat{x}))\right)\mathrm{d}u \\
&= \int\limits_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2 + n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)\right)\mathrm{d}u \\
&= \int\limits_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right)\left(1 + n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u) + \frac{\exp(c_n')}{24}\left(n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)\right)^2\right)\mathrm{d}u.
\end{aligned}$$

By splitting the integral into four terms, each of the terms can be assessed seperatly.

It follows by Lebesgue's dominated convergence theorem, utilizing that

$$\left|\mathbb{1}[u \in B_{,\delta}] \exp\left(-\frac{H}{2}u^2\right)\right| \leq \exp\left(-\frac{H}{2}u^2\right)$$

for all $n \in \mathbb{N}$, that the limit of the first term is given by

$$\lim_{n\to\infty} \int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) \mathrm{d}u = \int_{\mathbb{R}} \exp\left(-\frac{H}{2}u^2\right) \mathrm{d}u = \sqrt{2\pi H^{-1}}.$$

Furthermore, by the symmetry of the second term it follows that

$$\begin{aligned}
&\int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) \frac{h^3(\hat{x})}{6} u^3 \mathrm{d}u \\
&= \left(\int_{-\sqrt{n}\delta}^{0} \exp\left(-\frac{H}{2}u^2\right) \frac{h^3(\hat{x})}{6} u^3 \mathrm{d}u + \int_{0}^{\sqrt{n}\delta} \exp\left(-\frac{H}{2}u^2\right) \frac{h^3(\hat{x})}{6} u^3 \mathrm{d}u \right) \\
&= \left(- \int_{0}^{\sqrt{n}\delta} \exp\left(-\frac{H}{2}u^2\right) \frac{h^3(\hat{x})}{6} (-u)^3 (-1) \mathrm{d}u + \int_{0}^{\sqrt{n}\delta} \exp\left(-\frac{H}{2}u^2\right) \frac{h^3(\hat{x})}{6} u^3 \mathrm{d}u \right) \\
&= 0.
\end{aligned}$$

Multiplying the third term by n yields

$$\begin{aligned}
&\lim_{n\to\infty} \int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) n R_n(u) \mathrm{d}u \\
&= \lim_{n\to\infty} \int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) \frac{h^{(4)}(c)}{24} u^4 \mathrm{d}u \\
&\leq \lim_{n\to\infty} \int_{B_{n,\delta}} \left| \exp\left(-\frac{H}{2}u^2\right) \frac{h^{(4)}(c)}{24} u^4 \right| \mathrm{d}u \\
&\leq \lim_{n\to\infty} \int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) \frac{C'}{24} u^4 \mathrm{d}u \\
&= \int_{\mathbb{R}} \exp\left(-\frac{H}{2}u^2\right) \frac{C'}{24} u^4 \mathrm{d}u < \infty
\end{aligned}$$

where condition 3 is used in the second inequality and Lebesgue's dominated convergence theorem is used in the second equality. Since the third term muliplied by n is $O(1)$ the third term itself must be $O(n^{-1})$.

For the fourth term, recall that c_n' is between zero and $n^{-1/2}\frac{h^3(\hat{x})}{6}u^3 + R_n(u)$.

Furthermore, for $u \in B_{n,\delta}$

$$\left| n^{-1/2} \frac{h^{(3)}(\hat{x})}{6} u^3 \right| \leq \frac{h^{(3)}(\hat{x})}{6} \delta u^2, \quad |R_n(u)| = \left| n^{-1} \frac{h^{(4)}(c)}{24} u^4 \right| \leq \frac{h^{(4)}(c)}{24} \delta^2 u^2$$

such that

$$c_n' \leq \left| n^{-1/2} \frac{h^3(\hat{x})}{6} u^3 \right| + |R_n(u)| \leq \frac{h^{(3)}(\hat{x})}{6} \delta u^2 + \frac{h^{(4)}(c)}{24} \delta^2 u^2$$

and hence $\exp(-\frac{H}{2}u^2 + c_n')$ is dominated by an unnormalized normal density, say $\phi(u)$, on $B_{n,\delta}$ for sufficiently small δ.

