Multidimensional Modeling and Inference of Dichotomous Item Response Data

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Mia Johanna Katharina Kornely, M. Sc.
geborene Müller-Platz

aus

Bergisch Gladbach

Berichter: Universitätsprofessorin Dr. Maria Kateri
Professor Dr. Irini Moustaki

Tag der mündlichen Prüfung: 14. Mai 2021

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Eidesstattliche Erklärung

Mia Johanna Katharina Kornely, geborene Müller-Platz,

erklärt hiermit, dass diese Dissertation und die darin dargelegten Inhalte die eigenen sind und selbstständig, als Ergebnis der eigenen originären Forschung, generiert wurden.

Hiermit erkläre ich an Eides statt

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorandin dieser Fakultät und Universität angefertigt;
2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;
3. Wenn immer andere eigene- oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;
4. Wenn aus anderen eigenen- oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Diese Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;
5. Alle wesentlichen Quellen von Unterstützung wurden benannt;
6. Wenn immer ein Teil dieser Dissertation auf der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;
7. Kein Teil dieser Arbeit wurde vor deren Einreichung veröffentlicht.

13. Juli 2021

Unterschrift der Doktorandin

Acknowledgments

- First and foremost, I want to express my deepest gratitude to my first supervisor, Maria Kateri, and my second supervisor, Irini Moustaki, for continuously supporting my research in this exciting field. With your helpful comments and ideas, you have contributed substantially to the quality of this thesis. Especially, I want to thank you, Maria Kateri, for providing me with the opportunity to work and research in the Institute of Statistics and for making it possible for me to write this thesis in the first place. I also appreciated the chance to visit you, Irini Moustaki, at the London School of Economics and Political Science and research in London twice during this project.

- Ich möchte insbesondere meiner Frau Isabel für all ihre Unterstützung über die gesamte Zeit dieses Projektes danken. Ohne ihr unermüdliches Verständnis in allen Lebenslagen und ihr Engagement dabei, mich zu motivieren, wäre diese Dissertation niemals fertig gestellt worden. Zusätzlich möchte ich ihr für viele anregende Diskussionen über empirische Studien danken, die einen wertvollen Beitrag zu meinem Verständnis der Perspektive von Wissenschaftlern empirischer Fächer und der zugehörigen Erweiterung meines Horizontes beigetragen haben. Von nicht zu unterschätzender Bedeutung ist ihr Einsatz gewesen, diese Arbeit gründlich gegenzulesen und mit wertvollen Anmerkungen zu versehen, die es mir erlaubt haben die sprachliche Qualität der Arbeit deutlich zu erhöhen. Du bist der wunderbarste Mensch in meinem Leben!

- Mein herzlicher Dank gilt der Heinrich-Böll-Stiftung. Ihre finanzielle Förderung hat das Projekt über weite Strecke essentiell getragen und somit meine Forschung in diesem Zeitraum überhaupt ermöglicht. Darüber hinaus hat die ideelle Förderung sehr dazu beigetragen, meinen Horizont zu erweitern und ermutigt mich, stets als Multiplikatorin meinen Beitrag für freiheitliche demokratische Werte, Gerechtigkeit und Nachhaltigkeit zu leisten. Über das Stipendium bin ich Teil einer wundervollen Gemeinschaft geworden und auch über das Stipendium hinaus geblieben. Insbesondere möchte ich Sevilay Karaduman für ihre wertvollen Ratschläge und Rückmeldungen während ihrer Zeit als Promotionsreferentin des Studienwerks und meiner Zeit als Stipendiatin danken.

- Allen meinen ehemaligen Kollegen am Institut für Statistik und Wirtschaftsmathematik gebührt mein Dank für die vielen schönen gemeinsamen Momente, die kollegiale Arbeit und die spannenden Gespräche über Statistik aber auch die gegenseitige Unterstützung in anstrengenden Phasen. Dabei möchte ich Christian Kohl, Anastasia Gaponik, Benjamin Laumen, Max Döring, Fabian Mies, Simon Pitzen, Christian Schmitz, Michael Espendiller und Christina Empacher besonders danken. Zusätzlich bin ich Anastasia Gaponik, Christian Kohl und Michael Espendiller für das Gegenlesen von Teilen der Arbeit und ihre Rückmeldung wertvoller Kommentare in besonderem Maße dankbar. Weiterhin möchte ich insbesondere Benjamin Laumen aber

auch Anastasia Gaponik für ihre Tipps zu LaTeX-Paketen und -Befehlen danken, die mir bei der Erstellung eines Layouts geholfen haben, mit dem ich zufrieden bin.

- Ich möchte allen meinen Freunden und meiner Familie danken, die mich emotional unterstützt und mir verziehen haben, wenn ich in Gesprächen auch das eine oder andere mal zu häufig auf mein Projekt abgedriftet bin. Dabei möchte ich meinen Eltern besonders danken, die mich mit ihrem unendlichen Vertrauen in mich stets auf meinem Weg unterstützt haben. Zusätzlich möchte ich mich bei Patrick Eggermann für Tipps zur C++ Programmierung bedanken.

- Ebenfalls möchte ich Gerhard Dikta danken. Er hat mit seinen hervorragenden Vorlesungen besonders zu meinem Wissen über Stochastik und meiner Begeisterung für Statistik beigetragen und somit eine wichtige Grundlage für meine Forschung gelegt.

- Ich danke weiterhin dem Aachener Forum für Doktorandinnen. Neben all den wunderbaren Frauen, die ich dort kennen gelernt habe, und den vielen schönen Momenten und unterstützenden Gesprächen, möchte ich insbesondere Sandra Kiefer und Michaela Hiller danken. Es war mir eine Ehre und Freude, den Brigitte-Gilles-Preis für das Doktorandinnenforum mit euch zusammen entgegen zu nehmen.

- Ebenfalls möchte ich allen an den beiden Schreibwerkstattswochen Beteiligten danken: Kirsten Kemna, Friedrich Gabel, Antonia Martin-Sanabria, Gloria Gessinger, Christian Kohl und Cora Carmesin. Neben den allgemein schönen Momenten hatten wir eine super Atmosphäre und haben uns stets gegenseitig motiviert, fleißig an den Dissertationen zu arbeiten und somit große Fortschritte in der Zeit zu machen.

- Abschließend möchte ich meinen Kollegen am Werkzeugmaschinenlabor für die schöne gemeinsame Zeit während der letzten Monate der Fertigstellung dieser Dissertation, ihr Verständnis und ihre moralische Unterstützung danken.

Contents

Introduction **1**

1 Models of Item Response Theory for Binary Data **5**
1.1 Basic Latent Trait Models . . . 6
1.1.1 Distinction of Measurement and Structural Models . . . 6
1.1.2 Measurement Models . . . 8
1.1.3 Structural Models: The Distribution of the Latent Vector . . . 17
1.2 Models for the Marginal Probabilities . . . 18
1.2.1 Dutch Identity Approach . . . 18
1.2.2 Graphical Model Approach . . . 21
1.2.3 Conditionally Defined Model Approach . . . 24
1.2.4 The Distribution of the Latent Vector . . . 27
1.2.5 Interpretation as an Association Model . . . 28
1.2.6 Connection to the Ising Model . . . 29
1.3 Parameter Estimation for Item Response Models . . . 30
1.3.1 Fixed Effect Approach . . . 31
1.3.2 Random Effect Approach . . . 32
1.3.3 Estimation for the Rasch Model . . . 36
1.3.4 Analysis of the Structural Model and Estimation of the Latent Variables 50

2 Extended Generalized Linear Latent Variable and Structural Models **57**
2.1 Model Definition . . . 59
2.2 Example: Modeling of Guessing Behavior . . . 62
2.2.1 Heavy Tailed Link Functions for Ability Based Guessing . . . 64
2.2.2 Application to a Simulated and a Real Data Set . . . 65
2.3 Marginal Maximum Likelihood Estimation . . . 70
2.3.1 Preliminaries and Fisher's Identity . . . 70
2.3.2 Asymptotic Theory . . . 74
2.4 Asymptotic Tests, Confidence Regions and Model Selection . . . 77
2.5 Simulation Studies . . . 79
2.5.1 Precision of Skewness Parameter Estimates . . . 80
2.5.2 Reliability of ICC and Latent Trait Estimates under Skewness Uncertainty 86
2.5.3 Model Selection . . . 94
2.6 Summary, Discussion and Outlook . . . 99

3 The Asymptotic Posterior Distribution of Latent Vectors for Binary Items **103**
3.1 Preliminaries . . . 105
3.2 Review of APN for univariate latent traits . . . 107
3.3 A framework for asymptotic posterior distributions of multivariate latent traits . 109

3.4 Semiproper Centering Theory . . . 112
3.4.1 Regularity Conditions . . . 112
3.4.2 Main Results . . . 114
3.4.3 Summary, Discussion and Outlook . . . 116
3.4.4 Example: Regularity Conditions for Model (3.21) . . . 116
3.4.5 Proofs . . . 118
3.5 Proper Centering Theory . . . 138
3.5.1 Regularity Conditions . . . 138
3.5.2 Main Results . . . 140
3.5.3 Summary, Discussion and Outlook . . . 142
3.5.4 Example: Regularity Conditions for Model (3.21) . . . 143
3.5.5 Proofs . . . 144

4 Estimation for Multidimensional Logit IRT Models for Many Binary Items 171
4.1 Review of the Component Statistics and Linear EAPs . . . 173
4.2 Approximate Linear Expected A-Posteriori Estimators . . . 175
4.2.1 Derivation of the Estimator . . . 175
4.2.2 Simulation Study . . . 178
4.3 A new EM-Algorithm for MMLE in High Dimensional Settings . . . 182
4.3.1 Preliminaries . . . 182
4.3.2 A Posterior Normal Approximated MCEM-Algorithm . . . 185
4.3.3 Application to two Simulated Example Data Sets . . . 188
4.3.4 Comparison of the aMCEM-Algorithm to other approaches for EM-Algorithms for MMLE in the case of large q and large d . . . 191
4.4 Summary and Outlook . . . 195

5 Pseudo Likelihood Estimation for Binary Logit IRT Models 197
5.1 Revisit of the Approach for Generalized Rasch and Confirmatory Models . . . 198
5.2 Pseudo Likelihood Estimation for Exploratory Logit IRT Models . . . 202
5.2.1 Fitting the Manifest Probabilities Model . . . 202
5.2.2 Transformation of the Estimates to the Corresponding Logit IRT Model 208
5.3 Analysis of the Bias and the Efficiency of the Estimates . . . 214
5.4 Comparison to Marginal Maximum Likelihood Estimation . . . 217
5.5 Summary and Outlook . . . 220

Conclusions and Future Research 225

Appendix 227

A Proof of the Asymptotics of the Marginal Maximum Likelihood Estimator 227
A.1 The Constrained Rasch Model . . . 227
A.2 Extended Generalized Linear Latent Variable and Structural Models . . . 232

B Heuristic Iterative Simulation Based Bias Reduction of Pseudo Likelihood Estimates 235

C Aspects of the Implementation in R 243
C.1 Chapter 2: EAP and MMLE for EGLLVM and Generation of Data . . . 243
C.2 Chapter 4: aEAP and MMLE for Logit Models by aMCEM . . . 244

C.3 Chapter 5 and Appendix B: Pseudo Likelihood Estimation 246

List of Figures **251**

List of Tables **252**

Notation and Acronyms **253**

Bibliography **259**

Introduction

To analyze the fairness of an educational system of a country and to help with development of pedagogical concepts, large international surveys such as the *Programme for International Student Assessment (PISA)*[1] are important tools. An essential challenge in conducting such surveys is the measurement of not directly observable traits such as the ability of students in different subjects. These traits are known as *latent traits* or *latent variables*. A psychometric theory for measuring latent traits is the *Item response theory (IRT)*. The students respond to a set of items in a test where the response to each item can be categorized in a set of options (e.g. "correct" and "incorrect"). The items are constructed in such a way that the probability of a correct response depends on the ability in the analyzed subject. The collection of the responses of several students provides a data set for the analysis of their abilities. Some tasks, such as word problems in math exams (where the mathematical problem is given in ordinary language), might require several abilities, e.g. in this case mathematical and linguistic abilities. A vector of several latent traits is called a *latent vector*, a statistical model for categorical response data with latent vectors is called a *multidimensional IRT model* (MIRT model).

In the application of IRT, the estimates of latent vectors can have a considerable influence on outcomes (e.g. decide about diagnoses or admission to a school). Thus, the precise and reliable statistical analysis is essential. If a latent vector with a high dimension is supposed, the analysis can become computationally challenging. One of the goals of research regarding IRT is to lessen the computational effort of model fitting. Another goal is the development of methods for model checking. To obtain reliable estimates of the latent vectors in the application of a fitted IRT model, the theory and methods of their estimation have a clear importance as a third goal. In this thesis, we contribute to the statistical area of latent variable modeling for categorical response data and aim to contribute to the three previously stated goals. We restrict this work to binary data. In the following we briefly summarize the five chapters of this thesis.

Chapter 1: We give an introduction to IRT from a mathematical and statistical point of view here. We briefly discuss the model class of generalized linear latent variable models (GLLVM) under the restriction to binary response variables and with focus on continuously distributed latent vectors. Next, we summarize important approaches for increasing the model flexibility by using parametric link functions in order to improve the fit in the case of a skew dependence of the success probabilities on the latent vectors. Such skewness can be caused by guessing behavior of test participants in multiple choice tests.

A major issue of MIRT models is the need for numerical quadrature to evaluate the marginal probabilty mass function of the response vector – this is especially critical for high latent dimensions (i.e. the number of components of the latent vector) where correspondingly high dimensional integrals have to be solved. These integrals can be solved analytically in special cases. One of them is the GLLVM with logit link function if the pos-

[1] Cf. e.g. OECD (2019).

terior distribution of the latent vector is a normal distribution given the response vector. We review the literature regarding this approach in the second section of Chapter 1, which reveals the need for further research regarding the asymptotic posterior normality (APN) of latent vectors to allow an asymptotic justification of this approach for logit MIRT models.

Finally, we review frequentist fitting approaches for MIRT models and discuss properties of the estimators for the Rasch model in more detail. One of the most important and commonly used estimators is the marginal maximum likelihood estimator which is the maximum likelihood estimator based on the marginal distribution of the response vector. For this approach only a limited amount of theoretical results was published. The asymptotic normality of the estimators is a very important property in practical data analysis but has not yet been proved for the common IRT models. Moreover, computationally efficient algorithms for high dimensional settings are derived from claimed APN of the latent vectors, underlining the importance on results regarding APN of latent vectors.

We extend our presentation of the results regarding the Rasch model to the asymptotic normality of the marginal maximum likelihood estimator (MMLE) under mild conditions on the distribution of the latent variable. We also conducted a new simulation study for a comparison of the different estimation approaches for the Rasch model and present the results. In the analyzed setting, the estimators are similarly precise which implies that there is no clear best method and computational aspects become more important. Because many estimation approaches for (M)IRT models focus on the estimation of the model parameters, the estimation of the latent scores is commonly done in a second step. A review of the established approaches for this second step shows that the asymptotic properties for many important estimators are either have only been proved under restrictive conditions or not at all to date.

Chapter 2: In this chapter, we introduce an extension of the model class GLLVM for binary items with link functions and structural models that depend on parameters for their shape (EGLLVM). We point out how this is connected to several independent model approaches in the literature which are unified in this class. Next, we formulate a new model for guessing behavior using this class and discuss its properties and interpretation in relation to the established approach (the 3PL model). The application to two data sets illustrates the difficulty of estimating link function parameters, which is similar to classic guessing models. Especially, this underlines the necessity of a proof of the consistency and asymptotic normality and efficiency of the MMLE. In this chapter, we prove these for the new model class – which includes many classic models, thus contributing to the MMLE theory in general.

Using the asymptotic efficiency of the MMLE for EGLLVM we also proved the asymptotic chi-square distribution of Wald, score and likelihood-ratio(LR) test-statistics. In particular using the LR test for nested models, we derive a new tool to test the link functions of GLLVM against broad parametric alternatives. This is a conditional goodness of link test for GLLVM, conditional on the validity of the alternative model. We also discuss the selection of link functions for (E)GLLVM using classic information criteria.

In three simulation studies we provide evidence that parameters of link functions require a relatively large sample size for precise estimates. The skewness parameter for structural models on the other hand is already acceptably precise for a comparably small number of observed responses. For sample sizes of several thousand responses, the estimation of the skewness parameters for single items is sufficiently precise (and due to the consistency, it

improves further for increasing sample sizes, e.g. the sample size of several hundred thousand in the PISA studies). Another result of the simulation studies is that models with a common skewness parameter for all items and a skewness parameter for the structural model provide a good trade-off between parsimony and flexibility and yield good estimates of the ICCs and the latent traits in various settings of skewness of link functions and/or structural models. The third simulation study also reveals that the Akaike information criterion (AIC) as well as the LR tests can be used reliably to select the correct model. Especially, AIC and LR tests can be used to analyze if there is a lack of fit caused by an inappropriate link function. This implies that misspecifications in the link functions in surveys with a few thousand responses are not critical for the conclusions when they are not detected by a LR test or if a wrong model is selected using the AIC.

Chapter 3: We discuss the asymptotic theory of the estimation of the latent vectors here, which can be interpreted as a (empirical) Bayesian point estimation with previous estimation of the (M)IRT model parameters. As a primary target of this chapter, we investigate variants of a *Bernstein–von Mises* theorem of latent vectors, i.e. the asymptotic posterior normality of latent vectors. This chapter is a central part of the thesis and provides a comprehensive analysis of questions related to Bernstein–von Mises theorems and the asymptotics of latent vector estimation for binary IRT.

Especially, we generalize the current results for the asymptotic posterior normality of a single latent variable and the consistency of its maximum likelihood estimator (MLE) to latent vectors under similar conditions. Additionally, we expand the theory with the general consistency of the posterior distribution, the expected a-posteriori estimator (EAP) and maximum a-posteriori estimators (MAP) of the latent vector. By restricting the conditions, we prove strengthened results of the asymptotic posterior normality in the form of uniform convergence of the posterior distribution to a normal distribution instead of a point-wise convergence (i.e. a convergence for each single set). This second theory of asymptotic posterior normality also implies results of the asymptotic normality of several point estimators of the value of the latent vector and the locally uniform convergence of the moment generating function of the posterior to that of a multivariate normal distribution in probability – the latter is a generally strong theoretical result which is a further step in clarifying the Dutch Identity Conjecture (cf. Holland, 1990, Section 3.2 and 3.3). By also deriving a new asymptotic result on the conditional distribution of a response variable given all other response variables from this theory, we provide an asymptotic justification of a heuristic approach to derive an integral free representation of logit models, which is also connected to the Dutch Identity Conjecture. This underlines how this theory can be used in future theoretical research in IRT.

Overall, in this chapter we provide answers to open questions of the asymptotic posterior normality of latent vectors and general results on point estimators of the underlying value of the latent vector.

Chapter 4: In this chapter, we use the APN theory and the component statistics (the sufficient statistics for latent vectors in logit models) to develop a linear approximation of the EAP (aEAP) for a set of latent vectors behind a set of binary response vectors. Of special interest is that this approximation can be evaluated without computing multidimensional integrals. Using the APN theories further, we derive a new EM-algorithm to compute the MMLE for logit models with a large number of items that uses one sort of adaptive

Monte Carlo integration for the multidimensional integrals. This new algorithm has the specific benefit that it avoids iterative procedures during the E-step by applying the aEAP in contrast to classic adaptive procedures for high dimensional IRT models that require iterative procedures during the E- and the M-steps. The computationally intensive parts of the M-step can moreover be implemented by using logistic regression software so that the overall implementation is simplified. Moreover, using MC-integration allows a more flexible specification of the number of quadrature points used in comparison to the established deterministic adaptive quadrature EM-algorithms for high dimensional settings. Overall, we develop a procedure for computing the MMLE for logit MIRT models that is easy to implement for an arbitrary latent dimension and computationally efficient for high latent dimensions and many items.

Chapter 5: We can observe that even computationally efficient algorithms for MMLE potentially require a large amount of time if the data set is very large while a high latent dimension is supposed. An alternative approach is using integral free representations of logit models under the assumption of posterior normality (or its asymptotic justification) of latent vectors, which we call a manifest probabilities model. However, the computational cost of the apporach grow exponentially with the number of items, making it unattractive for large numbers of items. This illustrates why we need methods with substantially shorter computation times with an acceptable loss of efficiency – especially in cases with a very large number of items and high latent dimensions. The pseudo likelihood (PL) approach provides a foundation to solve this problem due to its vastly faster computation. We develop a PL method from the literature further so that a fitted model can always be interpreted as an IRT model. Then we derive two algorithms for our new method. Our implementations solve computational issues which were documented for the method that our method is based upon. Then we compare the computational time of our implementation of the PL method to an implementation of MMLE in several simulation studies. The results show an advantage of the PL method over the MMLE that becomes greater for increasing settings in each of the three dimensions (sample size, number of items, latent dimension). To obtain estimates for IRT models, a transformation of the parameter estimates is required. Our simulation studies reveal a notable bias of the PL estimates for the IRT models which decreases for an increasing number of items. This bias can be explained by the transformation used. Because the PL estimator (PLE) is a classic composite likelihood estimator for the manifest probabilities model with actual posterior normality of the latent vector, we expect that improved transformations of the estimates can lead to a fast and still acceptably precise estimation procedure in future research.

Chapter 1

Models of Item Response Theory for Binary Data

Questionnaires and tests are popular tools in data collection in social sciences, psychometrics, educational research and medicine. Multivariate data are often used in social sciences for measuring unobserved constructs such as ability, attitudes, state of disease, fear, etc. Latent variable models and structural equation models are used for measuring the associations among a set of manifest/observed variables (questionnaire items, tests) using a set of latent variables/factors (i.e. unobservable random variables). Those relationships define the measurement part of the model. However, often the interest is on addressing specific hypothesis about the latent variables and covariates, which defines the structural part of the model. Item response theory (IRT), e.g. latent trait and latent class models, focus on the measurement model in which item and subject properties are modeled together. The items scale can be dichotomous (right/wrong) as well as polytomous (nominal or mostly ordinal) and are thus treated by methods for categorical variables.

Modelling and measuring unobservable constructs such as the subjects' ability is a quite abstract approach. Classically the ability was modeled as a real valued parameter (cf. e.g. Rasch, 1960). Currently it is more common to work with an empirical Bayesian motivation and treat each subject's ability as the realization of an underlying unobserved random vector. Many models of this kind are covered by the generalized linear latent variable model (GLLVM) framework, which provides an unified methodology for the analysis and definition of new parametric item response models (cf. Moustaki and Knott, 2000; Bartholomew et al., 2011).

The maximum likelihood estimation (MLE) of item parameters of the Rasch model can be calculated with direct maximisation, when conditioning on sufficient statistics of the subject parameters (Andersen, 1973). But for more general models this is not possible. A common approach is to integrate out latent variables for maximization (cf. Bartholomew et al., 2011, Section 2.9). Although it is in theory unproblematic, in practice the associated estimation problem is computational intensive and estimation is no longer feasible for larger number of latent variables.

After integrating out the latent variables, the marginal distribution of the responses is derived. The probabilities of single response patterns from the marginal distribution are also known as manifest probabilities. A possible way to overcome the estimation problem is to reformulate the models of the manifest probabilities in such a manner targeting to avoid integrals in maximum likelihood estimation. A first step towards this direction was done by Tjur (1982), who derived a log-linear model representation for the dichotomous-data Rasch model. As a next step Cressie and Holland (1983) gave an alternative representation of the manifest probabilites from specific response patterns for dichotomous data. Thereby they obtained a log-linear model for the manifest probabilities of the Rasch model

similar to Tjur's, but they derived additionaly technical conditions for meaningful models. Later, Holland (1990) derived *The Dutch Identity* and generalized the result of Cressie and Holland (1983) under suitable conditions. Independend of Cressie and Holland (1983) and Holland (1990), Agresti (1993) generalized Tjur's work to a Rasch model for ordinal response variables.

In newer research, Anderson and Vermunt (2000) derived representations of latent trait models as association models (special ordinal log-linear models and generalizations), adopting a graphical model approach. For the dichotomous case they are identical to the model derived from *The Dutch Identity* (Anderson and Yu, 2007). Recently, the model class became a further research focus in the framework of network psychometrics and its relation to the Ising model of theoretical physics (cf. for example Marsman et al., 2018). Note that special estimation procedures have been developed for this class of models; for example within the R-package **plRasch** by Anderson et al. (2007). It is interesting to study how far this procedure can be extended in targeting general item response models.

In this chapter, we discuss item response models for dichotomously scored items. In the first section we discuss the basic concepts of models for continuous latent variables. In the second section we discuss model approaches, which lead to closed form expressions of the marginal probabilities. In this sense, the discussed models can be considered as models for the marginal distributions itself. Finally, we discuss some common, frequentist and likelihood based estimation approaches for parametric latent variable models and general common approaches for the factor scoring itself. We compare the different model fitting approaches in a new simulation study.

1.1 Basic Latent Trait Models

In Subsection 1.1.1, we introduce the main concepts and the distinction between the model components *measurement model* and *structural model*.[1] Additionally, we discuss the important relation to the empirical Bayesian estimation if we consider the latent variables as Bayesian parameters. After this general considerations, we will introduce in Subsection 1.1.2 the most important parametric model classes for the measurement model we need in this work and discuss some of the most important properties briefly. Last, we define the most important structural models in Subsection 1.1.3.

1.1.1 Distinction of Measurement and Structural Models

Models from item response theory can be applied on data sets which arised from the response of $n \in \mathbb{N}$ subjects to $d \in \mathbb{N}$ binary coded items in order to analyze the underlying values of $q < d$, $q \in \mathbb{N}$, latent constructs, represented by a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\intercal$ of latent variables, where $\intercal$ denotes *transpose*. In particular, the n subjects are considered to be drawn independently from a single large population. Hence, in the purely statistical setting we consider n independent and identically distributed responses, where each response is a d-variate response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\intercal$ of binary entries. In our setting, $\boldsymbol{\eta}$ and $\boldsymbol{Y}$ are defined over the same probability space $(\Omega, \mathcal{A}, \mathsf{P})$. The distribution of the response vector is given by integrating out the latent vector from the joint distribution of $(\boldsymbol{Y}, \boldsymbol{\eta})$, hence by

[1] Cf. e.g. Rose et al. (2017) for the use of these names.

the marginal distribution of $\boldsymbol{Y}$ with

$$P(\boldsymbol{y}) = \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}) = \int_{\Theta} \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\eta})\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}), \qquad \boldsymbol{y} \in \mathbb{Y} := \{0,1\}^d, \tag{1.1}$$

where $\mathcal{H}$ denotes the absolutely continuous distribution of $\boldsymbol{\eta}$ and $\Theta := \operatorname{supp}(\mathcal{H}) \subseteq \mathbb{R}^q$ the support of $\mathcal{H}$. We denote the probability density function (pdf) of $\mathcal{H}$ by $\mathfrak{h}$. In particular, (1.1) can be recognized as a continuous mixture of discrete distributions.

Remark 1.1.1
For simplicity of notation, we use $\boldsymbol{\eta}$ to denote the random latent vector as well as a realization of it. If not clear from the context, we write explicitly $\boldsymbol{\eta} \in \Theta$ for a realization or $\boldsymbol{\eta} \sim \mathcal{H}$ for the random vector with values in Θ. ◇

On the other hand, our goal is to interpret the model. Beside the context of application, where we interpret a particular value of $\boldsymbol{\eta}$ as ability, aptitude, etc., we suppose in the model formulation that $Y_1, \ldots, Y_d$ are conditional independent given $\boldsymbol{\eta}$, i.e.

$$\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\eta}) = \prod_{i=1}^{d} \mathsf{P}(Y_i = y_i \mid \boldsymbol{\eta}), \qquad \boldsymbol{y} \in \mathbb{Y}. \tag{1.2}$$

So, we try to set the dimension q of $\boldsymbol{\eta}$ as the smallest number, such that the conditional independence assumption is valid. When the d binary responses correspond to the correctness of the corresponding items of a test, then the number q is also known as the *test dimension*. Further, we interpret $\boldsymbol{\eta}$ as the information for which the conditional independence of $Y_1, \ldots, Y_d$ holds. In this sense, $\boldsymbol{\eta}$ explains $\boldsymbol{Y}$ beyond the probability of responding one. This coincides with the statistical modeling goal of a parsimonious model, where we select q as hyperparameter as the smallest value, such that the model has still a valid fit. However, in many applications, there is a hypothesis about q, which can thus be tested.

So, we observe an independent and identically distributed (iid) sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ with $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{dk})^\intercal$, $k = 1, \ldots, n$ and distribution given in equation (1.1). But contextually, we suppose a two step origin: First, values $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ were drawn from $\mathcal{H}$. Then the response variables $Y_{1k}, \ldots, Y_{dk}$ are drawn from the conditional distribution given that value of $\boldsymbol{\eta}_k$, i.e. are conditional independent with

$$(Y_{ik} \mid \boldsymbol{\eta}_k) \sim \mathscr{B}\big(1, P_i(\boldsymbol{\eta}_k)\big), \qquad i = 1, \ldots, d,\ k = 1, \ldots, n,$$

where $\mathscr{B}(1, P_i(\boldsymbol{\eta}_k))$ denotes the Bernoulli distribution with success probability $P_i(\boldsymbol{\eta}_k) := \mathsf{P}(Y_i = 1 \mid \boldsymbol{\eta}_k)$. The function $P_i \colon \Theta \to (0,1)$ is known as the *item response function* of item i. In the case of $q = 1$, P_i is known as *item characteristic curve (ICC)*.

Thus, $\boldsymbol{Y}$ acts as manifest indicator of $\boldsymbol{\eta}$ and the relation of $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ through (1.2) allows an estimation of the particular realized value of $\boldsymbol{\eta}$. This estimation is named *measurement* due to its application. To shorten the notation, we write $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ if (1.2) holds for $\boldsymbol{\eta}$ and $(Y_i \mid \boldsymbol{\eta}) \sim \mathscr{B}\big(1, P_i(\boldsymbol{\eta})\big)$, $i = 1, \ldots, d$ for given item response functions $P_1, \ldots, P_d$. Hence, the model $\mathcal{P}(\boldsymbol{\eta})$ is known as *measurement model*, which is distinctly described from the distribution $\mathcal{H}$ of $\boldsymbol{\eta}$.

On the other side, the distribution $\mathcal{H}$ contains the information about the population laying behind the test/survey participants. Through equation (1.1), $\mathcal{H}$ has influence on the distribution of $\boldsymbol{Y}$, but primarily it describes the structure behind the measurement object $\boldsymbol{\eta}$ and is thus denoted as *structural model*.

Remark 1.1.2
From a frequentist perspective, we simply consider a mixed model with mixture given by $\mathcal{H}$ such that the analyzed distribution is given by (1.1). But since our final target is $\boldsymbol{\eta}$, there is a strong connection to an empirical Bayes setting where model parameters are considered as hyperparameters, whose (frequentist) estimation leads to a final specification of the distribution of the model parameters, which are random variables in the Bayesian setting. Actually, as discussed by Morris (1983), from the empirical Bayes perspective there is only a small – if even any – difference between frequentist mixture and empirical Bayes models. As later discussed by Chang and Stout (1993), the estimation of $\boldsymbol{\eta}$ in IRT fits into the empirical Bayes setting. ◇

Due to Remark 1.1.2, we call $\mathcal{H}$ the *prior* distribution of $\boldsymbol{\eta}$, and the conditional distribution $\mathcal{H}(\bullet \mid \boldsymbol{Y} = \boldsymbol{y})$ the *posterior* distribution after observing the response pattern $\boldsymbol{Y} = \boldsymbol{y} \in \mathbb{Y}$.

An important area of application and research is related to making inference about the distribution from which the n survey participants are drawn. A fully specified model with estimated parameter values can further be used to judge (for example in the sense of diagnosis or assessment) single further subjects of the same population. In particular, if $P_1, \ldots, P_d$ are fitted, then the conditional independence allows a proper selection of a subset of the d items corresponding to $P_1, \ldots, P_d$ to estimate $\boldsymbol{\eta}$ without need for updating the model and its parameters. Especially in the situation of a computerized test, after responding to a basis set of items, the following items are selected adaptive to measure/estimate $\boldsymbol{\eta}$ with a given precision using a set of items as minimal as possible. These tests are called *computerized adaptive tests* (CAT). Due to this disjoint fitting of the measurement model and its following (partial) use to test single subjects, the model fitting is often called *calibration*. This coincides again with the empirical Bayes concept mentioned in Remark 1.1.2. Throughout this chapter, we consider the case of a fixed set of d items. But it is sometimes useful to have the testing situation of CAT in mind.

The concept of IRT models is strongly related to *structural equation models*, which aims for a detailed modeling of the relation between manifest and latent variables to allow strong interpretation. In special cases the modeling approaches are identical.

1.1.2 Measurement Models

In this section, we discuss properties of the response functions $P_1, \ldots, P_d$ for a test of $d \in \mathbb{N}$ items $Y_1, \ldots, Y_d$, which are typically assumed. Generally, it is meaningful to assume all P_is to be continuous on Θ or, moreover, continuously differentiable. In many applications it is meaningful to define an ordering in the probability of $Y_1 = 1$ in relation to the values of the latent vector $\boldsymbol{\eta}$. In particular, it is commonly assumed that P_i is strictly monotonic in η_j if η_v, $v = 1, \ldots, q$, $v \neq j$, are held fixed at any valid point (cf. e.g. Luzardo and Rodríguez, 2015).

1.1.2.1 Generalized Linear Latent Variable Models for Binary Data

Even if the historical development did not take this course, considering latent variables as unobserved explanatory variables opens the door to an adaption of the important model class of generalized linear models. We formulate the model class in the following definition.

Definition 1.1.3 (Generalized Linear Latent Variable Model for Binary Data)

(a) Suppose that $\boldsymbol{Y}$ is the vector of $d \in \mathbb{N}$ binary response variables with response functions $P_1, \dots, P_d$ and $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ for $\boldsymbol{\eta} \in \Theta \subset \mathbb{R}^q$ for $q \in \mathbb{N}$. If there are d strictly increasing continuous and bijective mappings $\mathfrak{g}_1, \dots, \mathfrak{g}_d \colon (0,1) \to \mathbb{R}$, numbers $\alpha_{10}, \dots, \alpha_{d0} \in \mathbb{R}$ and vectors $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_d \in \mathbb{R}^q$, such that

$$\mathfrak{g}_i(\mathsf{P}(Y_i = 1 \mid \boldsymbol{\eta})) = \mathfrak{g}_i(P_i(\boldsymbol{\eta})) = \alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}, \qquad i = 1, \dots, d, \tag{1.3}$$

holds, then we say a *generalized linear latent variable model* (GLLVM) holds for $\boldsymbol{Y}$. The term $\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}$ is called *predictor* for item i given $\boldsymbol{\eta}$ and $\mathfrak{g}_i$ is called the *link function* of item i.

(b) If $\mathfrak{g}_i = \text{logit}$ for $i = 1, \dots, d$, where

$$\text{logit}(x) := \log\left(\frac{x}{1-x}\right), \qquad x \in (0,1),$$

then we say a *logit model* holds for $Y_1, \dots, Y_d$ (cf. Bartholomew et al., 2011, Section 4.2).

(c) If $\mathfrak{g}_i = \Phi^{-1}$ for $i = 1, \dots, d$, where Φ^{-1} is the standard normal quantile function, i.e. the inverse of the standard normal cumulative distribution function (cdf) Φ, then we say a *probit model* holds for $Y_1, \dots, Y_d$ (cf. Bartholomew et al., 2011, Section 4.3). ◇

Remark 1.1.4

In contrast to classical definitions of a GLLVM for a binary response, such as in Moustaki and Knott (2000) or Bartholomew et al. (2011), we explicitly allow the use of different link functions $\mathfrak{g}_i \neq \mathfrak{g}_k$ for any two different items $i \neq k$. This allows for some further modeling flexibility while conserving the generalized linear structure. ◇

In principle, this model class can be viewed as some kind of generalized linear models with unobservable continuous explanatory variables $\boldsymbol{\eta}$. However, in categorical data analysis, there is a further motivation for this type of models, which we briefly illustrate in the following remark (cf. for example Bartholomew et al. 2011, Section 4.3 and 4.4).

Remark 1.1.5 (Underlying Variables)

Suppose $\boldsymbol{\eta} \in \Theta$ is given. Consider $\xi_i = \alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta} + \varepsilon_i$, where $\alpha_{i0} \in \mathbb{R}$, $\boldsymbol{\alpha}_i \in \mathbb{R}^q$ for $i = 1, \dots, d$, and $\varepsilon_1, \dots, \varepsilon_d$ are independent random variables, where ε_i has the cdf F_i. Then the indicator variables $Y_i = \mathbb{1}_{(-\infty,0]}(\xi_i)$, $i = 1, \dots, d$, satisfies a GLLVM of the form in Definition 1.1.3 with the link functions $\mathfrak{g}_i = F_i^{-1}$, $i = 1, \dots, d$. ◇

Especially through the underlying variable approach (cf. Remark 1.1.5), we recognize a close relation between GLLVM for binary data and the models of *factor analysis*. Due to this, GLLVM for binary data can be considered as a model of *binary factor analysis*. Therefore, the latent vector $\boldsymbol{\eta}$ is also known as a vector of (common) *factors*. Correspondingly, we also call the coefficients $\alpha_{1j}, \dots, \alpha_{dj}$ of η_j (cf. Definition 1.1.3) *factor loadings* or *loadings of factor* j, $j = 1, \dots, q$. Related to linear models, $\alpha_{10}, \dots, \alpha_{d0}$, are the *intercepts*.

The underlying variable approach leads to a direct interpretation: We can consider response variables $\boldsymbol{Y}$ from a GLLVM as indicator variables for a continuous response, where we observe a response of 1 if some threshold is (not) passed. In the notion of psychological

measurement, the variables $\varepsilon_1, \ldots, \varepsilon_d$ are the measurement errors and the link functions $\mathfrak{g}_1, \ldots, \mathfrak{g}_d$ are the continuous quantile functions of the distributions of $\varepsilon_1, \ldots, \varepsilon_d$. Thus, it is intuitive that in many applications $\mathfrak{g}_1 = \mathfrak{g}_i$ for all $i = 1, \ldots, d$. In particular, the often well fitting linear regression model with normally distributed error motivates the popular probit model.

Particularly, the measurement of a single latent trait (such as an ability) has an important status in IRT and its history. Also the special case of a single latent trait allows another interpretation of the model parameters. We now discuss this briefly and introduce some special cases of GLLVM for one dimensional models which we need in this work. For this case, $q = 1$ and thus vectors η and $\boldsymbol{\alpha}_i$ become scalars denoted by η and α_i. Using further the reparametrization $\beta_i = -\alpha_{i0}/\alpha_i$, $i = 1, \ldots, d$, the probit model in this case becomes

$$\Phi^{-1}(P_i(\eta)) = \alpha_i(\eta - \beta_i), \qquad i = 1, \ldots, d, \tag{1.4}$$

known also as the *Lawley-Lord model* (Andersen, 1977).[2] The similar model

$$\operatorname{logit}(P_i(\eta)) = \alpha_i(\eta - \beta_i), \qquad i = 1, \ldots, d, \tag{1.5}$$

is known as the *Birnbaum–Rasch model* or *two parameter logistic model (2PL model)* (cf. Birnbaum, 1968; Andersen, 1977). If we consider (1.5) with the restriction $\alpha := \alpha_1 = \ldots = \alpha_d$, then the model is the *Rasch model* or *one parameter logistic model (1PL model)* (cf. Rasch, 1960; Birnbaum, 1968). In the case of

$$\mathfrak{g}_i(P_i(\eta)) = \alpha(\eta - \beta_i), \qquad i = 1, \ldots, d, \tag{1.6}$$

for link functions $\mathfrak{g}_1, \ldots, \mathfrak{g}_d$, we speak of a *Rasch-type model*. In all one dimensional models, we suppose $\alpha_i > 0$ or $\alpha > 0$ to ensure that P_i is strictly increasing.