Therefore, since

$$\left(n^{-1/2} \frac{h^{(3)}(\hat{x})}{6} u^3 + R_n(u) \right)^2 \leq n^{-1} \frac{h^{(3)}(\hat{x})^2}{36} u^6 + n^{-2} \frac{C'^2}{576} u^8 + n^{-3/2} \frac{h^{(3)}(\hat{x})^2 C'}{144} u^7$$

it follows that

$$\begin{aligned}
&\int_{B_{n,\delta}} \exp\left(-\frac{H}{2}u^2\right) \frac{\exp(c_n')}{24} \left(n^{-1/2} \frac{h^3(\hat{x})}{6} u^3 + R_n(u) \right)^2 \mathrm{d}u \\
&\leq \int_{B_{n,\delta}} \frac{\phi(u))}{24} \left(n^{-1/2} \frac{h^3(\hat{x})}{6} u^3 + R_n(u) \right)^2 \mathrm{d}u \\
&\leq n^{-1} \frac{h^{(3)}(\hat{x})^2}{36} \frac{1}{24} \int_{B_{n,\delta}} \phi(u) u^6 \mathrm{d}u + n^{-2} \frac{C'^2}{576} \frac{1}{24} \int_{B_{n,\delta}} \phi(u) u^8 \mathrm{d}u + n^{-3/2} \frac{h^{(3)}(\hat{x})^2 C'}{144} \int_{B_{n,\delta}} \phi(u) u^7 \mathrm{d}u
\end{aligned}$$

and hence the fourth term is $O(n^{-1})$ since any order moments of a normal distribution is finite.

In conclusion, it is clear that

$$\frac{\sqrt{n} J_n}{\sqrt{2\pi H^{-1}}} = 1 + O(n^{-1}).$$

□

D | Monte Carlo Methods

This appendix regarding computations of the likelihood function for GLMMs using Monte Carlo methods is based on [24] and is meant as a supplement to Section 4.2.

The integral given by Equation (4.8) can be written as the expectation

$$\int_{\mathbb{R}} f(y \mid u; \beta) f(u; \sigma^2) \mathrm{d}u = \mathbb{E}_{\sigma^2}[f(y \mid U; \beta)],$$

and hence it follows that integral can be approximated using monte carlo methods. First *simple Monte Carlo* is considered, which estimates the expectation as

$$\mathbb{E}_{\sigma^2}[f(y \mid U; \beta)] \approx \frac{1}{M} \sum_{k=1}^{M} f(y \mid U_k; \beta)$$

where $U_k \sim N(0, \sigma^2)$ for $k = 1, \ldots, M$. The variance of the simple Monte Carlo estimate is given as

$$\mathbb{V}\mathrm{ar}\left[\frac{1}{M} \sum_{k=1}^{M} f(y \mid U_k; \beta)\right] = \frac{1}{M} \mathbb{V}\mathrm{ar}\left[f(y \mid U; \beta)\right]$$

and hence it follows that the *Monte Carlo error*, which is the standard deviation of the Monte carlo estimate, is of order $\frac{1}{\sqrt{M}}$.

In application $\mathbb{V}\mathrm{ar}\left[f(y \mid U; \beta)\right]$ can be estimated using the usual emperical variance estimate

$$\mathbb{V}\mathrm{ar}\left[f(y \mid U; \beta)\right] \approx \frac{1}{M-1} \sum_{k=1}^{M} \left(f(y \mid U_k; \beta) - \frac{1}{M} \sum_{k=1}^{M} f(y \mid U_k; \beta) \right)^2.$$

It should be noted that for a particular dataset y based on a specific realization, say u_0, of $U \sim N(0, \sigma^2)$, the density $f(y \mid U; \beta)$ would usually be heavily concentrated around u_0 such that $f(y \mid U; \beta)$ will have a large variance. This problem is particularly impactfull in higher dimensions because of the curse of dimensionality, effectively making the approach unusable in such settings.