The formulation of the one dimensional models with predictor $\alpha_i(\eta - \beta_i)$ helps us further to interpret the model parameters. If the link function corresponds to a distribution with median zero (e.g. logit and probit), then

$$P_i(\beta_i) = \frac{1}{2}, \qquad i = 1, \ldots, d.$$

Hence, β_i is the threshold value, η has to exceed to get

$$\mathsf{P}(Y_i = 1 \mid \eta) \geq \frac{1}{2}. \tag{1.7}$$

Due to its applications in educational testing, where $Y_i = 1$ symbolizes a correct response to item i, β_i is known as the *difficulty parameter* of item i. However, as its role in the statistical model is the same even if (1.7) does not hold, the intercept parameter α_{i0} in a general GLLVM for binary data can be thought of as a *difficulty* indicator for the item i, $i = 1, \ldots, d$. Furthermore, α_i can be interpreted as the *discrimination parameter* of item i. Thereby a high discrimination value allows to conclude relatively sure from $y_i = 1$ that the latent variable takes a higher value than β_i. Also, for higher α_i it is easier to discriminate between subjects with different values of η since for higher α_i it is more likely that subjects with $y_i = 0$ have $\eta < \beta_i$ and subjects with $y_i = 1$ have $\eta > \beta_i$ and therefore it is likely that these subjects have some distance on the latent scale. Generally, the bigger the vectors $\boldsymbol{\alpha}_i$ in a GLLVM are, the bigger the probability that two subjects with different values of $\boldsymbol{\eta}$ respond differently is and thus, we can discriminate them.

[2] Note also the contributions by Lawley (1943, 1944), Lord (1952) and Lord and Novick (1968) for establishing that model.

1.1.2.2 Other Link Functions and Models with Flexible Shape

The most important models in practice are logit and probit models and, in particular, the Lawley–Lord and the Birnbaum–Rasch model. As probit and logit correspond to a symmetric distribution, we always get this symmetry in the model. So denote by F a logistic or normal cdf. Then it holds for any $x \in \mathbb{R}$

$$F(x) = 1 - F(-x).$$

Note that even if the latent dimension and the generalized linear structure are appropriate, this property may be wrong and unsuitable for a specific real situation. So, considering alternatives is important from a statistical point of view to get the possibility of analyzing the link function. In particular, considering skewness might be a solution in data analysis, where the classical symmetric IRT models do not have a good fit. In the generalized linear model (GLM) framework, the two common choices (for right and left skewness) in binary data modeling are the log-log link

$$\operatorname{loglog}(x) := -\log(-\log(x)), \qquad x \in (0,1), \tag{1.8}$$

and the complementary-log-log link

$$\operatorname{cloglog}(x) := \log(-\log(1-x)), \qquad x \in (0,1), \tag{1.9}$$

which corresponds to the *Type I extreme value distribution* (TIEV distribution) (cf. e.g. Agresti, 2015, Subsection 5.6.3). In particular, denote by G its cdf defined by

$$G\colon \mathbb{R} \to (0,1), \qquad x \mapsto \exp(-\exp(-x)), \tag{1.10}$$

then $\operatorname{loglog}^{-1}(\bullet) = G(\bullet)$ and $\operatorname{cloglog}^{-1}(\bullet) = 1 - G(-\bullet)$. Of course, these link functions can also be used for GLLVM for binary data. As von Davier (2009) discussed, models depend generally on assumptions and should not only be used due to tradition. So, having alternatives in mind is meaningful. The following example illustrates a further conceptual background for possible skewness in an ICC.

Example 1.1.6
Consider the task of running a marathon. A test participant needs to have a certain level of fitness to have a reasonable chance of succeeding. Beginning at that level, we expect the chance of succeeding to grow relatively strongly. However, the level of fitness at which there is nearly no chance of failure is much higher. In Figure 1.1, we give a corresponding example ICC (black). However, the closest ICC of a probit model in terms of the Kolmogorov–Smirnov distance still provides a qualitative departure. In contrast, the closest loglog ICC (blue) is remarkably close. If a sport assessment test contains several athletic tests and additionally a marathon as an explicitly difficult task to select the best participants, then it is reasonable that the marathon task is the only item having this skew characteristic curve. Hence, using a loglog link for this item and probit or logit for the other items might provide a reasonable model for a precise measurement of latent traits. ◇

Hence, skew link functions are a basis for analyzing the link function of an IRT model and provide alternatives in the case of a bad fit of symmetric ones. But if there is little or even no prior knowledge regarding the skewness of particular items, one has to consider all 3^d combinations of cloglog, logit and loglog. Clearly, this is problematic for relatively

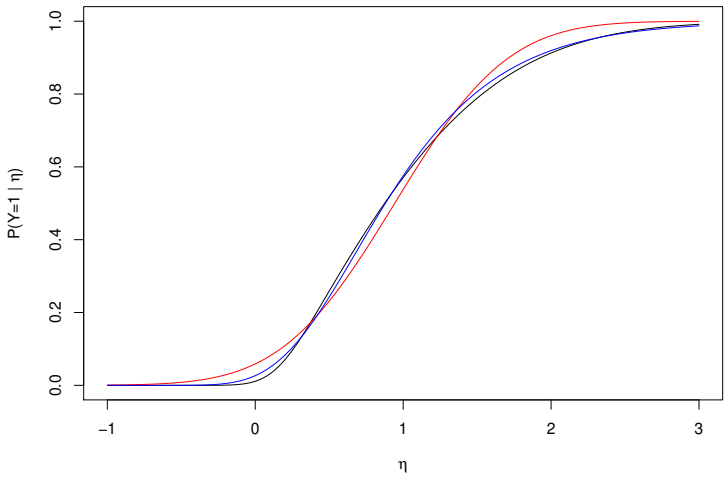

Figure 1.1: *A skew ICC (black) and the probit (red) and loglog (blue) ICCs with the smallest Kolmogorov–Smirnov distance.*

large d (e.g. for 10 items, we already get 59049 combinations). Moreover, the actual skewness might stay between logit and (c)loglog, or be even more skew (as in Example 1.1.6). Therefore, for large data sets considering a parametric link function $\mathfrak{g}(\bullet;\boldsymbol{\gamma}_i)$ with parameter vector $\boldsymbol{\gamma}_i$ can be an interesting alternative to $\mathfrak{g}_i$ in a GLLVM for binary data (cf. Definition 1.1.3). Of course, such a model is no longer a GLLVM.

In principal, any continuous parametric distributional family with support $\mathbb{R}$ and at least one parameter for the shape provides a basis for a link function (location and scale parameters can be fixed arbitrarily since they are absorbed into the linear predictor). Suppose that $\{F_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma}\in\tilde{\Theta}}$ is an appropriate parametric family with $\boldsymbol{\gamma} \in \tilde{\Theta} \subseteq \mathbb{R}^z$, $z \in \mathbb{N}$ (in many cases will hold $\tilde{\Theta} = \mathbb{R}$; then we write γ instead of $\boldsymbol{\gamma}$). We denote the corresponding family of quantile functions by $F_{\boldsymbol{\gamma}}^{-1}$. Then there are four primary options to specify $\mathfrak{g}_1, \ldots, \mathfrak{g}_d$ by this family by setting $\mathfrak{g}_i = F_{\gamma_i}^{-1}$ for $i = 1, \ldots, d$ and considering

1. A common γ parameter (i.e. $\gamma = \gamma_1 = \ldots \gamma_d$) with $\gamma \in \tilde{\Theta}$, which can be (i) known or (ii) unknown.

2. Possibly different $\gamma_1, \ldots, \gamma_d \in \tilde{\Theta}$, which can be (i) known or (ii) unknown.

In the process of model selection based on nested models or selecting parsimonious models between a single skewness parameter and a parameter for each item, grouping the items or fixing the parameters of a group can be further useful options. The main area of developments of parametric links is the wish for a flexible alternative between loglog, logit/probit and cloglog with a parameter to select between them. Commonly, a family is specified which contains the logit or probit as special case with a controlled deviation in the direct of left and/or right skewed links (which are often similar to loglog or cloglog). A very easy way to define such a family is given by defining the following family of densities:

$$f_\gamma \colon \mathbb{R} \to (0, \infty), \qquad x \mapsto 2g(x)K(\gamma x), \qquad \gamma \in \mathbb{R}, \tag{1.11}$$

where g is an arbitrary continuous density and K is an arbitrary continuous cdf of possibly different distributions with support $\mathbb{R}$. If g is symmetric around 0 and K has a density

which is symmetric around 0, then the density f_0 is obviously symmetric and a member of the family $\{f_\gamma\}_{\gamma\in\mathbb{R}}$, which itself behaves symmetrically around 0 in the sense

$$f_{-\gamma}(x) = f_\gamma(-x), \qquad \text{for all } x, \gamma \in \mathbb{R}.$$

By standardizing all f_γ, $\gamma \in \mathbb{R}$, a new family can be obtained, where every member has expectation zero and variance one as anchor but is differently skew. This anchor makes the interpretation of the parameters of corresponding IRT models easier: If we use the corresponding quantile functions as link functions, then the scales of the parameters do not change by much for different values of γ. We expect that this also helps the numerical procedures used for fitting because γ has a weaker influence on the values of other parameters. One of the most important examples is the family of *skew normal distributions*, which can be obtained by setting $g = \phi_1$ and $K = \Phi$, where ϕ_q denotes the density of the q-variate standard normal distribution. Its quantile function allows the definition of skew probit links. Bazán et al. (2010) discussed a representation of the family of standard skew normal distributions, which is ideal for our purpose.[3] For its definition, we need the definition of Gaussian measures. Denote by $\mathcal{B}^q$ the Borel-σ-algebra and by $(\mathbb{R}^q, \mathcal{B}^q)$ the Borel measure space of $\mathbb{R}^q$. Additionally, denote by $\phi_q(\,\bullet\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$ the probability density function (pdf) of the q-variate normal distribution $\mathcal{N}_q(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^q$ positive-definite covariance matrix $\boldsymbol{\Sigma}$. Then we define the Gaussian measure, which corresponds to $\mathcal{N}_q(\boldsymbol{\mu},\boldsymbol{\Sigma})$, by

$$\tilde{\Phi}_q(A;\boldsymbol{\mu},\boldsymbol{\Sigma}) := \int_A \phi_q(\boldsymbol{\eta};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{\eta}, \qquad A \in \mathcal{B}^q, \tag{1.12}$$

where the integral has to be understand with respect to the Lebesgue measure on $(\mathbb{R}^q, \mathcal{B}^q)$. For simplicity, we set $\tilde{\Phi}_q(A) := \tilde{\Phi}_q(A;\mathbf{0},\mathrm{I}_q)$, $A \in \mathcal{B}^q$, where $\mathbf{0}$ denotes the null-vector and I_q the $q \times q$ identity matrix.

Definition 1.1.7 (Standard Skew Normal Distribution, cf. Bazán et al., 2010)
Suppose X is a random variable defined over the probability space $(\Omega, \mathcal{A}, \mathsf{P})$ and the cdf Φ_γ of X is given by

$$\Phi_\gamma(x) = 2\tilde{\Phi}_2\left(\boldsymbol{r}(x);\mathbf{0},\boldsymbol{\Sigma}\right), \qquad x \in \mathbb{R}, \tag{1.13}$$

where

$$\boldsymbol{r}\colon \mathbb{R} \to \mathcal{B}^2, \qquad x \mapsto \boldsymbol{r}(x) = \left(\begin{pmatrix}-\infty\\-\infty\end{pmatrix}, \begin{pmatrix}\sigma^{-1}(x-\mu)\\0\end{pmatrix}\right],$$
$$\boldsymbol{\Sigma} = \begin{pmatrix}1 & -\delta\\-\delta & 1\end{pmatrix}, \quad \delta = \frac{\gamma}{\sqrt{1+\gamma^2}}, \mu = -\frac{\sqrt{2}\delta}{\sqrt{\pi-2\delta^2}}, \sigma = \sqrt{\frac{\pi}{\pi-2\delta^2}},$$

then we say X has a *standard skew normal distribution* with skewness parameter $\gamma \in \mathbb{R}$ and write $X \sim \mathcal{SN}(\gamma)$. ◇

For an illustration of the family of standard skew normal distributions, compare Figure 1.2. Denote by Φ_γ^{-1} the quantile function of the skew normal distribution of Definition 1.1.7

[3] Further consider the contributions of Bazán et al. (2004, 2006, 2014) and Venturelli (2015) regarding the IRT modeling using skew probit link functions.

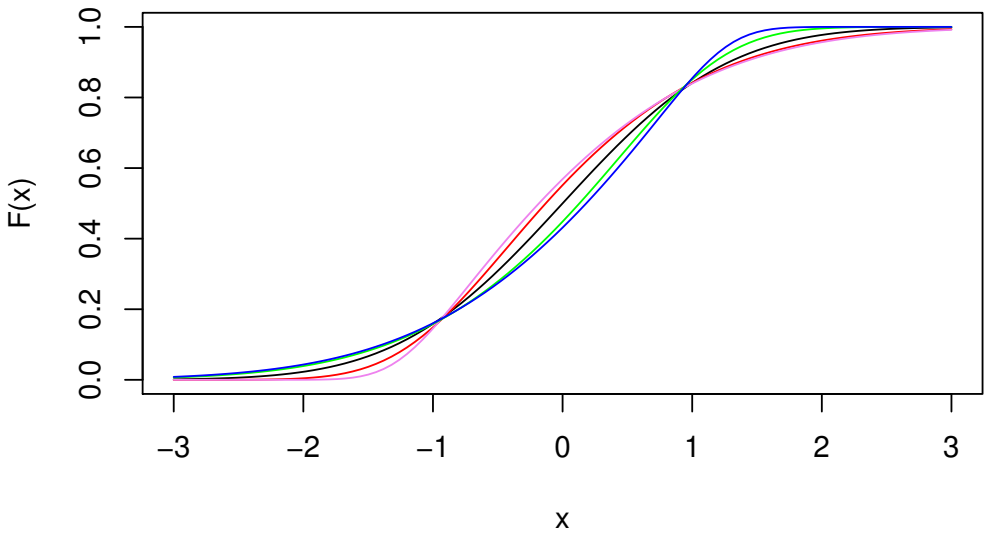

Figure 1.2: *Standard skew normal distributions with skewness parameters -6 (blue), -3 (green), 0 (black), 3 (red) and 6 (violet).*

with skewness parameter $\gamma \in \mathbb{R}$, and consider the *(standard) skew probit* link as

$$\mathrm{ssp}_\gamma(x) := \Phi_\gamma^{-1}(x), \qquad x \in (0,1). \tag{1.14}$$

The standard skew normal distribution of Definition 1.1.7 has some nice properties. First, ssp_0 equals the probit link. Due to the close relation between logistic and normal distributions, skew probit models can be very close to logit models. Moreover, it holds with the skew standard normal cdf $F_\gamma(\bullet)$:

$$\sup_{x\in\mathbb{R}} |F_{4.5658}(0.8238 \cdot x - 0.4642) - \exp(-\exp(-x))| \approx 0.008, \tag{1.15}$$

which is an even slightly better approximation of the TIEV distribution than a logistic distribution approximates a normal distribution (the error is ≈ 0.0095). Compare also Figure 1.3. The cdf $x \mapsto 1 - \exp(-\exp(x))$ can be approximated analogously. So, loglog and cloglog models can be very well approximated by skew probit models. So overall, the family of skew probit links is a flexible tool which contains some of the most important link functions, at least approximately. The skewness is a real valued parameter which enables the possibility of a continuous adjustment from the model on the skewness in the data. This skewness can be lie between cloglog/loglog and probit or even more extreme than cloglog/loglog.

Remark 1.1.8
The family skew normal distribution in Definition 1.1.7 can be parameterized equivalently by $\delta = \gamma(1+\gamma^2)^{-1/2} \in (-1,1)$. Bazán et al. (2010) expect that this can be helpful in computations. ◇

In a completely analogous way, a skew logit link can be defined. However, the closed form expression of the standard logistic distribution provides two alternative ways to define skew families of link functions. They are known as *scobit* and *power logit* (cf. the work of Samejima (1997, 2000) and Bolfarine and Bázán (2010) for detailed discussions).

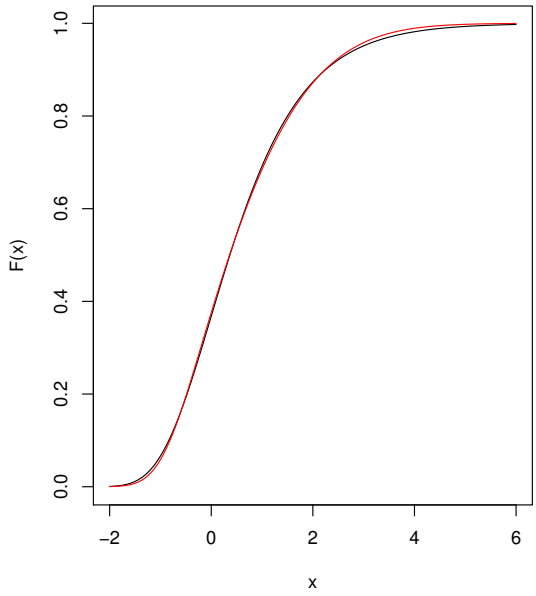

Figure 1.3: *Aproximation of the TIEV distribution (black) by a skew normal distribution (red)*

Another interesting parametric family of link functions is the log-gamma link, which was used by Genter and Farewell (1985) for goodness of link testing in GLM. This family of link functions captures the loglog, probit and cloglog link.

Besides these important cases of parametric link functions for binary data models, there are methods and concepts to derive specific links, such that they satisfy specific optimality criteria or provide specific interpretations, in a statistical information theory framework, but are beyond our scopes (cf. e.g. Kateri and Agresti, 2010).

1.1.2.3 Guessing Models

In addition to the mentioned development of general flexible skew item characteristic curves, there is also a phenomenon in multiple-choice data which led to a long tradition of IRT modeling: In particular for difficult items, the left tail of the empirical item characteristic curves can often be observed to be greater then zero (cf. e.g. Baker and Kim, 2004, p. 18). In particular, this is commonly greater than the estimated value of a 2PL or Lawley–Lord model. In multiple-choice tests, this can be well explained by guessing of test participants. So, we could expect

$$\lim_{\eta \to -\infty} \mathsf{P}(Y_i = 1|\eta) \geq c_i > 0 \tag{1.16}$$

and if item i has k answer options $c_i \approx \frac{1}{k}$. But the item response models we discussed until now yield $P(Y_i = 1|\eta) \to 0$ as $\eta \to -\infty$. To deal with guessing, Birnbaum (1968) introduced the *three parameter logistic model* (3PL model), which is an extension of the 2PL model with a further guessing parameter for a lower asymptote above zero. The model

is defined by

$$\operatorname{logit}\left(\frac{P_i(\eta)-c_i}{1-c_i}\right)=\alpha_i(\eta-\beta_i),\qquad i=1,\dots,d, \tag{1.17}$$

for $\eta\in\Theta$, with difficulty parameters $\beta_1\dots,\beta_d\in\mathbb{R}$, discrimination parameters $\alpha_1\dots,\alpha_d>0$ and guessing parameters $c_1,\dots,c_d\in[0,1)$. The concept is neither restricted to the logit link nor to the case of $q=1$. We formulate the corresponding model classes in the following definition.

Definition 1.1.9
Suppose $d\in\mathbb{N}$ items and $\boldsymbol{Y}\sim\mathcal{P}(\boldsymbol{\eta})$ for $\boldsymbol{\eta}\in\Theta\subset\mathbb{R}^q$ for $q\in\mathbb{N}$. If there exist strictly increasing continuous and bijective mappings $\mathfrak{g}_1,\dots,\mathfrak{g}_d\colon(0,1)\to\mathbb{R}$, scalars $\alpha_{10},\dots,\alpha_{d0}\in\mathbb{R}$, $c_1,\dots,c_d\in[0,1)$ and vectors $\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_d\in\mathbb{R}^q$, such that

$$\mathfrak{g}_i\left(\frac{P_i(\boldsymbol{\eta})-c_i}{1-c_i}\right)=\alpha_{i0}+\boldsymbol{\alpha}_i^{\intercal}\boldsymbol{\eta},\qquad i=1,\dots,d, \tag{1.18}$$

holds, then we say a *3PL-type model* holds.[4] ◇

Remark 1.1.10

1. Similar to the 3PL model, there is a 4PL model which further contains a parameter for the upper bound of the probability of correct response (cf. Barton and Lord, 1981). The definition is straight forward, but the model is practically less important.

2. A kind of "Rasch-3PL-type" model can directly be defined by considering the left and the right-hand side of the model equation correspondingly. As both modeling approaches target to different aims, such a model can be useful. We call the corresponding model *Rasch-type-guessing*

3. Similar to the models in the last subsection, the 3PL and 3PL-type models can be considered GLLVM with a parametric link function. ◇

Even if the 3PL model is a popular model for the inclusion of guessing by test participants, it has serious technical and practical problems. Problems with estimating c_i, $i=1,\dots,d$, or constructing simultaneous confidence bands are often reported in the literature (cf. e.g. Pashley, 1991, p. 4). Holland (1990, Section 3.2) even conjectured that it might be impossible to estimate $c_1,\dots,c_d$ consistently for long tests. Another – more practical – aspect is: The guessing in the 3PL model is incorporated by a single parameter which determines the lower asymptote. As for example San Martín et al. (2006) discussed: The lower asymptote is often smaller than the reciprocal of the number of response options, which indicates the attractiveness of a particular wrong response option or otherwise intelligent test design. Tn other cases, the lower asymptote is greater than the reciprocal of the number of response options. Then, some options can be eliminated before guessing (cf. e.g. San Martín et al., 2006, p. 187). In both cases, the selection depends on the ability of test participants so

[4] On a first view, the name seems not to be intuitive, because the L in 3PL refers to the logit link. However, as Baker and Kim (2004, p. 21) noted, the 3PL model is actually not a logit model and should better be called *Birnbaum's three parameter model*. On the other side, through the established name of the 3PL model, we think that *3PL-type* captures the primary aspects of its formulation to allow an easy recognition. Similarly, for example Fu et al. (2009) called the logit-version of (1.18) *multidimensional 3PL model*, even if the model has in truth $q+2$ parameters for each item. In this sense, the *fist* parameter is the intercept, *any second* a factor loading and *the third* the guessing parameter.

that it is reasonable to incorporate the dependence from the ability instead of taking the same lower asymptote for all abilities as property of the ICC.

Such aspects led to a development of alternative models to deal with guessing (cf. e.g. Pashley, 1991; San Martín et al., 2006; Cao and Stokes, 2008; Lee and Bolt, 2018). The 3PL model still is the most popular guessing model. This indicates the ongoing need for an interesting alternative. From our point of view, another aspect to discuss is the shape of the 3PL model and current alternatives. The current guessing models either cannot be interpreted by the underlying variable approach, or they decrease to zero in the same order as the common logit and probit models. We will discuss this in more detail in Section 2.2.

1.1.3 Structural Models: The Distribution of the Latent Vector

The structural model is the prior distribution $\mathcal{H}$ of $\boldsymbol{\eta}$. In this work, we focus on continuously distributed latent vectors. Generally, $\mathcal{H}$ is a part of the statistical model which should not be neglected. But it should also not be overrated as Bartholomew et al. (2011) discussed. In particular, we will discuss in detail that the influence of the prior distribution vanishes with $d \to \infty$ in Chapter 3.

One of the most important structural models is the (multivariate) normal distribution. As discussed by Bartholomew et al. (2011, Section 2.8 and Section 2.14), we need identification constraints if using a GLLVM with a normal distribution. As consequence, we can set without loss of generality $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$, which implies $\boldsymbol{\eta} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ as fixed structural model. To simplify the notation, we define some important combinations of measurement and structural models in the following definition.

As pointed out by Samejima (2000), latent variables do not have to be normally distributed. With further reference to Micceri (1989), Azevedo et al. (2011) and Bazán et al. (2014) proposed the assumption that a latent vector $\boldsymbol{\eta}$ has a parametric skew normal distribution. With the restriction of a standardized family, the resulting model parameters are identifiable, while the skewness in the distribution of $\boldsymbol{\eta}$ can be estimated jointly with the further model parameters. As the standard normal distribution is a special case of the family of standard skew normal distributions, the common assumption is a special case of the proposed model. This joint incorporation of possible skewness in the structural model allows an analysis of the specific sources of skewness behind the response data. Alternatively, using a discrete mixing distribution as a non-parametric model was discussed in the literature (cf. Butler and Louis, 1997).

Definition 1.1.11

1. Suppose $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ follows a GLLVM and all link functions $\mathfrak{g}_i$ are from a generic class M. Further suppose the distribution of $\boldsymbol{\eta}$ is from the family D. Then we say a *M-D model* holds. ◇

We illustrate this abstract definition on a few examples.

Example 1.1.12

1. Suppose $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ follows a logit model and $\boldsymbol{\eta} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ holds. Then we say a *logit-normal model* holds.

2. Suppose $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ follows a probit model and $\boldsymbol{\eta} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ holds. Then we say a *probit-normal model* holds.

3. Suppose $q = 1$ and $\boldsymbol{Y} \sim \mathcal{P}(\eta)$ with a GLLVM for $d \in \mathbb{N}$ items, where $\mathfrak{g}_i = \mathrm{ssp}_{\alpha_i} = \Phi_{\alpha_i}^{-1}$ (cf. (1.14)) for some $\alpha_1, \ldots, \alpha_d \in \mathbb{R}$. Further suppose η has a standardized exponential

distribution, that is $\eta := \xi - 1$ for $\xi \sim \text{Exp}(1)$. Then we say a *skew probit-exponential model* holds. ◇

The structural model has a more important role if there are further covariates, which should be taken into account in the analysis. Such covariates can be indicators for a group membership (like gender), corresponding to factors, or continuous variables as age or height. An important approach to include covariates in the model is given by formulating a latent linear model for the dependence of $\boldsymbol{\eta}$ on the covariates (cf. e.g. Antal, 2007).

1.2 Models for the Marginal Probabilities

To work with item response models, it is often important to evaluate the integral of the marginal probabilities of $\boldsymbol{Y}$. But for many important cases as the logit-link model with normal prior, it is not possible to evaluate the integral analytically. In these cases it is necessary to use numerical integral approximation techniques. In the Subsections 1.2.1, 1.2.2 and 1.2.3, we discuss three approaches of the IRT literature to avoid numerical integration and thereby get a model for the marginal distribution of $\boldsymbol{Y}$. These approaches focus on logit models and lead to the same resulting model for the marginal distribution. The approaches have implications on the distribution $\mathcal{H}$ of $\boldsymbol{\eta}$, which we discuss in Subsection 1.2.4. The resulting model can be interpreted as an association model. We discuss this in the Subsection 1.2.5. Simultaneously, the model is closely related to the Ising model of theoretical physics and the modern branch of network psychometrics. We will briefly sketch the relation in Subsection 1.2.6.

1.2.1 Dutch Identity Approach

In this section we present the derivation of a closed form expression of $P(\boldsymbol{y})$ under some requirements on the IRT model. The approach and the model derivation we discuss here were derived by Holland (1990, Corollary 1). The important tool is following *The Dutch Identity*.

Theorem 1.2.1 (Dutch Identity, Holland, 1990, Theorem 1)
Suppose $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ for $d \in \mathbb{N}$ items, some response functions $P_1, \ldots, P_d$ and $\boldsymbol{\eta} \in \Theta \subset \mathbb{R}^q$ for $q \in \mathbb{N}$ and with $\boldsymbol{\eta} \sim \mathcal{H}$ for any distribution $\mathcal{H}$. Then for every $\boldsymbol{y}_0 \in \mathbb{Y} = \{0,1\}^d$ holds

$$\frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_0)} = \mathsf{E}\Big(\exp\Big((\boldsymbol{y} - \boldsymbol{y}_0)^\intercal \boldsymbol{\lambda}(\boldsymbol{\eta}) \Big) \,\Big|\, \boldsymbol{Y} = \boldsymbol{y}_0 \Big), \tag{1.19}$$

for all $\boldsymbol{y} \in \mathbb{Y}$, where $\boldsymbol{\lambda}\colon \Theta \to \mathbb{R}^d$ is vector of item logit functions defined by

$$\boldsymbol{\lambda}(\boldsymbol{\eta}) := (\lambda_1(\boldsymbol{\eta}), \ldots, \lambda_d(\boldsymbol{\eta}))^\intercal, \qquad \boldsymbol{\eta} \in \Theta,$$

where $\lambda_i(\boldsymbol{\eta}) := \text{logit}(P_i(\boldsymbol{\eta}))$ is the item logit of item i, $i = 1, \ldots, d$. ◇

The moment generating function (MGF) $\text{M}_{\boldsymbol{\xi}}$ of a random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^\intercal$ is given by

$$\boldsymbol{t} \mapsto \text{M}_{\boldsymbol{\xi}}(\boldsymbol{t}) := \mathsf{E}\Big(\exp\big(\boldsymbol{t}^\intercal \boldsymbol{\xi} \big) \Big).$$

Therefore, we notice that the right-hand side of (1.19) is the posterior moment generating function of the random vector $\boldsymbol{\lambda}(\boldsymbol{\eta})$ given $\boldsymbol{Y} = \boldsymbol{y}_0$ and evaluated at $\boldsymbol{y} - \boldsymbol{y}_0$. For the effective use of this fact we make two further assumptions about $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$:

(A1) There is a vector $\boldsymbol{y}_0 \in \mathbb{Y}$ such that the posterior distribution of $\boldsymbol{\eta}$, given $\boldsymbol{Y} = \boldsymbol{y}_0$, is q-dimensional multivariate normal distributed with mean $\boldsymbol{\mu}_{\boldsymbol{y}_0} = \mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0) \in \mathbb{R}^q$ and positive definite covariance-matrix $\boldsymbol{\Sigma} = \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0)$, that is

$$(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0) \sim \mathcal{N}_q(\boldsymbol{\mu}_{\boldsymbol{y}_0}, \boldsymbol{\Sigma}).$$

(A2) $\boldsymbol{\lambda}$ is an affine map, that means there is a $d \times q$ matrix $\boldsymbol{A}$ and a $d \times 1$ vector $\boldsymbol{\beta}$, such that for all $\boldsymbol{\eta}$

$$\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} - \boldsymbol{\beta}.$$

Remark 1.2.2

1. By the definition of the logit link, assumption (A2) is equivalent to the assumption that a logit model holds for $\boldsymbol{Y}$. In this case, the matrix $\boldsymbol{A}$ is formed by the parameters $(\alpha_{ik})_{i=1,k=1}^{d,q}$ with $(\alpha_{i1}, \ldots, \alpha_{iq}) = \boldsymbol{\alpha}_i^\intercal$, $i = 1, \ldots, d$, and the vector $\boldsymbol{\beta}$ by $(-\alpha_{10}, \ldots, -\alpha_{d0})^\intercal$.

2. Even if the assumption (A1) about the posterior distribution is useful (as we will see), it is quite curious. We will discuss this assumption in Section 1.2.4 and related questions in Chapter 3.

3. One might wonder how critical the pattern $\boldsymbol{y}_0$ in assumption (A1) is: If there is such a $\boldsymbol{y}_0$, does a similar result hold for a different $\boldsymbol{y}_1 \in \mathbb{Y}$, too? In particular, does $\boldsymbol{\Sigma} = \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_1)$ hold? Overall, how much attention do we have to pay to $\boldsymbol{y}_0 \in \mathbb{Y}$? The following proposition shows that $\boldsymbol{y}_0 \in \mathbb{Y}$ is actually arbitrary. ◇

Proposition 1.2.3
Suppose $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$. Assume that the conditions (A1) and (A2) hold. Then it holds for all $\boldsymbol{y} \in \mathbb{Y}$:

$$(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}) \sim \mathcal{N}_q\big(\boldsymbol{\mu}_{\boldsymbol{y}_0} + \boldsymbol{\Sigma}\boldsymbol{A}^\intercal(\boldsymbol{y} - \boldsymbol{y}_0), \boldsymbol{\Sigma}\big). \qquad \diamond$$

Proof.
This was first proved by Knott and Albanese (1993) for the restriction to $q = 1$ (cf. alternatively Anderson and Yu, 2007). The case of a general $q \in \mathbb{N}$ follows directly from the arguments of Hessen (2012, Theorem 2) with the restriction to binary items. Alternatively, it can easily be derived from Theorem 6.7 of Li (2010). □

Returning to the derivation of a closed form expression of $P(\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{Y}$, consider a random vector $\boldsymbol{\xi} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then we know that for any affine transformation $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b}$ with $k \times d$ matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ and vector $\boldsymbol{b} \in \mathbb{R}^k$ it holds $\boldsymbol{\eta} \sim \mathcal{N}_k(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal)$. Therefore assumption (A2) ensures that $\boldsymbol{\lambda}$ is an affine transformation of the random vector $\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0$ and with assumption (A1) we get

$$(\boldsymbol{\lambda}(\boldsymbol{\eta}) \mid \boldsymbol{Y} = \boldsymbol{y}_0) \sim \mathcal{N}_d(\boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}), \boldsymbol{B}),$$

with $\boldsymbol{B} := \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal$.

For a multivariate normally distributed random vector $\boldsymbol{\xi} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the moment generating function is explicitly given by

$$\mathrm{M}_{\boldsymbol{\xi}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\right).$$

Thus, it is possible to evaluate the right-hand side of (1.19) and we get

$$\frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_0)} = \exp\left((\boldsymbol{y} - \boldsymbol{y}_0)^\top \boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_0)^\top \boldsymbol{B}(\boldsymbol{y} - \boldsymbol{y}_0)\right). \tag{1.20}$$

Let $\boldsymbol{b}'$ be the diagonal of $\boldsymbol{B}$, $\mathrm{diag}(\boldsymbol{b}')$ the corresponding diagonal matrix and $\boldsymbol{\Gamma} := \boldsymbol{B} - \mathrm{diag}(\boldsymbol{b}')$. Since $\boldsymbol{y}_0$ is arbitrary, we can set it to $\boldsymbol{y}_0 = \boldsymbol{0}$ for simplicity. Notice that $y_i^2 = y_i$, since $y_i \in \{0, 1\}$, so (1.20) becomes

$$\begin{aligned}
\frac{P(\boldsymbol{y})}{P(\boldsymbol{y}_0)} &= \exp\left(\boldsymbol{y}^\top \boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{B} \boldsymbol{y}\right) \\
&= \exp\left(\boldsymbol{y}^\top \boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}\boldsymbol{y}^\top (\boldsymbol{\Gamma} + \mathrm{diag}(\boldsymbol{b}'))\boldsymbol{y}\right) \\
&= \exp\left(\boldsymbol{y}^\top \boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}\boldsymbol{y}^\top \mathrm{diag}(\boldsymbol{b}')\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Gamma} \boldsymbol{y}\right) \\
&= \exp\left(\boldsymbol{y}^\top \left(\boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}\boldsymbol{b}'\right) + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Gamma} \boldsymbol{y}\right).
\end{aligned}$$

With setting $a = \log(P(\boldsymbol{0}))$ and $\boldsymbol{b} = \boldsymbol{\lambda}(\boldsymbol{\mu}_{\boldsymbol{y}_0}) + \frac{1}{2}\boldsymbol{b}'$, the marginal distribution (1.1) satisfies

$$\begin{aligned}
\log(P(\boldsymbol{y})) &= a + \boldsymbol{y}^\top \boldsymbol{b} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Gamma} \boldsymbol{y} \\
&= a + \sum_{i=1}^{d} b_i y_i + \sum_{i=1}^{d} \sum_{k=i+1}^{d} \gamma_{ik} y_i y_k,
\end{aligned} \tag{1.21}$$

since $\boldsymbol{\Gamma} = (\gamma_{ik})$ is symmetric. Note that $\log(P(\boldsymbol{y}))$ is independent of the latent vector $\boldsymbol{\eta}$, so it only depends on observed data. In particular, formula (1.21) leads to the direct definition of a model for the response vector $\boldsymbol{Y}$. To allow for easy distinction between the model formula (1.21) and the abstract marginal probabilities (1.1), we define it as its own model and summarize the derivation.

Theorem/Definition 1.2.4 (Manifest Prob. Model, cf. Holland, 1990, Corollary 1)
Consider a random vector $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$. Suppose that conditions (A1) and (A2) hold. Then it exist scalars a, $b_1, \ldots, b_d$, $\gamma_{ij} \in \mathbb{R}$, $j > i$, such that the distribution $\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}) = P(\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{Y}$ satisfies (1.21) and we say that a *manifest probabilities model* holds.[5] ◇

We illustrate the closed form marginal distribution with a small example.

Example 1.2.5
In the case of the Rasch model we get with $\boldsymbol{\Sigma} = \sigma^2 > 0$ and $\boldsymbol{A} = \boldsymbol{\alpha} = (\alpha, \ldots, \alpha)^\top$ with $\alpha > 0$:

$$\boldsymbol{B} = \sigma^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top = \sigma^2 \alpha^2 \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}.$$

[5] We call it manifest probabilities after Holland (1990), who called the marginal probabilities generally manifest probabilities as they can theoretically be estimated directly as relative frequencies.

So the manifest probabilities model in (1.21) becomes

$$\log(P(\boldsymbol{y})) = a + \sum_{i=1}^{d} \left(\alpha\mu_{\boldsymbol{y}_0} - \alpha\beta_i + \frac{\alpha^2\sigma^2}{2} \right) y_i + \alpha^2\sigma^2 \sum_{i=1}^{d} \sum_{k=i+1}^{d} y_i y_k.$$

In this set-up, the more complex Birnbaum-Rasch model with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\intercal \in \mathbb{R}^d$ becomes

$$\log(P(\boldsymbol{y})) = a + \sum_{i=1}^{d} \left(\alpha_i\mu_{\boldsymbol{y}_0} - \alpha_i\beta_i + \frac{\alpha_i^2\sigma^2}{2} \right) y_i + \sum_{i=1}^{d} \sum_{k=i+1}^{d} (\alpha_i\sigma y_i)(\alpha_k\sigma y_k). \qquad \diamond$$

Notice that we can only make relative comparison of items and subjects. In particular, the data does not contain any information about a precise value of $\boldsymbol{\mu}_{\boldsymbol{y}_0}$ (especially for arbitrary $\boldsymbol{y}_0$). So $\boldsymbol{\mu}_{\boldsymbol{y}_0}$ itself is arbitrary and we can set $\boldsymbol{\mu}_{\boldsymbol{y}_0} = \mathbf{0}$ as an identification constraint.

1.2.2 Graphical Model Approach

Random variables and their dependencies can be represented by graphs. In such a graph, knots represent random variables and edges direct dependences. In a directed graph, the direction symbolizes the direction of dependence. In our case, items and latent variables are knots. Since we assume conditional independence of the items given the ability, there are no edges between item knots. In item response theory, we model the response probabilities under the given latent variable, so we get directed edges from the latent variable knots to item knots. A possible graph of the dependence-structure of four items and one tested latent variable is given in Figure 1.4 a). For complex dependence structures with more than one latent variable, graphs give an intuitive way of defining corresponding models and set restrictions to parameter values in the obtained manifest probabilities model. Graphical models are useful for creation of models for manifest probabilities. Using the dependencies, we derive a model formulation for these probabilities following Anderson and Vermunt (2000); Anderson et al. (2007); Anderson and Yu (2007, 2014). First, we take the reverse direction and model the latent variable given the responses: We start with posterior-modelling. The corresponding graph for four items and one latent variable is given in Figure 1.4 b).

To start the derivation, we make a first assumption:

(B1) For all $\boldsymbol{y} \in \mathbb{Y}$ holds: The latent vector $\boldsymbol{\eta}$ is multivariate normal distributed conditional on $\boldsymbol{Y} = \boldsymbol{y}$ with

$$(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}) \sim \mathcal{N}_q(\boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}\colon \mathbb{Y} \to \mathbb{R}^q$ with

$$\boldsymbol{y} \mapsto \boldsymbol{\Sigma}\boldsymbol{A}^\intercal\boldsymbol{y} =: \boldsymbol{\mu}_{\boldsymbol{y}}$$

for a matrix $(\alpha_{ik})_{i,k=1}^{d,q} = \boldsymbol{A} \in \mathbb{R}^{d\times q}$ and a further symmetric positive definite matrix $(\sigma_{jr})_{j,r=1}^{q,q} = \boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$, which are equal for all $\boldsymbol{y} \in \mathbb{Y}$.

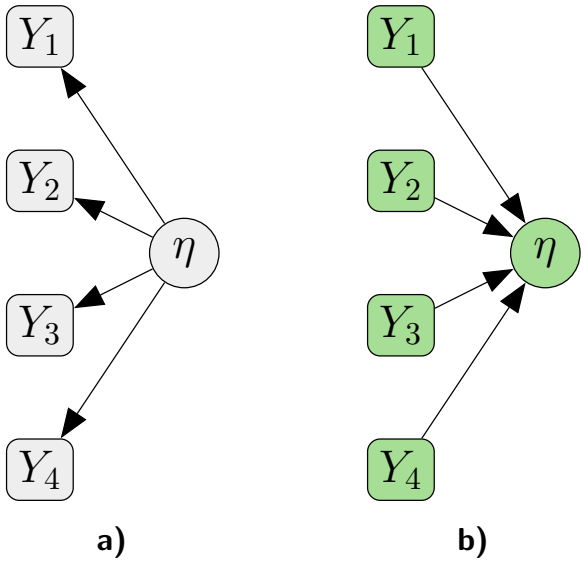

Figure 1.4: *Example graphs representing item response models.*
a) Reflective model: Model directed influence of latent structure on items
b) Formative model: Infer from item responses the latent structure.

Denoting the joint density of $\boldsymbol{\eta}$ and $\boldsymbol{Y}$ by f, we get from the rules of conditional probabilities and densities

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{\eta}) &= \mathsf{P}(\boldsymbol{Y}=\boldsymbol{y})\mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y}=\boldsymbol{y}) \\ &= P(\boldsymbol{y})\mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y}=\boldsymbol{y}) \\ &= P(\boldsymbol{y})(2\pi)^{-\frac{q}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\eta}-\boldsymbol{\mu}_{\boldsymbol{y}})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}-\boldsymbol{\mu}_{\boldsymbol{y}})\right) \\ &= P(\boldsymbol{y})(2\pi)^{-\frac{q}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{y}}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}}\right) \exp\left(-\frac{1}{2}\boldsymbol{\eta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right) \exp\left(\boldsymbol{\mu}_{\boldsymbol{y}}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right). \quad (1.22) \end{aligned}$$

Our aim is to derive a log-multiplicative (and for special cases in particular a log-linear) model for the marginal probabilities $P(\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{Y}$. The equation (1.22) gives a relation between P and f which can be used to derive a model formulation. In particular, we notice that (1.22) yields a multiplicative decomposition of f into three parts. The first part is

$$P(\boldsymbol{y})(2\pi)^{-\frac{q}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{y}}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}}\right), \quad (1.23)$$

depending only on $\boldsymbol{y} \in \mathbb{Y}$. The second part is

$$\exp\left(-\frac{1}{2}\boldsymbol{\eta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right),$$

depending only on $\boldsymbol{\eta} \in \Theta = \mathbb{R}^q$. And, finally, the third part is

$$\exp\left(\boldsymbol{\mu}_{\boldsymbol{y}}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right),$$

depending on both, $\boldsymbol{y} \in \mathbb{Y}$ and $\boldsymbol{\eta} \in \Theta$, but without a further possibility to decompose it into a product of two functions only depending on $\boldsymbol{y} \in \mathbb{Y}$ and $\boldsymbol{\eta} \in \Theta$, respectively. An essential step in the model derivation is the insertion of further structure into the first part. Hence, we make a further crucial assumption regarding the first part in term (1.23).

(B2) There is a linear map $g\colon \mathbb{Y} \to \mathbb{R}$, $\boldsymbol{y} \mapsto \boldsymbol{z}^\intercal \boldsymbol{y}$ for a coefficient vector $\boldsymbol{z} = (z_1, \dots, z_d)^\intercal \in \mathbb{R}^d$ such that

$$g(\boldsymbol{y}) = \log(P(\boldsymbol{y})) - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{y}}| - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{y}}^\intercal \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}} \tag{1.24}$$

for all $\boldsymbol{y} \in \mathbb{Y}$.

Using the definition of $\boldsymbol{\mu}$ in assumption (B1), we get the following representation of the joint density f of $\boldsymbol{\eta}$ and $\boldsymbol{Y}$:

$$f(\boldsymbol{y}, \boldsymbol{\eta}) = \exp\left(g(\boldsymbol{y}) + \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\eta}^\intercal \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right), \qquad (\boldsymbol{y}, \boldsymbol{\eta}) \in \mathbb{Y} \times \Theta.$$

Therefore, the coefficients of g must take such values, for which f is a density, while P is restricted by (1.24). Now, to derive a model for P, we solve the equation (1.24) for $P(\boldsymbol{y})$ and get, using again the definitions of $\boldsymbol{\mu}_{\boldsymbol{y}}$ in (B1) and $g(\boldsymbol{y})$ in (B2),

$$\begin{aligned} P(\boldsymbol{y}) &= \exp\left(-\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \boldsymbol{z}^\intercal \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal \boldsymbol{y}\right) \\ &= \exp\left(a + \boldsymbol{z}^\intercal \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal \boldsymbol{y}\right), \end{aligned} \tag{1.25}$$

with $a := -\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}|$. Notice that since P itself is a density, it is sufficient to ensure $\sum_{\boldsymbol{y}\in\mathbb{Y}} P(\boldsymbol{y}) = 1$ for reasonable parameter values. In particular, using a as normalization constant, there is no need to take f into account any longer. So there is up to now no practical estimation restriction on the parameters.

We achieve the manifest probabilities model from Theorem/Definition 1.2.4 by taking logarithms and using a reparametrization. First,

$$\begin{aligned} \log(P(\boldsymbol{y})) &= a + \sum_{i=1}^{d} z_i y_i + \frac{1}{2}\sum_{i=1}^{d}\sum_{k=1}^{d}\sum_{j=1}^{q}\sum_{r=1}^{q} \alpha_{ij} y_i \sigma_{jr} \alpha_{kr} y_k \\ &= a + \sum_{i=1}^{d} z_i y_i + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{q}\sum_{r=1}^{q} \alpha_{ij}\sigma_{jr}^2\alpha_{ir} y_i^2 + \frac{1}{2}\sum_{i=1}^{d}\sum_{k\neq i}\sum_{j=1}^{q}\sum_{r=1}^{q} \alpha_{ij} y_i \sigma_{jr} \alpha_{kr} y_k \\ &= a + \sum_{i=1}^{d}\left(z_i + \frac{1}{2}\sum_{j=1}^{q}\sum_{r=1}^{q}\alpha_{ij}\sigma_{jr}^2\alpha_{ir}\right) y_i \\ &\qquad + \sum_{i=1}^{d}\sum_{k=i+1}^{d}\sum_{j=1}^{q}\sum_{r=1}^{q} \alpha_{ij} y_i \sigma_{jr} \alpha_{kr} y_k, \end{aligned} \tag{1.26}$$

since $y_i \in \{0, 1\}$ and so $y_i^2 = y_i$. Finally, we get the same formula as (1.21) if we summarize in the last sum

$$\begin{aligned} b_i &:= z_i + \frac{1}{2}\sum_{j=1}^{q}\sum_{r=1}^{q}\alpha_{ij}\sigma_{jr}\alpha_{ir}, \quad i = 1, \dots, d \\ \gamma_{ik} &:= \begin{cases} \sum_{j=1}^{q}\sum_{r=1}^{q}\alpha_{ij}\sigma_{jr}\alpha_{kr}, & k \neq i, \\ 0, & k = i, \end{cases} \qquad i, k = 1, \dots, d. \end{aligned} \tag{1.27}$$

Remark 1.2.6

1. Actually, the conditions of the Dutch Identity approach and of the graphical model approach are equivalent, as first discussed by Anderson and Yu (2007) for a single latent variable and by Anderson and Yu (2014) for the case considered here. From the modeling perspective, an advantage of conditions (A1) and (A2) (cf. p. 19) of the Dutch Identity approach is that we start with a logit IRT model and the additional assumption of the posterior normal distribution of the latent vector for a single arbitrary response pattern. These conditions imply
   a) the posterior normality for all response patterns with linear posterior expected value and equal covariance matrix as in condition (B1) due to Proposition 1.2.3 with $\boldsymbol{y}_0 = \mathbf{0}$ and $\boldsymbol{\mu_0} = \mathbf{0}$ and
   b) that the manifest probabilities model holds as we discussed in Section 1.2.1.

   The graphical model approach gives insight into the backward direction, as condition (B2) together with the assumed form of the linear posterior expected values actually contains the profile of the manifest probabilities model. But these assumptions imply that a logit model for $\boldsymbol{Y}$ conditional on $\boldsymbol{\eta}$ holds (cf. Anderson and Yu 2014, equation (11)).
2. Anderson and co-authors mainly considered so called *confirmatory factor models*, where it is a-priori known (or assumed) on which items a latent variable has any effect. In contrast, we focus on the case where is unknown, the so called *exploratory factor model*. While this makes no difference in the model formulation, it influences the corresponding estimation procedures and identification constraints. A strength of the graphical model approach is the clear representation of these dependencies. ◇

1.2.3 Conditionally Defined Model Approach

Anderson and Yu (2007) and Anderson et al. (2007) (later generalized by Anderson et al. (2010) and Anderson (2013)) discussed a further approach of deriving the manifest probabilities model (cf. Theorem/Definition 1.2.4) using the concept of conditionally defined models (cf. Arnold et al., 1999). We will now briefly present a derivation of the manifest probabilities model which is based on the arguments of Anderson et al. (2007).

Suppose again that $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ satisfies a logit model for $d \in \mathbb{N}$ items and a $q \in \mathbb{N}$ dimensional latent vector $\boldsymbol{\eta}$. Thus, we suppose that $\alpha_{ij} \in \mathbb{R}$, $i = 1, \ldots, d$, $j = 0, \ldots, q$ exist such that

$$\lambda_i(\boldsymbol{\eta}) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij}\eta_j, \qquad i = 1, \ldots, d.$$

Under the stated assumptions

$$T_j := \sum_{i=1}^{d} \alpha_{ij} Y_i$$

would be a sufficient statistic for η_j, $j = 1, \ldots, q$, if we would knew the values of α_{ij}, $i = 1, \ldots, d$, $j = 1, \ldots, q$ (cf. Bartholomew et al., 2011, Section 2.5). Further recall that if we additionally assume that $\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0$ is normally distributed for any $\boldsymbol{y}_0 \in \mathbb{Y}$, we get

$$\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}) = \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0)\boldsymbol{A}^\intercal \boldsymbol{y} + \big(\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0) - \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0)\boldsymbol{A}^\intercal \boldsymbol{y}_0\big),$$

for all $\boldsymbol{y} \in \mathbb{Y}$, where $\boldsymbol{A} = (\alpha_{ij})_{i,j=1}^{d,q}$, using Proposition 1.2.3. Hence, the posterior expected value, which is the Bayes efficient estimator for $\boldsymbol{\eta}$ under the squared error loss, is an affine transformation of $\boldsymbol{T} = (T_1, \ldots, T_q)^\intercal$. Now suppose that we want to predict Y_i for an arbitrary item $i \in \{1, \ldots, d\}$ for a subject having the unknown value $\boldsymbol{\eta} = \boldsymbol{\eta}_0 \in \mathbb{R}^q$. Further suppose that we already observed the responses $Y_k = y_k$ of the stated subject to the items $k = 1, \ldots, d$, $k \neq i$. The previous notes on the statistic $\boldsymbol{T}$ lead to the following *rest-score estimator*[6] of η_j for the prediction of Y_i

$$\tilde{\eta}_j^{(-i)} = c_j + \sum_{r=1}^{q} \psi_{jr} \left(\sum_{k=1}^{i-1} \alpha_{kj} Y_i + \sum_{k=i+1}^{d} \alpha_{kj} Y_i \right), \qquad j = 1, \ldots, q,$$

for some parameters $c_1, \ldots, c_q$, $\psi_{jr} \in \mathbb{R}$, $j, r = 1, \ldots, q$. The parameters ψ_{jr} are weighting parameters and reflect the strength of the relation between the factors η_j and η_r. Using these estimators, we get as estimate of $\mathsf{P}(Y_i = 1 \mid \boldsymbol{\eta} = \boldsymbol{\eta}_0)$ with $\tilde{\boldsymbol{\eta}}^{(-i)} = (\tilde{\eta}_1^{(-i)}, \ldots, \tilde{\eta}_q^{(-i)})^\intercal$

$$\begin{aligned}
\widehat{P_i(\boldsymbol{\eta}_0)} :=& P_i\left(\tilde{\boldsymbol{\eta}}^{(-i)}\right) \\
=& \frac{\exp\left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} c_j + \sum_{j=1}^{q} \alpha_{ij} \sum_{r=1}^{q} \psi_{jr} \left(\sum_{k=1}^{i-1} \alpha_{kr} Y_k + \sum_{k=i+1}^{d} \alpha_{kr} Y_k \right)\right)}{1 + \exp\left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} c_j + \sum_{j=1}^{q} \alpha_{ij} \sum_{r=1}^{q} \psi_{jr} \left(\sum_{k=1}^{i-1} \alpha_{kr} Y_k + \sum_{k=i+1}^{d} \alpha_{kr} Y_k \right)\right)} \\
=& \frac{\exp\left(b_i + \sum_{k=1}^{d} \gamma_{ik} Y_k\right)}{1 + \exp\left(b_i + \sum_{k=1}^{d} \gamma_{ik} Y_k\right)},
\end{aligned} \tag{1.28}$$

with

$$\begin{aligned}
b_i :=& \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} c_j, \\
\gamma_{ik} :=& \begin{cases} \sum_{j=1}^{q} \sum_{r=1}^{q} \alpha_{ij} \psi_{jr} \alpha_{kr}, & k = 1, \ldots, i-1, i+1, \ldots, d, \\ 0 & k = i, \end{cases}
\end{aligned} \tag{1.29}$$

for any $i = 1, \ldots, d$. Alternatively to a prediction of Y_i, one might want to estimate $\mathsf{P}(Y_i = 1 \mid \boldsymbol{\eta} = \boldsymbol{\eta}_0)$ directly after observing the complete data set. As Anderson et al. (2007) argued by referring to classical test theory of psychometric and non-parametric IRT, it is still preferable to use the rest-score instead of the total test-score. We want to add a further remark: As we see in equation (1.28), we get a regression model with Y_i as response and Y_k, $k \neq i$ as explanatory variables. Generally, it is highly questionable to use the response variable itself as an explanatory variable.

As a result from this heuristical discussion, we get an estimate for the following conditional probability:

$$\mathsf{P}(Y_i = 1 \mid Y_k,\, k = 1, \ldots, i-1, i+1, \ldots, d).$$

Moreover, as the discussion does not depend on a specific i, we can repeat it for all items.

[6] The statistics $T_1, \ldots, T_d$ are weighted sums of all responses, weighted by their discrimination value regarding the latent variable η_j. Bearing the impact of these statistics in IRT in mind (cf. also the *components,* Bartholomew et al., 2011, Section 2.5), the name *(total) test-score* seams reasonable. Slightly adapted, Anderson et al. (2007) called the statistics $\tilde{T}_{ij} := \sum_{k=1}^{i-1} \alpha_{kj} Y_k + \sum_{k=i+1}^{d} \alpha_{kj} Y_k$, $i = 1, \ldots, d$, $j = 1, \ldots, q$, *rest-scores*, as they are the remaining test-scores after leaving a single item out.

The key is now to define the following set of conditional probabilities for the possibly depending Bernoulli random variables $Y_i \sim \mathscr{B}(1, \pi_i)$, $\pi_i \in (0,1)$, $i = 1, \ldots, d$:

$$\text{logit}\left(\mathsf{P}(Y_i = 1 \mid Y_k,\, k = 1, \ldots, i-1, i+1, \ldots, d)\right) := b_i + \sum_{k=1}^{d} \gamma_{ik} Y_k, \qquad i = 1, \ldots, d, \tag{1.30}$$

with $b_i, \gamma_{ik} \in \mathbb{R}$, $i, k = 1, \ldots, d$, while $\gamma_{ii} = 0$, $i = 1, \ldots, d$. Liu (1994) proved (cf. also Joe and Liu, 1996) that these conditional probabilities are jointly compatible if and only if $\gamma_{ik} = \gamma_{ki}$ for all $i, k = 1, \ldots, d$ and that they uniquely imply the following joint distribution of the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\intercal$:

$$P(\boldsymbol{y}) := \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}) = \exp\left(a + \sum_{i=1}^{d} b_i y_i + \sum_{i=1}^{d} \sum_{k=i+1}^{d} y_i \gamma_{ik} y_k\right), \qquad \boldsymbol{y} \in \mathbb{Y}, \tag{1.31}$$

for a normalization constant $a \in \mathbb{R}$. The condition $\gamma_{ik} = \gamma_{ki}$ is consistent to the interpretation of the underlying weighing parameters ψ_{jr}, which reflect the strength of relation between η_j and η_r – that makes only sense if $\psi_{jr} = \psi_{rj}$, $r, j = 1, \ldots, q$. The distribution (1.31) is exactly the manifest probabilities model. With model parameters given in (1.29) as they are motivated using the rest-score estimator, the proof of Liu (1994) implies that the set of conditional probabilities (1.30) are compatible if and only if $\psi_{ir} = \psi_{rj}$, $r, j = 1, \ldots, q$. As discussed by Anderson et al. (2007), if the distribution (1.31) is derived from the Dutch Identity or using graphical models, we get $(\psi_{jr})_{j,r=1}^{q,q} = \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0)$ which is necessarily symmetric. In contrast to $\mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0)$, $(\psi_{jr})_{j,r=1}^{q,q}$ is allowed to be negative definite to define a proper joint distribution. But in this case we lose the interpretation as an IRT model with posterior normal distributed latent vector.

Interestingly, if the joint distribution of a binary response vector $\boldsymbol{Y}$ is given by (1.31), then (1.30) is always true. Thus, it is impossible to find a set of probabilities for $Y_i = 1$ given Y_k, $k \neq i$ different from (1.30), such that they imply the joint distribution (1.31). Moreover, if we suppose a logit model for $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0$ is normally distributed for any $\boldsymbol{y}_0 \in \mathbb{Y}$ (that means the conditions (A1) and (A2) of Section 1.2.1 are true), then the conditional probability of $Y_i = 1$ given Y_k, $k = 1, \ldots, i-1, i+1, \ldots, d$, also satisfies logit models for all $i = 1, \ldots, d$.

To see this, we denote a response pattern without observation i by

$$\boldsymbol{Y}^{(-i)} := (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_d)$$

, $i = 1, \ldots, d$. Further, we define the index set $I^{(-i)|k} := \{\ell \in \mathbb{N} \,|\, k < \ell < d, \ell \neq i\}$, of indexes larger then k without i. With this, we get the decomposition

$$\sum_{i=1}^{d} \sum_{k=i+1}^{d} y_i \gamma_{ik} y_k = \sum_{k=1}^{d} y_v \gamma_{vk} y_k + \sum_{k \neq v} \sum_{\ell \in I^{(-v)|k}} y_k \gamma_{k\ell} y_\ell,$$

for each $v \in \{1, \ldots, d\}$, since $(\gamma_{ik})_{i,k=1}^{d,d}$ is symmetric with the diagonal equal to zero. Therefore, the probability (1.31) can also be written in the form

$$P(\boldsymbol{y}) = \exp\left(b_i y_i + \sum_{k=1}^{d} y_i \gamma_{ik} y_k\right) \exp\left(a + \sum_{k \neq i} b_i y_i + \sum_{k \neq i} \sum_{\ell \in I^{(-i)|k}} y_k \gamma_{k\ell} y_\ell\right). \tag{1.32}$$

Of course $P(\boldsymbol{Y}=\boldsymbol{y}) = P(Y_i = y_i\,,\,\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)})$ for each $i = 1,\ldots,d$. By definition, the searched conditional probability is given by

$$\begin{aligned}\mathsf{P}(Y_i = y_i \,|\, \boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)}) =& \frac{\mathsf{P}(Y_i = y_i\,,\,\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)})}{\mathsf{P}(\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)})} \\ =& \frac{\mathsf{P}(Y_i = y_i\,,\,\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)})}{\mathsf{P}(Y_i = 0\,,\,\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)}) + \mathsf{P}(Y_i = 1\,,\,\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)})}.\end{aligned}$$

Using (1.32) with

$$\begin{aligned}\kappa^{(-i)} :=& \exp\left(a + \sum_{k\neq i} b_i y_i + \sum_{k\neq i}\sum_{\ell\in I^{(-i)|k}} y_k \gamma_{k\ell} y_\ell\right) \\ \zeta^{(i)}(x) :=& \exp\left(b_i x + \sum_{k=1}^{d} x\gamma_{ik} y_k\right), \qquad x \in \{0,1\},\end{aligned}$$

we receive

$$\begin{aligned}\mathsf{P}\left(Y_i = y_i \,|\, \boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)}\right) &= \frac{\kappa^{(-i)}\zeta^{(i)}(y_i)}{\kappa^{(-i)}\left(\zeta^{(i)}(0) + \zeta^{(i)}(1)\right)} \\ &= \frac{\exp\left(b_i y_i + \sum_{k=1}^{d} y_i\gamma_{ik}y_k\right)}{1 + \exp\left(b_i + \sum_{k=1}^{d}\gamma_{ik}y_k\right)},\end{aligned} \tag{1.33}$$

since $\exp\left(b_i x + \sum_{k=1}^{d} x\gamma_{ik}y_k\right) = 1$ if $x = 0$. This is what we had to show.

1.2.4 The Distribution of the Latent Vector

The two very important assumptions for the derivation of the manifest probabilities model are the conditions (A1) and (A2) of Section 1.2.1 (cf. p. 19). Assumption (A2) enforces a logit model for the binary responses and is a modeling assumption about the data, which can be true or false and should be analyzed. But (A1) is an assumption about the distribution of $\boldsymbol{\eta}$ *after* the data acquisition. This is highly questionable practice. Hence, we have to discuss this assumption.

As first discussed by Li (2010) and later by Epskamp et al. (2018), from the total probability theorem it follows that (A1) holds if the Lebesgue density $\mathfrak{h}$ of $\boldsymbol{\eta}$ is given by

$$\mathfrak{h}(\boldsymbol{\eta}) = \sum_{\boldsymbol{y}\in\mathbb{Y}} \phi_q(\boldsymbol{\eta}; \boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{\Sigma}) P(\boldsymbol{y}), \qquad \boldsymbol{\eta}\in\mathbb{R}^q, \tag{1.34}$$

with $\boldsymbol{\mu}_{\boldsymbol{y}} \in \mathbb{R}^q$ and $\mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}_0) = \phi_q(\boldsymbol{\eta}; \boldsymbol{\mu}_{\boldsymbol{y}_0}, \boldsymbol{\Sigma})$, for $\boldsymbol{y}$, $\boldsymbol{y}_0 \in \mathbb{Y}$, i.e. a mixture of covariance homogenous normal distributions where the mixing distribution is the marginal of $\boldsymbol{Y}$. As both stated assumptions imply the posterior normality for all response patterns with homogeneous covariances (cf. Proposition 1.2.3) and the manifest probabilities model, the latent vector $\boldsymbol{\eta}$ must have the Lebesgue density (1.34), where $P(\boldsymbol{y})$ is given by (1.21) and all $\boldsymbol{\mu}_{\boldsymbol{y}}$ by Proposition 1.2.3. Since the resulting prior distribution is always proper, this theoretically is a valid choice for a prior density and can be assumed (cf. Li, 2010, for further discussion).

Unfortunately, the questions regarding this assumption are not answered exhaustively. Besides the possible desire for modeling flexibility regarding the distribution of $\boldsymbol{\eta}$, the most

important prior distribution is the multivariate (standard) normal distribution as noted in Section 1.1.3. A strong argument for the normal prior is its theoretical and empirical foundation in many psychological applications. As an even more important problem in considering the prior density (1.34), we can see that the density $\mathfrak{h}$ depends on the marginal distribution of $\boldsymbol{Y}$. While this is theoretically always the case in the sense of the Bayes theorem, (1.34) can only be specified if P is known. Especially, it depends on the number of items d. Thus, assuming the density (1.34) independently of the data as a valid prior density is difficult and could in principal only be a random result from specifying some arbitrary prior. Of course, this is practically impossible. So, prior (1.34) can only be justified in an empirical Bayesian sense.[7]

A possibility to still consider a normal prior or to leave the prior unspecified is given by an asymptotic reasoning of the posterior normality for a large number of items. A first asymptotic justification in the IRT literature was given by Chang and Stout (1991, 1993), when they showed one kind of asymptotic posterior normality for a broad class of IRT models with a single latent variable. As far as we know, up to now there is no generalization to multiple latent variables (even if Chang and Stout (1993) stated its possibility). In Chapter 3, we will discuss the asymptotics of the posterior distribution of $\boldsymbol{\eta}$ for $d \to \infty$. In particular, we will first give a generalization of the result by Chang and Stout (1993) in Section 3.4. Based on general results regarding the posterior asymptotics from the Bayesian literature (so called *Bernstein–von Mises theorems*), we will discuss a stronger form of asymptotic posterior normality in Section 3.5.

The idea of an asymptotic justification of (A1) and (A2) is a part of *Holland's Dutch Identity Conjecture* (cf. Holland, 1990, Section 3.2 and 3.3). A setback in regards to the goal of proving the conjecture was given by Zhang and Stout (1997), who indicated that the results of Chang and Stout (1993) are not sufficient to prove the asymptotic equivalence of logit models and manifest probability models even for a single latent variable in the presence of more or less arbitrary structural models. However, we want to remark that it is not obvious in which sense exactly these models can be equivalent. The theory of Section 3.5 also implies an asymptotic justification of the heuristically derived conditionally defined models approach of the last subsection (cf. Subsection 3.5.3). This is a further foundation to tackle the Dutch Identity Conjecture in future research.

1.2.5 Interpretation as an Association Model

The d items are binary and can therefore be represented as a 2^d contingency table. This interpretation opens the toolbox of classical categorial data analysis techniques, in particular log-linear and association models for binary and ordinal classification variables (cf. for example Kateri, 2014, Chapter 5 and 6). A main area of application of association models is given in modeling the ordinal structure of variables and the interaction between them. The mainly interesting quantities of association models are the distances between the scores of successive categories. Since a binary variable only has one distance, it offers no benefit to model a 2^d contingency table by association models. Moreover, association models coincide with classical log-linear and log-multiplicative models for binary data. The connection between item response models and association models is still useful and interesting.

[7] As Li (2010, p. 29–33) analyzed via simulations, this mixture of normal distributions is flexible and can fit various distributions quite well. If some fixed prior like $\mathcal{N}(\mathbf{0}, \mathrm{I}_q)$ is wrong, prior (1.34) therefore yields a more precise model.

First, consider all main effects and all two-factor interactions and get:

$$\log(P(\boldsymbol{y})) = \lambda + \lambda_{j_1}^{Y_1} + \cdots + \lambda_{j_d}^{Y_d} + \lambda_{j_1 j_2}^{Y_1 Y_2} + \cdots + \lambda_{j_1 j_d}^{Y_1 Y_d} + \cdots + \lambda_{j_{d-1} j_d}^{Y_{d-1} Y_d}, \tag{1.35}$$

where $j_i \in \{1, 2\}$ for all i to indicate the stages (in our case $Y_i = 1$ or $Y_i = 0$). Next, we consider some identification constraints. As in the classical item response models, we treat each incorrect response as *zero point* and set $\lambda_1^{Y_i} = 0$, for all i. Analogously we deal with the interaction terms, if one of the items has an incorrect response: $\lambda_{1 j_k}^{Y_i Y_k} = \lambda_{j_k 1}^{Y_i Y_k} = 0$, for all $i, k = 1, \ldots, d$, with $i < k$. Consequently the remaining parameters are

$$\lambda_2^{Y_i},\ i = 1, \ldots, d \quad \text{and} \quad \lambda_{22}^{Y_i Y_k},\ i, k = 1, \ldots, d.$$

By renaming $\lambda = a$, $\lambda_2^{Y_i} = b_i$ and $\lambda_{22}^{Y_i Y_k} = \gamma_{ik}$, model (1.35) becomes the manifest probabilities model (1.26). Therefore the parameters b_i can interpreted as main and γ_{ik} as two-factor effects. Additionally γ_{ik} is a sum of products of α_{ij}, α_{kr} and σ_{jr} and represents the influence of latent variable η_j on response variable Y_i. If Y_i does not depend directly on η_j, $\alpha_{ij} = 0$ should hold. Notice that η_j and η_r may be dependent and η_j has indirect influence on Y_i via η_r. In our model this will be ensured through the sum in (1.27). Model (1.35) is a classical hierarchical log-linear model of no three-factor interactions, i.e. homogeneous pairwise associations.

Recall that setting structural restrictions, model (1.26) and thus (1.35) can be associated with classical item response models. For example, the Rasch model ($\gamma_{ik} = c$) corresponds to a uniform association model and the Birnbaum-Rasch model holds for $\gamma_{ik} = c_i c_k$, which corresponds to a generalizationg of RC(1) models to d-way tables. Notice that the Birnbaum-Rasch model is a special association model and log-multiplicative and therefore not captured by classical log-linear models. Compare also the work of Anderson and Vermunt (2000) for the relation of the manifest probabilities model with association models.

1.2.6 Connection to the Ising Model

The so called Ising model in statistical mechanics that describes ferromagnetism in crystals. The model consists of discrete variables with two possible states (-1 and 1) representing the state of atomic spins. These variables are arranged in graphs. In the research area of *network psychometrics*, models such as the Ising model are used for psychometric modeling. As discussed by Epskamp et al. (2018) and Marsman et al. (2018), the Ising model can be interpreted as an IRT model and is identical to the manifest probabilities model we discussed earlier in this section. Differences primarily occur in the interpretation and discussion of assumptions. As noted in the previous section, we generally suppose a fixed prior distribution $\mathcal{H}$ and a resulting posterior distribution $\mathcal{H}(\bullet \mid \boldsymbol{Y})$, which is approximately a normal distribution for a large number of items d. In this context, as it is related to statistical inference for $\boldsymbol{\eta}$ with increasing information with an increasing number of items, we suppose that $\|\mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y})\| \xrightarrow{\mathsf{P}} 0$ for $d \to \infty$ and for an arbitrary norm. In contrast, in network psychometrics, $\mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y})$ is assumed to stay constant for a larger number of items (cf. Marsman et al., 2018, p. 17). This is related to the notion of *thermodynamic limits*, where we consider an increasing number of particles and, formally, the limit of a network (cf. also Marsman et al., 2018, p. 17). In particular, as noted by Marsman et al. (2018, p. 20), if we interpret the complete framework of the Ising model as a Bayesian estimation procedure, we get a sequence of prior distributions for $d \to \infty$, which become increasingly more informative. So overall, the relation to the Ising model is interesting to note but not relevant for the statistical setting we are discussing.

1.3 Parameter Estimation for Item Response Models

In any practical situation, after the data acquisition, one has to fit a model for further analysis. As in any field in statistics, there are several model fitting approaches. This is true especially for the parametric models we primarily consider here. The most common approaches are likelihood based methods and Bayesian estimation techniques.

Overall, we suppose that we observe the responses of $n \in \mathbb{N}$ subjects to a test of $d \in \mathbb{N}$ items with test dimension $q \in \mathbb{N}$, $q < d$. Thus, we observe stochastically independent realizations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ with values in $\mathbb{Y} = \{0,1\}^d$ and $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{dk})^\intercal$ and $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{dk})^\intercal$ for $k = 1, \ldots, n$. Our primary target in this section is the likelihood based estimation of the parameters in the response functions $P_1, \ldots, P_d$ of the measurement model.

We suppose in this section that a GLLVM (cf. Definition 1.1.3) holds. We collect all parameters in a vector $\boldsymbol{\vartheta}$ and denote the parameter set by Θ'. Thus, under the general GLLVM for binary data from Definition 1.1.3 $\boldsymbol{\vartheta} = (\alpha_{10}, \ldots, \alpha_{dq})^\intercal \in \Theta' = \mathbb{R}^{d(q+1)}$. Further, we set $\boldsymbol{\eta}_k = (\eta_{1k}, \ldots, \eta_{qk})^\intercal$ for $k = 1, \ldots, n$ for the values of the latent vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ of the n responding subjects. We indicate the dependence of response functions on the parameter vector as a further argument. Thus, we now write $P_i(\boldsymbol{\eta}; \boldsymbol{\vartheta})$ for $\boldsymbol{\eta} \in \Theta$ and $\boldsymbol{\vartheta} \in \Theta'$. Analogously, we write

$$\mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1 \mid \boldsymbol{\eta} = \boldsymbol{\eta}_1; \boldsymbol{\vartheta})$$

for the probability that the response vector $\boldsymbol{Y}_1 \sim \mathcal{P}(\boldsymbol{\eta}_1)$ has realization $\boldsymbol{y}_1 \in \mathbb{Y}$ conditional on the latent vector $\boldsymbol{\eta}_1$ and the parameter vector $\boldsymbol{\vartheta}$.

As mentioned in Section 1.1.3, we can distinguish between models with presence or absence of a structural model. Consequently, both approaches lead to different estimation techniques. In the absence of a structural model, we speak of a fixed effect model. First, we give an overview of estimation techniques for fixed effect IRT models. Next, we consider the counterparts for random effect models, which denote IRT models with a distributional assumption for the latent vectors as structural model. As it is one of the most important item response models, we will discuss the estimation for the Rasch model in detail. In particular, we are able to state precise results regarding the Rasch model due to its simple structure, its use of the canonical link function and the long research history regarding it. Thus, some methods are primarily described for this model. Due to the notion of psychological measurement, the estimation of a test participant's value of the latent vector plays an important role. Finally, we will give a brief overview of the estimation of latent variables.[8]

In the most common case, we are not able to give a closed form representation of point estimators. Since global optimization is quite difficult, local optimization is commonly considered (which can be combined with testing several starting values to increase the likelihood of getting a global optimum). We will not go into detail of nonlinear numerical optimization, see Nocedal and Wright (2006) for an overview.

In the setting of Bayesian estimation, the model parameters and the latent variables are jointly considered as parameters in the Bayesian sense.

Thus, equipped with an additional prior distribution for all model parameters (or a joint prior distribution for all parameters and latent variables, respectively), we make inference

[8] Bear in mind that it is not common to suppose different link functions for different items. Hence, most references can be applied under restrictions on the model formulation. However, the concepts itself are similar.

based on the joint posterior distribution of the parameters and latent variables. As the construction of conjugate priors in this setting is quite difficult (if even possible), the practical estimation can be made by applying MCMC algorithms. We will not discuss this approach here more in detail. Compare for example the book by Fox (2010).

1.3.1 Fixed Effect Approach

Latent variables traditionally were considered *person parameters*. From our modern perspective, a model considering latent variables as parameters is known as a fixed effect model. Thus, we suppose that there are unknown constant vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \in \Theta$, such that $\boldsymbol{Y}_i \sim \mathcal{P}(\boldsymbol{\eta}_i)$, $i = 1, \ldots, n$ for a measurement model with response functions $P_1, \ldots, P_d$. So, we treat the vector $\boldsymbol{\eta}_i$ as the parameter value such that $Y_{1k}, \ldots, Y_{dk}$ are conditionally independent and this parameter can theoretically be estimated directly for $k = 1, \ldots, n$. Together with the independence of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ we can formulate a likelihood like for an independent but not identically distributed model for $Y_{ik} \sim \mathscr{B}(1, P_i(\boldsymbol{\eta}_k))$, $i = 1, \ldots, d$, $k = 1, \ldots, n$.

Then, we can formulate a likelihood function as

$$\begin{aligned} \mathcal{L}_{\text{joint}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) &:= \prod_{k=1}^{n} \mathsf{P}(\boldsymbol{Y}_k = \boldsymbol{y}_k \mid \boldsymbol{\eta} = \boldsymbol{\eta}_k; \boldsymbol{\vartheta}) \\ &= \prod_{k=1}^{n} \prod_{i=1}^{d} \mathfrak{g}_i^{-1}\left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij}\eta_{jk}\right)^{y_{ik}} \\ &\quad \times \left[1 - \mathfrak{g}_i^{-1}\left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij}\eta_{jk}\right)\right]^{1-y_{ik}}, \end{aligned} \tag{1.36}$$

for $(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \in \mathbb{R}^{d(q+1)+nq}$. We obtain estimates of $\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ by maximizing $\mathcal{L}_{\text{joint}}$ over $\mathbb{R}^{d(q+1)+nq}$. As we get estimates of the measurement model parameters and the latent variables/person parameters simultaneously, this method is known as *joint maximum likelihood estimation (JMLE)*.

The number of arguments of $\mathcal{L}_{\text{joint}}$ depends on the number of items d, the number of subject parameters per subject q and the number of subjects n itself. To derive asymptotic theory, we have to analyze a sequence of tests of growing complexity. Even if it is always an artificial situation, we assume that these tests have the same testing subject and therefore we usually assume q to be constant. Also, in the common sense of asymptotic theory, we let the number of subjects increase and consider the limit as $n \to \infty$. But to decide whether we keep the test itself fixed and test only an increasing group of subjects, or if we test more subjects with an also increasing number of items, is a difficult decsion. Mathematically, we have to decide if d is constant or if we have a sequence $\{d_n\}_{n\in\mathbb{N}}$ with $d_n \to \infty$ for $n \to \infty$. For the latter, we also have to assume the increasing rate of $\{d_n\}_{n\in\mathbb{N}}$. In the literature, both approaches are discussed and the behavior is quite different. In special cases of an increasing number of items, the joint maximum likelihood method can be consistent, but in the case of a constant number of items, the method usually is not (cf. for example Agresti (2013, Subsection 13.1.5) and Haberman (1977)). However, to the best of our knowledge, there are no general (in-)consistency results. A major cause for inconsistency of JMLE is the increasing number of parameters as $n \to \infty$. A possible approach to handle the consistency problem is to constrain the maximization as a regularization approach (cf. e.g. Tutz and

Gertheiss, 2016, especially Section 5). Recently, Chen et al. (2019) discussed a *constrained JMLE* in the setting of IRT models, which is a regularized JMLE approach and leads to a computationally efficient procedure as well as similar good or even better estimates in high dimensional settings than a standard method (marginal MLE, see the next section; analyzed via a simulation study) – for $n, d \to \infty$, they also discussed consistency.

An alternative approach to obtain proper estimates in fixed effect models is the *conditional maximum likelihood estimation (CMLE)*. To obtain the conditional likelihood function, the probability of $\boldsymbol{Y}_k = \boldsymbol{y}_k$ conditional on a sufficient statistic for the person parameters $\boldsymbol{\eta}_k$ is considered for $k = 1, \dots, n$. Through the condition on a sufficient statistic, $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$ vanish from the probability expression. In consequence, if a likelihood function based on these probabilities is used, the parameters of the measurement model are the only remaining parameters and we stay in a common setting regarding asymptotic estimation theory. But since an appropriate sufficient statistic for $\boldsymbol{\eta}$ exists only if the discrimination parameters are known and we use the logit-link, the Rasch model and its generalization to multiple latent variables are the only item response models whose parameters can be estimated due to CMLE.

We will discuss the JMLE and CMLE in the special case of the Rasch model in more detail in Subsection 1.3.3.2 and 1.3.3.1, respectively.

An advantage of fixed effect approaches is the independence from a distributional assumption of the latent variables, as a further assumption is always a possible source of miss-specification. A further, clear advantage of the JMLE is the simultaneous obtainment of estimates for the measurement model and the latent variables, while the likelihood itself has a simple structure. A disadvantage of the JMLE is the difficult handling with a large number of parameters, while the CMLE is restricted to a small number of models. Generally, modern psychology is a strongly theory driven research area, which often implies a distributional assumption of the latent variables. The additional information gained from that assumption might be useful. In the following we discuss the random effect approach for estimating the measurement model parameters, which differ from the fixed effect approach by considering a structural model for the latent variables.

1.3.2 Random Effect Approach

In the random effect approach, we again suppose that there are unknown vectors $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n \in \Theta$, such that $\boldsymbol{Y}_i \sim \mathcal{P}(\boldsymbol{\eta}_i)$, $i = 1, \dots, n$ for a measurement model with response functions $P_1, \dots, P_d$. But this time, we further suppose that $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$ are realizations of independent and identically distributed random vectors with a distribution $\mathcal{H}$, which is the structural model.