Clearly if $f(y \mid U; \beta)$ has a high variance then M should be chosen very large which can be computational infesible.

Therefore, other approaches such as *importance sampling* is needed. Importance sampling can be used in a more general setup but will now be introduced for our context for GLMMs.

Suppose there exists a denisty g such that

$$\frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)} \approx \text{constant}, \quad f(y \mid u;\beta)f(u;\sigma^2) > 0 \implies g(u) > 0.$$

Then

$$\begin{aligned}\int_{\mathbb{R}} f(y \mid u;\beta)f(u;\sigma^2)\mathrm{d}u &= \int_{\mathbb{R}} \frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)} g(u)\mathrm{d}u \\ &= \mathbb{E}\left[\frac{f(y \mid V;\beta)f(V;\sigma^2)}{g(V)}\right] \end{aligned} \tag{D.1}$$

where V denotes a random variable with denisty g. Furthermore, it follows that applying the simple Monte Carlo approximation on Equation (D.1) yields

$$\int_{\mathbb{R}} f(y \mid u;\beta)f(u;\sigma^2)\mathrm{d}u \approx \frac{1}{M}\sum_{k=1}^{M} \frac{f(y \mid V_k;\beta)f(V_k;\sigma^2)}{g(V_k)}$$

where V_k for $k = 1, \ldots, M$ has density g. It follows by the choice of g that $\frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)}$ has low variance and hence the Monte Carlo error is small even though $f(y \mid U;\beta)$ might have a high variance.

The question at hand of course becomes how g should be chosen. In order to ensure a finite Monte Carlo error it is evident that $\frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)}$ ought to be bounded and therefore g should have heavy tails.

Similar to the justification for adaptive Gauss-Hermite quadrature, recall that Equation (4.9) implies that $f(y \mid u;\beta)f(u;\sigma^2)$ is approximately proportional to a normal density with mean μ_{LP} and variance σ^2_{LP}. Therefore if g is chosen to be said normal density, then $\frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)}$ is approximately constantly equal to the normalizing constant $f(y;\beta,\sigma^2)$. Alternatively, in order to obtain heavy tails g could be chosen as the density for the t-distribution with the same parameters. Care needs to be taken when choosing the degrees of freedom, because if it is chosen too small then the approximation might not work well and reversely if the degrees of freedom is chosen too big then the tails dont become significantly heavier than the tailes of the normal distribution.

An another option also utilizing that

$$\frac{f(y \mid u;\beta)f(u;\sigma^2)}{f(u \mid y;\beta,\sigma^2)} = f(y;\beta,\sigma^2)$$

would be as follows: Fix $\beta_0 \in \mathbb{R}^p$ and $\sigma_0^2 \in \mathbb{R}_+$ and define

$$g(u) = f(u \mid y;\beta_0,\sigma_0^2) = \frac{f(y \mid u;\beta_0)f(u;\sigma_0^2)}{f(y;\beta_0,\sigma_0^2)}$$

such that

$$\begin{aligned}\int_{\mathbb{R}} f(y \mid u;\beta)f(u;\sigma^2)\mathrm{d}u &= \int_{\mathbb{R}} \frac{f(y \mid u;\beta)f(u;\sigma^2)}{g(u)} g(u)\mathrm{d}u \\ &= f(y;\beta_0,\sigma_0^2)\int_{\mathbb{R}} \frac{f(y \mid u;\beta)f(u;\sigma^2)}{f(y \mid u;\beta_0)f(u;\sigma_0^2)} f(u \mid y;\beta_0,\sigma_0^2)\mathrm{d}u \\ &= f(y;\beta_0,\sigma_0^2)\mathbb{E}_{(\beta_0,\sigma_0^2)}\left[\frac{f(y \mid U;\beta)f(U;\sigma^2)}{f(y \mid U;\beta_0)f(U;\sigma_0^2)} \mid Y = y\right].\end{aligned}$$