We consider a response vector $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$, $\boldsymbol{\eta} \sim \mathcal{H}$ and suppose that the response functions $P_1, \dots, P_d$ depend on a parameter vector $\boldsymbol{\vartheta}$. Further, let us suppose that $\mathcal{H}$ is absolutely continuous with respect to the counting or Lebesgue measure, such that $\mathfrak{h}$ denotes the Radon–Nikodym density of $\mathcal{H}$. Then the joint density f of $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ is given by

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\eta}; \boldsymbol{\alpha}) &= P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\eta}; \boldsymbol{\vartheta}) \mathfrak{h}(\boldsymbol{\eta}) \\
&= \left(\prod_{i=1}^{d} \mathfrak{g}^{-1} \left(\alpha_{i0} + \sum_{k=1}^{q} \alpha_{ik} \eta_k \right)^{y_i} \left[1 - \mathfrak{g}^{-1} \left(\alpha_{i0} + \sum_{k=1}^{q} \alpha_{ik} \eta_k \right) \right]^{1 - y_i} \right) \mathfrak{h}(\boldsymbol{\eta}).
\end{aligned}
$$

This can directly be used to define an analogous form of the JMLE by maximizing

$$\prod_{k=1}^{n} f(\boldsymbol{y}_k, \boldsymbol{\eta}_k; \boldsymbol{\vartheta}) = \prod_{k=1}^{n} \left(\prod_{i=1}^{d} \mathfrak{g}_i^{-1} \left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} \eta_{jk} \right)^{y_{ik}} \times \left[1 - \mathfrak{g}_i^{-1} \left(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} \eta_{jk} \right) \right]^{1-y_{ik}} \right) \mathfrak{h}(\boldsymbol{\eta}_k)$$

in $(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$ over $\Theta' \times \Theta^n$. The prior density now performs as a regularization because it shrinks the estimates of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ toward modal points of $\mathfrak{h}$. This can help with the problems which occur in JMLE. However, there is a popular and successful alternative random effect approach, which can be obtained intuitively from the mixed model approach. The marginal distribution of $\boldsymbol{Y}_1$ is given by

$$P(\boldsymbol{y}; \boldsymbol{\vartheta}) := \mathsf{P}^{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\vartheta}) := \mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}; \boldsymbol{\vartheta}) = \int_{\Theta} P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\eta}; \boldsymbol{\vartheta}) \, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}), \qquad \boldsymbol{y} \in \mathbb{Y}.$$

Hence we observe realizations of an iid sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{\text{iid}}{\sim} \mathsf{P}^{\boldsymbol{Y}}(\,\bullet\,; \boldsymbol{\vartheta})$ with the probability mass function $P(\,\bullet\,; \boldsymbol{\vartheta})$. Hence, the likelihood is given by

$$\begin{aligned} \mathcal{L}_{\text{marginal}}(\boldsymbol{\vartheta}) :=& \mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n; \boldsymbol{\vartheta}) \\ &= \prod_{k=1}^{n} P(\boldsymbol{y}_k; \boldsymbol{\vartheta}) = \prod_{k=1}^{n} \int_{\Theta} \mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}) \, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}), \qquad \boldsymbol{\vartheta} \in \Theta'. \end{aligned} \tag{1.37}$$

The corresponding estimation is called *marginal maximum likelihood estimation (MMLE)*, as we only consider the marginal distribution of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ if considering the latent variables as an unobserved part of a hypothetical full data set. In principal, maximizing $\mathcal{L}_{\text{marginal}}$ and analyzing the properties of the maximal points is a subject of classical statistics with a given likelihood function. However, this proves to be difficult in theory and in practice. Butler and Louis (1997) proved the consistency of MMLE for a general class of random effect models for binary data, which include item response models for various prior distributions. However, their approach to prove the consistency is remarkably complex. As we will see when discussing MMLE for the Rasch model, the asymptotic can also be analyzed by more common approaches when considering restricted settings. But even if the estimates are commonly assumed to be asymptotically normal, to the best of our knowledge there is no proof of a corresponding general result in the literature. Due to this, we will discuss an existence, consistency and asymptotic normality result for a relatively general class of IRT models in Chapter 2.

The concept of MMLE for item response models was first discussed by Bock and Lieberman (1970), who focused on the Lawley–Lord model. But in contrast to (1.37), they formulated the likelihood for the representation as a 2^d contingency table with multinomial sampling scheme. Further, they used a Gauss–Hermite quadrature for the integral approximation and the Newton-Raphson algorithm for optimization. Unfortunately, this approach becomes unfeasible for growing d and critical values of d are reached relatively quickly. Bock and Aitkin (1981) presented the first feasible approach for MMLE by applying the *expectation-maximization (EM)*-algorithm and using a Gauss–Hermite quadrature (i.e. treating the latent vectors as *missing data*). A nice property of the EM-algorithm for

many IRT models is that the measurement model parameters can be updated item-wise and separately from the structural model parameters during the M-step. Commonly, the EM-algorithm is stable but converges relatively slowly. To overcome this, they suggested switching to Newton-Raphson when one reaches a neighborhood of the optimal point. Until then, this approach is established for MMLE of IRT models and provides a basis for the implemented version in freely available software packages (e.g. the R-package **ltm** of Rizopoulos (2006) uses a hybrid EM-BFGS-algorithm for a latent vector of up to dimension 2).

As mentioned by Bock and Aitkin (1981), this approach can be extended to arbitrary latent dimensions from a conceptual point of view. However, as Huber et al. (2004) discussed, applying a Gauss–Hermite quadrature to more than two latent variables can be difficult and problematic. A brief formulation of the reason for this problem is the *curse of dimension*: The required quadrature points for the same approximation error increase exponentially with the dimension of the integral, i.e. the latent dimension. As an alternative, they suggest the use of the Laplace approximation. A disadvantage of using the Laplace approximation is that the approximation quality depends on the number of items, which cannot be increased easily during the model fitting. A principally good approach in numerical analysis to ensure a specific quality of the integral approximation is *adaptive quadrature*, which summarizes approaches of local refinement of the approximation grid. However, as Lyness (1983) discussed, adaptive quadrature is incompatible with iterative procedures. In particular, even if the actual integral is quite smooth, the approximation may even be discontinuous. This leads to the conclusion that a fixed number of quadrature points is necessary for good application of numerical quadrature methods in iterative procedures. A possible approach suggested by Lyness (1983) was to first optimize with respect to a relatively low number of quadrature points, then refine the grid and restart the optimization procedure. If the new optimal point is in the neighborhood of the old one, break. Else, make a further refinement.

Alternative approaches to increase the approximation quality without increasing the number of quadrature points were presented by Rabe-Hesketh et al. (2002) and Schilling and Bock (2005). These approaches are based on the assumption that $\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}$ is approximately normal[9] and combine ideas of Laplace approximation and numerical quadrature. In particular, the quadrature points of Gauss–Hermite are transformed by efficiently applying the approximate posterior normality.[10] A disadvantage of this approach is that in every iteration, a new transformation has to be computed, which leads to increased computational effort. For example, Huber et al. (2004) needed about 20 hours to fit a model containing ten items, two latent variables and 60 observed samples, which is infeasibly long. However, the implemented approach used finite differences for approximating the gradients, which can be an explanation for the needed time and the bad performance is possibly an artefact. In the last years, further approaches were published which focus on an acceleration of this approach (cf. e.g. Rizopoulos, 2012) or combine it with other approaches like the Laplace approximation (cf. e.g. Bianconcini et al., 2017).

An alternative which completely avoids the curse of dimension is given by *Monte Carlo*

[9] As we will discuss in Chapter 3, this is true for common IRT models for sufficiently many items.

[10] The authors also call the approach *adaptive Gauss–Hermite quadrature*. But we want to remark that this approach is completely different from that of numerical analysis. Therefore bear in mind that the adaptive quadrature in the literature regarding the fitting of IRT models means the adaptive transformation of a fixed number of quadrature points and not the local refinement of the adaptive quadrature in the numerical analysis literature.

(MC) integration. One of the first applications of MC integration in the EM-algorithm (MCEM-algorithm) for fitting of latent variable models was published by Sammel et al. (1997). They used the straight forward approach, where values of the latent vector are sampled from the structural model, and reported the low speed of the method. The latter can be explained by the selection of a sample size of 10 000. A strong decrease in the needed sampling size can be obtained using variance reduction techniques. In particular, one sample generated in the first iteration can be reused in all following iterations, which makes the approach faster. One such approach was published by Schilling and Bock (2005), who adapted their approach for transforming the Gauss–Hermite quadrature points to construct an importance sampling distribution. In their simulation study, they achieved a comparable quality to Gauss–Hermite quadrature using transformed quadrature points. So, for a high latent dimension, this MC approach is advantageous because the sampling size can be selected appropriately and it is possible to increase the number slightly if it seems appropriately (e.g. by using the approach of Lyness (1983)). After analyzing the effect of a combination with further variance reduction approaches, for example antithetic variables, the efficiency of the approach can possibly be increased further. Other successful approaches use Gibbs sampling for the MC integration (cf. e.g. Schilling and Bock, 2005, Section 2.3) or a stochastic Gauss–Hermite (cf. e.g. Chen, 2006). In Section 4.3, we will discuss a further MCEM approach for tests having high d and q and without the need of an explicit assumption regarding $\mathcal{H}$. This new algorithm is moreover easier to implement in comparison to the MCEM-algorithm discussed by Schilling and Bock (2005).

One of the main motivations behind the derivation of the manifest probabilities model in Section 1.2 was to gain more insight into the structure lying behind the distribution of $\boldsymbol{Y}$ if a logit model instead of a general GLLVM is supposed. However, this was always closely related to get a closed form expression of the marginal distribution. Of course, this can be used to define a likelihood function for estimation, as first described by Holland (1990) for the manifest probabilities model and later discussed by Anderson and co-authors (cf. for example Anderson and Vermunt (2000) and Anderson et al. (2007)).

Every response pattern $\boldsymbol{y}$ is an element of $\mathbb{Y} = \{0,1\}^d$, the set of all response patterns. We denote by $n(\boldsymbol{y})$ the observed frequency of response pattern $\boldsymbol{y}$ and by $n = \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y})$ the total number of observations. Let $\mu(\boldsymbol{\vartheta},\boldsymbol{y}) := n\,P(\boldsymbol{y};\boldsymbol{\vartheta})$ be the expected frequencies of $\boldsymbol{y}$ under the model, where $\boldsymbol{\vartheta} = (z_1,\ldots,z_d,\alpha_{11},\ldots,\alpha_{dq},\sigma_{11},\sigma_{12},\ldots,\sigma_{qq})$ is the vector of all parameters of the manifest probabilities model in its form in equation (1.26). Assuming a Poisson distribution for the count of a single $\boldsymbol{y}\in\mathbb{Y}$, the likelihood function is given by[11]

$$\mathcal{L}_{\text{manifest}}(\boldsymbol{\vartheta}) := \prod_{\boldsymbol{y}\in\mathbb{Y}} \frac{\mu(\boldsymbol{\vartheta},\boldsymbol{y})^{n(\boldsymbol{y})}}{n(\boldsymbol{y})!}\exp(-\mu(\boldsymbol{\vartheta},\boldsymbol{y})), \qquad \boldsymbol{\vartheta}\in\mathbb{R}^{d(q+1)+q(q+1)/2}. \tag{1.38}$$

Further identification constraints can eliminate the parameters $\sigma_{11},\sigma_{12},\ldots,\sigma_{qq}$ and reduce the parameter space to $\Theta' = \mathbb{R}^{d(q+1)}$. We call the estimation of $\boldsymbol{\vartheta}$ obtained by maximizing $\mathcal{L}_{\text{manifest}}$ *manifest (probabilities) maximum likelihood estimation (MPMLE)*. Principally, maximum likelihood estimates obtained by maximizing (1.38) are also marginal maximum likelihood estimates as they can be obtained by maximizing (1.37). But there are conceptual and algorithmic differences. First, in the MMLE given by maximizing (1.37), we suppose a single fixed prior distribution $\mathcal{H}$, which can be freely selected as long as it is possible

[11] As usual, the multinomial, the product binomial and the Poisson sampling scheme (conditional on n) of the corresponding contingency table can be used to formulate likelihood functions having the same kernel. We consider the Poisson sampling scheme.

to evaluate the integral. In contrast, for maximizing (1.38), we suppose a prior, which is a mixture of homogeneous normal distributions, and the mixture weights are given by $P(\bullet;\boldsymbol{\vartheta})$, as discussed in Section 1.2.4. As already discussed in Section 1.2.4, such a prior can theoretically be used for MMLE. Then it is a fixed prior. In contrast, by estimating parameters using MPMLE, the mixture weights of the prior are estimated, too. Since different mixture weights lead to different distributions, MPMLE determines the shape of the prior. Also, (1.38) is of a different form compared to (1.37), which strongly influences the implementation. In (1.38), one has to deal with numerical quadrature, while (1.38) is a closed form expression containing a product over all possible response patterns $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, not only the observed ones. Finally, using an asymptotic justification for a large number of items d, one can apply a maximization of $\mathcal{L}_{\text{manifest}}$ to get parameter estimates for a general logit model, even if not supposing the proposed prior family. So, from our point of view, it is important to distinguish between both kinds of marginal maximum likelihood estimation.

Since all 2^d response pattern have to be taken into account[12] in an iterative maximization procedure, the computational expenditure of calculating a value of $\mathcal{L}_{\text{manifest}}$ grows exponentially with increasing d. A first specialized method to deal with this problem was presented by Anderson et al. (2007) for a generalized Rasch model (or rather for the corresponding manifest probabilities model). Their procedure is also based on the manifest probabilities model and is a member of the class of composite likelihood methods (cf. e.g. Varin et al., 2011, for a general overview). Recently, Paek (2016, cf. also Paek and Anderson (2017)) presented an iterative two-step procedure for the generalization[13] of the method of Anderson et al. (2007) for a more general logit model (a confirmatory factor model, where it is a-priori known on which items latent variables have an effect). In Chapter 5, we will discuss further developments of the pseudo likelihood method.

Similar to the manifest probabilities model, other combinations of measurement and structural models can lead to a closed form marginal distribution – e.g. a loglog measurement model with an exponential-gamma structural model (cf. Conaway, 1990). The approach of Conaway (1990) is interesting for loglog models, but – analogously to MPMLE – becomes infeasible for larger d.

We will discuss the MMLE, the MPMLE and the method of Anderson et al. (2007) in the special case of the classical Rasch model in more detail in Subsection 1.3.3.3, 1.3.3.4 and 1.3.3.5, respectively.

1.3.3 Estimation for the Rasch Model

Regardless of its strong restrictions, the Rasch model is still widely used and, therefore, an important case of item response models. Due to its simplicity, a wider theory of its model fitting has been developed in the literature. This provides a deeper understanding of the structure behind response models and interesting results regarding methods for estimation. In this section, we will discuss some methods for the Rasch model in detail.

Additionally, the Rasch model is the only item response model for which minimal sufficient statistics exist (cf. Andersen, 1977), thus providing the applicability of the conditional maximum likelihood estimation method. Furthermore, new fitting approaches are first de-

[12] One might wonder if this can avoided simply by using the expression of the manifest probabilities model in the marginal maximum likelihood function. This is not a solution as all response pattern still have to be taken into account for the computation of the normalization constant.

[13] The idea of this algorithm was previously sketched by Anderson and Yu (2007) for the 2PL model, but not discussed in detail.

veloped for the Rasch model or its multidimensional or polytomous extensions (cf. e.g. Anderson et al., 2007). In this section, we take a look at the different estimation methods for the Rasch model and then compare them via a small simulation study. We consider a generic response vector $\boldsymbol{Y} \sim \mathcal{P}(\eta)$ for a test with $d \in \mathbb{N}$ items and dimension $q = 1$. Further, we consider the Rasch model in the formulation

$$\text{logit}(\mathsf{P}(Y_i = 1 \mid \eta)) = \eta - \beta_i, \qquad i = 1, \ldots, d, \tag{1.39}$$

for model parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$. This implies that $\mathsf{Var}(\eta) = \sigma^2 > 0$ is a parameter and not fixed to one as identification constraint in the random effect case (notice that in the literature exists also the notion of considering the Rasch model in formulation (1.39) under the constraint $\sigma^2 = 1$, cf. e.g. Rizopoulos, 2006). To simplify the notation, in this section we denote by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ the vector of all latent variables for the n responding subjects. The n observed response vectors are still denoted by $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. Notice that model (1.39) still holds if we add any arbitrary constant to η and β_i. In random effect models, we get identifiable parameters by anchoring the latent variables with the expected value of $\mathcal{H}$. In fixed effect models, we need an additional identification constraint, for example $\sum_{i=1}^{d} \beta_i = 0$ or $\beta_1 = 0$.

1.3.3.1 Conditional Maximum Likelihood Estimation

Since minimal sufficient statistics for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ exist for the Rasch model, a conditional likelihood method for the fixed effect Rasch model has been developed by Andersen (1973). The basis is a conditional probability which is independent of $\boldsymbol{\eta}$. Using these probabilities, a conditional likelihood function only depending on the parameters $\boldsymbol{\beta}$ can be formulated.

First we define the number of correct responses

$$|\boldsymbol{y}| := \sum_{i=1}^{d} y_i, \tag{1.40}$$

with $\boldsymbol{y} \in \mathbb{Y}$. We take a look at the probability for the event $\boldsymbol{Y} = \boldsymbol{y}$ with number of correct responses $|\boldsymbol{Y}| = t$, with $\boldsymbol{y} \in \mathbb{Y}$ and $t \in \mathbb{N}$, for a given η. Denote by $\mathbb{1}$ the indicator function, then we get for any $\boldsymbol{y} \in \mathbb{Y}$ and $t \in \mathbb{N}$

$$\begin{aligned}
\mathsf{P}\left(\boldsymbol{Y} = \boldsymbol{y}, |\boldsymbol{Y}| = t \,|\, \eta\right) = \mathsf{P}\left(\boldsymbol{Y} = \boldsymbol{y}, |\boldsymbol{y}| = t \,|\, \eta\right) &= \mathbb{1}(|\boldsymbol{y}| = t) \prod_{i=1}^{d} \frac{\exp(y_i(\eta - \beta_i)))}{1 + \exp(\eta - \beta_i)} \\
&= \mathbb{1}(|\boldsymbol{y}| = t) \frac{\exp(t\eta)}{\prod_{i=1}^{d} 1 + \exp(\eta - \beta_i)} \exp\left(\sum_{i=1}^{d} -y_i\beta_i\right).
\end{aligned}$$

Hence we get for the conditional probability of $\boldsymbol{Y} = \boldsymbol{y}$ given $|\boldsymbol{Y}| = t$ and η

$$\begin{aligned}
\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid |\boldsymbol{Y}| = t, \eta) &= \frac{\mathsf{P}\left(\boldsymbol{Y} = \boldsymbol{y}, |\boldsymbol{Y}| = t | \, \eta\right)}{\mathsf{P}\left(|\boldsymbol{Y}| = t | \, \eta\right)} \\
&= \mathbb{1}(|\boldsymbol{y}| = t) \frac{\exp(\sum_{i=1}^{d} -y_i\beta_i)}{\sum_{\boldsymbol{y}' \in \mathbb{Y} \,:\, |\boldsymbol{y}'| = t} \exp(\sum_{i=1}^{d} -y_i'\beta_i)},
\end{aligned}$$

which is independent of η as we expect since $|\boldsymbol{Y}|$ is a sufficient statistic for η. With this conditional probability, we can define a likelihood function for the *conditional maximum*

likelihood estimation(CMLE) for the Rasch model by

$$\begin{aligned}\mathcal{L}_{\text{cond}}(\boldsymbol{\beta}) := \prod_{j=1}^{n} \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_j \mid |\boldsymbol{Y}| = |\boldsymbol{y}_j|) &= \prod_{j=1}^{n} \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_j \mid |\boldsymbol{Y}| = |\boldsymbol{y}_j|, \eta_j) \\ &= \prod_{j=1}^{n} \frac{\exp(\sum_{i=1}^{d} -y_{ij}\beta_i)}{\sum_{\boldsymbol{y}' \in \mathbb{Y} \,:\, |\boldsymbol{y}'| = |\boldsymbol{y}_j|} \exp(\sum_{i=1}^{d} -y'_i\beta_i)} \\ &= \frac{\exp(\sum_{j=1}^{n}\sum_{i=1}^{d} -y_{ij}\beta_i)}{\prod_{j=1}^{n} \sum_{\boldsymbol{y}' \in \mathbb{Y} \,:\, |\boldsymbol{y}'| = |\boldsymbol{y}_j|} \exp(\sum_{i=1}^{d} -y'_i\beta_i)}\end{aligned}$$

and hence the log-likelihood is

$$\ell_{\text{cond}}(\boldsymbol{\beta}) = \left(\sum_{j=1}^{n}\sum_{i=1}^{d} -y_{ij}\beta_i\right) - \sum_{j=1}^{n} \log\left(\sum_{\boldsymbol{y}' \in \mathbb{Y} \,:\, |\boldsymbol{y}'| = t} \exp\left(\sum_{i=1}^{d} -y'_i\beta_i\right)\right).$$

As Andersen (1973) showed, the derivatives are given by

$$\frac{\partial \ell_{\text{cond}}(\boldsymbol{\beta})}{\partial \beta_k} = \left(\sum_{j=1}^{n} -y_{kj}\right) + \exp(-\beta_k) \sum_{t=1}^{d} \nu(t) \frac{\gamma_{d-1}(t-1, \boldsymbol{\beta}^{(-k)})}{\gamma_d(t, \boldsymbol{\beta})}, \qquad k = 1 \ldots, d,$$

where $\nu(t)$ is the number of observed response patterns with $|\boldsymbol{y}| = t$, $\boldsymbol{\beta}^{(-k)} := (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_d)$ and γ_k is defined for general $k \in \mathbb{N}$ by

$$\begin{aligned}\gamma_k \,&: \{0, \ldots, k\} \times \mathbb{R}^k \to \mathbb{R}, \\ &(t, \boldsymbol{\beta}) \mapsto \sum_{\boldsymbol{y}' \in \{0,1\}^k \,:\, |\boldsymbol{y}'| = t} \exp\left(-\sum_{i=1}^{k} \beta_i y'_i\right).\end{aligned}$$

Therefore, the likelihood equations are given by

$$\sum_{j=1}^{n} y_{kj} = \exp(-\beta_k) \sum_{t=1}^{d} \nu(t) \frac{\gamma_{d-1}(t-1, \boldsymbol{\beta}^{(-k)})}{\gamma_d(t, \boldsymbol{\beta})}, \qquad k = 1 \ldots, d. \tag{1.41}$$

Notice that we again obtain the sufficient statistics for the item parameters β_k, $k = 1 \ldots, d$ on the left hand side. As Andersen (1973) mentioned, no inference is possible if (nearly) all subjects respond with $|\boldsymbol{y}| = d$ or $|\boldsymbol{y}| = 0$, but he also proved the consistency under a very reasonable condition, given below.

Theorem 1.3.1 (Consistency of CMLE for the Rasch model, Andersen, 1973)
Consider a sequence $\{\boldsymbol{Y}_k\}_{k \in \mathbb{N}}$, where $\boldsymbol{Y}_k \sim \mathcal{P}(\eta_k)$ satisfies the Rasch model (1.39) *for some $\eta_k \in \mathbb{R}$. Denote by $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{1\,n}, \ldots, \hat{\beta}_{d\,n})$ the CMLE for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ based on the sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, $n \in \mathbb{N}$ and by*

$$\nu_n(t) := \sum_{j=1}^{n} \mathbb{1}(|\boldsymbol{Y}_j| = t), \qquad 0 \le t \le d,$$

the number of response patterns with $t \in \{0, \ldots, d\}$ correct responses.
If $\nu_n(t) \xrightarrow{\mathsf{P}-\text{a.s.}} \infty$ for at least one integer $1 \le t \le d-1$ as $n \to \infty$, then $\hat{\beta}_{i\,n} \xrightarrow{\mathsf{P}} \beta_i$ for $i = 1, \ldots, d$. ◇

Remark 1.3.2
If the number of response patterns, where not all items are either correctly or incorrectly responded, tend to infinity for $n \to \infty$, then the condition is fulfilled. If instead the condition is not fulfilled, then the number of response pattern with all items responded correctly or incorrectly tends to infinity, while even for $n \to \infty$ the number of other patterns stay finite. In this case, we expect difficulties for any statistical method. ◇

A disadvantage of the CMLE is that the mappings γ_k, $k \in \mathbb{N}$ need the evaluation of all 2^k possible response patterns. For example for 40 items we get $2^{40} \approx 10^{12}$ possible response patterns. Thus, the method becomes computationally unfeasible for a large number of items d.

1.3.3.2 Joint Maximum Likelihood Estimation

For the Rasch model, the likelihood function for JMLE (cf. equation (1.36)) is given by

$$\mathcal{L}_{\text{joint}}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{j=1}^{n} \prod_{i=1}^{d} \frac{\exp(y_{ij}(\eta_j - \beta_i))}{1 + \exp(\eta_j - \beta_i)}, \qquad (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathbb{R}^{d+n},$$

and therefore the log-likelihood is given by

$$\ell_{\text{joint}}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{j=1}^{n} \sum_{i=1}^{d} y_{ij}(\eta_j - \beta_i) - \log(1 + \exp(\eta_j - \beta_i)), \qquad (\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathbb{R}^{d+n}.$$

From standard calculus follows for the derivatives

$$\frac{\partial \ell_{\text{joint}}(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \beta_k} = \sum_{j=1}^{n} -y_{kj} + \frac{\exp(\eta_j - \beta_k)}{1 + \exp(\eta_j - \beta_k)}, \quad k = 1, \ldots, d,$$

$$\frac{\partial \ell_{\text{joint}}(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \eta_k} = \sum_{i=1}^{d} \quad y_{ik} - \frac{\exp(\eta_k - \beta_i)}{1 + \exp(\eta_k - \beta_i)}, \quad k = 1, \ldots, n,$$

and so the likelihood equations become

$$\sum_{j=1}^{n} y_{kj} = \sum_{j=1}^{n} \frac{\exp(\eta_j - \beta_k)}{1 + \exp(\eta_j - \beta_k)} = \sum_{j=1}^{n} P(Y_k = 1 | \eta_j), \quad k = 1, \ldots, d, \tag{1.42}$$

$$\sum_{i=1}^{d} y_{ik} = \sum_{i=1}^{d} \frac{\exp(\eta_k - \beta_i)}{1 + \exp(\eta_k - \beta_i)} = \sum_{i=1}^{d} P(Y_i = 1 | \eta_k), \quad k = 1, \ldots, n. \tag{1.43}$$

Notice that the sums on the left hand side are the minimal sufficient statistics for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ and are called the *raw item response score* and the *raw subject response score*, respectively.

A solution of these equations does not always exist. Since the sums on the right-hand side take values strictly between 0 and n (and d) for equations (1.42) and (1.42), respectively, we get infinite estimates if y_{ij} are always one or always zero for one i and all j or all i and one j. The set-ups for which the JMLEs do not exist are investigated by Haberman (1977).

Theorem 1.3.3 (Existence of JMLEs for the Rasch model, Haberman, 1977)
Consider a sample of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{Y}$ for $n \in \mathbb{N}$ response vectors on $d \in \mathbb{N}$ items. The JMLE for the Rasch model (1.39) *does* ***not*** *exist if and only if there are sets A, B, C, $D \subset \mathbb{N}$ such that*

(i) $(A \neq \emptyset \text{ and } C \neq \emptyset) \quad$ *or* $\quad (B \neq \emptyset \text{ and } D \neq \emptyset)$,

(ii) $A \cap B = C \cap D = \emptyset$,

(iii) $A \cup B = \{1, \dots d\}$,

(iv) $C \cup D = \{1, \dots n\}$,

(v) $y_{ik} = 0,\ i \in A,\ k \in C, \qquad y_{ik} = 1,\ i \in B,\ k \in D.$ ◇

A typical statement for joint maximum likelihood estimators of the parameters of item response models and in particular the Rasch model is that they are inconsistent. Before we take a look at this statement, we recall from the discussion in 1.3.1 that it is not clear which limit we have to compute. A typical choice is to fix d while n approaches infinity. In this situation one can show that the estimates are not consistent (Agresti, 2013, Section 13.1.5). Moreover, for fixed d and $n \to \infty$ the probability that the JMLE exists approaches zero (Haberman, 1977). But in reality, neither d nor n approaches infinity but we have finite sample sizes. Therefore, to make general statements about this method (particularly for large d and n), it is appropriate to analyze the limit as both approach infinity.

Theorem 1.3.4 (Consistency of JMLE for the Rasch model, Haberman, 1977)
Let $\{n_m\}_{m\in\mathbb{N}}$ be a strictly increasing sequence, $\{d_m\}_{m\in\mathbb{N}}$ be a increasing sequence with $d_m \to \infty$ as $m \to \infty$ and $n_m \geq d_m$ for each $m \in \mathbb{N}$.
Let $\{\beta_m\}_{m\in\mathbb{N}}$ and $\{\eta_m\}_{m\in\mathbb{N}}$ be bounded sequences. Consider the family of independent random variables $\{Y_{ik}\}_{(i,k)\in\mathbb{N}^2}$ with

$$Y_{ik} \sim \mathscr{B}\Big(1,\ \big(1 + \exp(-\eta_k + \beta_i)\big)^{-1}\Big) \qquad i, k \in \mathbb{N}.$$

Denote by $\hat{\beta}_{i\,m}$ the JMLE of β_i, $1 \leq i \leq d_m$, and by $\hat{\eta}_{j\,m}$ the JMLE of η_j, $1 \leq j \leq n_m$, based on the sample Y_{ik}, $1 \leq i \leq d_m$, $1 \leq k \leq n_m$, $m \in \mathbb{N}$. If $d_m^{-1} \log(n_m) \to 0$, as $m \to \infty$, then

$$\hat{\beta}_{i\,m} \xrightarrow{\mathsf{P}} \beta_i \quad \text{ and } \quad \hat{\eta}_{j\,m} \xrightarrow{\mathsf{P}} \eta_j \quad 1 \leq i \leq d_m,\ 1 \leq j \leq n_m.$$ ◇

However, for finite d, the estimator is biased. In the 1970's and 1980's correcting the bias was a research topic, compare Baker and Kim (2004, Section 5.6.2) for the rest of this paragraph. If the estimators are rescaled to mean zero, the bias for $\hat{\beta}_i$, $i = 1, \dots, d$ can be quantified to be of order $d/(d-1)$ compared to the bias of the conditional maximum likelihood estimator. This leads to a simple correction method and we define the *corrected JMLE (cJMLE)* of β_i, $i = 1, \dots, d$ by

$$\hat{\beta}_i^{(\text{cJMLE})} := \frac{d-1}{d}\hat{\beta}_i, \quad i = 1, \dots, d,$$

where $\hat{\beta}_i$ denotes JMLE of β_i with sum-to-zero constraint:

$$\sum_{i=1}^{d} \hat{\beta}_i = 0.$$

Even if we do not know the finite sample bias of the CMLE, if existing, this correction is an improvement since the CMLE is consistent for finite d (cf. Theorem 1.3.1). In general, the

relative bias depends on the number of items and on the "distribution" of $\boldsymbol{\beta}$. In particular, for $d \leq 10$, an asymmetric distribution of $\boldsymbol{\beta}$ can have a higher influence on the bias than the number of items and, therefore, the correction factor could be inappropriate. But empirical research showed that for $d \geq 20$ the correction factor leads to a noteworthy bias decrease.

Additionally, a correction factor for the estimators of η_j, $j = 1, \ldots, n$, was suggested to be $(d-2)/(d-1)$, but in the literature there is no theoretical or empirical motivation. Hence, the quality of this correction of the JMLE of the latent variables is questionable.

1.3.3.3 Marginal Maximum Likelihood Estimation

For the Rasch model (1.39), the likelihood function for MMLE (cf. equation (1.37)) takes the form

$$\mathcal{L}_{\text{marginal}}(\boldsymbol{\beta}) = \prod_{j=1}^{n} \int_{\mathbb{R}} \prod_{i=1}^{d} \frac{\exp(y_{ij}(\eta - \beta_i))}{1 + \exp(\eta - \beta_i)} \, \mathcal{H}(\mathrm{d}\eta), \qquad \boldsymbol{\beta} \in \mathbb{R}^d, \tag{1.44}$$

where $\mathcal{H}$ denotes an arbitrary distribution on $\mathbb{R}$, commonly $\mathcal{H} = \mathcal{N}_1(0, \sigma^2)$ with parameter $\sigma^2 > 0$ (which is identical to the discrimination parameter in this case).

But even for the Rasch model, with the common prior assumptions like $\mathcal{H} = \mathcal{N}(0, \sigma^2)$, the integral in (1.44) cannot be evaluated analytically. Therefore the integrals have to be evaluated with numerical approximation methods and the development of efficient methods for calculating (a possibly not existing) maximum of $\mathcal{L}$ was a subject of intensive research in the last decades.

An early work considering the consistency of maximum likelihood estimators for mixing distributions was done by Kiefer and Wolfowitz (1956). As Butler and Louis (1997) discussed, the results of Kiefer and Wolfowitz (1956) are not appropriate for a general handling of MMLE in random effect models for binary data. For their general theory, Butler and Louis (1997) introduced a specialized concept of convergence of the marginal distribution for the handling of identifiability and consistency. Of course, the consistency of MMLE for the Rasch models follows from the general result of Butler and Louis (1997) under appropriate conditions, too. The work of Butler and Louis (1997) shows how difficult the handling of mixture models is in general. However, it is interesting to note that the MMLE for the Rasch model can be treated in the common framework of Lehmann and Casella (1998) if the distribution $\mathcal{H}$ is fixed to any arbitrary distribution. An example could be $\mathcal{H} = \mathcal{N}_1(0, 1)$ as constrained version of $\mathcal{H} = \mathcal{N}_1(0, \sigma^2)$ with parameter $\sigma^2 > 0$. But even the Cantor distribution would be a valid choice for $\mathcal{H}$. We formulate the resulting statement about consistency and asymptotic normality in the following theorem, which is new in the IRT context to the best of our knowledge.

Theorem 1.3.5 (Asymptotics of the constrained MMLE for the Rasch Model)
Consider a sequence of iid random vectors $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}}$ with $\boldsymbol{Y}_k \sim \mathcal{P}(\eta_k)$ for $d \in \mathbb{N}$ items and an iid sequence $\{\eta_k\}_{k\in\mathbb{N}}$ with $\eta_k \sim \mathcal{H}$, $k \in \mathbb{N}$, for a probability measure $\mathcal{H}$. Further suppose that $\boldsymbol{Y}_1$ satisfies the Rasch model (1.39) for item difficulty parameters in the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\intercal \in \mathbb{R}^d$.

Then the probability is tending to one as $n \to \infty$ such that solutions $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^\intercal$ of the likelihood equations

$$\sum_{k=1}^{n} Y_{ik} = \sum_{k=1}^{n} \int_{\mathbb{R}} \frac{\exp(\eta - \beta_i)}{1 + \exp(\eta - \beta_i)} \, \mathcal{H}(\mathrm{d}\eta \mid \boldsymbol{Y}_k; \boldsymbol{\beta}) = \sum_{k=1}^{n} \mathsf{E}(Y_{ik} \mid \boldsymbol{Y}_k), \qquad i = 1, \ldots, d, \tag{1.45}$$

exist with $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ *for* $n \to \infty$, *where* $\hat{\boldsymbol{\beta}}_n := \mathbf{0}$ *if no solution of* (1.45) *exist.*

If the information matrix $\mathcal{I} \in \mathbb{R}^{d\times d}$ *with elements*

$$\mathcal{I}_{k\ell} = \mathsf{E}\left(-\frac{\partial^2 \log P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\beta_k \partial\beta_\ell}\right), \qquad k,\ell = 1,\ldots,d,$$

is positive definite, then the sequence $\{\hat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ *is asymptotically normal distributed and efficient in the sense that*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \boldsymbol{Z} \sim \mathcal{N}_d(0, \mathcal{I}^{-1}). \qquad \diamond$$

Proof.
The result follows from checking the usual regularity conditions (cf. for example Lehmann and Casella, 1998, Chapter 6, Theorem 5.1). Details can be found in Appendix A. □

Notice the similarity of the likelihood equations (1.45) to the likelihood equations of the item parameters of JMLE in (1.42): In JMLE, we equate the total score of each item with its expectation under the given values of the latent variables. Here, we equate the item totals with its conditional expectation under the observed response patterns. Further notice that if the data contains problematic cases like items with no (in-)correct responses, MMLE will fail similarly to the previous maximum likelihood methods.

Of course, the Rasch model with parametric variance of a continuous latent trait can also be fitted efficiently by MMLE under some further conditions. We elaborate the details in the following remark.

Remark 1.3.6
In addition to the notation and conditions of Theorem 1.3.5, suppose:

(i) $\mathcal{H}$ is a member of a parametric family $\{\mathcal{H}_{\sigma^2}\}_{\sigma^2>0}$ of probability measures with $\mathcal{H} \equiv \mathcal{H}_{\sigma_0^2}$ for some $\sigma_0^2 > 0$.

(ii) The family $\{\mathcal{H}_{\sigma^2}\}_{\sigma^2>0}$ is dominated by the Lebesgue measure, all members have the same support $\Theta \subseteq \mathbb{R}$, finite third moments and σ^2 equals the variance of the distribution $\mathcal{H}_{\sigma^2}$.

(iii) Denote by $\mathfrak{h}(\,\bullet\,, \sigma^2)$ the pdf of $\mathcal{H}_{\sigma^2}$, $\sigma^2 > 0$. $\mathfrak{h}(\eta,\,\bullet\,) \in C^3((0,\infty),\mathbb{R})$ for all $\eta \in \Theta$. Denote by $\mathfrak{h}^{(v)}$ the derivative of $\mathfrak{h}$ of vth order with respect to σ^2, $v = 1,\ldots,3$. For some $\varepsilon > 0$

$$\int_\Theta \sup_{\sigma^2\in(\sigma_0^2-\varepsilon,\sigma_0^2+\varepsilon)} |\mathfrak{h}^{(v)}(\eta,\sigma^2)|\,\mathrm{d}\eta < \infty$$

for $v = 1,\ldots,3$.

Then the parameter vector $(\beta_1,\ldots,\beta_d,\sigma_0^2)$ can be consistently estimated by maximizing

$$\prod_{j=1}^n \int_{\mathbb{R}} \left(\prod_{i=1}^d \frac{\exp(Y_{ij}(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right) \mathfrak{h}(\eta,\sigma^2)\mathrm{d}\eta$$

in $(\beta_1,\ldots,\beta_d,\sigma^2) \in \mathbb{R}^d \times (0,\infty)$. If the Fisher information is positive definite, then the estimator is asymptotically normal distributed and efficient. The stated conditions are satisfied if $\mathcal{H}_{\sigma^2} = \mathcal{N}(0,\sigma^2)$, $\sigma^2 > 0$. $\diamond$

Proof.
Follows directly from Remark 2.3.1 and Theorem 2.3.2. □

Generally, we will see in Chapter 2 that the asymptotic efficiency for binary mixed models can be shown under some restrictions.

1.3.3.4 Manifest Probabilities Maximum Likelihood Estimation

As we discussed in Subsection 1.3.2, the MPMLE can recognized as MMLE under the parametric family of prior distributions given by 1.34 for $\mathcal{H}$. However, the method itself is quite different. Hence, we point out some mayor properties for the MPMLE for the Rasch model.

We saw in Example 1.2.5, for the Rasch model (1.39), the manifest probabilities model 1.2.4 becomes

$$\log(\mu(\boldsymbol{\vartheta},\boldsymbol{y})) = \log(nP(\boldsymbol{y};\boldsymbol{\vartheta})) = a' + \sum_{i=1}^{d} b_i y_i + c\sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k, \qquad \boldsymbol{y} \in \mathbb{Y},$$

where $a' = \log(n) + a$ and $\boldsymbol{\vartheta} = (a', b_1, \ldots, b_d, c)^\intercal \in \mathbb{R}^{d+2}$. So the log-likelihood becomes

$$\begin{aligned}
\ell_{\text{manifest}}(\boldsymbol{\vartheta})) =& \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y})\log(\mu(\boldsymbol{\vartheta},\boldsymbol{y})) - \mu(\boldsymbol{\vartheta},\boldsymbol{y}) - \log(n(\boldsymbol{y})!) \\
=& na' + \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y})\sum_{i=1}^{d} b_i y_i + c\sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y})\sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k \\
& - \sum_{\boldsymbol{y}\in\mathbb{Y}} \exp\left(a' + \sum_{i=1}^{d} b_i y_i + c\sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k\right) - \sum_{\boldsymbol{y}\in\mathbb{Y}} \log(n(\boldsymbol{y})!).
\end{aligned}$$

This implies for the partial derivatives

$$\begin{aligned}
\frac{\partial \ell_{\text{manifest}}(\boldsymbol{\vartheta}))}{\partial a'} &= n - \sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta},\boldsymbol{y}), \\
\frac{\partial \ell_{\text{manifest}}(\boldsymbol{\vartheta}))}{\partial b_i} &= \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y}) y_i - \sum_{\boldsymbol{y}\in\mathbb{Y}} y_i \mu(\boldsymbol{\vartheta},\boldsymbol{y}), \qquad i = 1,\ldots,d, \\
\frac{\partial \ell_{\text{manifest}}(\boldsymbol{\vartheta}))}{\partial c} &= \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y})\sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k - \sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta},\boldsymbol{y})\sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k,
\end{aligned}$$

for $\boldsymbol{\vartheta} = (a', b_1, \ldots, b_d, c)^\intercal \in \mathbb{R}^{d+2}$. To simplify the notation, we set:

- $n(i) := \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y}) y_i$ is the number of observations with $y_i = 1$, $i = 1,\ldots,d$,
- $\mathbb{Y}(i)$ is the set of response patterns with $y_i = 1$, $i = 1,\ldots,d$,
- $I(\boldsymbol{y}) := \sum_{i=1}^{d}\sum_{k=i+1}^{d} y_i y_k = \sum_{k=1}^{|\boldsymbol{y}|-1} k = \frac{|\boldsymbol{y}|\cdot(|\boldsymbol{y}|-1)}{2}$, the number of interactions, where the $|\boldsymbol{y}|$ is the number of correct responses (cf. (1.40)).

And so finally the likelihood equations are

$$n = \sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}), \tag{1.46}$$

$$n(i) = \sum_{\boldsymbol{y}\in\mathbb{Y}(i)} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}), \qquad i = 1, \ldots, d \tag{1.47}$$

$$\sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{x}) I(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}) I(\boldsymbol{y}), \tag{1.48}$$

for a', $\boldsymbol{b}$ and c, respectively. These likelihood equations have to be solved numerically. Note that equation (1.46) is redundand given (1.47). In particular, a is only a normalization constant without interpretational value.

for the Rasch model

$$b_i = -\alpha\beta_i + \frac{\alpha^2\sigma^2}{2}, \qquad c = \alpha^2\sigma^2. \tag{1.49}$$

Because of (1.49), we can transform the estimates $(\hat{a}, \hat{b}_1, \ldots, \hat{b}_d, \hat{c})$ obtained from the likelihood equations back to the original representation of Rasch model item parameters. This takes the form

$$\hat{\beta}_i = -\hat{b}_i + \frac{\hat{c}}{2}, \qquad \widehat{\mathsf{Var}(\eta)} = \hat{c} + \hat{c}^2 \mathsf{Var}_{\hat{\boldsymbol{\vartheta}}}\left(\sum_{i=1}^{d} Y_i\right),$$

under the identification constraint $\alpha = 1$, where $\mathsf{Var}_{\boldsymbol{\vartheta}}(\bullet)$ denotes the variance under the manifest probabilities model with parameter vector $\boldsymbol{\vartheta}$. The fact that the manifest probabilities model for the Rasch model is a log-linear model implies that the classical estimation theory for those models can be applied. If the actual prior is a member of the family specified by (1.34), then the estimator is consistent and asymptotically normal distributed (Agresti, 2013). However, because in typical settings a prior distribution as $\mathcal{H} = \mathcal{N}(0, \sigma^2)$ is assumed (and often justified from the psychological theory), we have a slight miss-specification which may cause worse results in comparison to the previously discussed methods.

Similar to the previous methods, there are some critical cases where no real valued parameter estimates exist. We want to discuss some important cases briefly.

1. At least one item has only incorrect responses

Suppose that none of the subjects have responded correctly to at least one item. Without loss of generality, we assume that this is the case for exactly one item, the I-th. This means that $y_I = 0$ in each observed response pattern and so $n(I) = 0$. Thus for $i = I$, the likelihood equation in (1.47) becomes

$$n(I) = 0 = \sum_{\boldsymbol{y}\in\mathbb{Y}(I)} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}) = \sum_{\boldsymbol{y}\in\mathbb{Y}(I)} \exp\left(a' + \sum_{i=1}^{d} b_i y_i + c \sum_{i=1}^{d} \sum_{k=i+1}^{d} y_i y_k\right).$$

Since we sum on the right-hand side over each possible response pattern with $y_I = 1$, we sum over 2^{d-1} strictly positive terms. The only parameter which is in each term but not in each term of other likelihood equations is b_I. This implies $b_I = -\infty$. If the Rasch model itself is used,

$$\mathsf{P}(Y_I = 1|\eta) = \frac{1}{1 + \exp(-\eta + \beta_I)}.$$

To get $y_I = 0$ in each observed pattern, the probability for $Y_I = 0$ has to be (nearly) one. This implies consequently $\beta_I = \infty$.

One regularization method for sampling zeros in contingency tables is to add $\frac{1}{2}$ to each cell, compare Agresti (2013, Section 10.6.5). But this method manipulates the data and has more influence on the model than just ensuring existing estimates. From a more practical point of view, we make a relative comparison inside a subpopulation when interpreting the fit. With a similar argument, we typically set $E(\eta) = 0$ in the prior assumptions. Item I was simply too difficult but provides no further information about the value of difficulty (β_I) or the abilities. So it is appropriate to delete item I from the data instead of adding $\frac{1}{2}$ to all cells or at least all zero cells.

2. At least one item has only correct responses

In this case we get $n(I) = n$ for item I. It follows from the likelihood equations

$$\sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}) = \sum_{\boldsymbol{y}\in\mathbb{Y}(I)} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}).$$

This leads analogously to $b_I = \infty$ or $\beta_I = -\infty$.

3. No observed response pattern has more than one correct response

We assume that each response pattern has exactly one correct response, else we encounter the problems of case one. Then we get the likelihood equation

$$\sum_{\boldsymbol{y}\in\mathbb{Y}} \mu(\boldsymbol{\vartheta}, \boldsymbol{y}) I(\boldsymbol{y}) = \sum_{\boldsymbol{y}\in\mathbb{Y}} n(\boldsymbol{y}) I(\boldsymbol{y}) = 0.$$

This leads to the same problem as in case one, but we can interprete it differently. First, we can conclude that the test dimension is not one. So the assumptions of the Rasch model are not satisfied. Second we observe a form of *complete ability separation*[14], i.e.: Each item tests exactly one ability and we test disjoint populations with one strong ability and otherwise weak abilities. Therefore it seems more appropriate to classify the subjects in groups and in this case it seems that $q = d$. Instead of estimating difficulties and abilities, it may be suitable to estimate the membership of the subjects to particular classes. In such a case we choose a discrete prior distribution $\mathcal{H}$ for $\boldsymbol{\eta}$, respectively a *latent class model*. We illustrate this problem as an example item response data.

Example 1.3.7

We assume a test on n_1 physicists, n_2 foresters and n_3 sailors and let them respond to items from their area of expertise. If all subjects have a high ability in their area of expertise but low abilities in the other two areas, e.g. $\eta_1 \in [10, 15]$ for physicists but $\eta_2, \eta_3 \in [-15, -10]$, a conceivable data set is given in Table 1.1. Then a possible "truth" may be reflected by the item response model

$$\operatorname{logit}(P(Y_i = 1|\boldsymbol{\eta})) = \eta_i, \quad i = 1, 2, 3.$$

So a fit of a Rasch model may not be adequate. A more appropriate model is

$$\operatorname{logit}(P(Y_i = 1|\boldsymbol{\eta})) = \alpha_i \eta_i - \beta_i$$

[14] Compare also the related topic of *complete separation* or *perfect discrimination* in logistic regression, i.e. the cases in which there is a hyperplane that perfectly separates all explanatory variables for both possible values of the response variable and causing non-existence of the MLE (cf. e.g. Agresti, 2013, Subsection 5.5.5).

subject j	y_1	y_2	y_3
1	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	1	0	0
n_1 +1	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_1 + n_2$	0	1	0
$n_1 + n_2 + 1$	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_1 + n_2 + n_3$	0	0	1

Table 1.1: *Ability separated data*

or in a slightly more unclear situation the model

$$\text{logit}(P(Y_i = 1|\boldsymbol{\eta})) = \sum_{k=1}^{3} \alpha_{ik}\eta_k - \beta_i.$$

◇

1.3.3.5 Pseudo Likelihood Estimation

Recall that one aim of developing the manifest probabilities model in Section 1.2 was to avoid numerical quadrature for the marginal distribution of $\boldsymbol{Y}$ (and consequently for the MMLE), which is problematic for high latent dimension q. The avoidance of non-closed integrals was successful, but now we have to calculate expected frequencies for each possible response pattern. Since we have dichotomous data, we can generally specify the number of possible patterns to be 2^d. For few items, the number can be handled and the method is feasible. But for growing numbers of items, this is not computationally feasible. For example for 40 items we get $2^{40} \approx 10^{12}$ possible response patterns. So, instead of a high q, now a large number of items d is very problematic. This implies the necessity of further specific research.

Anderson et al. (2007) suggested using a pseudo-type likelihood method. The method presented by them is a composite likelihood method (cf. Varin et al. (2011)) of a *fully conditional type*. For the sake of consistency, we will also call it a *pseudo likelihood method*. Especially, Anderson et al. (2007) presented a method for a generalized Rasch model in the sense that model (1.26) is generalized to ordinal response variables and the parameters $\alpha_{11}, \ldots, \alpha_{dq}$ (the factor loadings) are known. This estimation approach is based mainly on the idea behind the conditional specified models, which we discussed briefly in Subsection 1.2.3. We will now give a brief exposition of the method presented by Anderson et al. (2007) for the Rasch model (cf. equation (1.39))

$$\text{logit}(P(Y_i = 1|\eta)) = \eta - \beta_i,$$

with unknown item parameters β_i and $\mathsf{Var}(\eta)$. Instead of integrating out the unknown latent variable η, Anderson et al. (2007) suggested replacing η with an estimate obtained from other responses of the same subject. Practically, we condition the response probability

of item i on the items $k \neq i$ instead of η and set

$$\text{logit}(P(Y_i = 1|Y_k,\, k \neq i)) = \sigma^2 \sum_{k \neq i} Y_k - \beta_i, \quad i = 1, \ldots, d \tag{1.50}$$

where the weighting parameter σ^2 is the unknown conditional variance of η given $\boldsymbol{Y}$. The rest score $\sum_{k \neq i} Y_k$ becomes a categorial explanatory variable and is known when starting the estimation. In this setup we can use the classical maximum likelihood estimation for ordinary logistic regression.

To simplify the notation, we write $T_i := \sum_{k \neq i} Y_k$ and $P(Y_i = 1|Y_k,\, k \neq i) =: \pi_i$. Then the conditional probability (1.50) becomes

$$\text{logit}(\pi_i) = \sigma^2 T_i - \beta_i, \quad i = 1, \ldots, d.$$

We write the realization of T_i for subject j as $t_{ij} = \sum_{k \neq i} y_{kj}$, and since π_i depends on T_i, we define for subject j, $j = 1, \ldots, n$,

$$\text{logit}(\pi_{ij}) = \sigma^2 t_{ij} - \beta_i, \quad i = 1, \ldots, d. \tag{1.51}$$

Therefore, we define the pseudo likelihood for one response pattern $\boldsymbol{y} \in \mathbb{Y}$ as

$$\begin{aligned}
P(\boldsymbol{Y} = \boldsymbol{y}) &= \prod_{i=1}^{d} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} = \prod_{i=1}^{d} \exp\left[\log\left(\left[\frac{\pi_i}{1 - \pi_i}\right]^{y_i}\right)\right](1 - \pi_i) \\
&= \exp\left[\sum_{i=1}^{d} \log\left(\left[\frac{\pi_i}{1 - \pi_i}\right]^{y_i}\right)\right] \prod_{i=1}^{d} (1 - \pi_i) \\
&= \exp\left[\sum_{i=1}^{d} y_i \,\text{logit}\,(\pi_i)\right] \prod_{i=1}^{d} \frac{1}{1 + \exp(\sigma^2 T_i - \beta_i)} \\
&= \exp\left[\sum_{i=1}^{d} y_i (\sigma^2 T_i - \beta_i)\right] \prod_{i=1}^{d} \frac{1}{1 + \exp(\sigma^2 T_i - \beta_i)}.
\end{aligned}$$

Then the pseudo likelihood function for n observed responses and the parameters $(\sigma^2, \beta_1, \ldots, \beta_d) = \boldsymbol{\vartheta}$ is given by

$$\begin{aligned}
\mathcal{L}_{\text{pseudo}}(\boldsymbol{\vartheta}) &= \prod_{j=1}^{n} \exp\left(\sum_{i=1}^{d} y_{ij} (\sigma^2 t_{ij} - \beta_i)\right) \prod_{i=1}^{d} \frac{1}{1 + \exp(\sigma^2 t_{ij} - \beta_i)} \\
&= \exp\left(\sum_{j=1}^{n} \sum_{i=1}^{d} y_{ij} (\sigma^2 t_{ij} - \beta_i)\right) \prod_{j=1}^{n} \prod_{i=1}^{d} \frac{1}{1 + \exp(\sigma^2 t_{ij} - \beta_i)}.
\end{aligned}$$

The log likelihood function thus becomes

$$\ell_{\text{pseudo}}(\boldsymbol{\vartheta}) = \sum_{j=1}^{n} \sum_{i=1}^{d} y_{ij} (\sigma^2 t_{ij} - \beta_i) - \sum_{j=1}^{n} \sum_{i=1}^{d} \log\left(1 + \exp(\sigma^2 t_{ij} - \beta_i)\right).$$

Using some calculus we get for the derivatives

$$\begin{aligned}
\frac{\partial \ell_{\text{pseudo}}(\boldsymbol{\vartheta})}{\partial \sigma^2} &= \sum_{j=1}^{n}\sum_{i=1}^{d} y_{ij}t_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{d} t_{ij}\frac{\exp(\sigma^2 t_{ij} - \beta_i)}{1+\exp(\sigma^2 t_{ij} - \beta_i)} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{d} y_{ij}t_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{d} t_{ij}\pi_{ij}, \\
\frac{\partial \ell_{\text{pseudo}}(\boldsymbol{\vartheta})}{\partial \beta_k} &= -\sum_{j=1}^{n} y_{kj} + \sum_{j=1}^{n}\frac{\exp(\sigma^2 t_{kj} - \beta_k)}{1+\exp(\sigma^2 t_{kj} - \beta_k)} \\
&= -\sum_{j=1}^{n} y_{kj} + \sum_{j=1}^{n}\pi_{kj}, \qquad\qquad k = 1,\ldots,d.
\end{aligned}$$

This finally implies the likelihood equations

$$\begin{aligned}
\sum_{j=1}^{n}\sum_{i=1}^{d} y_{ij}t_{ij} &= \sum_{j=1}^{n}\sum_{i=1}^{d} t_{ij}\pi_{ij} \\
\sum_{j=1}^{n} y_{kj} &= \sum_{j=1}^{n}\pi_{kj}, \qquad\qquad k = 1,\ldots,d,
\end{aligned} \tag{1.52}$$

with $\pi_{ij} = \pi_{ij}(\boldsymbol{\vartheta})$ defined in equation (1.51). The equations (1.52) have to be solved numerically for obtaining estimates for $\boldsymbol{\vartheta} = (\sigma^2, \beta_1, \ldots, \beta_d)$. We call the obatined estimator *pseudo likelihood estimator (PLE)*.

The probabilities (1.51) imply the manifest probabilities model (cf. Subsection 1.2.3). Hence, if a Rasch model with a prior of the class (1.34) holds, then the PLE is a classcial composite likelihood method. Anderson et al. (2007) discussed that the method is consistent if the manifest probabilities model holds. However, we can expect worse estimation results because it is a limited-information method. Moreover, the manifest probabilities model may also be an inadequate approximation (especially for few items under an asymptotic justification). Then we get a cumulative approximation error. On the other side, classical methods like MMLE always use some approximations. Additionally, the estimation equations (1.52) themselves are easy to evaluate. Overall, further research regarding properties of the PLE is required.

1.3.3.6 Comparison of the Methods

We discussed several estimation methods for IRT models and focussed on methods for the Rasch model in the last section. In this subsection, we want to compare the methods briefly with focus on the Rasch model to allow a justified decision for a particular method in a practical data analysis.

One aspect to consider is the presence or absence of prior knowledge about η, which may lead to the selection of a fixed (JMLE, CMLE) or random effect approach (MMLE, MPMLE, PLE). As in classical Bayesian analysis, we expect less impact of the particular prior distribution for a large number of items (cf. also Bartholomew et al., 2011). Especially, for a large number of items, the posterior normality could be justified asymptotically. Hence, the MPMLE and PLE are also meaningful if we have a relatively large number of items. However, for a large number of items, the MPMLE is difficult to apply. In that case, the PLE

Item Difficulty Parameter Values

Item i	Value of β_i	Item i	Value of β_i
1	−0.905807	11	−0.918887
2	−1.336721	12	0.216187
3	−0.913367	13	1.899601
4	0.386846	14	0.510270
5	0.675736	15	−1.109430
6	−0.083869	16	−0.104312
7	0.138496	17	0.458432
8	1.608452	18	−1.882087
9	−0.882378	19	0.801041
10	1.240412	20	1.953393

Table 1.2: *The values of the item difficulty parameters in the simulation study. For $d = 5$ and $d = 10$ items, the first five/ten values are taken.*

is a fast alternative. When applying PLE, one should always have the status as a composite likelihood method in mind, which leads to a larger standard error. In the presence of prior knowledge, the MMLE is the method of choice as it can be used directly. Generally, one needs computational knowledge about numerical quadrature to apply MMLE. But for the classical Rasch model with one latent dimension, there are good standard approaches.

In the absence of prior information, one can apply a fixed effect method. For a small number of items, the CMLE is a well established method. Difficulties arise for a large number of items due to the need of an evaluation of all possible response patterns like the MPMLE. For a relatively large number of items, the JMLE (in particular the bias corrected cJMLE) can be a fixed effect alternative. An advantage of applying JMLE is the simultaneous estimation of item parameters and values of the latent variables. Alternatively, one could apply shrinkage methods for the JMLE to handle the consistency problems.

Because there are conceptual differences, we will perform a small simulation study to get an impression of the differences in precision between the methods.

Simulation Setup

We consider the Rasch model in the formulation

$$\text{logit}(P_i(\eta)) = \eta - \beta_i, \qquad i = 1, \ldots, d,$$

with $\eta \sim \mathcal{N}(0, 1)$ for $n \in \{50, 100, 500\}$ subjects and $d \in \{5, 10, 20\}$ items.[15] The considered values for $\beta_1, \ldots, \beta_d$ are given in Table 1.2. For each pair of d and n, we simulate new values of $\eta_1, \ldots, \eta_n$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 250 times. We check all data sets to ensure finite estimates of the proposed methods. In critical cases, we generate new data sets.

We consider the methods for Rasch model as discussed in Subsection (1.3.3). For the MMLE, we suppose $\mathcal{H} = \mathcal{N}(0, \sigma^2)$. Additionally, we consider the MMLE with the nested model $\mathcal{H} = \mathcal{N}(0, 1)$ and denote it by MMLEn, since we discussed it further theoretically and it allows an impression of the estimation quality on nested models.

We use the implementation of the methods of the following R-packages: **sirt** for JMLE (Robitzsch, 2020), **eRm** for CMLE (Mair et al., 2019), **ltm** for MMLE (Rizopoulos, 2006), **plRasch** for MPMLE and PLE (Anderson et al., 2007).

[15] For more than 20 items, the MPMLE was no longer feasible.

Simulation Procedure and Evaluation

The MMLE and MMLEn anchor the parameters using $\mathsf{E}(\eta) = 0$. The other methods do not have this anchor. One possibility to anchor their estimates is an approximation using $\widehat{\mathsf{E}(\eta)} \stackrel{!}{=} 0$. A more appropriate option is to transform/identify all estimators using a criterion independent of the distribution $\mathcal{H}$. Hence, we consider the identification constraint $\beta_1 = 0$ for all estimators. For all combinations of d and n, we compute for all considered methods the following measures for the estimators of $\beta_2, \ldots, \beta_d$, where $\hat{\beta}_i^{(k)}$ denotes the estimation of β_i based on the kth simulated data set using the particular method:

1. The mean absolute bias:
$$\text{maBias} = \frac{1}{d-1} \sum_{i=2}^{d} \left| \frac{1}{250} \sum_{k=1}^{250} \left(\hat{\beta}_i^{(k)} - \beta_i \right) \right|.$$

2. The mean variance:
$$\text{mVar} = \frac{1}{d-1} \sum_{i=2}^{d} \frac{1}{249} \sum_{k=1}^{250} \left(\hat{\beta}_i^{(k)} - \bar{\beta}_i \right)^2$$
where $\bar{\beta}_i$ is the mean estimate of β_i with
$$\bar{\beta}_i := \frac{1}{250} \sum_{k=1}^{250} \hat{\beta}_i^{(k)}.$$

3. The average root mean squared error:
$$\text{armse} := \frac{1}{d-1} \sum_{i=2}^{d} \sqrt{\frac{1}{250} \sum_{k=1}^{250} \left(\hat{\beta}_i^{(k)} - \beta_i \right)^2}.$$

The resulting computed values are given in the Tables 1.3, 1.4 and 1.5. It is obvious that the JMLE is decisively worse in comparison to any other method. Further, the cJMLE constitutes a large improvement in all three criteria to a degree that it is only slightly worse in comparison to any of the other methods. All other methods yield practically identical values. The MMLEn produces the best values, however, for its application strong prior knowledge is necessary. As compensation for the slightly worse quality, the cJMLE provides estimates for $\eta_1, \ldots, \eta_n$, which have to be estimated further by applying the other methods if necessary. Overall, there is no clear best method in terms of armse, mBias and mVar. Thus, one can select the method based on a consideration of conceptual and computational properties. In particular, the PLE method seems to be performing quite well, even if it is just an approximate composite likelihood estimation due to the wrong prior distribution. This motivates to work further on the analysis and development of the pseudo likelihood methods, as we do in Chapter 5.

1.3.4 Analysis of the Structural Model and Estimation of the Latent Variables

With the exception of variants of joint maximum likelihood estimation and fully Bayesian methods, all previously discussed methods exclusively estimated parameters of the distribution of $\boldsymbol{Y}$ while the concrete values of $\boldsymbol{\eta}$ remained unknown. In many research situations, this or an estimation of the structural model can be sufficient. We illustrate this with the following example.

ARMSE Estimation Methods Rasch Model

n subjects	method	5 items	10 items	20 items
50	JMLE	0.739756	0.599038	0.556264
	cJMLE	0.524588	0.505998	0.513650
	MMLE	0.497696	0.493945	0.506705
	MMLEn	0.483851	0.495164	0.504601
	CMLE	0.502699	0.494003	0.507355
	MPMLE	0.497638	0.493887	0.505732
	PLE	0.498569	0.493735	0.506279
100	JMLE	0.533206	0.432926	0.389106
	cJMLE	0.353261	0.353241	0.353746
	MMLE	0.337630	0.345319	0.349292
	MMLEn	0.332248	0.342329	0.348308
	CMLE	0.339957	0.345357	0.349387
	MPMLE	0.337547	0.345336	0.348571
	PLE	0.337679	0.345092	0.348871
500	JMLE	0.322342	0.232632	0.184635
	cJMLE	0.152848	0.151976	0.152604
	MMLE	0.147076	0.148064	0.150513
	MMLEn	0.146585	0.146921	0.150466
	CMLE	0.147394	0.148069	0.150526
	MPMLE	0.147121	0.147962	0.150188
	PLE	0.147030	0.148044	0.150314

Table 1.3: *The computed values of the average root mean squared error based on* 250 *replications for different estimation methods of the difficulty parameters of* $\beta_2, \ldots, \beta_d$ *with identification constraint* $\beta_1 = 0$.

Example 1.3.8

Consider the situation of an educational survey in a school with the question: Are there factors (such as gender, migration background, income of parents et cetera) which lead to a structural (dis-)advantage of the corresponding groups of students in a specific subject? Supposing homogeneity inside of all groups, we can apply a latent ANOVA model on η. Hence, we suppose the structural model $(\eta \mid \boldsymbol{X} = \boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2)$, where $\boldsymbol{X}$ is an explanatory variable containing the factor levels of $p \in \mathbb{N}$ factors, where each factor v as $m_v \in \mathbb{N}$ levels, $v = 1, \ldots, p$. Thus, $\mu(\boldsymbol{x})$ takes the form

$$\mu(\boldsymbol{x}) = \gamma_0 + \sum_{v=1}^{p} \sum_{h=1}^{m_v} \gamma_{vh} \mathbb{1}(x_v = h)$$

with model parameters $\gamma_0, \gamma_{11}, \ldots, \gamma_{pm_p} \in \mathbb{R}$. Notice that we can set $\gamma_0 = 0$ and $\sigma^2 = 1$ as identification constraints without loss of generality. The parameters $\gamma_0, \gamma_{11}, \ldots, \gamma_{pm_p}$ can be estimated straight with the parameters of the measurement model in a straight forward manner with the previously discussed methods. After the estimation of $\gamma_0, \gamma_{11}, \ldots, \gamma_{pm_p}$, the research question can be analyzed statistically using these estimates by testing if the factors have a significant influence. Thus, obtaining estimates of the value of η for the $n \in \mathbb{N}$ responding subjects is not necessary. ◇

Mean Absolute Bias Estimation Methods Rasch Model

n subjects	method	5 items	10 items	20 items
50	JMLE	0.290517	0.184394	0.129431
	cJMLE	0.079348	0.062256	0.063269
	MMLE	0.058559	0.048035	0.053283
	MMLEn	0.055345	0.048226	0.050983
	CMLE	0.056736	0.047909	0.053685
	MPMLE	0.058074	0.048242	0.053555
	PLE	0.058457	0.048455	0.054135
100	JMLE	0.252005	0.156338	0.104422
	cJMLE	0.035970	0.032503	0.041902
	MMLE	0.012915	0.019130	0.033079
	MMLEn	0.012209	0.020144	0.031611
	CMLE	0.015262	0.019883	0.033481
	MPMLE	0.013358	0.019566	0.033429
	PLE	0.013300	0.019118	0.033634
500	JMLE	0.230929	0.139391	0.081030
	cJMLE	0.019110	0.015979	0.016216
	MMLE	0.005674	0.006645	0.008092
	MMLEn	0.005089	0.006622	0.007956
	CMLE	0.005565	0.006656	0.008142
	MPMLE	0.006165	0.006642	0.008427
	PLE	0.006154	0.006722	0.008653

Table 1.4: *The computed values of the mean absolute bias based on* 250 *replications for different estimation methods of the difficulty parameters of* $\beta_2, \ldots, \beta_d$ *with identification constraint* $\beta_1 = 0$.

Bock and Aitkin (1981) sketched an iterative approach for a nonparametric estimation of the prior distribution by updating the item parameters. Shortly later, Mislevy (1984, 1985) published detailed estimation techniques for structural models, which can be done particularly well by an EM-algorithm. Mislevy (1984, 1985) discussed the situation where the structural model parameters are the exclusive parameters to be estimated and the measurement model parameters are treated as known (due to precise previous estimation). Thus, Mislevy (1984, 1985) suggested a first complete calibration of the measurement model by using a fixed distribution for $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, e.g. the commonly used $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. However, this approach can be inserted into the EM-algorithm for the measurement model parameters as discussed by Adams et al. (1997a) and Adams et al. (1997b) for logit models. While it does note make a large computational difference, this has two major advantages: The standard error estimates can be expected to be more reliable as they do not ignore the standard errors of the measurement model parameter estimates and also, we can expect more precise point estimates as well.

Principally, the fit of parametric structural models can be included in all above mentioned random effect estimation methods by replacing the fixed prior distribution with a parametric one and maximizing the resulting new likelihood function with respect to all parameters. In the particular case of MPMLE and PLE, the selection of structural models is more restricted. However, the model can still extended for covariates, for example by formulating

Mean Variance Estimation Methods Rasch Model

n subjects	method	5 items	10 items	20 items
50	JMLE	0.417721	0.312615	0.289169
	cJMLE	0.267342	0.253218	0.260975
	MMLE	0.244639	0.243041	0.255413
	MMLEn	0.231666	0.244406	0.253604
	CMLE	0.249893	0.243209	0.256011
	MPMLE	0.244629	0.242942	0.254376
	PLE	0.245467	0.242776	0.254844
100	JMLE	0.193944	0.153370	0.137375
	cJMLE	0.124124	0.124230	0.123981
	MMLE	0.114949	0.119661	0.121657
	MMLEn	0.111275	0.117365	0.121068
	CMLE	0.116490	0.119645	0.121686
	MPMLE	0.114910	0.119643	0.121083
	PLE	0.114977	0.119485	0.121281
500	JMLE	0.035950	0.028106	0.025617
	cJMLE	0.023008	0.022766	0.023119
	MMLE	0.021563	0.021993	0.022724
	MMLEn	0.021712	0.021652	0.022711
	CMLE	0.021808	0.021994	0.022728
	MPMLE	0.021718	0.021962	0.022616
	PLE	0.021691	0.021984	0.022648

Table 1.5: *The computed values of the mean variance based on 250 replications for different estimation methods of the difficulty parameters of $\beta_2, \ldots, \beta_d$ with identification constraint $\beta_1 = 0$.*

the main effects of the resulting associtation model as a linear term of the covariates (cf. Li, 2010; Anderson et al., 2010; Anderson, 2013). Note that the effect of covariates differ from that of latent linear models such as that of Example 1.3.8 if the covariate effect is formulated as a linear term for the main effect (cf. e.g. Li, 2010, equation (6.151) on p. 112).

In other practical situations, we are strongly interested in estimates of the latent vectors. Examples are psychological and psychiatric diagnoses or assessment tests. The property of solely estimating the parameters of the (marginal) distribution of $\boldsymbol{Y}$ leads to a general 2-step scheme for estimating the latent vectors. We denote the estimate of $\boldsymbol{\eta}$ that explains the observed response pattern $\boldsymbol{Y} = \boldsymbol{y} \in \mathbb{Y}$ by $\hat{\boldsymbol{\eta}}_{\boldsymbol{y}}$. Denote further by $\boldsymbol{\vartheta}$ the set of all parameters of the marginal distribution of $\boldsymbol{Y}$ in the considered model and by $\hat{\boldsymbol{\vartheta}}$ an estimate of $\boldsymbol{\vartheta}$. Then the procedure is the following (cf. e.g. Mislevy, 1985):

1. Estimate the parameters $\boldsymbol{\vartheta}$ with $\hat{\boldsymbol{\vartheta}}$ based on a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $n \in \mathbb{N}$.
2. Estimate $\boldsymbol{\eta}_k$ which lies behind the response pattern $\boldsymbol{y}_k$ by $\hat{\boldsymbol{\eta}}_{\boldsymbol{y}_k}$ for $k = 1, \ldots, n$ by plugging in $\hat{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta}$ and treating it as the true value of $\boldsymbol{\vartheta}$.

For step 1 we can use any of the previously discussed methods. Of course, we can estimate $\hat{\boldsymbol{\eta}}_{\boldsymbol{y}}$ for any $\boldsymbol{y} \in \mathbb{Y}$ and not only for the observed patterns $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. In particular, the

estimation of latent vectors can be part of the analysis subsequent to the estimation of the corresponding structural model.

As discussed for example by Booth and Hobert (1998) in the setting of mixed effect models, this two-step-procedure may lead to underestimated standard errors of $\hat{\boldsymbol{\eta}}_{\boldsymbol{y}}$. Hence, one has to pay attention in the statistical analysis and this procedure is not completely uncritical. It has also advantages, however, since we avoid the questionable consistency of the JMLE if the methods in both steps are consistent. This procedure also fits perfectly into common applied settings of psychological measurement, where the measurement model is calibrated at the beginning and then later used for estimating $\hat{\boldsymbol{\eta}}_{\boldsymbol{y}}$ for some $\boldsymbol{y} \in \mathbb{Y}$.

Under the restriction to one dimensional measurement models, Lord (1980) used the *maximum likelihood estimator (MLE)* for η based on a response pattern $\boldsymbol{y} \in \mathbb{Y}$, which is a maximal point of the mapping

$$\eta \mapsto \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \eta).$$

Soon afterwards, Lord (1983) discussed the MLE for η more in detail. Based on the theory of Bradley and Gart (1962), he proved the consistency of the MLE for the 3PL model under the not formally defined condition to a test lengthening using *strictly parallel forms*[16]. Later, Klauer (1990) discussed consistency and asymptotic efficiency of the MLE under milder conditions on the measurement model, but still under the restriction to one dimensional models. Using a counter example, Klauer (1990) showed that even if the 2PL model holds and the Fisher information approaches infinity, the MLE can be inconsistent and in particular not asymptotically normal distributed. Hence, the discussion of asymptotic properties under a formal framework is essential (for example Samejima, 1977, p. 166, supposed the condition on the Fisher information to be sufficient). A few years later, Chang and Stout (1993) presented an alternative proof of the consistency of the MLE under slightly different conditions than Klauer (1990) on a one dimensional measurement model. A clear disadvantage of the MLE is: If $\boldsymbol{y} = (0, \ldots, 0)^\intercal$ (or $\boldsymbol{y} = (1, \ldots, 1)^\intercal$), then the MLE is $\hat{\eta} = -\infty$ (or $\hat{\eta} = \infty$, respectively). An alternative is given in the *Bayes modal estimator* (also known as *maximum a-posteriori estimator, MAP*). The MAP is given as a maximal point of the posterior density $\mathfrak{h}(\,\bullet \mid \boldsymbol{Y} = \boldsymbol{y})$ and thus equivalently of the more convenient mapping

$$\eta \mapsto \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \eta)\mathfrak{h}(\eta), \tag{1.53}$$

for a prior density $\mathfrak{h}$ of η. The use of this method is especially justified under the random effect approach since we already use the density $\mathfrak{h}$ in the estimation of the model parameters. The use of the MAP can lead to a larger bias in finite sample cases in comparison to the MLE[17], but the prior performs as a shrinkage and avoids commonly infinite estimates (cf. Samejima (1980), Lord (1983) and Samejima (1969)). As a further alternative, Warm (1989) introduced the *weighted likelihood estimator (WLE)*. Here, instead of the density $\mathfrak{h}$ from a structural model, a weighing function based on the Fisher information matrix is used (i.e. Jeffeys prior). Warm (1989) proved that the WLE is unbiased to order d^{-1}, where d is the number of items. Especially he analyzed that the WLE usually has a smaller bias in comparison to the MLE and MAP.

[16] Lord (1983) does not give a formal definition of *strictly parallel forms*. We suppose that he means: Two mappings $f, g\colon X \to \mathbb{R}$, for some vector space X, are *strictly parallel* if and only if there is an element $v \in X$ such that $f(w) = g(w - v)$ for all $w \in X$. A sequence $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of *strictly parallel* mappings, if and only if f_1 and f_n are strictly parallel for all $n \in \mathbb{N}$.

[17] In many cases, the MLE is biased for a finite sample size.

Recently, Sinharay (2015) discussed MLE, WLE and MAP under mixed-format and multidimensional IRT models and generalized results on the MLE, WLE and MAP to mixed-format tests and results on the MLE to multidimensional IRT models. However, as Sinharay (2015, p. 524–525) noted, his generalization of the results on the MLE to multidimensional models is based on the work Lord (1983) which leads to strong conditions on the measurement model. A problem of the work of Sinharay (2015) is that the condition of strictly parallel forms for multidimensional models is not defined in the work[18]. It is questionable if the average Fisher information matrix will have full rank for $d \to \infty$ (and thus, cannot be inverted). As Klauer (1990, p. 24) already noted in his discussion of the one dimensional case: It is questionable if the average information can even be assumed to converge. Hence, there are still open questions, even if we expect the MLE to be consistent under standard conditions. Similar to Chang and Stout (1993), we will receive alternative results regarding the consistency of the MLE for multidimensional models as a by-product of our analysis in Chapter 3. We further extend them to the asymptotic normality of the MLE to answer another open question under multidimensional models.

Under the random effect approach, there is a large class of estimators as further alternative to the MAP estimator. Because the estimation of the latent vectors fit into the empirical Bayes approach under the random effect approach (cf. Remark 1.1.2), we directly obtain the full class of Bayes efficient estimators for $\boldsymbol{\eta}$ as Bayesian parameters. The most important and popular estimator aside from the MAP is the *expected a-posteriori (EAP) estimator*:

$$\hat{\boldsymbol{\eta}}_{\text{EAP}} := \mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y}). \tag{1.54}$$

Bock and Mislevy (1982) analyzed the EAP in the IRT setting and concluded several advantages, like for example the prevention of infinite estimators, the minimization of the posterior mean squared error[19], independence of derivatives and its inclusion in the EM-algorithm for estimating the actual model parameters. Notice that under the manifest probabilities model, we can use Proposition (1.2.3) and obtain the EAP as a linear function of the observed pattern $\boldsymbol{y}$ (cf. also Li, 2010; Hessen, 2012). We will further discuss this as a basis for estimation in Section 4.1.

Of course, other Bayes efficient estimators can also be formulated this way. One example is the *median a-posteriori (MeAP) estimator* for a unidimensional η, which is defined as the number $\hat{\eta}_{\text{median}}$ such that

$$\mathsf{P}(\eta \leq \hat{\eta}_{\text{median}} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{2}.$$

The MeAP estimator is the Bayes efficient estimator under the Laplace-loss function $(\hat{\eta}, \eta) \mapsto a|\eta - \hat{\eta}|$ for any $a > 0$ (cf. for example Rüschendorf, 2014).

Take notice of the fact that in one dimensional models, the theory of Chang and Stout (1993) allows the consideration of the MLE as a Bayesian estimator for a large number of items. We will discuss this in detail, together with further asymptotic properties of Bayes efficient estimators and the already stated discussion of the consistency of the MLE in, Chapter 3.

Till this day, theoretical and computational research regarding estimators for the latent vectors has an active research community (consider for example Tsutakawa and Soltys

[18] Cf. footnote 16.

[19] This is the Bayes efficiency with respect to the *squared* or *Gaussian loss*, $(\hat{\eta}, \eta) \mapsto (\eta - \hat{\eta})^2$ (cf. for example Rüschendorf, 2014).

(1988), Junker (1991), Kim and Nicewander (1993), Ogasawara (2012, 2013), Wang (2015), Sinharay (2015)).

Chapter 2

Extended Generalized Linear Latent Variable and Structural Models

In Section 1.1 of the previous chapter we have discussed how important the assumption of the link function in the model class GLLVM for binary data is. We have seen that several models have been discussed to increase the model flexibility and the option to capture skewness in the ICC. Several of such models can be considered a GLLVM with a parametric link function. A similar aspect is the distribution of the latent variable which we call the structural model. The structural model is often supposed to be a normal distribution. This assumption can be critical and led to the extension of the structural model to a parametric distribution like the family of skew normal distributions. In this chapter we discuss an extension of the GLLVM framework with focus to the considered binary data with respect to skewed link functions and skewed latent distributions that depend on parameters for their skewness/shape. By supposing different skewness parameters for all link functions, we obtain a model in that each item can have a different measurement model. The derived multidimensional model is formulated in a way to obtain a unified estimation and testing framework of IRT models with skew link functions and structural models.

The logical next step from estimating the skewness of a link function or of the structural model is to select the best one and to test it. In this chapter we focus on a framework to select and test the link function and the structural model. By formulating a unifying model class for GLLVM with parametric link functions and parametric distributions for the latent variable – the extended GLLVM (EGLLVM) –, we open the toolbox of nested model testing for the purpose of checking the link function and the structural model. In this chapter, we focus on a skew-probit model with skew-normal distribution for the latent variable, which is an important skew IRT model, but until now it was primarily analyzed by Bayesian methods (Bazán et al., 2014). Additionally, we discuss the formulation of a guessing model using the EGLLVM and illustrate properties of the class and the fitting via this model.

An important issue which arises by the formulation of well-defined nested model tests is the formulation of a unifying model class that allows an asymptotic efficient (marginal maximum likelihood) estimation of the model parameters – an asymptotic property which is essential for good asymptotic tests. This problem is serious: The research focus in IRT was the development of computationally efficient procedures; theoretical properties like the asymptotic normality are commonly assumed but a proof of the asymptotic normality and efficiency was never published (cf. Subsection 1.3.2). Another issue arises from testing the skewness parameters by nested models: The fit of the more complex model has to be checked, too. A general issue in the application of these tests arises from practical estimation in finite sample sizes: An often reported difficulty in the frequentist estimation of guessing and skewness parameters (i.e. some “third” parameter of an item) is the difficulty

to estimate these parameters. The consistency of the MMLE ensures the valid estimation for sufficiently large data sets, but for practical sizes it might be problematic and could affect the tests for skewness.

First of all, we formulate the EGLLVM in Section 2.1 and define all important terms of a EGLLVM in a way which enables us to prove the necessary results for asymptotic estimation and test theory. In addition to a brief discussion of the formal benefits of this formulation, we show how to formulate important parametric link models or skew parametric versions of other important models as EGLLVM. An important property of the EGLLVM is that every model can be interpreted through the underlying variable approach (cf. Remark 1.1.5). Hence, the 3PL model and many other guessing models are not included in the EGLLVM.

In Section 2.2 we show how to formulate guessing models using EGLLVM and discuss advantages of this model formulation. We also illustrate the application of one of these example models to a simulated and a real data set and compare it to a classic guessing model. These two examples show that difficulties of estimating the "third" item parameter also occur when applying EGLLVM (cf. e.g.the the general remark on "third" item parameters by Samejima, 1999, Section V and the discussion regarding the 3PL model in Subsection 1.1.2.3). This further indicates the need of a proof of the consistency and general asymptotic properties of the MMLE to show that it is eventually possible to estimate these parameters for an EGLLVM.

In Section 2.3 we then formulate precise conditions on all model components and prove the asymptotic existence, consistency, asymptotic normality and asymptotic efficiency of the MMLE (further details are given in Appendix A.2). As a by-product, we prove Fisher's identity for EGLLVM and the asymptotic normality and efficiency for the MMLE for many IRT models (the usual IRT model components satisfy the conditions directly). A direct conclusion of the asymptotic normality and efficiency of the MMLE is the asymptotic χ^2-distribution of the common asymptotic test statistics. We formulate the corresponding theorem in Section 2.4. Wald, score and in particular likelihood ratio (LR) tests can be used for the desired tests for skewness parameters.