Since

$$f(y;\beta,\sigma^2) = \int_{\mathbb{R}} f(y \mid u;\beta)f(u;\sigma^2)\mathrm{d}u$$

it follows that

$$\frac{f(y;\beta,\sigma^2)}{f(y;\beta_0,\sigma_0^2)} = \mathbb{E}_{(\beta_0,\sigma_0^2)}\left[\frac{f(y \mid U;\beta)f(U;\sigma^2)}{f(y \mid U;\beta_0)f(U;\sigma_0^2)} \mid Y = y\right]. \tag{D.2}$$

Since $f(y;\beta_0,\sigma_0^2)$ is constant wrt. β and σ^2 it follows that the maximum likelihood estimate can be found as

$$\begin{aligned}(\hat{\beta},\hat{\sigma}^2) &= \arg\max_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+} \frac{f(y;\beta,\sigma^2)}{f(y;\beta_0,\sigma_0^2)} \\ &\approx \arg\max_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+} \frac{1}{M}\sum_{k=1}^{M} \frac{f(y \mid U_k;\beta)f(U_k;\sigma^2)}{f(y \mid U_k;\beta_0)f(U_k;\sigma_0^2)}\end{aligned}$$

where the approximation follows from the Monte Carlo estimate of Equation (D.2) and $U_1,\dots U_M$ are i.i.d. with density $f(u \mid y;\beta_0,\sigma_0^2)$.

However, considerations have to be made regarding how to simulate from $U|(Y = y)$.

Although $f(u \mid y;\beta,\sigma^2)$ is a non-standard density

$$f(u \mid y;\beta,\sigma^2) = \frac{f(y \mid u;\beta_0)f(u;\sigma^2)}{f(y;\beta,\sigma^2)}$$

since $f(y \mid u;\beta)f(u;\sigma^2)$ is well known by the definition of a GLMM, it follows that $f(u \mid y;\beta,\sigma^2)$ is known up to proportionality. Therefore, *rejection sampling* can be utilized.

Proposition D.0.1. Rejection Sampling

Let $f(x) \propto h(x)$ be a density and assume there exists a density g and $K \in \mathbb{R}$ such that

$$h(x) \leq Kg(x).$$

Generate X with density g, $W \sim \text{unif}[0,1]$ and accept X if $W \leq \frac{h(X)}{Kg(X)}$. Then the conditional density of X given that it has been accepted is f and the probability of accept is given by $\int_{\mathbb{R}} h(x)\mathrm{d}x/K$.

Proof. The probability of accept is given by

$$\begin{aligned}
\mathbb{P}\left(W \leq \frac{h(X)}{Kg(X)}\right) &= \int_{\mathbb{R}} \int_0^{\frac{h(x)}{Kg(x)}} g(x)\mathrm{d}w\mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{h(x)}{Kg(x)} g(x)\mathrm{d}x \\
&= \frac{\int_{\mathbb{R}} h(x)\mathrm{d}x}{K}
\end{aligned}$$

Furthermore, the conditional distribution function of X given accept is given by

$$\begin{aligned}
\mathbb{P}\left(X \leq y \mid W \leq \frac{h(X)}{Kg(X)}\right) &= \frac{\mathbb{P}\left(X \leq y, W \leq \frac{h(X)}{Kg(X)}\right)}{\mathbb{P}\left(W \leq \frac{h(X)}{Kg(X)}\right)} \\
&= K \frac{\int_{-\infty}^{y} \int_0^{\frac{h(x)}{kg(x)}} g(x)\mathrm{d}w\mathrm{d}x}{\int_{\mathbb{R}} h(x)\mathrm{d}x} \\
&= \frac{\int_{-\infty}^{y} h(x)\mathrm{d}x}{\int_{\mathbb{R}} h(x)\mathrm{d}x} \\
&= \int_{-\infty}^{y} f(x)\mathrm{d}x
\end{aligned}$$

□

Simply recognizing $f(u \mid y; \beta, \sigma^2)$, $f(y \mid u; \beta) f(u; \sigma^2)$ and e.g. $t_d(u; \mu_{\text{LP}}, \sigma^2_{\text{LP}})$ for $d \in \mathbb{N} \backslash \{0\}$ as respectively $f(z), h(z)$ and $g(z)$ in Proposition D.0.1 immediately yields a method to simulate from the conditional distribution of U given y.

It should be noted that the density for any t-distribution would work for this purpose, but by choosing it such that it is centered at the Laplace mean μ_{LP} and scaled by the σ^2_{LP} the K is kept as small as possible so that the probability of accept is as large is possible.

Being able to simulate from the conditional distribution of U given y has other usages than just the one mentioned above.

For instance, the conditional mean described in Proposition (4.1.3) and Example 4.2.4 can be estimated using Monto Carlo approximation simply by

$$\mathbb{E}\left[U \mid Y = y\right] \approx \frac{1}{M} \sum_{k=1}^{M} U_k$$

where $U_1, \ldots, U_M$ are i.i.d with density $f(u \mid y; \beta, \sigma^2)$.

Furthermore, in Section 6.2 simulations are made from $U \mid Y = y$ in order to estimate the marginal score and observed information.

E | The EM Algorithm

This appendix containing the proof of Theorem 5.3.2 and a result regarding the convergence of the EM algorithm is meant a supplement to Section 5.3.

Proof of Theorem 5.3.2

Proof. In the proof of Proposition 5.2.7 it was shown that

$$s_{\text{obs}}(\theta \mid r, y_{(r)}) = \frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)}.$$

Taking the partial derivative wrt. $\theta^\top$ yields

$$\begin{aligned}
&\frac{\partial}{\partial \theta^\top} s_{\text{obs}}(\theta \mid r, y_{(r)}) \\
&= \frac{\frac{\partial}{\partial \theta^\top} \left(\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \right) p_{Y_{(r)}}(y_{(r)}; \theta)}{p_{Y_{(r)}}(y_{(r)}; \theta)^2} && \text{(E.1)} \\
&\quad - \frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \frac{\partial}{\partial \theta^\top} \left(\int p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \right)}{p_{Y_{(r)}}(y_{(r)}; \theta)^2}. && \text{(E.2)}
\end{aligned}$$

Equation (E.1) is given by

$$\begin{aligned}
&\frac{\frac{\partial}{\partial\theta^\top}\left(\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})\right)}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= \frac{\int -J_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) + s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) \frac{\partial}{\partial\theta^\top} p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= \int \frac{-J_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta)}{p_{Y_{(r)}}(y_{(r)}; \theta)} + \frac{s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})})^\top p_Y(y_{(r)}, y_{(\bar{r})}; \theta)}{p_{Y_{(r)}}(y_{(r)}; \theta)} \mathrm{d}\nu(y_{(\bar{r})}) \\
&= \int -J_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_{Y \mid Y_{(r)}}(y_{(r)}, y_{(\bar{r})} \mid y_{(r)}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \\
&\quad + \int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})})^\top p_{Y \mid Y_{(r)}}(y_{(r)}, y_{(\bar{r})} \mid y_{(r)}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \\
&= -\mathbb{E}\left[J_{\text{full}}(\theta \mid Y) \mid R = r, Y_{(R)} = y_{(r)}\right] + \mathbb{E}\left[s_{\text{full}}(\theta \mid Y) s_{\text{full}}(\theta \mid Y)^\top \mid R = r, Y_{(R)} = y_{(r)}\right]
\end{aligned} \tag{E.3}$$

where the first equality follows by interchanging the integral and differentiation, the second equality follows since $\frac{\partial}{\partial\theta^\top} p_Y(y;\theta) = s_{\text{full}}(\theta \mid y)^\top p_Y(y;\theta)$ and the fourth equality follows by Lemma 5.2.6. By similar arguments it follows that Equation (E.2) is given by

$$\begin{aligned}
&- \frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})}) \frac{\partial}{\partial\theta^\top}\left(\int p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})\right)}{p_{Y_{(r)}}(y_{(r)}; \theta)^2} \\
&= -\frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})}) p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)} \frac{\int s_{\text{full}}(\theta \mid y_{(r)}, y_{(\bar{r})})^\top p_Y(y_{(r)}, y_{(\bar{r})}; \theta) \mathrm{d}\nu(y_{(\bar{r})})}{p_{Y_{(r)}}(y_{(r)}; \theta)} \\
&= -\mathbb{E}\left[s_{\text{full}}(\theta \mid Y) \mid R = r, Y_{(R)} = y_{(r)}\right] \mathbb{E}\left[s_{\text{full}}(\theta \mid Y)^\top \mid R = r, Y_{(R)} = y_{(r)}\right].
\end{aligned} \tag{E.4}$$