Finally, we discuss three simulation studies regarding the frequentist fit of skew-probit-skew-normal IRT in Section 2.5. The first study focuses on the estimation of the skewness parameters of single items and a skewness parameter of the structural model. The study reveals that the estimation for skewness parameters of single items has issues which are similar to other models with a "third" item parameter. The second study examines the estimation of the ICCs and latent traits using models of different parsimony. This study reveals that a model with a common skewness parameter for all items and another skewness parameter for the structural model (i.e. a skew-probit-skew-normal model) provides a good trade-off between increased flexibility for incorporating skewness and model parsimony. In cases where a probit-normal model holds, the skew-probit-skew-normal model performs nearly identically well to the probit model such that the skew-probit-skew-normal model can be favorable. The third study examines model selection and LR tests. It povides evidence that the Akaike information criterion (AIC) and the LR test are reliable approaches for the link function and structural model selection. Hence, LR tests are appropriate for testing the skewness in an IRT model by comparing to parametric alternatives.

2.1 Model Definition

In the definition of GLLVM for binary variables (cf. Definition 1.1.3), we supposed that the d items may have d different link functions $\mathfrak{g}_1, \ldots, \mathfrak{g}_d$. To extend the class of models to parametric link functions, we now suppose that $\mathfrak{g}_i = \mathfrak{g}(\bullet\,; \boldsymbol{\gamma}_i)$ for some family of quantile functions with parameter vector $\boldsymbol{\gamma}_i \in \mathbb{R}^z$, $z \in \mathbb{N}$. Hence, we can interpret the parametric link functions as quantile functions of the error distributions in the underlying variable approach. Additionally, the distribution $\mathcal{H}$ can depend on a parameter vector $\boldsymbol{\delta} \in \mathbb{R}^b$, $b \in \mathbb{N}$. Hence, $\mathcal{H}$ can be a normal, a skew normal and another distribution depending on some shape parameter, too. Although we develop the model for a general family of forms and distributions of the latent vectors, our focus will be the skew normal distribution. Note that the model can be extended for the case of $\boldsymbol{\eta}_k \sim \mathcal{H}_k$, $k = 1, \ldots, n$, with different $\mathcal{H}_1, \ldots, \mathcal{H}_n$ which depend on the same parameter vector.

For the development of the asymptotic estimation and test theory, it is absolutely required to define the model and every component precisely. To prove the asymptotic normality and efficiency of the MMLE and the asymptotic χ^2-distribution of the Wald, score and LR statistics, the model is parameterized in a way such that the results of Lehmann and Casella (1998) and Lehmann and Romano (2005) can be applied in our proofs. Moreover, one of our main objectives of this chapter is to use the LR test for nested model testing to perform a conditional goodness of link test. In order to apply the results by Lehmann and Romano (2005) to prove the asymptotic distribution for the statistics of nested model comparisons, we need a specialized definition of nested models. Using this set-up, we can test link functions and latent variable distributions or select among models with different link functions and different latent variable distributions by formulating a more general alternative model, that contains all candidate models as special cases with specific model parameters. Therefor it is not sufficient to require that a nested model is defined on a parameter space which is a subset of the parameter space of the more complex model but also regularity regarding the embedding of the nested model into the complex model. With the following definition, we formulate these terms precisely. Subsequent to the definition, we briefly discuss advantages of these definitions and illustrate that these formal terms are appropriate to define many commonly used models or generalizations of them.[1]

Definition 2.1.1

(a) Let a latent vector $\boldsymbol{\eta} \in \Theta \subseteq \mathbb{R}^q$ and parameters $\boldsymbol{\delta} \in \Theta_1 \subseteq \mathbb{R}^b$, $\alpha_{i0} \in \mathbb{R}$, $\boldsymbol{\alpha}_i \in \mathbb{R}^q$, $\boldsymbol{\gamma}_i \in \Theta_2 \subseteq \mathbb{R}^z$, $i = 1, \ldots, d$, be given for $q, b, d, z \in \mathbb{N}$. The full parameter vector is set to

$$\begin{aligned} \boldsymbol{\vartheta} &= (\vartheta_1, \ldots \vartheta_u)^\intercal \\ &=: (\alpha_{10}, \ldots, \alpha_{d0}, \alpha_{11}, \ldots, \alpha_{dq}, \gamma_{11}, \ldots, \gamma_{1z}, \gamma_{21}, \ldots, \gamma_{dz}, \delta_1, \ldots, \delta_b)^\intercal \in \Theta^* \end{aligned} \tag{2.1}$$

with $\Theta^* := \mathbb{R}^{(q+1)d} \times \Theta_2^d \times \Theta_1$ and $u := (q+1)d + b + dz$. Suppose that a b-parametric distribution family $\mathcal{H}(\bullet\,; \boldsymbol{\delta})$, $\boldsymbol{\delta} \in \Theta_1$, is given with $\operatorname{supp}(\mathcal{H}(\bullet\,; \boldsymbol{\delta})) = \Theta$ independently of $\boldsymbol{\delta}$. Let mappings $P_1, \ldots, P_d \colon \Theta_0 \times \mathbb{R}^{q+1} \times \Theta_2 \to (0,1)$ be given by

$$P_i(\boldsymbol{\eta}) := P_i(\boldsymbol{\eta}; \alpha_{i0}, \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i) := F\big(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i\big), \qquad \boldsymbol{\eta} \in \Theta, \qquad i = 1, \ldots, d, \tag{2.2}$$

[1] Note that although we need the specific definitions for parameterization and nested models in order to apply the results by Lehmann and Casella (1998) and Lehmann and Romano (2005) they did not name the terms. Since these terms are important and allow a precise formulation, we now choose new names for them.

for a mapping $F\colon \mathbb{R} \times \Theta_2 \to (0,1)$, where $F(\,\bullet\,;\boldsymbol{\gamma})$ is a cdf on $\mathbb{R}$ for all $\boldsymbol{\gamma} \in \Theta_2$. Then we call the family of probability measures on $\big(\{0,1\}^{d\times n}, \mathsf{Pot}(\{0,1\}^{d\times n})\big)$, where Pot denotes the power set, defined by the following counting densities for all $\boldsymbol{\vartheta} \in \Theta^*$

$$P(\boldsymbol{y}_1,\dots,\boldsymbol{y}_n;\,\boldsymbol{\vartheta}) := \prod_{k=1}^{n} \int_{\Theta} \prod_{i=1}^{d} P_i(\boldsymbol{\eta})^{y_{ik}} \big(1 - P_i(\boldsymbol{\eta})\big)^{1-y_{ik}}\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta},\boldsymbol{\delta}),$$

for all $\boldsymbol{y}_k \in \{0,1\}^d$, $k = 1,\dots,n$, an *extended generalized linear latent variable model* (for binary data), *EGLLVM* for short. We denote the complete family by $\mathcal{M} = \mathcal{M}(\Theta^*)$ for the *parameter space* Θ^* and a single member by $\mathcal{M}(\boldsymbol{\vartheta})$ for a $\boldsymbol{\vartheta} \in \Theta^*$.

(b) Suppose $\mathcal{M}_1(\Theta_1^*)$ and $\mathcal{M}_2(\Theta_2^*)$ are two EGLLVMs for the same $d, n, u \in \mathbb{N}$ and two parameter spaces $\Theta_1^*, \Theta_2^* \subseteq \mathbb{R}^u$. We then write $\mathcal{M}_1(\Theta_1^*) \subseteq \mathcal{M}_2(\Theta_2^*)$ if for every $\boldsymbol{\vartheta}_1 \in \Theta_1^*$ there is a $\boldsymbol{\vartheta}_2 \in \Theta_2^*$ such that $\mathcal{M}_1(\boldsymbol{\vartheta}_1) = \mathcal{M}_2(\boldsymbol{\vartheta}_2)$. If $\mathcal{M}_1(\Theta_1^*) \subseteq \mathcal{M}_2(\Theta_2^*)$ and $\mathcal{M}_2(\Theta_2^*) \subseteq \mathcal{M}_1(\Theta_1^*)$, we say the models are *equivalent* and write $\mathcal{M}_1(\Theta_1^*) = \mathcal{M}_2(\Theta_2^*)$. If $\mathcal{M}_1(\Theta_1^*) \subseteq \mathcal{M}_2(\Theta_2^*)$ but **not** $\mathcal{M}_2(\Theta_2^*) \subseteq \mathcal{M}_1(\Theta_1^*)$, we say $\mathcal{M}_1(\Theta_1^*)$ is *nested* to $\mathcal{M}_2(\Theta_2^*)$ and write $\mathcal{M}_1(\Theta_1^*) \subset \mathcal{M}_2(\Theta_2^*)$.

(c) Suppose there are two EGLLVMs $\mathcal{M}_1(\Theta_1^*)$ and $\mathcal{M}_2(\Theta_2^*)$ with $\mathcal{M}_1(\Theta_1^*) \subset \mathcal{M}_2(\Theta_2^*)$. Then we say $\mathcal{M}_1(\Theta_1^*)$ is *properly nested* to $\mathcal{M}_2(\Theta_2^*)$ if there is a $u_0 \in \{1,\dots,u-1\}$ and $\boldsymbol{g} \in C^1(\Theta_2^*, \mathbb{R}^{u-u_0})$, i.e. a continuously differentiable $\boldsymbol{g}\colon \Theta_2^* \to \mathbb{R}^{u-u_0}$, such that

$$\Theta_1^* = \big\{\boldsymbol{\vartheta} \in \Theta_2^*\colon \boldsymbol{g}(\boldsymbol{\vartheta}) = \boldsymbol{0}\big\}$$

and the Jacobian of $\boldsymbol{g}$ has rank $u - u_0$ for all $\boldsymbol{\vartheta} \in \Theta_1^*$.

(d) Suppose $\mathcal{M}(\Theta^*)$ is an EGLLVM for $\Theta^* \subset \mathbb{R}^u$, $u \in \mathbb{N}$. Let $C^k(M_1, M_2)$ be the space of k-times continuously differentiable mappings $\mathbb{R}^{k_1} \supseteq M_1 \to M_2 \subseteq \mathbb{R}^{k_2}$, $k_1, k_2 \in \mathbb{N}$. We say $\mathcal{M}(\Theta^*)$ has a *regular parameterization of class* C^k for $k \in \mathbb{N}$ on $\Theta' \subseteq \mathbb{R}^{u_0}$ for a $u_0 \in \{1,\dots,u\}$ if there is a $\boldsymbol{\varphi} \in C^k(\Theta', \Theta^*)$ such that $\Theta^* = \boldsymbol{\varphi}(\Theta')$, Θ' has dimension u_0 and $\mathcal{J}\boldsymbol{\varphi}(\boldsymbol{\vartheta}) \neq \boldsymbol{0}$ for all $\boldsymbol{\vartheta} \in \Theta'$. We write $\mathcal{M}(\Theta', \boldsymbol{\varphi})$ for $\mathcal{M}(\Theta^*)$ and say that it has a *C^k-parameterization* for short.

(e) Suppose two EGLLVM $\mathcal{M}_1(\Theta_1^*)$ and $\mathcal{M}_2(\Theta_2^*)$ with $\mathcal{M}_1(\Theta_1^*) \subset \mathcal{M}_2(\Theta_2^*)$. We say $\mathcal{M}_1(\Theta_1^*)$ is a *C^k-regular nested* model of $\mathcal{M}_2(\Theta_2^*)$ if $\mathcal{M}_1(\Theta_1^*)$ is properly nested to $\mathcal{M}_2(\Theta_2^*)$ and has a C^k-parameterization where $\boldsymbol{\varphi}$ and $\boldsymbol{g}$ are defined for the same u_0.

(f) Suppose $\{\mathcal{M}_n(\Theta^*)\}_{n\in\mathbb{N}}$ is a sequence of EGLLVM for some $\Theta^* \subseteq \mathbb{R}^u$ and $u \in \mathbb{N}$ and $d \in \mathbb{N}$. If a sequence of independent random vectors $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}}$ over $\{0,1\}^d$ can be well defined for every $\boldsymbol{\vartheta} \in \Theta^*$, such that $(\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n) \sim \mathcal{M}_n(\boldsymbol{\vartheta})$, then we say $\{\mathcal{M}_n(\Theta^*)\}_{n\in\mathbb{N}}$ is an *increasing sequence of EGLLVM* and write $\mathcal{M}(\Theta^*, \mathbb{N}) := \{\mathcal{M}_n(\Theta^*)\}_{n\in\mathbb{N}}$. For a particular sequence $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}}$ for some $\boldsymbol{\vartheta} \in \Theta^*$, we write $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}} \sim \mathcal{M}(\boldsymbol{\vartheta}, \mathbb{N})$. If $\mathcal{M}_1(\Theta^*)$ has a C^k-parameterization φ on Θ', we further write $\mathcal{M}(\Theta', \boldsymbol{\varphi}, \mathbb{N})$ for $\mathcal{M}(\Theta^*, \mathbb{N})$. ◇

The stated concept of parameterizations allows the expression of nested models in the model formulation of (a). Due to this, GLLVM of Definition 1.1.3 with fixed structural models (e.g. $\mathcal{H} = \mathcal{N}_q(\boldsymbol{0}, \mathrm{I}_q)$) are nested models to certain EGLLVM. This is what the model formulation already intends. Moreover, we can examine both types of models or other nested models (like Rasch-type models, models with parametric link but fixed structural

models or vice versa) in this chapter with the same model formulation. Compare also Remark 1.3.6 on the asymptotics of the Rasch model with parametric variance of the latent variables. This simplifies the notation in the following sections a lot and enables a unified theoretical foundation.

In particular, for the asymptotic theory we consider here, the supposed *true* parameter vector has to have an open neighborhood in Θ. Suppose for example that Θ_1 and Θ_2 are open, $\mathbf{1} \in \Theta_2$ and we restrict

$$\Theta' := \left\{\boldsymbol{\vartheta} \in \Theta^* \colon \boldsymbol{\gamma}_i = \mathbf{1},\, i = 1, \ldots, d\right\}.$$

Then such open neighborhoods no longer exist with respect to the standard euclidean norm denoted by $\| \bullet \|$. But a C^k-parameterization, for any $k \in \mathbb{N}$, for the alternative restricted parameter space and open subset of $\mathbb{R}^{(q+1)d+b}$, $\Theta'' = \mathbb{R}^{(q+1)d} \times \Theta_1$, is given by the mapping $\boldsymbol{\varphi} \colon \Theta'' \to \Theta^*$,

$$\boldsymbol{\vartheta} \mapsto (\vartheta_1, \ldots, \vartheta_{(q+1)d}, 1, \ldots, 1, \vartheta_{(q+1)d+1}, \ldots, \vartheta_{(q+1)d+b})^\intercal.$$

Further notice that these parameterizations include identification constraints. This leads to equivalent models on a parameter space which is a subset of a lower dimensional vector space. Of course, they still have the same Hausdorff-dimension which means that regardless of the shape of Θ'' and Θ^* and if they are linear spaces, they have the same geometric distribution. Hence, they always behave as we intuitively expect from reparameterizations. Before we consider some illustrative examples, we will give some technical remarks.

Remark 2.1.2

(i) If there is a C^k-parametrization $\mathcal{M}(\Theta', \boldsymbol{\varphi})$ for a model $\mathcal{M}(\Theta^*)$ and some $k \in \mathbb{N}$ and $\boldsymbol{\varphi}$ is an homeomorphism, then it practically is a re-parameterization of the same model as an equivalent model and not a new parameterization of a model as EGLLVM in Definition 2.1.1 (a). A trivial example of such a re-parameterization is the change of α_{i0} to $\alpha'_{i0} := 2\alpha_{i0}$, for $i = 1, \ldots, d$. This is especially true if we select $\boldsymbol{\varphi} = \mathrm{id}$, which leads to exactly the same model. However, it is important to include the option of a re-parameterization due to the following reasons:

  a) In practical settings, a re-parameterization corresponds to the switch of equivalent identification constraints.

  b) The case $\mathcal{M}(\Theta^*, \mathrm{id})$ enables the unified handling of the most complex model considered and all its nested models by writing $\mathcal{M}(\Theta', \boldsymbol{\varphi})$. We therefore do not need to distinguish between $\mathcal{M}(\Theta^*)$ and $\mathcal{M}(\Theta', \boldsymbol{\varphi}) \subset \mathcal{M}(\Theta^*)$ in the notation.

(ii) Due to Definition 2.1.1 (b), we can restrict the model class to equivalence classes of models and do not distinguish explicitly between $\mathcal{M}_1(\Theta_1^*)$ and $\mathcal{M}_2(\Theta_2^*)$ if $\mathcal{M}_1(\Theta_1^*) = \mathcal{M}_2(\Theta_2^*)$. ◇

To get a better understanding of all model components, we consider two special cases.

Example 2.1.3

Denote by Φ_γ, $\gamma \in \mathbb{R}$, the CDF of the standard skew normal distribution of Definition 1.1.7 and by ϕ_γ its pdf.

(a) Let $\mathcal{H}$ be given by the family

$$\mathsf{P}(\eta_1 < a_1, \ldots, \eta_q < a_q) = \prod_{j=1}^{q} \Phi_{\gamma_j}(a_j), \qquad \text{for all } \boldsymbol{a} \in \mathbb{R}^q,$$

of distributions. Then an IRT model with standard skew probit link and multivariate standard skew normal distributed latent vector is given by

$$\mathsf{P}(Y_{ik} = 1 \mid \boldsymbol{\eta}) = \Phi_{\gamma_i}\left(\alpha_{01} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}\right), \qquad \boldsymbol{\eta} \in \mathbb{R}^q,\ i = 1, \ldots, d,$$

$$\mathsf{P}(\boldsymbol{Y}_k = \boldsymbol{y}_k) = \int_{\mathbb{R}^q} \left(\prod_{i=1}^{d} \Phi_{\gamma_i}\left(\alpha_{01} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}\right)^{y_{ik}} \left(1 - \Phi_{\gamma_i}\left(\alpha_{01} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}\right)\right)^{1-y_{ik}} \right) \prod_{j=1}^{q} \phi_{\delta_j}(\eta_j)\, \mathrm{d}\boldsymbol{\eta},$$

for all $\boldsymbol{y}_k \in \{0,1\}^d$ and $k = 1, \ldots, n$. In particular, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are iid. The frequentist version of the model by Bazán et al. (2014) is obtained by restricting the model of this example to $q = 1$. The model discussed by Samejima (2000) is obtained by restricting to $q = 1$, setting $\delta_1 = 0$ and replacing the skew probit link in all items by a power logit link. If we replace all link functions by a scobit link, we get the frequentist version of a model by Bolfarine and Bázán (2010).

(b) Consider the notation of (a) with $q = 2$. Suppose that we have the following for a $d_1 \in \{1, \ldots, d-1\}$ with $d_2 := d_1 + 1$ for all $k = 1, \ldots n$:

(i) $Y_{1k}, \ldots, Y_{d_1 k}$ are conditionally independent given η_{1k} with

$$(Y_{ik} \mid \eta_1 = u) \sim \mathscr{B}\left(1, \Phi_{\gamma_1}\left(\alpha_{i0} + \alpha_{i1} u\right)\right), \qquad u \in \mathbb{R},\ i = 1, \ldots, d_1,$$

for $\gamma_1, \alpha_{10}, \ldots, \alpha_{d_1,0} \in \mathbb{R}$, $\alpha_{11}, \ldots, \alpha_{d_1 1} \in (0, \infty)$.

(ii) $Y_{d_2 k}, \ldots, Y_{dk}$ are conditionally independent given η_{2k} with

$$(Y_{ik} \mid \eta_2 = u) \sim \mathscr{B}\left(1, \Phi_{\gamma_2}\left(\alpha_{i0} + \alpha_{i1} u\right)\right), \qquad u \in \mathbb{R},\ i = d_2, \ldots, d,$$

for $\gamma_1, \alpha_{d_2 0}, \ldots, \alpha_{d,0} \in \mathbb{R}$, $\alpha_{d_2 1}, \ldots, \alpha_{d1} \in (0, \infty)$.

(iii) η_{1k} and η_{2k} are conditional independent given a further latent variable η_{3k} with

$$\left(\begin{pmatrix} \eta_{1k} \\ \eta_{2k} \end{pmatrix} \,\middle|\, \eta_{3k} = u \right) \sim \mathcal{N}_2(u\boldsymbol{b}_1, \boldsymbol{\Sigma}_1), \qquad u \in \mathbb{R},$$

for $\boldsymbol{b}_1 \in \mathbb{R}^2$ and diagonal matrix $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{2\times 2}$ with positive diagonal.

(iv) η_{3k} has a standard skew normal distribution with cdf Φ_{b2} for $b_2 \in \mathbb{R}$.

The complete hierarchical model consists of lower dimensional sub-models which are typically easy to interpret. In this case, there is a group of two items which depend on a single, but different latent variable. Furthermore, both groups are supposed to have a common error distribution in this case. The two latent variables are conditionally independent given a further latent variable, which has a parametric skew distribution. ◇

2.2 Example: Modeling of Guessing Behavior

In Subsubsection 1.1.2.3, we discussed that possible guessing behavior of test participants in a multiple choice test is a serious issue in the modeling of an ICC. In particular, we discussed that the most popular model for this setting, the 3PL model, is problematic. Due to this, several alternatives were discussed so far. However, from our point of view, there is a further critical aspect which was not taken into account in the discussion so far.

Since an ICC of the 3PL model with guessing parameter > 0 has a lower asymptote which is greater than zero, the 3PL model cannot be interpreted by the underlying variable approach. In particular, the link function cannot be considered as the quantile function of the error distribution. That is a notable disadvantage of the 3PL model in comparison to the (E)GLLVM because they always possess this convenient attribute. Some alternatives, such as that of Pashley (1991) share this property of the 3PL. Other alternatives, such as that of San Martín et al. (2006) and Lee and Bolt (2018), are generally difficult to interpret from this point of view due to their nonlinear formulation.[2] However, the model of Lee and Bolt (2018) is a flexible skew but light tailed model, i.e. we reach a success probability of zero (or one, respectively) by the same order as the usual GLLVM. San Martín et al. (2006) presented a model where a parameter decides whether it is a Rasch-type guessing model (i.e. the 3PL model with equal discrimination) or a model with more mass on the left tail of the ICCs than a usual Rasch model has. However, it still has light tailed ICCs, i.e. even if guessing, test participants with arbitrarily small ability do not have a success probability which is greater than that provided by a Rasch or 2PL model. Another alternative model which was published by Cao and Stokes (2008) captures interesting response strategies, but the resulting ICCs are of the 2PL model structure. These types of models provide qualitatively different structures. All of the models above can be very useful in specific contexts. However, there is no model in the literature of models with guessing which

(i) captures the idea of the 3PL model that even test participants with arbitrarily small ability have a success probability which is significantly above zero (due to guessing),

(ii) incorporates the idea of ability based guessing in the sense that differently skilled test participants have different success probabilities even if they guess,

(iii) can be interpreted for all parameter values by the underlying variable approach and the ICC does not change drastically in its qualitative properties if the model parameters differ slightly (i.e. the ICC is always a cdf and not only for some parameter values).

From our point of view, a heavy (or, at least "heavier") left tail of the error distribution captures the notion of the possibility to answer correctly by a comparably high probability by guessing exactly. Simultaneously, as the tail is still decreasing to zero, the chance of guessing correctly decreases, so the ability still influences the success probability. It is of course mandatory that there is information in the data to capture the properties of a model such that not arbitrary general properties are formulated for a model. Next we discuss an approach for such a model.

Sometimes in binary regression, we observe response probabilities significantly different from 0 and 1 at extreme values of the linear predictor. The Cauchy distribution is an appropriate candidate for the link function in such cases because it is heavy tailed. We will now present some simple new constructions for link functions, which combine the curvatures of the Cauchy and the normal distribution to obtain links. This enables us to define new models that are similar to 3PL-type probit models but have the theoretical properties (i)–(iii). Subsequently, we will formulate a corresponding IRT model and illustrate its application to a simulated and a real data set. We will compare the fit by the new model with a Rasch-type guessing model to get an impression of its statistical properties. The

[2] San Martín et al. (2006) and Lee and Bolt (2018) discuss the interpretation of the model parameters, such as the item difficulty.

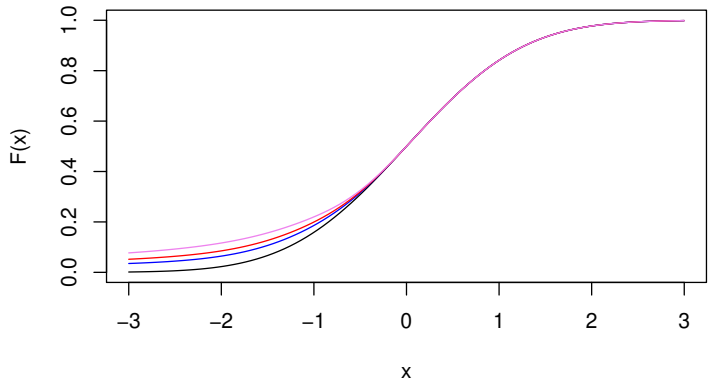

Figure 2.1: *Members of the family of distributions defined by* $F_4(\bullet\,; 1\,, \gamma)$ *for* $\gamma = 0$ *(black),* $\gamma = \frac{1}{3}$ *(blue),* $\gamma = \frac{1}{2}$ *(red) and* $\gamma = \frac{3}{4}$ *(violet).*

examples show difficulties of estimating the guessing parameters. This indicates the strong need for a theoretical discussion of the estimation theory, which we will discuss in Section 2.3.

2.2.1 Heavy Tailed Link Functions for Ability Based Guessing

First, we define the cdf of the Cauchy distribution:

$$F_{\mathcal{C}} \colon \mathbb{R} \to (0,1), \qquad x \mapsto \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Using $F_{\mathcal{C}}$ and Φ, $\Phi(\bullet\,; -0.5, 0.4)$ – the cdfs of $\mathcal{N}(0,1)$ and $\mathcal{N}(-0.5, 0.4)$ – , we can define the following two families $\{F_3(\bullet\,; \gamma_1)\}_{\gamma_1 \in (0,\infty)}$ and $\{F_4(\bullet\,; \boldsymbol{\gamma})\}_{\boldsymbol{\gamma} \in (0,\infty)\times[0,\infty)}$ of cdfs on $\mathbb{R}$:

$$\begin{aligned} F_3(x; \gamma_1) &\coloneqq \Big(F_{\mathcal{C}}(x)(1 - \Phi(x; -0.5, 0.4)) + \Phi(x)\Phi(x; -0.5, 0.4)\Big)^{\gamma_1}, \\ F_4(x; \boldsymbol{\gamma}) &\coloneqq F_4(x; \gamma_1, \gamma_2) \coloneqq \gamma_2 F_3(x; \gamma_1) + (1 - \gamma_2)\Phi(x), \end{aligned}$$

for all $x \in \mathbb{R}$, $\gamma_1 \in (0, \infty)$, $\gamma_2 \in [0, 1]$. The cdf $F_3(\bullet\,; 1)$ for all practical purposes behaves like the standard normal distribution for $x \geq 0$ and like the standard Cauchy distribution for $x \leq -1$ with a smooth transition on $(-1, 0)$ – meaning it has a heavy left tail. The cdfs $F_3(\bullet\,; \gamma_1)$ have arbitrarily heavy left tails for $\gamma_1 < 1$ and $\gamma_1 \to 0$. For $\gamma_1 > 1$ the distributions get lighter left tails than the Cauchy distribution which become arbitrarily light tailed for $\gamma_1 \to \infty$. If γ_1 is fixed to an arbitrary value in $(0, \infty)$ then the cdf $F_3(\bullet\,; \gamma_1, \gamma_2)$ equals the standard normal cdf for $\gamma_2 = 0$ and $F_3(\bullet\,; \gamma_1)$ for $\gamma_2 = 1$. If in particular $\gamma_1 = 1$, then F_4 changes practically only on $(-\infty, 0)$. Compare Figure 2.1 and Figure 2.2 for an impression of various members of the parametric families of distributions.

We propose the following family of parametric link functions as an alternative basis to the 3PL model for the modeling of the increased success probability for low latent traits based on guessing:

$$\operatorname{lhtg}_c(x; \gamma) \coloneqq F_4^{-1}(x; c, \gamma), \qquad x \in \mathbb{R},\ \gamma \in [0, 1], \tag{2.3}$$

for a fixed $c \in (0, \infty)$, where $F_4^{-1}(\bullet\,; c, \gamma)$ denotes the quantile function of $F_4(\bullet\,; c, \gamma)$ and lhtg stands for *left-heavy-tailed-guessing*. Principally, c can be handled as a parameter as

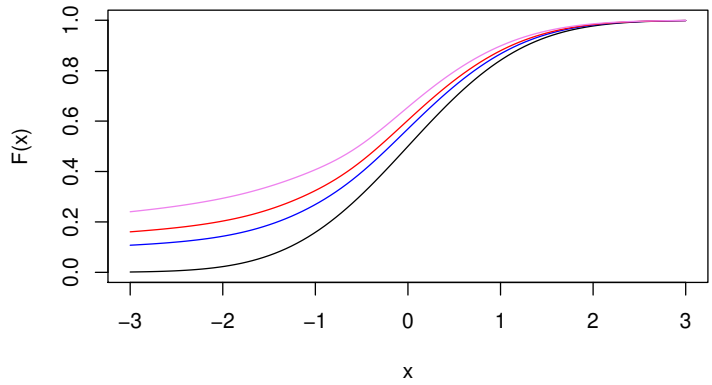

Figure 2.2: *Members of the family of distributions defined by* $F_4(\bullet\,; 0.5\,, \gamma)$ *for* $\gamma = 0$ *(black),* $\gamma = \frac{1}{3}$ *(blue),* $\gamma = \frac{1}{2}$ *(red) and* $\gamma = \frac{3}{4}$ *(violet).*

well. However, we prefer the setting of a fixed $c \in (0, 1]$ for the handling of a reasonable upper bound of the heavy-tailedness. Thereby we increase the model parsimony and decrease the space of computational difficulties while fitting.

2.2.2 Application to a Simulated and a Real Data Set

In this section, we will briefly illustrate the application of an IRT model using the lhtg link with fitting by marginal maximum likelihood. We examine the Rasch-type model

$$F_4^{-1}\big(\mathsf{P}(Y_i = 1 \mid \eta); 1, \gamma_i^{\text{lhtg}}\big) = \alpha^{\text{lhtg}}(\eta - \beta_i^{\text{lhtg}}), \qquad i = 1, \ldots, d, \tag{2.4}$$

with the parameterization of the Lawley–Lord and 2PL model for $d \in \mathbb{N}$ items and a single latent variable $\eta \in \mathbb{R}$. Hence, each item has its own parameter $\gamma_i^{\text{lhtg}} \in [0, 1]$ indicating the heavy tailedness in the sense of the proportion of a left Cauchy-tail. For this model, we set $c = 1$ for the link function (cf. equation (2.3)). Due to this, the median of the error distribution is zero for each $\gamma_i^{\text{lhtg}} \in \mathbb{R}$. Hence, β_i^{lhtg} can be interpreted as a difficulty parameter as usual. We still consider the common α^{lhtg} as a discrimination parameter and assume $\alpha^{\text{lhtg}} > 0$ but treat it as unknown. For simplicity, we consider $\eta \sim \mathcal{N}(0, 1)$. We fit the model using MMLE with the Gauss-Hermite quadrature with 30 quadrature points and a quasi-Newton method which incorporates bounds to ensure the constraints of the parameter set of the R function `optim` (the limited-memory Broyden–Fletcher–Goldfarb–Shanno method for bound constraints (L-BFGS-B), cf. Byrd et al., 1995). Further aspects of our implementation in R are discussed in Section C.1 in the appendix. For comparison, we also fit a Rasch-type-guessing (rtg) model with

$$\mathsf{P}(Y_i = 1 \mid \eta) = c_i^{\text{rtg}} + \frac{1 - c_i^{\text{rtg}}}{1 + \exp(-1.7017 \cdot \alpha^{\text{rtg}}(\eta - \beta_i^{\text{rtg}}))}, \qquad i = 1, \ldots, 5,\ \eta \in \mathbb{R}, \tag{2.5}$$

using the R-package **ltm** (cf. Rizopoulos, 2006).

First, we consider one simulated data set of $d = 5$ items with $n = 2000$ observed responses. This data set is generated by model (2.4), where we consider that 2 items have reasonable influence of guessing and the remaining items follow a Rasch-type probit model. For both models, the observed and estimated expected frequencies are displayed in Table 2.3. The

i	β_i^{lhtg}	γ_i^{lhtg}	$\hat{\beta}_i^{\text{lhtg}}$	$\hat{\gamma}_i^{\text{lhtg}}$	$\hat{\beta}_i^{\text{rtg}}$	$\hat{c}_i^{\text{rtg}}$
1	1.0	0.0	1.033129	0.136665	1.033067	0.012333
2	0.8	0.0	0.775286	0.021855	0.770888	0.000626
3	0.0	0.5	0.019101	0.011561	0.030899	0.007110
4	1.0	0.8	0.951486	0.734675	1.045507	0.089912
5	−0.5	0.0	−0.412455	1.000000	−0.269525	0.130233

Table 2.1: *True and estimated item parameters for one simulated data set. True discrimination:* $\alpha^{lhtg} = 1$. *Estimated discrimination:* $\hat{\alpha}^{lhtg} = 1.043011$ *and* $\hat{\alpha}^{rtg} = 1.056772$.

model	AIC	X^2	G^2	df	$p^{X^2}_{\text{asymptotic}}$	$p^{G^2}_{\text{asymptotic}}$	$p^{X^2}_{\text{bootstrap}}$	$p^{G^2}_{\text{bootstrap}}$
lhtg	11357.13	13.307	13.303	20	0.864	0.864	0.908	0.914
rtg	11358.73	14.789	14.894	20	0.788	0.782	0.851	0.854

Table 2.2: *The AIC and the goodness of fit statistics with their corresponding p-values for the simulated data set. For the bootstrap-p-values, a parametric bootstrap with 1000 simulated data sets is used.*

AIC and the goodness of fit (GoF) statistics X^2 and G^2 for the corresponding 2^5 contingency table are given in Table 2.2. Due to the estimated expected frequencies of both models, we can approximate the p-values using an asymptotic χ^2-distribution. Additionally, we computed p-values using a parametric bootstrap. Both models provide a good fit and are similarly good in terms of the AIC. The true parameter and the estimated parameter values are given in Table 2.1. We see a good recovery of the parameter values by the fit of the lhgt model with the exception of γ_3 and γ_5 – to discuss this aspect, we compare the parameter estimates of both models model. Due to the formulation of the rtg model in (2.5), we can compare the estimates of α^{rtg} and $\beta_1^{\text{rtg}}, \ldots, \beta_5^{\text{rtg}}$ with those of α^{lhtg} and $\beta_1^{\text{lhtg}}, \ldots, \beta_5^{\text{lhtg}}$. We can see that they are similar except $\hat{\beta}_5^{\text{rtg}}$, which is slightly worse compared to $\hat{\beta}_5^{\text{lhtg}}$. The direct comparison of the γ_i^{lhtg} values with the c_i^{rtg} values is not reasonable. But we can still see that $\hat{c}_3^{\text{rtg}}$ and $\hat{c}_5^{\text{rtg}}$ have a similar error to $\hat{\gamma}_3^{\text{lhtg}}$ and $\hat{\gamma}_5^{\text{lhtg}}$. This indicates that the realized data are a bit misleading in order to estimate the guessing structure or there is not enough information to allow a reliable estimation of the guessing. For a better comparison, consider the visualization of the true ICCs and their estimates in Figure 2.3. We see that the rtg model performs overall similarly to the lhtg model. In particular, the guessing of item 3 is under- and that of item 5 is overestimated. Interestingly, the estimated lower asymptote of the items 4 and 5 are even above the ICC of the lhtg model in the range of $\eta \in (-4, -2)$ and not at their average (to set this in relation, the realized values of η lie between ≈ -3.213 and ≈ 3.495).

We can summarize this example with the observation that the lhtg model provides a similar fit to the rtg model. In particular, we can observe that the lhtg model has fitting issues which are similar to the often reported fitting issues of the 3PL model and its special cases (like the rtg model we consider here).[3]

[3] Additional to the theoretical results that follow in the next section, increasing the sample size in this simulation setting makes the consistency of the MMLE for the parameters of the lhtg model visible. We can therefore conclude that a large sample size is necessary for good model fits but that they are still possible.

Response Pattern	TS	OF	lhtg EF	rtg EF	Response Pattern	TS	OF	lhtg EF	rtg EF
0 0 0 0 0	0	397	396.8	392.1	1 1 1 0 0	3	7	7.6	8.0
1 0 0 0 0	1	19	22.7	23.3	1 1 0 1 0	3	2	1.8	1.8
0 1 0 0 0	1	27	28.5	29.3	1 0 1 1 0	3	5	6.7	7.0
0 0 1 0 0	1	110	113.4	113.5	0 1 1 1 0	3	10	10.0	10.5
0 0 0 1 0	1	53	57.6	58.2	1 1 0 0 1	3	13	17.0	16.2
0 0 0 0 1	1	270	269.5	275.7	1 0 1 0 1	3	71	63.4	62.0
1 1 0 0 0	2	5	4.4	4.3	0 1 1 0 1	3	96	96.1	94.5
1 0 1 0 0	2	22	16.5	16.5	1 0 0 1 1	3	15	14.9	14.4
0 1 1 0 0	2	30	23.9	24.0	0 1 0 1 1	3	21	22.2	21.6
1 0 0 1 0	2	6	5.3	5.2	0 0 1 1 1	3	82	83.3	82.5
0 1 0 1 0	2	7	7.2	7.1	1 1 1 1 0	4	4	5.5	6.9
0 0 1 1 0	2	31	27.6	27.2	1 1 1 0 1	4	61	69.5	69.1
1 0 0 0 1	2	40	35.9	35.5	1 1 0 1 1	4	14	13.4	13.8
0 1 0 0 1	2	53	51.5	51.0	1 0 1 1 1	4	43	53.3	52.4
0 0 1 0 1	2	190	193.0	195.4	0 1 1 1 1	4	80	82.6	81.5
0 0 0 1 1	2	69	61.6	61.3	1 1 1 1 1	5	147	137.0	138.2

Table 2.3: *Observed frequencies (OF), total scores (TS) and estimated expected frequencies (EF) by both models of a simulated example data set.*

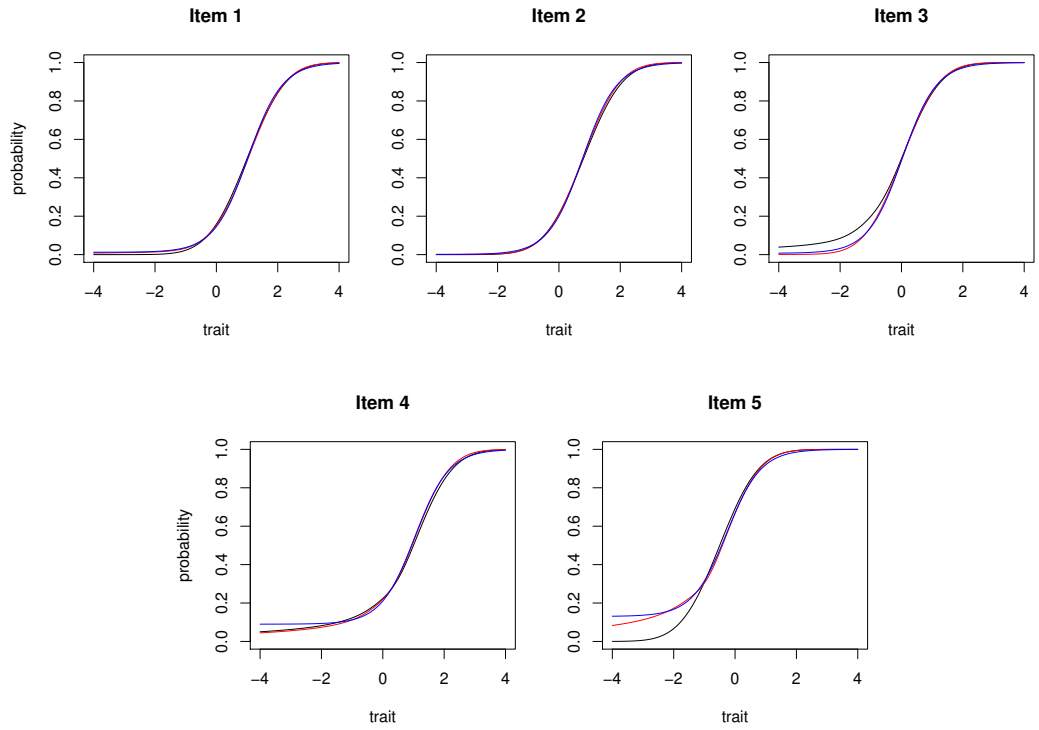

Figure 2.3: *The true ICCs (black) and their estimates by the rtg model (blue) and the lhtg model (red) for the simulated data.*

model	AIC	X^2	G^2	df	$p^{X^2}_{\text{asymptotic}}$	$p^{G^2}_{\text{asymptotic}}$	$p^{X^2}_{\text{bootstrap}}$	$p^{G^2}_{\text{bootstrap}}$
lhtg	4956.32	18.143	22.242	20	0.578	0.328	0.824	0.704
rtg	4955.461	17.933	21.383	20	0.592	0.375	0.63	0.533
Rasch	4945.875	18.345	21.798	25	0.827	0.647	0.639	0.509

Table 2.4: *The AIC and the goodness of fit statistics with their corresponding p-values for the LSAT data set. For the bootstrap-p-values, a parametric bootstrap with 1000 simulated data sets is used.*

Response Pattern	TS	OF	lhtg EF	rtg EF	Rasch EF	Response Pattern	TS	OF	lhtg EF	rtg EF	Rasch EF
0 0 1 1 1	0	3	2.4	2.1	2.4	1 1 1 0 0	3	11	8.8	8.6	8.6
1 0 0 0 0	1	10	9.8	9.3	10.3	1 1 0 1 0	3	21	25.0	25.6	25.0
0 1 1 1 0	1	1	2.0	1.9	1.8	1 0 1 1 0	3	15	11.9	11.7	11.7
0 0 0 0 1	1	1	0.9	0.7	0.9	0 1 1 1 0	3	2	2.1	2.0	2.1
1 1 0 0 1	1	2	2.6	2.5	2.5	1 1 0 0 1	3	56	55.1	56.1	55.2
0 0 0 1 0	1	6	5.5	5.8	5.5	1 0 1 0 1	3	28	26.2	25.5	25.8
1 1 1 1 0	2	16	11.5	11.4	11.4	0 1 1 0 1	3	3	4.6	4.4	4.6
0 0 1 1 0	2	3	5.5	4.9	5.3	1 0 0 1 1	3	81	74.6	76.1	75.1
1 0 1 1 1	2	0	1.0	0.9	0.9	0 1 0 1 1	3	16	13.1	13.4	13.3
1 1 0 1 0	2	14	15.4	15.6	15.5	0 0 1 1 1	3	4	6.3	6.0	6.2
1 0 1 1 0	2	0	2.9	2.9	2.7	1 1 1 1 0	4	28	26.9	28.5	27.7
1 1 1 0 0	2	3	1.4	1.2	1.3	1 1 1 0 1	4	61	61.4	61.3	61.2
1 0 1 0 0	2	29	33.1	34.7	34.2	1 1 0 1 1	4	173	179.3	174.1	177.9
0 0 0 1 1	2	8	6.3	6.5	6.1	1 0 1 1 1	4	80	84.1	83.1	83.3
0 1 0 0 0	2	1	3.0	2.8	2.8	0 1 1 1 1	4	15	13.6	13.8	14.7
0 1 1 0 1	2	11	8.4	8.9	8.2	1 1 1 1 1	5	298	295.4	297.4	295.8

Table 2.5: *Observed frequencies (OF), total scores (TS) and estimated expected frequencies (EF) by all three models of the Law School Admission Test (VI) data (LSAT).*

As a second example, we study the data of the law school admission test (VI) (LSAT for short), which is a popular example data set for IRT in educational applications. The data set consists of $d = 5$ items, too, and has $n = 1000$ observed responses. In this example, we again fit model (2.4) and a Rasch model with unknown discrimination and the same Rasch-type-guessing type model as before for comparison. We fit the Rasch and the Rasch-type-guessing model using the **ltm** package. The LSAT data can be fitted well by a Rasch model with parametric discrimination (cf. Rizopoulos, 2006). The parsimonious Rasch ICCs therefore are our benchmark. The observed frequencies and the estimated expected frequencies of all three models are given in Table 2.5. The AIC of the three models and their GoF statistics X^2 and G^2 for the underlying 2^5 contingency table are given in Table 2.4. We state the asymptotic p-values for the sake of completeness but they are of questionable reliability here. The obtained p-values using the parametric bootstrap support the good fit by the Rasch model as well as a similar and generally good fit by the lhtg and rtg models, again.