The result follows by summing over the n subjects. □

Convergence of the EM Algorithm

Theorem E.0.1. Convergence of the EM Aglorithm

Suppose that $\{\ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})\}_{k \in \mathbb{N}_0}$ is a bounded sequence and that

$$Q(\theta^{(k+1)}; \theta^{(k)}) - Q(\theta^{(k)}; \theta^{(k)}) \geq K \|\theta^{(k+1)} - \theta^{(k)}\|_2^2 \quad \text{for } k = 0, 1, \ldots \tag{E.5}$$

where $\|\cdot\|_2$ denotes the L^2-norm. Then for some θ^* in the closure of the Θ it follows that

$$\theta^{(k)} \underset{k \to \infty}{\longrightarrow} \theta^*.$$

Proof. It follows immdiatly from Theorem 5.3.4 that boundedness of $\ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})$ implies that the sequence converges to some $C < \infty$. In particular, it follows that $\left(\ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})\right)_{k \in \mathbb{N}_0}$ is a Cauchy sequence such that for any $\epsilon > 0$ there exists $k(\epsilon)$ s.t. $k \geq k(\epsilon), r \geq 1$ implies that

$$\sum_{j=1}^{r} \left(\ell_{\text{obs}}(\theta^{(k+j)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) - \ell_{\text{obs}}(\theta^{(k+j-1)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})\right) = \ell_{\text{obs}}(\theta^{(k+r)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) - \ell_{\text{obs}}(\theta^{(k)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})$$
$$< \epsilon. \tag{E.6}$$

Furthermore, in the proof of Theorem 5.3.4 it is shown that

$$0 \leq Q(\theta^{(k+j)}; \theta^{(k)}) - Q(\theta^{(k+j-1)}; \theta^{(k)}) \leq \ell_{\text{obs}}(\theta^{(k+j)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})}) - \ell_{\text{obs}}(\theta^{(k+j-1)} \mid \mathbf{r}, \mathbf{y}_{(\mathbf{r})})$$

for $j = 1, \ldots, r$ such that Equation (E.5) and (E.6) implies that

$$K \sum_{j=1}^{r} \|\theta^{(k+j)} - \theta^{(k+j-1)}\|_2^2 \leq \sum_{j=1}^{r} \left(Q(\theta^{(k+j)}; \theta^{(k)}) - Q(\theta^{(k+j-1)}; \theta^{(k)})\right)$$
$$< \epsilon.$$

Applying the triangle inequality implies that

$$K \| \sum_{j=1}^{r} \theta^{(k+j)} - \theta^{(k+j-1)} \|_2^2 < \epsilon$$

such that

$$\|\theta^{(k+r)} - \theta^{(k)}\|_2 < \sqrt{\frac{\epsilon}{K}}$$

showing that $\left(\theta^{(k)}\right)_{k \in \mathbb{N}_0}$ is a Cauchy sequence. Hence it follows by the completeness of the closure of Θ that the sequence converges to some θ^* in the closure of Θ. □

F | R Code Assuming Ignorable Missing Data Mechanism

This appendix contains the R code for the maximization of respectively the joint, conditional and marginal likelihood discussed in Section 6.1. In particular, the R code was converted to latex code using RMarkdown and the knit function.

Formatting the data

```
cfa=read.csv("cfa.csv")
response = rep(0,0,663*36)
 for (i in 1:663){
   for(j in 6:41){
     response[(i-1)*36+(j-5)] = cfa[i,j]
     }
   }
 item = factor(rep(1:36,663))
 subject = factor(rep(1:663, each=36))
 data = data.frame(subject,item,response)
#save(data, file = "data.RData")
#load("data.RData")
```

Joint maximum likelihood

```
jml <- glm( response ~ -1 + subject + item, data = data, family = binomial)


summary(jml)


beta_jml=c("item1"=0, -coef(jml)[(698-34):698])
theta_sd_jml=sqrt(var(coef(jml)[1:(698-35)]))
```

Conditional maximum likelihood

```
library(Epi)
cml <- clogistic(response ~ item, strata = subject, data)
beta_cml=c("item1"=0,-coef(cml))
```

Marginal maximum likelihood

```
library(lme4)
mml = glmer(response ~-1+(1|subject)+item, family = binomial, data = data)

mmlGH5 = glmer(response ~-1+(1|subject)+item, family = binomial, data = data,
            nAGQ = 5)
summary(mml)
summary(mmlGH5)
beta_mml = -(as.numeric(coef(mmlGH5)[[1]][2,][-1])
             - as.numeric(coef(mmlGH5)[[1]][2,][2]))
```

Comparison

```
round(rbind("jml"=beta_jml, "cml2"=beta_cml, "mml"=beta_mml),digits=2)
norm(beta_jml - beta_cml, type = "2")
norm(beta_mml - beta_cml, type = "2")
norm(beta_mml - beta_jml, type = "2")
order(beta_jml)
order(beta_cml)
order(beta_mml)
```

Goodness of fit test

```
subjectscore = rowSums(cfa[6:41], na.rm  = TRUE )
table(subjectscore)
vec = rep(0,35)
for(s in c(1:33,35)){
  cml_s<- clogistic(response ~ item, strata = subject,
```

```
            NormalTokdata[data$subject %in% which(subjectscore == s), ])
  vec[s] = cml_s$loglik[2]
}

Z =2*(sum(vec) - cml$loglik[2])
df = 35*34
1-pchisq(Z, df)
```

G | R Code for Parameter Estimation When Modelling Dropout Effect

This appendix contains the R code for the implementation and maximization of the marginal likelihood discussed in Section 5.4. In particular, the R code was converted to latex code using RMarkdown and the knit function.

Loading relevant data and packages needed in the following, and defining the initial value.

```
load("data.RData")
load("beta_mml.rda")
library(mvtnorm)
library(parallel)
init = c("beta"=-beta_mml, "tau"=-3, "eta"=0.1, "sigma_theta_log"=0,
       "sigma_xi_log"=0,"rho_tan"=pi/2*tan(0.1))
```

Defining the logarithm of the integrand, aka the g function, which is to be maximized in order to obtain Laplace approximation of integral.

```
g = function(par1, par2,i){

  y = data[data$subject == i,]$response

  diff_y = par1["theta"]-par2[1:36]
  diff_r = par1["xi"]-(par2["tau"]+(1:36)*par2["eta"])

  p_y_vec = exp(y*(diff_y))/(1+exp(diff_y))
```

```
  p_r_vec = exp(diff_r)/(1+exp(diff_r))

  p_y_log = sum(log(p_y_vec), na.rm = TRUE)

  k = max(which(!is.na(y)))+1

  if(k==37){
    p_r_log = sum(log(p_r_vec))
  }  else{
    p_r_log = sum(log(p_r_vec[1:(k-1)]))
          + unname(log(1/(1+exp(par1["xi"]-(par2["tau"]+k*par2["eta"])))))
  }

  p_y_log + p_r_log + log(dmvnorm(c(par1["theta"],par1["xi"]), mean = rep(0, 2),
       sigma = matrix(nrow= 2, ncol = 2, c( exp(par2["sigma_theta_log"])^2,
2/pi*atan(par2["rho_tan"])* exp(par2["sigma_theta_log"]) *exp(par2["sigma_xi_log"]),
2/pi*atan(par2["rho_tan"])* exp(par2["sigma_theta_log"]) *exp(par2["sigma_xi_log"]),
exp(par2["sigma_xi_log"])^2)), log = FALSE))

}
```