In Figure 2.4, we see the estimated ICCs of the three models. While the ICCs for item

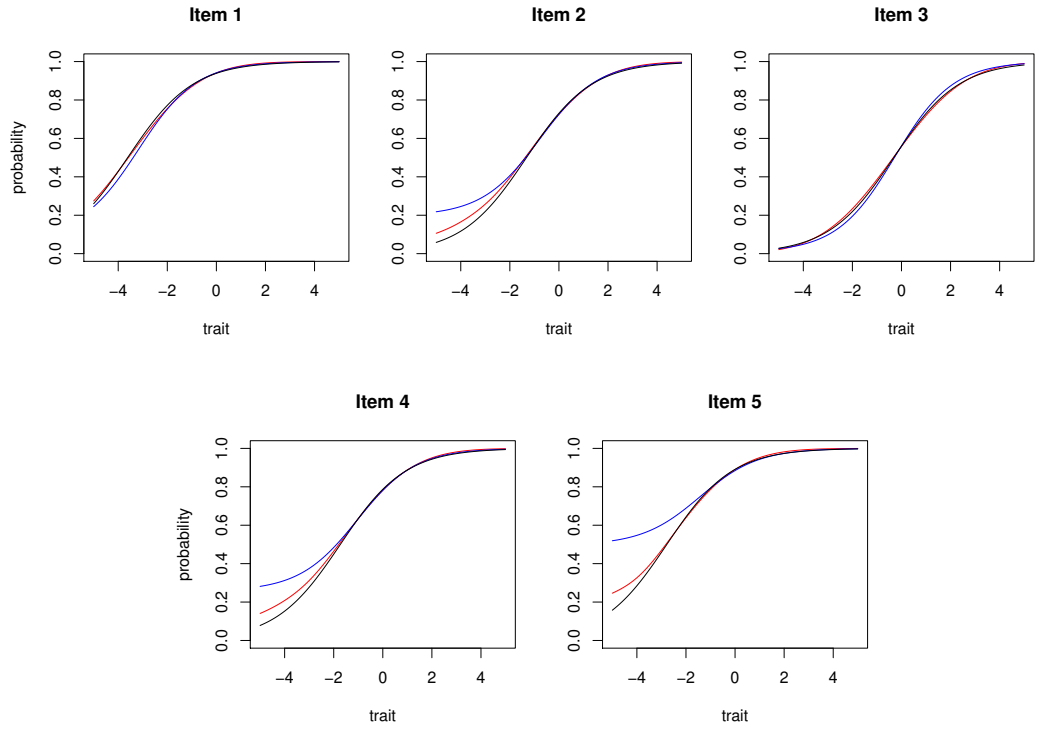

Figure 2.4: *Estimated ICCs of the Law School Admission Test (VI) data (LSAT) by the rtg (blue), the lhtg (red) and the Rasch model (black).*

1 and 3 do not differ greatly, we recognize a difference in the remaining items 2, 4 and 5. However, in all three cases, the Rasch-type-guessing model has a larger difference to the Rasch model than the lhtg model. In particular, the Rasch-type-guessing model has an estimated guessing parameter of ≈ 0.496 for item 5, which is questionablly large. For the same item, we obtained $\hat{\gamma}_5 \approx 0.992$ for the lhtg model, which is still below 1, but nearly as large as possible. It seems that item 5 is difficult to fit by guessing models and an upper bound can improve the fit. Overall, the estimated ICCs by the lhtg model are appropriate for the LSAT data.

Both examples generally illustrate that the left heavy tailed link function can be used for guessing models, which then provides the advantage of the interpretation through the underlying variable approach. However, in this example we see that the parameters of the link function can be difficult to estimate. This is an observation which is often reported for parametric links (cf. e.g. Samejima, 1999, who mentioned in Section V general difficulties when estimating some "third" item parameter when discussing her alternative to the 3PL model and cf. also the discussion regarding the 3PL model in Subsection 1.1.2.3). This circumstance shows the absolute necessity of proving the consistency of the MMLE to ensure that it is eventually possible to estimate the parameters if enough data is collected. We will tackle the asymptotic properties of the MMLE for EGLLVM next.

2.3 Marginal Maximum Likelihood Estimation

Following the principal concept of fitting of classical random effect IRT models with fixed link, we suggest the parameter estimation by MMLE as we have done in the examples in Subsection 2.2.2. In Subsection 1.3.2, we discussed that the focus of IRT research is mostly on computational aspects and further development of computationally efficient algorithms for estimation. Theoretical aspects are rarely focused on (one exception is e.g. Butler and Louis, 1997, who proved the consistency of MMLE for a large class of binary mixture models). In particular, there is no proof of the asymptotic normality and efficiency of the MMLE for any IRT model to the best of our knowledge. Since these asymptotic properties are some of the most important foundations for statistical inference for parametric IRT, they have to be proven.

Motivated by this fact, we do not focus on the computational aspects of MMLE but on theoretical ones here. In Subsection 2.3.1, we prove that *Fisher's identity* holds for a large class of EGLLVM for binary data. This well known identity is important for theoretical as well as for computational questions, including questions in relation to the EM-algorithm. In Subsection 2.3.2, we formulate some additional conditions which ensure that the MMLE for an EGLLVM is asymptotically normal distributed and efficient. The class of appropriate EGLLVM includes many well known IRT models (recall also Remark 1.3.6, which summarizes the asymptotic normality for the Rasch model with $\mathcal{H} = \mathcal{N}(0, \sigma^2)$ as a conclusion from the following theory). In our theory, $\mathcal{H}$ can be selected from a large class of distributions (e.g. arbitrary members of the exponential family).

2.3.1 Preliminaries and Fisher's Identity

The conditions to ensure that Fishers identity is true for a EGLLVM are:

(PLM1) The family of distributions $\{\mathcal{H}(\,\bullet\,;\boldsymbol{\delta})\}_{\boldsymbol{\delta}\in\Theta_1}$ is dominated by a σ-finite measure $\mathscr{G}$ on the measurable space $(\Theta, \mathcal{B}(\Theta))$, where Θ is the support of $\mathcal{H}(\,\bullet\,;\boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \Theta_1 \subseteq \mathbb{R}^b$ and $k = 1, \ldots, n$, for $n, b \in \mathbb{N}$. Furthermore,

$$\int_{\Theta} \|\boldsymbol{\eta}\|\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}, \boldsymbol{\delta}) < \infty, \qquad \text{for all } \boldsymbol{\delta} \in \Theta_1. \tag{2.6}$$

(PLM2) We denote by $\mathfrak{h}(\,\bullet\,;\boldsymbol{\delta})$ the Radon–Nikodym density of $\mathcal{H}(\,\bullet\,;\boldsymbol{\delta})$ with respect to $\mathscr{G}$. $\mathfrak{h}(\boldsymbol{\eta};\,\bullet\,) \in C^1(\Theta_1, \mathbb{R})$, for all $\boldsymbol{\eta} \in \Theta$, and $\partial_{\delta_g}\mathfrak{h}\colon \Theta \times \Theta_1 \to \mathbb{R}$ is defined by

$$(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) \mapsto \partial_{\delta_g}\mathfrak{h}(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) := \left.\frac{\partial \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})}{\partial \delta_g}\right|_{(\boldsymbol{\eta},\boldsymbol{\delta})=(\boldsymbol{\eta}_1,\boldsymbol{\delta}_1)}.$$

Furthermore,

$$\int_{\Theta} \sup_{\boldsymbol{\delta}\in O(\boldsymbol{\delta}_1,g,\varepsilon)} \left|\partial_{\delta_g}\mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\right| \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) < \infty, \tag{2.7}$$

for some $\varepsilon = \varepsilon(\delta, g) > 0$ such that $O(\boldsymbol{\delta}_1, g, \varepsilon) \subset \Theta_1$, for all $\boldsymbol{\delta}_1 \in \Theta_1$ and $1 \le g \le b$, where

$$O(\boldsymbol{\delta}, g, \varepsilon) := \{(\delta_1, \ldots, \delta_{g-1})\} \times (\delta_g - \varepsilon, \delta_g + \varepsilon) \times \{(\delta_{g+1}, \ldots, \delta_b)\}. \tag{2.8}$$

(PLM3) There is a $F(\bullet\,;\bullet) \in C^1(\mathbb{R} \times \Theta_2; (0,1))$ such that

$$\mathsf{P}(Y_{ik} = 1 \mid \boldsymbol{\eta}) = P_i(\boldsymbol{\eta}) = F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i), \qquad \boldsymbol{\eta} \in \Theta,\ i = 1, \ldots, d,$$

for $(\alpha_{i0}, \ldots, \alpha_{iq})^\intercal \in \mathbb{R}^{q+1}$, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d \in \Theta_2 \subset \mathbb{R}^z$ and $k = 1, \ldots, n$. Additionally, F satisfies the following properties:

(a) The mapping $F(\bullet\,;\boldsymbol{\gamma})$ is a cumulative distribution function for all $\boldsymbol{\gamma} \in \Theta_2$,

(b) For all $x_1 \in \mathbb{R}$ and all $\boldsymbol{\gamma}_1 \in \Theta_2$:

$$f(x_1; \boldsymbol{\gamma}_1) := \left.\frac{\mathrm{d}F(x;\boldsymbol{\gamma})}{\mathrm{d}x}\right|_{(x;\boldsymbol{\gamma})=(x_1;\boldsymbol{\gamma}_1)} > 0.$$

(c) For every $\boldsymbol{\gamma} \in \Theta_2$, there is a $c_1 = c_1(\boldsymbol{\gamma}) < \infty$ such that

$$f(x;\boldsymbol{\gamma}) < c_1, \qquad \text{for all } x \in \mathbb{R}. \tag{2.9}$$

(d) The mapping $\partial_{\gamma_g} F\colon \mathbb{R} \times \Theta_2 \to \mathbb{R}$ is defined for $g = 1, \ldots, z$ as

$$(x_1; \boldsymbol{\gamma}_1) \mapsto \partial_{\gamma_g} F(x_1; \boldsymbol{\gamma}_1) := \left.\frac{\partial F(x;\boldsymbol{\gamma})}{\partial \gamma_g}\right|_{(x;\boldsymbol{\gamma})=(x_1;\boldsymbol{\gamma}_1)}.$$

For every $\boldsymbol{\gamma} \in \Theta_2$ and $g \in \{1, \ldots, z\}$, there is a $c_2 = c_2(\boldsymbol{\gamma}, g) < \infty$ such that

$$\left|\partial_{\gamma_g} F(x;\boldsymbol{\gamma})\right| \leq c_2, \qquad \text{for all } x \in \mathbb{R}. \tag{2.10}$$

The measure $\mathscr{G}$ is typically a product of restricted Lebesgue and counting measures. In most applications considered by this work, $\mathscr{G}$ is the Lebesgue measure. However, the work of Bartholomew (1998) illustrates that distributions of $\boldsymbol{\eta}$ without Lebesgue density should be kept in mind, even if one is analyzing an item response data set. For this theory, we need properties of σ-finite measures but no specific additional properties of the Lebesgue measure. Therefore we formulate the theory in this section in a more general fashion.

First, we have to define some further mappings. Recall from Definition 2.1.1 part (a) that $u = (q+1)d + b + dz$, where q is the latent dimension, d is the number of items, b is the number of parameters of the structural model and z is the number of parameters of the link function. In particular, with $\Theta_1 \subseteq \mathbb{R}^b$, and $\Theta_2 \subseteq \mathbb{R}^z$, the complete parameter space (without identification constraints) is $\Theta^* = \mathbb{R}^{(q+1)d} \times \Theta_2^d \times \Theta_1 \subseteq \mathbb{R}^u$. We define

$$\begin{aligned}\mathcal{L}_{\text{complete}}\colon \Theta^* \times \Theta^n \times \{0,1\}^{dn} &\to \mathbb{R},\\ (\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) &\mapsto \prod_{k=1}^{n} \Bigg(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}_k;\, \boldsymbol{\gamma}_i)^{y_{ik}} \\ &\qquad \times \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}_k;\, \boldsymbol{\gamma}_i)\right)^{1-y_{ik}} \Bigg) \mathfrak{h}(\boldsymbol{\eta}_k, \boldsymbol{\delta}).\end{aligned} \tag{2.11}$$

It is important here to include $\mathfrak{h}$ in $\mathcal{L}_{\text{complete}}$ because $\mathfrak{h}$ depends on $\boldsymbol{\vartheta}$. In nested models where $\mathfrak{h}$ is independent from $\boldsymbol{\vartheta}$, the presence of $\mathfrak{h}$ in $\mathcal{L}_{\text{complete}}$ does not have any influence. We write for short

$$\mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) =: \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n),$$

if $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are clear in the context. Obtaining estimates for $\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ by maximizing $\mathcal{L}_{\text{complete}}$, for fixed $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \{0,1\}^d$, corresponds to the random-effect pendant of joint maximum likelihood estimation (cf. Subsection 1.3.2). If we had observed $(\boldsymbol{y}_1, \boldsymbol{\eta}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{\eta}_n)$ and considered $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ as explanatory variables for $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, then

$$\boldsymbol{\vartheta} \mapsto \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$$

would correspond to a likelihood function of a classical regression model. Due to this, $\mathcal{L}_{\text{complete}}$ of (2.11) is known as a *complete data likelihood* function. In this notion, the latent vectors are considered *missing data*. Next, we define

$$\begin{aligned} &P\colon \{0,1\}^d \times \Theta^* \to (0,1) \\ &(\boldsymbol{y}; \boldsymbol{\vartheta}) \mapsto \int_\Theta \left(\prod_{i=1}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}), \end{aligned} \tag{2.12}$$

i.e. the probability mass function (pmf) of the marginal distribution of $\boldsymbol{Y}_k$, $k = 1, \ldots, n$, for a fixed $\boldsymbol{\vartheta} \in \Theta^*$. Using this pmf, we can define a likelihood function for the actually observable $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, a so called *observed data likelihood* function:

$$\mathcal{L}_{\text{obs}}\colon \Theta^* \times \{0,1\}^{dn} \to \mathbb{R}, \qquad (\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \mapsto \prod_{k=1}^n P(\boldsymbol{y}_k; \boldsymbol{\vartheta}). \tag{2.13}$$

Then we directly get the following alternative representation:

$$\mathcal{L}_{\text{obs}}(\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \int_{\Theta^n} \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\, \mathscr{G}^n(\mathrm{d}\boldsymbol{\eta}_1, \ldots, \mathrm{d}\boldsymbol{\eta}_n),$$

for all $\boldsymbol{\vartheta} \in \Theta^*$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \{0,1\}^d$. As in Subsection 1.3.2, we call $\mathcal{L}_{\text{obs}}$ also *marginal likelihood* function. As a basis for the optimization problem, we define the *log-marginal likelihood* function as:

$$\ell\colon \Theta^* \times \{0,1\}^{dn} \to \mathbb{R}, \qquad (\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \mapsto \log\left(\mathcal{L}_{\text{obs}}(\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\right). \tag{2.14}$$

If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \{0,1\}^d$ are clear in a particular context, we write for short

$$\ell(\boldsymbol{\vartheta}) := \ell(\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n), \qquad \boldsymbol{\vartheta} \in \Theta^*.$$

Finally, we define the contribution of subject k to the log-marginal likelihood as the mapping

$$\ell_k\colon \Theta^* \times \{0,1\}^d \to \mathbb{R}, \qquad (\boldsymbol{\vartheta}; \boldsymbol{y}) \mapsto \log\left(P_k(\boldsymbol{y}; \boldsymbol{\vartheta})\right),$$

for all $k = 1, \ldots, n$. Again, we write $\ell_k(\boldsymbol{\vartheta}) := \ell_k(\boldsymbol{\vartheta}; \boldsymbol{y})$, $\boldsymbol{\vartheta} \in \Theta^*$, for short if $\boldsymbol{y} \in \{0,1\}^d$ is clear in the context. Obviously, we get the identity

$$\ell = \sum_{k=1}^n \ell_k.$$

Now, suppose the observation of a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ as a supposed realization of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, whose joint distribution is given by the pmf

$$(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \mapsto \mathcal{L}_{\text{obs}}(\boldsymbol{\vartheta}_0; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n),$$

for a supposed *true* $\boldsymbol{\vartheta}_0 \in \Theta^*$ and $n \in \mathbb{N}$. Analogously to the classic setting (cf. Subsection 1.3.2), we call any $\hat{\boldsymbol{\vartheta}} \in \Theta^*$ a *marginal maximum likelihood estimate* of the supposed true $\boldsymbol{\vartheta}_0 \in \Theta^*$, which satisfies

$$\ell(\hat{\boldsymbol{\vartheta}}) = \max_{\boldsymbol{\vartheta} \in \Theta^*} \ell(\boldsymbol{\vartheta}). \tag{2.15}$$

We now derive a representation of the gradient of ℓ. This derivation is similar to a part of the derivation of the marginal likelihood equations of the 3PL model by Baker and Kim (2004, equation (6.5) on p. 161)[4]. To allow a handy but clear notation, we define the gradient, the Hessian and the Jacobian with respect to $\boldsymbol{\vartheta}$ as operators by

$$\nabla_{\boldsymbol{\vartheta}} := \left(\frac{\partial}{\partial \vartheta_1}, \dots, \frac{\partial}{\partial \vartheta_u} \right)^\intercal, \qquad \nabla\nabla^\intercal_{\boldsymbol{\vartheta}} := \nabla_{\boldsymbol{\vartheta}} \nabla^\intercal_{\boldsymbol{\vartheta}} \tag{2.16}$$

$$\mathcal{J}_{\boldsymbol{\vartheta}} := \left(\begin{pmatrix} \frac{\partial}{\partial \vartheta_1} & \cdots & \frac{\partial}{\partial \vartheta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \vartheta_u} & \cdots & \frac{\partial}{\partial \vartheta_u} \end{pmatrix} \bullet \right)^\intercal. \tag{2.17}$$

First, we directly get

$$\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \vartheta_v} = \sum_{k=1}^{n} \frac{1}{P(\boldsymbol{y}_k; \boldsymbol{\vartheta})} \frac{\partial P(\boldsymbol{y}_k; \boldsymbol{\vartheta})}{\partial \vartheta_v} \tag{2.18}$$

for all $1 \leq v \leq (q+1)d + b + z^d$. Suppose $1 \leq v \leq d$, i.e. $\vartheta_v = \alpha_{v0}$ (cf. (2.1)). Then condition (2.9) allows the change of differentiation and integration due to Lebesgue's theorem of dominated convergence.[5] This implies that

$$\begin{aligned}
\frac{\partial P(\boldsymbol{y}_k; \boldsymbol{\vartheta})}{\partial \alpha_{v0}} &= \int_\Theta \frac{\partial}{\partial \alpha_{v0}} \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)^{y_{ik}} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)\right)^{1-y_{ik}} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta}) \, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\
&= \int_\Theta (-1)^{1-y_{vk}} f(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_v) \\
&\qquad \times \left(\prod_{\substack{i=1 \\ i \neq v}}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)^{y_{ik}} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)\right)^{1-y_{ik}} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta}) \, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\
&= \int_\Theta \frac{\partial}{\partial \alpha_{v0}} \log \left(F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_v)^{y_{vk}} \left(1 - F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_v)\right)^{1-y_{vk}} \right) \\
&\qquad \times \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)^{y_{ik}} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}; \boldsymbol{\gamma}_i)\right)^{1-y_{ik}} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta}) \, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}).
\end{aligned} \tag{2.19}$$

[4] Principally, Baker and Kim (2004, Section 6.3) present the approach of Bock and Lieberman (1970, Section 3) – who examined the Lawley–Lord model – for the 3PL model. However, some of the most interesting aspects are extensions by Baker and Kim (2004).

[5] Similar applications of Lebesgue's dominated convergence theorem are commonly discussed in introductory courses to calculus, cf. e.g. Forster (2012, p. 125, Satz 2). Compare Appendix A for further details.

Equations (2.18) and (2.19) further imply that

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \alpha_{v0}} &= \sum_{k=1}^{n} \frac{1}{P(\boldsymbol{y}_k;\boldsymbol{\vartheta})} \frac{\partial P(\boldsymbol{y}_k;\boldsymbol{\vartheta})}{\partial \alpha_{v0}} \\ &= \sum_{k=1}^{n} \mathsf{E}\left(\frac{\partial}{\partial \alpha_{v0}} \log\left(F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)^{y_{vk}} \left(1 - F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)\right)^{1-y_{vk}} \right) \,\middle|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta} \right).\end{aligned} \tag{2.20}$$

From the logarithm rules, it follows clearly that

$$\begin{aligned}&\log\left(\mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\right) \\ &\quad= \sum_{k=1}^{n} \log\left(F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)^{y_{vk}} \left(1 - F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)\right)^{1-y_{vk}} \right) \\ &\qquad+ \sum_{k=1}^{n} \log\left(\mathfrak{h}(\boldsymbol{\eta}_k, \boldsymbol{\delta}) \prod_{\substack{i=1 \\ i \neq v}}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_{ik}} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_{ik}} \right).\end{aligned} \tag{2.21}$$

This implies for equation (2.20), due to the functional independence of the second term on the right-hand side of (2.21) from α_{v0}, the linearity of conditional expectation together with the independence of the n random tuples $(\boldsymbol{Y}_1, \boldsymbol{\eta}_1), \ldots, (\boldsymbol{Y}_n, \boldsymbol{\eta}_n)$:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \vartheta_v} &= \mathsf{E}\Bigg(\sum_{k=1}^{n} \frac{\partial}{\partial \alpha_{v0}} \log\Bigg(F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)^{y_{vk}} \\ &\qquad\qquad \times \left(1 - F(\alpha_{v0} + \boldsymbol{\alpha}_v^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_v)\right)^{1-y_{vk}} \Bigg) \,\Bigg|\, \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta} \Bigg) \\ &= \mathsf{E}\left(\frac{\partial}{\partial \alpha_{v0}} \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \,\middle|\, \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta} \right).\end{aligned} \tag{2.22}$$

Analogous representations are possible for the further derivatives (see Appendix A.2 for details). Overall, we get:

$$\nabla_{\boldsymbol{\vartheta}} \ell(\boldsymbol{\vartheta}) = \mathsf{E}\left(\nabla_{\boldsymbol{\vartheta}} \log \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta} \right), \tag{2.23}$$

similar to (2.22) which is sometimes known as *Fisher's identity*[6]. While global maximization of ℓ is as commonly difficult, this representation of $\nabla_{\boldsymbol{\vartheta}} \ell$ can be well used for derivative based local optimization methods and often is a basis for the fitting routines for classical IRT models (cf. e.g. Baker and Kim, 2004, who did not call their version of (2.23) Fisher's identity).

2.3.2 Asymptotic Theory

Now, we will discuss the asymptotic existence, consistency and efficiency of the MMLE for EGLLVMs based on an iid sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ for $n \to \infty$. Additionally to the stated conditions (PLM1)–(PLM3) in Subsection 2.3.1, we require the following four further conditions for the application of classic asymptotic MLE theory (cf. e.g. Lehmann and Casella, 1998).

[6] Cf. for example Rose et al. (2017, p. 808).

(PLM4) For all $\boldsymbol{\delta} \in \Theta_1$

$$\int_{\Theta} \|\boldsymbol{\eta}\|^3 \, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}, \boldsymbol{\delta}) < \infty. \tag{2.24}$$

(PLM5) $\mathfrak{h}(\boldsymbol{\eta}; \bullet) \in C^3(\Theta_1, \mathbb{R})$ for all $\boldsymbol{\eta} \in \Theta$. The mappings $\partial_{\delta_g}\partial_{\delta_f}\mathfrak{h} \colon \Theta \times \Theta_1 \to \mathbb{R}$ and $\partial_{\delta_g}\partial_{\delta_f}\partial_{\delta_r}\mathfrak{h} \colon \Theta \times \Theta_1 \to \mathbb{R}$ are defined by

$$\begin{aligned}
(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) &\mapsto \partial_{\delta_g}\partial_{\delta_f}\mathfrak{h}(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) := \left.\frac{\partial^2 \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})}{\partial\delta_g \partial\delta_f}\right|_{(\boldsymbol{\eta},\boldsymbol{\delta})=(\boldsymbol{\eta}_1,\boldsymbol{\delta}_1)}, \\
(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) &\mapsto \partial_{\delta_g}\partial_{\delta_f}\partial_{\delta_r}\mathfrak{h}(\boldsymbol{\eta}_1, \boldsymbol{\delta}_1) := \left.\frac{\partial^3 \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})}{\partial\delta_g \partial\delta_f \partial\delta_r}\right|_{(\boldsymbol{\eta},\boldsymbol{\delta})=(\boldsymbol{\eta}_1,\boldsymbol{\delta}_1)},
\end{aligned}$$

for all $g, f, r = 1, \ldots, b$.

$$\begin{aligned}
\int_{\Theta} \sup_{\boldsymbol{\delta} \in O(\boldsymbol{\delta}_1, g, \varepsilon)} \left|\partial_{\delta_g}\partial_{\delta_f}\mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\right| \, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) < \infty, \\
\int_{\Theta} \sup_{\boldsymbol{\delta} \in O(\boldsymbol{\delta}_1, g, \varepsilon)} \left|\partial_{\delta_g}\partial_{\delta_f}\partial_{\delta_r}\mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\right| \, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) < \infty,
\end{aligned} \tag{2.25}$$

for some $\varepsilon = \varepsilon(\delta, g) > 0$ such that $O(\boldsymbol{\delta}_1, g, \varepsilon) \subset \Theta_1$ for all $\boldsymbol{\delta}_1 \in \Theta_1$ and $g, f, r = 1, \ldots, b$, where the set $O(\boldsymbol{\delta}, g, \varepsilon)$ is given in (2.8).

(PLM6) $F \in C^3(\mathbb{R} \times \Theta_2; (0,1))$ and all partial derivatives up to third order are (globally) bounded as mappings on $\mathbb{R}$ for every point in Θ_2.

(PLM7) The model is identified in the sense that there are no two $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2 \subset \Theta$ for which the corresponding EGLLVMs are equivalent.

Remark 2.3.1
If $\{\mathcal{H}(\bullet, \boldsymbol{\delta})\}_{\boldsymbol{\delta} \in \Theta_1}$ is a sub-family of the exponential family of distributions, then (2.6), (2.7), (PLM4) and (PLM5) are directly satisfied.[7] ◇

Theorem 2.3.2
Let $\mathcal{M}(\Theta', \boldsymbol{\varphi}, \mathbb{N})$ be a C^3-regular parameterization of an increasing sequence of EGLLVM over a parameter space $\Theta' \subseteq \mathbb{R}^{u_0}$, $u_0 \in \mathbb{N}$. Suppose that $\mathcal{M}(\Theta', \boldsymbol{\varphi}, \mathbb{N})$ satisfies the conditions (PLM1)–(PLM7). *Further suppose there is a $\boldsymbol{\vartheta}_0 \in \Theta' \setminus \partial\Theta'$ such that $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}} \sim \mathcal{M}(\varphi(\boldsymbol{\vartheta}_0), \mathbb{N})$ holds for an* iid *sequence $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}}$ of response vectors on $\{0,1\}^d$ (cf. Definition 2.1.1). Then a sequence $\{\hat{\boldsymbol{\vartheta}}_n\}_{n\in\mathbb{N}}$ exists, where $\hat{\boldsymbol{\vartheta}}_n \colon (\Omega, \mathcal{A}, \mathsf{P}) \to (\Theta', \mathcal{B}(\Theta'))$ for $n \in \mathbb{N}$, such that with*

$$\tilde{\ell}_k \colon \Theta' \times \{0,1\}^d \to \mathbb{R}, \qquad \tilde{\ell}_k(\boldsymbol{\vartheta}, \boldsymbol{y}) := \ell_k(\varphi(\boldsymbol{\vartheta}), \boldsymbol{y}), \qquad \boldsymbol{\vartheta} \in \Theta', \boldsymbol{y} \in \{0,1\}^d, k \in \mathbb{N}, \tag{2.26}$$

$$\lim_{n\to\infty} \mathsf{P}\left(\nabla_{\boldsymbol{\vartheta}} \sum_{k=1}^{n} \tilde{\ell}_k(\hat{\boldsymbol{\vartheta}}_n, \boldsymbol{Y}_k) = \boldsymbol{0}\right) = 1$$

[7] This follows directly from Theorem 2.7.1 of Lehmann and Romano, 2005, p. 49, and Theorem 5.10 of Lehmann and Casella, 1998, p. 28.

and $\hat{\boldsymbol{\vartheta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\vartheta}_0$ *for* $n \to \infty$. *If additionally*

$$\mathcal{I} := \mathsf{E}\left(-\nabla\nabla_{\boldsymbol{\vartheta}}^{\intercal}\tilde{\ell}_1(\boldsymbol{\vartheta}_0; \boldsymbol{Y}_1)\right)$$

is positive definite, then the sequence $\{\hat{\boldsymbol{\vartheta}}_n\}_{n\in\mathbb{N}}$ *is asymptotically efficient in the sense that*

$$\sqrt{n}\big(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0\big) \xrightarrow{\mathcal{D}} \boldsymbol{Z} \sim \mathcal{N}_{u_0}\left(\boldsymbol{0}, \mathcal{I}^{-1}\right). \tag{2.27}$$

◇

Proof.
The stated conditions imply the conditions of Theorem 5.1 in Chapter 6 of Lehmann and Casella (1998, p. 463), which directly implies the result.[8] Here, we briefly sketch the main steps. Refer to Appendix A.2 for further details of the proof.

The first step is to show that

$$\sqrt{n}\big(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0\big) = \left(\mathcal{I} + o_{\mathsf{P}}(1)\right)^{-1} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \nabla_{\boldsymbol{\vartheta}} \tilde{\ell}_1(\boldsymbol{\vartheta}_0, \boldsymbol{Y}_k), \tag{2.28}$$

where the $o_{\mathsf{P}}(1)$ term holds for $n \to \infty$. Additionally, from the multivariate central limit theorem follows that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \nabla_{\boldsymbol{\vartheta}} \tilde{\ell}_1(\boldsymbol{\vartheta}_0, \boldsymbol{Y}_k) \xrightarrow{\mathcal{D}} \boldsymbol{Z}' \sim \mathcal{N}_{u_0}\left(\boldsymbol{0}, \mathcal{I}\right), \qquad \text{for } n \to \infty.$$

Finally, from a version of Slutsky's theorem (cf. Lehmann and Casella, 1998, Chapter 6, Lemma 5.2 on p. 465) follows that

$$\left(\mathcal{I} + o_{\mathsf{P}}(1)\right)^{-1} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \nabla_{\boldsymbol{\vartheta}} \tilde{\ell}_1(\boldsymbol{\vartheta}_0, \boldsymbol{Y}_k) \xrightarrow{\mathcal{D}} \mathcal{I}^{-1}\boldsymbol{Z}' = \boldsymbol{Z}, \tag{2.29}$$

for $n \to \infty$ and $\boldsymbol{Z}$ given in (2.27). □

Notice that we always get the version

$$\sqrt{n}\big(\boldsymbol{\varphi}(\hat{\boldsymbol{\vartheta}}_n) - \boldsymbol{\varphi}(\boldsymbol{\vartheta}_0)\big) \xrightarrow{\mathcal{D}} \boldsymbol{Z}'' \sim \mathcal{N}_u\left(\boldsymbol{0}, \mathcal{J}\boldsymbol{\varphi}(\boldsymbol{\vartheta}_0)\mathcal{I}^{-1}\mathcal{J}\boldsymbol{\varphi}(\boldsymbol{\vartheta}_0)^{\intercal}\right)$$

of (2.27) on the parameter space $\boldsymbol{\varphi}(\Theta')$ from the delta method because $\mathcal{J}\boldsymbol{\varphi}(\boldsymbol{\vartheta}_0) \neq \boldsymbol{0}$ by Definition 2.1.1 (d). However, this asymptotic distribution can be singular due to a singular covariance matrix and, hence, does not have a pdf on $\mathbb{R}^u$.

[8] Lehmann and Casella (1998) define asymptotic efficiency of the single components of $\hat{\boldsymbol{\vartheta}}_n$. In contrast, we define the alternative variant of the complete vector as for example Witting and Müller-Funk (1995) do. Further, notice that the result still follows directly for our definition of efficiency because Lehmann and Casella (1998) prove the state asymptotic normality of the complete vector.

2.4 Asymptotic Tests, Confidence Regions and Model Selection

Model selection can be based on information criteria, which can be defined directly for the EGLLVM. Under the notation of Theorem 2.3.2 and for a data set with $n \in \mathbb{N}$ responses, the AIC is defined as

$$\mathrm{AIC} \coloneqq -2\left(\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n) - u_0\right)$$

where the MLE $\hat{\boldsymbol{\vartheta}}_n$ is computed for the observed responses $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and $\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n)$ is defined as

$$\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n) \coloneqq \sum_{k=1}^{n} \tilde{\ell}_1(\hat{\boldsymbol{\vartheta}}_n, \boldsymbol{y}_k)$$

for $\tilde{\ell}_1$ that is specified in (2.26). We then select the model with the smallest AIC value. By comparing several models with different link functions or link functions with a different number of parameters for the link functions, we can select the best link function. For example, we can compare EGLLVMs with the probit link for all items, a skew-probit link with the same skewness-parameter for all items and a skew-probit link with a skewness-parameter for each item to find the best trade-off between model parsimony and link function flexibility.

A more or less direct implication of the asymptotic normality and efficiency of the MMLE in Theorem 2.3.2 is the asymptotic χ^2-distribution of the standard Wald, score and LR statistics. Thus, we can construct asymptotic confidence regions and make asymptotic tests for EGLLVM using these statistics. Moreover, we can use these tests for testing the link function or the structural model of a GLLVM against parametric families because these tests can be used for nested model comparison. Hence, the following theorem is the basis for an analysis of the link function of IRT models.

Theorem 2.4.1
For $r \in \mathbb{N}$, let $\xi_r \sim \chi^2_r$.