Defining the gradient of the g function wrt. θ and ξ, known in this code as "par1". This function will be supplied to the BFGS method in optim in order to maximize g.

```
score_g = function(par1, par2, i){
  diff_y = par1["theta"]-par2[1:36]
  diff_r = par1["xi"]-(par2["tau"]+(1:36)*par2["eta"])

  y = data[data$subject == i,]$response
  k = max(which(!is.na(y)))+1
```

```
  p_vec = exp((diff_y))/(1+exp(diff_y))
  p_r_vec =c (exp((diff_r))/(1+exp(diff_r),0)

  g_dtheta = sum((y-p_vec)[-which(is.na(y))]) + (2/pi*atan(par2["rho_tan"]))
     *par1["xi"]/((1-(2/pi*atan(par2["rho_tan"]))^2)
     *exp(par2["sigma_theta_log"])*exp(par2["sigma_xi_log"]))- par1["theta"]
     /((1-(2/pi*atan(par2["rho_tan"]))^2)*exp(par2["sigma_theta_log"])^2)



  g_dxi =  sum((rep(1,k-1)-p_r_vec)[1:(k-1)])  -p_r_vec)[k]
    + (2/pi*atan(par2["rho_tan"]))*par1["theta"]/((1-(2/pi*atan(par2["rho_tan"]))^2)
    *exp(par2["sigma_theta_log"])*exp(par2["sigma_xi_log"]))
    - par1["xi"]/((1-(2/pi*atan(par2["rho_tan"]))^2)*exp(par2["sigma_xi_log"])^2)

  c(g_dtheta,g_dxi)
}
```

Defining the Laplace function which estimates the Laplace approximation of the integral of $exp(g)$ for a given subject given parameters β, δ, ψ or equivalently $\beta, \tau, \eta, \sigma_\theta, \sigma_\xi, \rho$, known in this code as "par2".

```
Laplace = function(i, par2){
  #Optimize g function
  fit = optim(par =c("theta"=0,"xi"=0), fn = g, gr = score_g, par2=par2, i=i,
              control = list(fnscale = -1), method = "BFGS")



  #Calculate derivatives for hessian
  diff_y = fit$par[1]-par2[1:36]
  diff_r = fit$par[2]-(par2["tau"]+(1:36)*par2["eta"])

  y = data[data$subject == i,]$response
  k = max(which(!is.na(y)))+1
```

```
  c1 = - sum((exp((diff_y))/(1+exp(diff_y))^2)[1:(k-1)])
       - 1/((1-(2/pi*atan(par2["rho_tan"]))^2)*exp(par2["sigma_theta_log"])^2)

  c23 = (2/pi*atan(par2["rho_tan"]))/((1-(2/pi*atan(par2["rho_tan"]))^2)
        *exp(par2["sigma_theta_log"])*exp(par2["sigma_xi_log"]))

  c4 = - sum((exp((diff_r))/(1+exp(diff_r))^2)[1:(k-1)])
        - 1/((1-(2/pi*atan(par2["rho_tan"]))^2)*exp(par2["sigma_xi_log"])^2)

  hessian = matrix(nrow= 2, ncol = 2,  unname(c(c1, c23, c23, c4)))
  exp(g(fit$par, par2,i))*2*pi*sqrt(1/det(hessian))
}
```

Setting up parallelised computations and calculating the loglikelihood given parameters "par2".

```
cl = makeCluster(8)
clusterExport(cl, c("g","dmvnorm", "score_g", "data"))

loglikelihood = function(par2){
  sum(log(unlist(parLapply(cl = cl, 1:663, Laplace, par2=par2))))

}
```

Using the optim function to maximize the loglikelihood wrt. "par2".

```
fit = optim(par = init, fn =loglikelihood, control = list(maxit = 100000,
            fnscale = -1))

par = fit$par
# save(par, file="parameters.Rda")
#load("parameters.Rda")
```

```
 beta = par[1:36]-par[1]
loglikelihood(init)
loglikelihood(fit$par)
```