(a) Let $\mathcal{M}(\Theta, \boldsymbol{\varphi}, \mathbb{N})$ be a C^3-regular parameterization of an increasing sequence of EGLLVM over a parameter space $\Theta \subseteq \mathbb{R}^{u_0}$, $u_0 \in \mathbb{N}$. Suppose $\mathcal{M}(\Theta, \boldsymbol{\varphi}, \mathbb{N})$ satisfies the conditions (PLM1)–(PLM7) *and with*

$$\tilde{\ell}_1 \colon \Theta \times \{0,1\}^d \to \mathbb{R}, \qquad \tilde{\ell}_1(\boldsymbol{\vartheta}, \boldsymbol{y}) \coloneqq \ell_1(\varphi(\boldsymbol{\vartheta}), \boldsymbol{y}), \qquad \boldsymbol{\vartheta} \in \Theta,\, \boldsymbol{y} \in \{0,1\}^d,$$

is

$$\mathcal{I}(\boldsymbol{\vartheta}) \coloneqq \mathsf{E}\left(-\nabla \nabla_{\boldsymbol{\vartheta}}^{\intercal} \tilde{\ell}_1(\boldsymbol{\vartheta}; \boldsymbol{Y}_1)\right)$$

positive definite for every $\boldsymbol{\vartheta} \in \Theta$. Further suppose the existence of $\boldsymbol{\vartheta}_0 \in \Theta \setminus \partial\Theta$ such that $\{\boldsymbol{Y}_k\}_{k\in\mathbb{N}} \sim \mathcal{M}(\boldsymbol{\varphi}(\boldsymbol{\vartheta}_0), \mathbb{N})$. Denote by $\hat{\boldsymbol{\vartheta}}_n$ the MLE of $\boldsymbol{\vartheta}_0$ based on $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, for each $n \in \mathbb{N}$ and set

$$\tilde{\ell}^{(n)}(\boldsymbol{\vartheta}) \coloneqq \sum_{k=1}^{n} \tilde{\ell}_1(\boldsymbol{\vartheta}, \boldsymbol{Y}_k), \qquad \boldsymbol{\vartheta} \in \Theta,\, n \in \mathbb{N}.$$

Then

$$2\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n) - 2\tilde{\ell}^{(n)}(\boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \xi_{u_0}, \qquad n \to \infty. \tag{2.30}$$

(b) *Under the conditions of (a) and for an increasing sequence of EGLLVM* $\mathcal{M}_0(\Theta_0, \mathbb{N})$ *over a parameter space* $\Theta_0 \subseteq \mathbb{R}^{u_0^*}$, $u_0^* \in \mathbb{N}$ *and* $u_0^* < u_0$. *Suppose* $\mathcal{M}_0(\Theta_0, \mathbb{N})$ *is* C^3*-regular nested to the model of (a) with parameterization* $\boldsymbol{\varphi}_0^*\colon \Theta_0 \to \Theta$. *Suppose further that there is* $\boldsymbol{\vartheta}_0^* \in \Theta_0 \setminus \partial\Theta_0$ *such that* $\boldsymbol{\varphi}_0^*(\boldsymbol{\vartheta}_0^*) = \boldsymbol{\vartheta}_0$ *and denote by* $\hat{\boldsymbol{\vartheta}}_n^*$ *the MLE of* $\boldsymbol{\vartheta}_0^*$ *based on* $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n)$ *for all* $n \in \mathbb{N}$. *Then*

$$2\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n) - 2\tilde{\ell}^{(n)}(\boldsymbol{\varphi}_0^*(\hat{\boldsymbol{\vartheta}}_n^*)) \xrightarrow{\mathcal{D}} \xi_{u_0 - u_0^*}, \qquad n \to \infty. \tag{2.31}$$

(c) *Under the conditions of (a) and for a* $\boldsymbol{g} \in C^1(\Theta, \mathbb{R}^{u_1})$ *with* $u_1 \in \{1, \dots, u_0 - 1\}$. *Denote by* $\mathcal{J}\boldsymbol{g}$ *the Jacobian of* $\boldsymbol{g}$ *and suppose that*

$$\boldsymbol{V}(\boldsymbol{\vartheta}) \coloneqq \mathcal{J}\boldsymbol{g}(\boldsymbol{\vartheta})\mathcal{I}(\boldsymbol{\vartheta})^{-1}\mathcal{J}\boldsymbol{g}(\boldsymbol{\vartheta})^\intercal$$

is positive definite for every $\boldsymbol{\vartheta} \in \Theta$ *and* $\boldsymbol{\vartheta} \mapsto \boldsymbol{V}(\boldsymbol{\vartheta})$ *is continuous. Then for* $n \to \infty$

$$n\left(\boldsymbol{g}\left(\hat{\boldsymbol{\vartheta}}_n\right) - \boldsymbol{g}\left(\boldsymbol{\vartheta}_0\right)\right)^\intercal \boldsymbol{V}\left(\boldsymbol{\vartheta}_0\right)^{-1}\left(\boldsymbol{g}\left(\hat{\boldsymbol{\vartheta}}_n\right) - \boldsymbol{g}\left(\boldsymbol{\vartheta}_0\right)\right) \xrightarrow{\mathcal{D}} \xi_{u_1}, \tag{2.32}$$

$$n\left(\boldsymbol{g}\left(\hat{\boldsymbol{\vartheta}}_n\right) - \boldsymbol{g}\left(\boldsymbol{\vartheta}_0\right)\right)^\intercal \boldsymbol{V}\left(\hat{\boldsymbol{\vartheta}}_n\right)^{-1}\left(\boldsymbol{g}\left(\hat{\boldsymbol{\vartheta}}_n\right) - \boldsymbol{g}\left(\boldsymbol{\vartheta}_0\right)\right) \xrightarrow{\mathcal{D}} \xi_{u_1}. \tag{2.33}$$

(d) *Under the conditions of (a) and for an* $r \in \{1, \dots, u_0\}$, *denote by* $\boldsymbol{D} \in \mathbb{R}^{u_0 \times r}$ *a matrix with* r *different columns of the* $u_0 \times u_0$ *identity matrix. Then*

$$\nabla\tilde{\ell}^{(n)}(\boldsymbol{\vartheta}_0)^\intercal \boldsymbol{D}\left(\boldsymbol{D}^\intercal\mathcal{I}(\boldsymbol{\vartheta}_0)\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\nabla\tilde{\ell}^{(n)}(\boldsymbol{\vartheta}_0) \xrightarrow{\mathcal{D}} \xi_r, \qquad n \to \infty. \tag{2.34}$$

(e) *Under the conditions of (d) and for a further increasing sequence of EGLLVM* $\mathcal{M}_0(\Theta_0, \mathbb{N})$ *over the parameter space*

$$\Theta_0 \coloneqq \{\boldsymbol{\vartheta} \in \Theta\colon \boldsymbol{D}^\intercal\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0^*\}$$

for $\boldsymbol{\vartheta}_0^* = \boldsymbol{D}^\intercal\boldsymbol{\vartheta}_0$. *Denote by* $\hat{\boldsymbol{\vartheta}}_n^*$ *the MLE of* $\boldsymbol{\vartheta}_0$ *in the parameter space* Θ_0 *based on* $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n)$, $n \in \mathbb{N}$. *Then*

$$\nabla\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n^*)^\intercal \boldsymbol{D}\left(\boldsymbol{D}^\intercal\mathcal{I}(\hat{\boldsymbol{\vartheta}}_n^*)\boldsymbol{D}\right)^{-1}\boldsymbol{D}^\intercal\nabla\tilde{\ell}^{(n)}(\hat{\boldsymbol{\vartheta}}_n^*) \xrightarrow{\mathcal{D}} \xi_r, \qquad n \to \infty. \tag{2.35}$$

◇

Proof.
The stated conditions imply the conditions of Theorem 12.2.1 of Lehmann and Romano (2005, p. 486), cf. Appendix A.2. Then (a) and (b) follow from Theorem 12.4.2 of Lehmann and Romano (2005, p. 515–516), part (c) follows from Section 12.4.2 of Lehmann and Romano (2005, p. 508–510) and parts (d) and (e) follow from Section 12.4.3 of Lehmann and Romano (2005, p. 511–513). □

Remark 2.4.2

(a) To compute the statistic in (2.35), it is not required to know $\boldsymbol{\vartheta}_0$ completely. It suffices to specify the value $\boldsymbol{\vartheta}_0^*$. As a result, (2.35) can be used if only a hypothesis regarding a subset of the parameters is made. This allows (2.35) to be used to test nested models if the nested model is specified by restricting several parameters to a fixed value.

(b) In contrast, (2.34) requires a specification of the complete vector $\boldsymbol{\vartheta}_0$ but still enables testing of only a subset of the components of $\boldsymbol{\vartheta}$ in a simple null hypothesis. An advantage of the score test is that no point estimator has to be computed, which makes the resulting test comparably simple.

(c) Similar to (2.35), the statistic in (2.33) does not require a completely specified vector $\boldsymbol{\vartheta}_0$ to construct a test. Actually, a specification of the value of $\boldsymbol{g}(\boldsymbol{\vartheta}_0)$ suffices. Particularly if $\boldsymbol{g}(\boldsymbol{\vartheta}_0) = \mathbf{0}$, then (2.35) can be used to test if the nested model holds, which is specified by

$$\Theta_0 = \{\boldsymbol{\vartheta} \in \Theta \colon \boldsymbol{g}(\boldsymbol{\vartheta}) = \mathbf{0}\}.$$

In contrast, (2.32) requires $\boldsymbol{\vartheta}_0$ even if $\boldsymbol{g}$ is not bijective. Hence, (2.32) can be used appropriately for tests if $\boldsymbol{g}$ is bijective, in particular for the case $\boldsymbol{g} = \mathrm{id}$. The statistic (2.32) can be used for testing for all $\boldsymbol{g}$, so it is sometimes preferable.

(d) The statistic in (2.30) can be used appropriately to test simple null hypotheses. The statistic (2.31) can be used for nested model testing. ◇

Using the LR-test in Theorem 2.4.1 part (b), we can test for example the one dimensional probit-normal model

$$\Phi^{-1}(\mathsf{P}(Y_i = 1 \mid \eta)) = \alpha_{i0} + \alpha_{i1}\eta, \quad i = 1, \dots, d, \quad \eta \sim \mathcal{N}(0, 1) \tag{2.36}$$

for $d \in \mathbb{N}$ items against the skew-probit-skew-normal model

$$\mathrm{ssp}_{\gamma_1}(\mathsf{P}(Y_i = 1 \mid \eta)) = \alpha_{i0} + \alpha_{i1}\eta, \quad i = 1, \dots, d, \quad \eta \sim \mathcal{SN}(\gamma_2), \tag{2.37}$$

where $\mathcal{SN}(\gamma_2)$ denotes the standard skew normal distribution with skewness parameter $\gamma_2 \in \mathbb{R}$ (cf. Definition 1.1.7). Using this test, we can test if the latent variable has a skew normal distribution with non-zero skewness and/or if all ICCs are skew with a common skewness. The alternative especially contains a loglog and a cloglog model because the ssp link function can approximate them really well. Another possible alternative is given by the model

$$\mathrm{ssp}_{\gamma_i}(\mathsf{P}(Y_i = 1 \mid \eta)) = \alpha_{i0} + \alpha_{i1}\eta, \quad i = 1, \dots, d, \quad \eta \sim \mathcal{N}(0, 1), \tag{2.38}$$

where the latent variable has a standard normal distribution and all items are modeled using a skew-probit link. But now, there is a skewness parameter for each item such that it is possible to test if any link function has a non-zero skewness. Of course, many other alternatives can be formulated, too. Of course, the skew-probit link function family cannot capture every possible skewness and every test is more powerful against specific departures from the hypothesis. An appropriate extension of the full procedure is therefore a GoF test of the more complex model to obtain with the nested model comparison a conditional goodness of link test.

2.5 Simulation Studies

From Theorem 2.3.2, we know that we can estimate model parameters of EGLLVM consistently by MMLE for $n \to \infty$, but the examples of Subsection 2.2.2 indicate that the

estimation of item skewness parameters can be difficult for a finite sample size. How well the skewness parameter of $\mathcal{H}$ can be estimated by frequentist methods is not well known because the IRT literature regarding models with skewness parameters focuses on Bayesian methods (cf. e.g. Bazán et al., 2014). One objective of this chapter is the selection of link functions for a GLLVM, which we only discussed theoretically in Subsection 2.4. In this section, we discuss three simulation studies which we conducted to extend the knowledge regarding the application of MMLE to fit EGLLVM to the data and the use of the EGLLVM framework for link selection by empirical results.

In the first study, we focus on the point estimation of skewness parameters. In the second study, we focus on the estimation of the ICC (considered a function) and the estimation of the underlying latent variables using the EAP. In the third study, we focus on model selection using the AIC and the LR test for nested model comparison. Overall, the simulation studies imply that a relatively large number of responses is needed to estimate skewness parameter of items precisely. For the skewness parameter of the structural model, a large number of responses is needed but comparably less than for the skewness parameters of the items. In contrast, the model selection and the tests for the link functions seem to be relatively unaffected by the issues of estimating the skewness parameters and yield good results in the examined settings.

In all three simulation studies we used MMLE with the BFGS quasi-Newton method implemented in the R function `optim` and Gauss–Hermite quadrature with 30 quadrature points to fit the models. For the computation of the EAP, we used Gauss–Hermite with 30 quadrature points, too. Further aspects of our implementation of the methods in R are discussed in Section C.1 of the appendix.

2.5.1 Precision of Skewness Parameter Estimates

We discussed in Example 1.1.6 that there are cases in practice where one can be uncertain about the skewness of single items or the skewness of the structural model. In the first simulation study of this chapter, we analyze the point estimation in such settings using the MMLE and focus on the properties of the estimates of the skewness parameters.

2.5.1.1 Questions, Setups and Methods

In this simulation study, we want to analyze how precise we can estimate skewness parameters of single items and the skewness parameter of a skew normal distribution for η. Additionally, we question how well these estimates perform as measures of the direction and strength of skewness. To this end, we analyze data sets with $n \in \{100, 1000, 5000, 20000\}$ and $d \in \{10, 20\}$. For each combination, we generate 100 data sets from model

$$\mathrm{ssp}_{\gamma_i}(\mathsf{P}(Y_i = 1 \mid \eta)) = \alpha_{i0} + \alpha_{i1}\eta, \quad i = 1, \ldots, d, \quad \eta \sim \mathcal{H}, \tag{2.39}$$

with the values of α_{i0}, α_{i1}, $i = 1, \ldots, d$, in Table 2.6. Additionally, we examine the following two setups:

1. The data are generated with $\gamma_1 = 3.3$ and $\gamma_i = 0$, $i = 2, \ldots, d$ and $\mathcal{H} = \mathcal{N}(0, 1)$. The parameters α_{i0}, α_{i1}, $i = 1, \ldots, d$ and γ_1 are considered unknown and will be estimated. The parameters $\gamma_i = 0$, $i = 2, \ldots, d$ and $\mathcal{H}$ are fixed.

2. The data are generated with $\gamma_1 = 3.3$, $\gamma_2 = -3.3$ and $\gamma_i = 0$, $i = 3, \ldots, d$ and $\mathcal{H} = \mathcal{SN}(\gamma_{\text{struct}})$ is a standard skew normal distribution with skewness parameter

i	α_{i0}	α_{i1}	γ_i	i	α_{i0}	α_{i1}	γ_i
1	-0.906	0.829	4.378	11	-0.919	1.186	2.108
2	-1.337	1.173	4.024	12	0.216	1.172	3.980
3	-0.913	1.198	3.059	13	1.900	0.837	3.232
4	0.387	1.017	3.639	14	0.510	0.978	1.874
5	0.676	1.050	3.051	15	-1.109	1.114	2.184
6	-0.084	0.822	3.401	16	-0.104	1.164	2.134
7	0.138	1.152	4.256	17	0.458	1.117	2.965
8	1.608	1.155	2.800	18	-1.882	1.012	2.930
9	-0.882	0.918	2.405	19	0.801	0.938	3.323
10	1.240	1.082	2.781	20	1.953	1.019	1.371

Table 2.6: *True item parameter values of α_{i0}, α_{i1} in all simulations in Section 2.5 and of γ_i, $i = 1, \ldots, 20$ in setup 4 of the simulation study in Subsection 2.5.2. The α_{i0}s are generated independently from $\mathcal{U}(\,(-2, 2)\,)$, the α_{i0}s from $\mathcal{U}(\,(0.8, 1.2)\,)$ and the γ_is from $\mathcal{N}(3.3, 0.64)$.*

$\gamma_{\text{struct}} = -3.5$. The parameters α_{i0}, α_{i1}, $i = 1, \ldots, d$ and γ_1, γ_2 and γ_{struct} are considered unknown and will be estimated. The parameters $\gamma_i = 0$, $i = 3, \ldots, d$ are held fixed.

Here, we focus on the parameters γ_1, γ_2 and γ_{struct}. Restricting our presentation of the results to these parameters enables us to provide a detailed analysis of the summary statistics of the estimates, reflecting their quality. Nevertheless the parameters α_{i0}, α_{i1}, $i = 1, \ldots, d$ are still estimated simultaneously. Estimating $2d + 1$ (or $2d + 3$) parameters instead of a single parameter (or 3, respectively) influences the estimation of γ_1 (and γ_2, γ_{struct}). If one were to estimate only γ_1 (and γ_2, γ_{struct}), the estimates could be expected to be more precise. Next, we present the results and then discuss and interpret them.

2.5.1.2 Results

A boxplot for all d of the estimates for γ_1 in setup 1 is displayed in Figure 2.5. The corresponding values are given in Table 2.7. Figure 2.6, 2.7 and 2.8 contain presentations of the boxplots of the estimates of γ_1, γ_2 and γ_{struct} in setup 2 – the values are given in Table 2.8. In some cases, the estimates were much larger than the true value in absolute value. To increase the visibility, we did not visualize the values which are larger/smaller than the lower/upper whisker in the boxplots but provided the proportion of these values in the tables. Notice that the skewness parameters have a nonlinear influence, which decreases for $|\gamma| \to \infty$, i.e. the difference of models with $\gamma_1 = 15$, $\gamma_1 = 150$ or even larger is negligible, while models with $\gamma_1 = 0$, $\gamma_1 = 2$ and $\gamma_1 = 4$ have comparably strong differences. Since in a clear majority of the cases the outlier occurred at the extreme side, this nonlinear influence of the skewness parameters explain the large spread of the estimated values – especially for small d.

In setup 1, we see that the estimates for $n = 100$ are spread over a large range. The sign of the estimate is also only correct in approximately 75% of the cases. In the case of $n = 1000$, the direction of the skewness is estimated correctly in all cases and the median of the estimates is close to the true value. However, the spread is still relatively large. For higher n, the distribution of the estimates centers its mass in a small region around the true value. When comparing $d = 10$ to $d = 20$, we see that the estimates for $n = 100$ are worse

d	n	Q1	mean	median	Q3	sd	PLW	PUW
10	100	0.043	89.536	8.064	66.653	200.114	0.02	0.17
	1000	1.688	11.602	3.246	6.679	28.060	0	0.15
	5000	2.539	3.496	3.149	4.471	1.583	0	0.01
	20000	2.927	3.372	3.336	3.658	0.686	0.01	0.02
20	100	0.073	721.615	21.274	126.175	2178.685	0.04	0.18
	1000	2.398	9.377	3.539	5.577	44.231	0	0.06
	5000	2.758	3.452	3.372	4.048	0.961	0	0.01
	20000	2.945	3.311	3.227	3.716	0.675	0.01	0.01

Table 2.7: *Summary statistics of the estimates for $\gamma_1 = 3.3$ of 100 simulated data sets in setup 1 of the simulation study in Subsection 2.5.1. Q1 and Q3 denote the 25% and 75% quantiles respectively, PLW and PUW denote the proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots.*

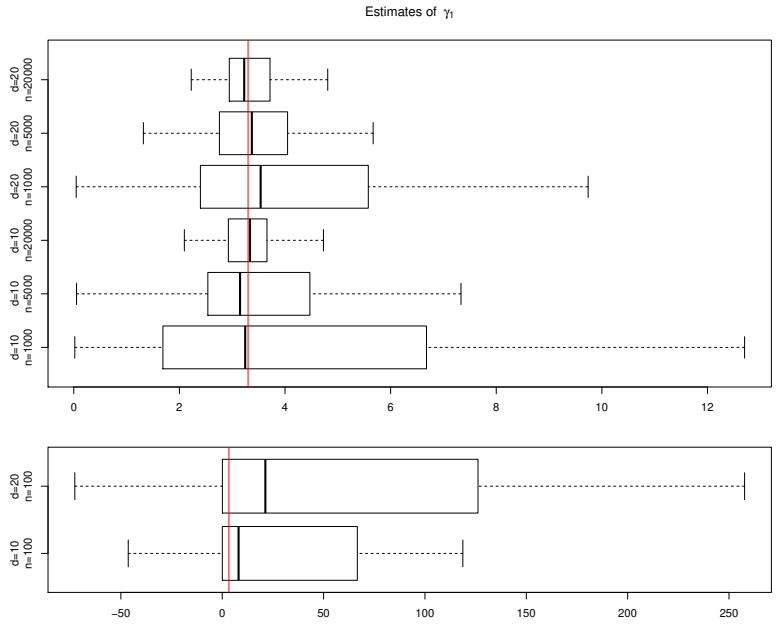

Figure 2.5: *Boxplots of the estimates for $\gamma_1 = 3.3$ (marked by a red line) of 100 simulated data sets in setup 1 of the simulation study in Subsection 2.5.1. The cases for n and d are displayed to the left of the boxplots. The proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots are given in Table 2.7. Due to the different spread of the boxplots, they are separated into two ranges of values to improve the clarity for all boxplots.*

parameter	d	n	Q1	mean	median	Q3	sd	PLW	PUW
$\gamma_1 = 3.3$	10	100	−0.287	95.354	22.913	99.602	210.763	0.02	0.12
		1000	1.881	36.574	4.939	25.859	125.103	0	0.14
		5000	2.591	4.530	3.737	5.250	4.076	0	0.05
		20000	2.845	3.551	3.470	4.059	1.100	0.01	0.01
	20	100	0.020	396.182	26.088	550.957	7128.565	0.08	0.23
		1000	2.051	16.474	3.251	6.598	68.002	0	0.11
		5000	2.602	4.175	3.215	4.431	5.083	0	0.07
		20000	2.960	3.337	3.327	3.576	0.561	0	0.02
$\gamma_2 = -3.3$	10	100	−12.513	66.122	−2.795	6.271	277.814	0.08	0.17
		1000	−13.273	−9.196	−2.331	−0.024	30.517	0.06	0.01
		5000	−12.123	−7.184	−3.409	−1.216	10.296	0.05	0
		20000	−7.994	−6.343	−5.507	−2.711	4.446	0.05	0
	20	100	−18.231	547.991	−1.383	3.890	4682.940	0.20	0.20
		1000	−14.320	−10.164	−3.209	−0.970	25.181	0.04	0
		5000	−13.539	−8.227	−2.765	−1.130	11.170	0.06	0
		20000	−8.960	−7.587	−4.463	−2.714	7.233	0.14	0
$\gamma_{\text{struct}} = -3.5$	10	100	−5.047	−2.773	−2.775	−1.395	2.455	0	0.01
		1000	−4.258	−3.541	−3.546	−2.713	0.888	0	0
		5000	−3.802	−3.303	−3.335	−3.023	0.767	0	0.03
		20000	−3.703	−3.484	−3.523	−3.351	0.438	0	0.01
	20	100	−5.242	−3.762	−4.301	−2.428	1.817	0	0.01
		1000	−3.964	−3.581	−3.624	−3.103	0.654	0	0
		5000	−3.732	−3.494	−3.509	−3.287	0.491	0.01	0.01
		20000	−3.655	−3.517	−3.492	−3.393	0.190	0	0

Table 2.8: *Summary statistics of the estimates of γ_1, γ_2 and γ_{struct} of 100 simulated data sets in setup 2 of the simulation study in Subsection 2.5.1. Q1 and Q3 denote the 25% and 75% quantiles respectively, PLW and PUW denote the proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots.*

for $d = 20$ than for $d = 10$. For the larger n considered, the skewness parameter estimates for $d = 20$ are more precise than for $d = 10$.

In setup 2, the situation of the estimates of γ_1 and γ_2 is comparable to setup 1. But as we expect for a model containing more parameters, the estimates are overall worse. Interestingly, the effect of an increasing n seems to be stronger on the estimator for γ_1 than on the estimator for γ_2.

2.5.1.3 Discussion

Setup 1: As the sign of the estimate is only correct in approximately 75% of the cases, the estimates are unreliable even as an indicator of the sign of skewness and questionable as a measure for the skewness of an ICC. An explanation is that the parameter only determines the skewness of a single ICC and, hence, it has only a small influence on the success probability for the corresponding 100 binary responses in comparison to the difficulty and the discrimination parameters of that item. For several thousand or more responses, the estimates become precise and models with single items containing skewness parameters seem to be applicable. The difference of the results for $d = 10$ and $d = 20$ can be explained with the too large number of parameters for $d = 20$ for the comparably small number of 100 responses. The observed quality of the estimates for $d = 10$ and $d = 20$ for the com-

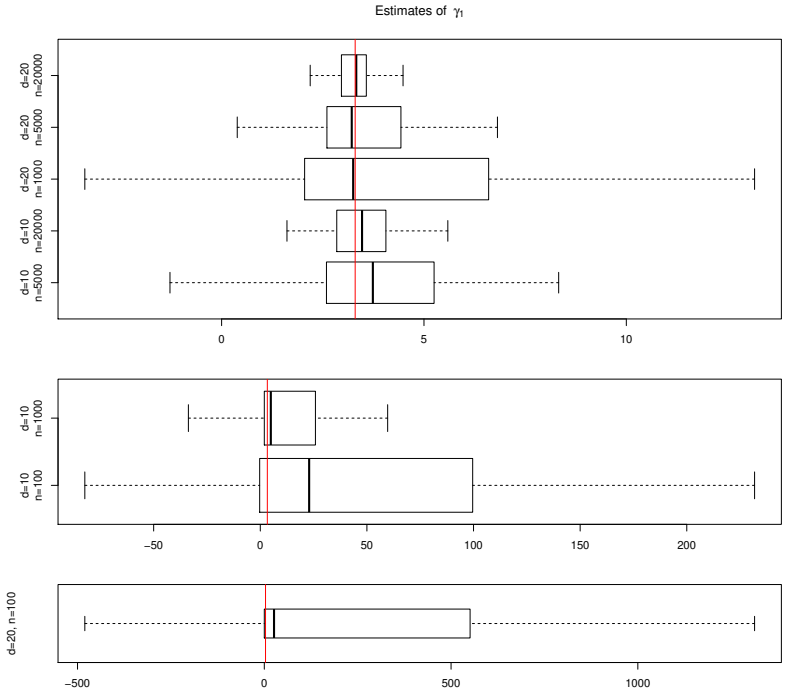

Figure 2.6: *Boxplots of the estimates for $\gamma_1 = 3.3$ (marked by a red line) of 100 simulated data sets in setup 2 of the simulation study in Subsection 2.5.1. The cases for n and d are displayed to the left of the boxplots. The proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots are given in Table 2.8. Due to the different spread of the boxplots, they are separated into three ranges of values to improve the clarity for all boxplots.*

parably small sample size of $n = 100$ is consistent with the fact that the larger number of parameters for $d = 20$ require generally more information. Hence, for a critically small sample size (for the models considered) the larger number of parameters results in worse results. But generally, responses to 20 items contain more information than for 10 items. More specifically, since the response patterns matter and we do not analyze all items individually, increasing the sample size for $d = 20$ results in more additional information than for $d = 10$, i.e. the additional 10 items have a positive effect to fitting the first ten item's ICC because we consider them together in the response vector's marginal distribution. This causes that there is a sample size at which the larger information of more items weights heavier than the larger number of parameters. We observe this switch in quality of the estimates in this study at $n = 1000$. Similar effects also occur in other IRT settings and for other methods, for example recall the joint maximum likelihood estimator for the Rasch model in Subsubsection 1.3.3.2, where we discussed that the bias for estimating the item parameters decreases when the number of items is increased. Similarly, the pseudo likelihood estimator in the current state of development has a bias that decreases for a larger number of items (we will discuss this in Section 5.3 in detail). Therefore, if the sample size is large enough to exceed the point where quality changes to be beneficial for longer tests, these methods will benefit greatly from a longer test. We have seen that MMLE is

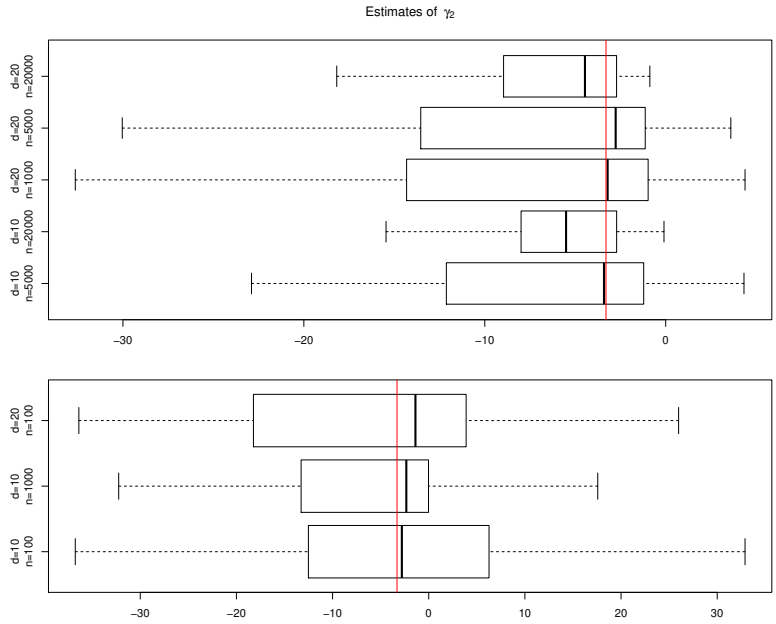

Figure 2.7: *Boxplots of the estimates for $\gamma_2 = -3.3$ (marked by a red line) of 100 simulated data sets in setup 2 of the simulations study in Subsection 2.5.1. The cases for n and d are displayed to the left of the boxplots. The proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots are given in Table 2.8. Due to the different spread of the boxplots, they are separated into two ranges of values to improve the clarity for all boxplots.*

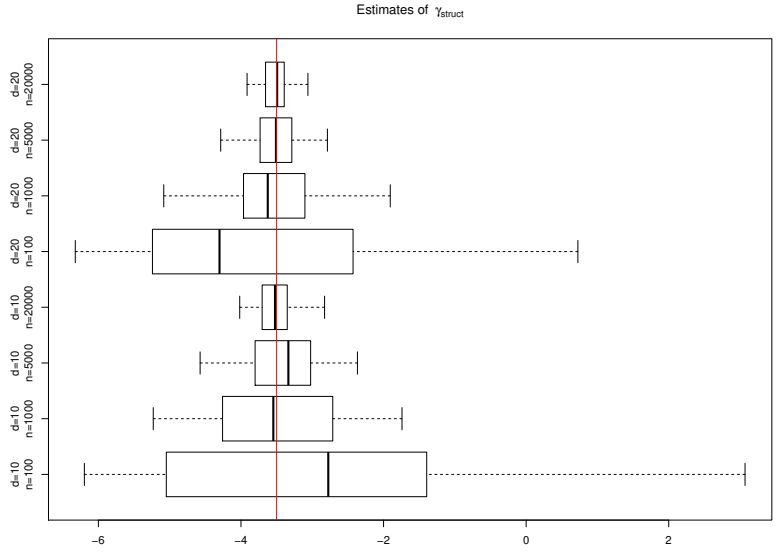

Figure 2.8: *Boxplots of the estimates for $\gamma_{struct} = -3.5$ (marked by a red line) of 100 simulated data sets in setup 2 of the simulation study in Subsection 2.5.1. The cases for n and d are displayed to the left of the boxplots. The proportion of estimates which are smaller/larger than the lower/upper whisker in the boxplots are given in Table 2.8.*

consistent – and therefore not biased – but longer tests still have a positive impact on the speed at which the quality of the estimation improves.

Setup 2: A possible cause for the differences of the estimates for γ_1 and γ_2 is the higher difficulty of item 2. While we get acceptable estimates of γ_1 for $n = 5000$ and good estimates for $n = 20000$, the estimates of γ_2 seem to be merely an indicator of the direction in all cases. However, it seems that the direction of the skewness can be estimated reliably.

Overall, we see that skewness parameters of single ICCs are difficult to estimate and need a relatively large sample size. It seems meaningful rather to decide for loglog, cloglog or probit depending on the value of the estimates rather than the point estimate for the considered items if the standard deviation of the estimator is comparably high. Hence, skew link models for binary data should be applied with caution. For data sets with a few houndred responses we do not recommend to model the ICCs of single items with skewness parameters because the estimators do not seem to be reliable for data sets of these sample sizes. But for large data sets such as that resulting from the PISA surveys with several 100,000 responses the skewness estimates can be expected to be reliable.

In comparison to γ_1 and γ_2, γ_{struct} can be estimated well in all cases. For $n \geq 1000$, the point estimates can be considered sufficiently precise for applications. This coincides with the observation of the simulation study of Section 2.5.2 in setup 5, where the considered skewness of the structural model provides a benefit.

2.5.2 Reliability of ICC and Latent Trait Estimates under Skewness Uncertainty

In the last simulation study, we focused on the estimation of the skewness parameters of a few items or $\mathcal{H}$ by MMLE for a relatively wide range of numbers of responses. We will now focus on models which have skew links for all items and/or a skew $\mathcal{H}$ and their comparison to models which have fixed ones with different parsimony. While problems regarding for example the 3PL model are intensively analyzed from different perspectives, several skew parametric but not on guessing focused models (like the skew-probit models) are not as well understood. The literature on parametric link models focuses on Bayesian techniques. Since the prior distributions can be a strong regularizing influence on comparably small data sets, it is open how well we can fit parametric skew link models by frequentist methods in comparison to fixed link models. Due to the results of the last study, we focus on $n = 1000$ and $n = 2000$ as range of a sample size which seems to be relatively critical for the skew link estimation but not too small for any skewness be estimated.

2.5.2.1 Questions, Setups and Quality Measures

In this simulation study, we aim to shed light on the following questions:

1. If there were be some skewness in the data, would a more complex parametric model which incorporates the skewness provide a better fit than a more parsimonious one?

2. If the direction of the skewness is correctly estimated a priori, but the strength is still unknown, then how good are loglog/cloglog, symmetric fixed link and parametrically skewed models in comparison?

3. If the type of skewness in the data is wrongly considered, can the considered one compensate the misspecification?

4. If there is actually no skewness, does this harm the fit of more complex models? Does a wrongly considered fixed skew link model yield a remarkably worse fit?

All these questions are closely related. Generally, we want to investigate the tradeoff between flexible and parsimonious models under some uncertainty and missspecification. In this simulation study, we restrict to one dimensional models (i.e. $q = 1$) to focus on the primary aspects of the questions. We also study general skewness aside of guessing.

Quality Measures In this study, we examine several models which are in some (or even all) setups misspecified. It therefore is not reasonable to compare the estimates for the different models using the common parametric measures such as the root mean squared error. However, using the parameter estimates for all models, we estimate the ICCs of the models. Especially due to the fact, that several models either have a wrong link function that is held fix (e.g. a probit link when skew-probit links with non-zero parameters are true) or that the link functions parameters are estimated and always contain some error, a direct comparison of the model parameters with the true parameters does not provide an quantification of the actual quality of the estimated ICCs. We introduce two nonparametric measures to simultaneously compare the estimates of all ICCs with the true underlying ICCs because of this common aspect. Suppose that $d, n \in \mathbb{N}$ are given; then we denote by $P_1, \dots, P_d \colon \mathbb{R} \to (0,1)$ the true ICCs in the examined setting and by $\hat{P}_1, \dots, \hat{P}_d \colon \mathbb{R} \to (0,1)$ their estimates by the parametric models. To measure the quality of the estimated ICCs, we compute the following two measures of the distance between the estimated ICCs and the true ICCs:

$$\text{mKSD} := \max_{1 \le i \le d} \max_{\eta \in \mathbb{R}} \left| P_i(\eta) - \hat{P}_i(\eta) \right| \tag{2.40}$$

$$\text{aKSD} := \frac{1}{d} \sum_{i=1}^{d} \max_{\eta \in \mathbb{R}} \left| P_i(\eta) - \hat{P}_i(\eta) \right|. \tag{2.41}$$

We call these two measures *maximum Kolmogorov–Smirnov distance* (mKSD) and *average Kolmogorov–Smirnov distance* (aKSD) due to the close relation to the Kolmogorov–Smirnov statistic for comparing an estimated cdf to the supposed true one. Notice that both mKSD and aKSD always take values in $(0, 1)$. The structural models have always expectation zero and variance one. Additionally, the difficulty parameters are always in $(-1.9, 2)$. Therefore, the greatest departure will be between -3 and 3, which is also the range with the greatest influence of errors in the ICCs. So, to simplify the computation of aKSD and mKSD, we restrict (2.40) and (2.41) to $\eta \in [-4, 4]$.

To examine the final estimation of latent traits, we also measure the quality of point estimates of the true underlying latent traits. Denote by $\eta_1, \dots, \eta_n$ the true values of the latent variables, which are generated for each data set. As an estimator, we use the EAP and denote the estimate of η_k by $\hat{\eta}_k$, $k = 1, \dots, n$. The EAP combines several general aspects which make it good choice in a practical application as well as in a simulation study: For each possible response pattern and all estimated model parameters it is guaranteed that the EAP exists[9] and is then unique by its definition. Moreover, the EAP can easily be computed using the quadrature points for MMLE. The median and maximum a-posteriori estimator (MeAP, MAP, respectively) as well as the MLE require additional iterative procedures (or alternatively a sampling from the posterior for all unique response pattern for MeAP).

[9] This is directly implied from Theorem 3.4.5 in the next chapter by inserting the models considered here.

Additionally, it cannot be simply guaranteed that MAP and MLE are unique while it cannot even be guaranteed that the MLE exists at all for a finite d. Since there are n estimates for each data set, we compute the following measures

$$\text{MAE} := \text{median}\left(\left\{\left|\hat{\eta}_k - \eta_k\right| : k = 1, \ldots, n\right\}\right), \tag{2.42}$$

$$\text{AAE} := \frac{1}{n}\sum_{k=1}^{n}\left|\hat{\eta}_k - \eta_k\right|, \tag{2.43}$$

$$\text{SdAE} := \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\left|\hat{\eta}_k - \eta_k\right| - \text{AAE}\right)^2}, \tag{2.44}$$

i.e. the *median absolute error* (MAE), the *average absolute error* (AAE) and the *standard deviation of the absolute error* (SdAE). We do not compute further measures of the departure in the structural model. We select these measures because they provide an intuitive notion of how well two test participants $k_1, k_2 \in \{1, \ldots, n\}$, $k_1 \neq k_2$, can be compared using the difference $\hat{\eta}_{k_2} - \hat{\eta}_{k_2}$ and how well one participant's latent variable can be compared with a constant c by comparing $\hat{\eta}_{k_2}$ with c (e.g. some threshold in an assessment). The structural model will affect the estimates of $\eta_1, \ldots, \eta_n$, but because several models assume a fixed structural model, measuring its departure from the true one is less reasonable than measuring its influence.

For each data setup, we will generate 100 data sets and compute the statistics for each one. As a final quality measure, we compute the average value of each statistic over the 100 data sets.

Data Setups We study five setups of the data generation. For each setup, we select $d \in \{10, 20\}$ and $n \in \{1000, 2000\}$ (we do not consider $n = 100$ due to the results of the first study). In each case, we generate n values of the latent variables by a structural model $\mathcal{H}$ and then a corresponding response pattern for each latent variable. The general model is as follows in all cases:

$$\text{ssp}_{\gamma_i}(P_i(\eta)) = \alpha_{i0} + \alpha_{i1}\eta, \qquad i = 1, \ldots, d, \qquad \eta \sim \mathcal{H}, \tag{2.45}$$

where ssp is the standard skew probit link of equation (1.14). In the five setups, the model values are specified as follows:

1. A common probit-normal holds, i.e. $\mathcal{H} = \mathcal{N}(0, 1)$, $\gamma_i = 0$, $i = 1, \ldots, d$. The values of α_{i0}, α_{i1}, $i = 1, \ldots, d$ are given in Table 2.6.

2. A skew-probit-normal holds with the same skewness parameter $\gamma_i = 3.3$, $i = 1, \ldots, d$. The values of the remaining parameters are the same as in the first case.

3. A skew-probit-skew-normal model holds with $\gamma_i = 3.3$, $i = 1, \ldots, d$ and $\mathcal{H} = \mathcal{SN}(\gamma_{\text{struct}})$ is a standard skew normal distribution with skewness parameter $\gamma_{\text{struct}} = -3.5$, cf. Definition 1.1.7. The remaining parameters are the same as in the first case.

4. A skew-probit-skew-normal model holds. In this case, all item parameters are selected as in Table 2.6 and $\mathcal{H}$ again has skewness parameter -3.5.

5. A probit-skew-normal model holds with $\gamma_i = 0$, $i = 1, \ldots, d$ and $\mathcal{H}$ again has skewness parameter -3.5.

Models

(i) Rasch-type probit-normal model, i.e. $\gamma_i = 0$ is fixed and $\alpha_{i1} = \alpha > 0$ for $i = 1, \dots, d$, and $\mathcal{H} = \mathcal{N}(0, 1)$ and α is unknown

(ii) probit-normal model, i.e. $\gamma_i = 0$, for $i = 1, \dots, d$, is fixed and $\mathcal{H} = \mathcal{N}(0, 1)$

(iii) (quasi) loglog-normal model, i.e. $\gamma_i = 4.5658$, for $i = 1, \dots, d$, is fixed (cf. equation (1.15)) and $\mathcal{H} = \mathcal{N}(0, 1)$.

(iv) skew probit-normal model with $\gamma_i = \gamma$, for $i = 1, \dots, d$, and γ is unknown and $\mathcal{H} = \mathcal{N}(0, 1)$

(v) probit-skew normal model with $\gamma_i = 0$, for $i = 1, \dots, d$, is fixed and $\mathfrak{h} = f_\gamma$ for the density f_γ of the standard skew normal distribution in Definition 1.1.7

(vi) skew probit-skew normal model with $\gamma_i = \gamma_1 \in \mathbb{R}$, for $i = 1, \dots, d$, and γ_1 is unknown and $\mathfrak{h} = f_\gamma$ for the density f_{γ_2}, $\gamma_2 \in \mathbb{R}$ of the standard skew normal distribution

(vii) skew probit-skew normal model with $\gamma_i \in \mathbb{R}$, for $i = 1, \dots, d$, and $\gamma_1, \dots, \gamma_d$ is unknown and $\mathfrak{h} = f_\gamma$ for the density f_{γ_2}, $\gamma_2 \in \mathbb{R}$ of the standard skew normal distribution

In the following, we will illustrate some thoughts on how the models were selected. We chose models that cover a great variety of approaches and that represent classic models and models incorporating skewness of different sort and parsimony. On one hand, we chode model (i) and (ii) because they are commonly used models and their skewness is zero. Model (ii) is correct in the first setup and the most parsimonious model in there. Model (i) instead is always misspecified because it is never utilized for data generation in any setup. Examining model (i) in this study however is interesting because the Rasch model is at least a very parsimonious model and the spread of the actual parameters α_{i1} is bounded to values between 0.8 and 1.2. It is therefore interesting to see if a common discrimination for all items can be sufficient to obtain good estimates for the ICCs and the latent traits. Additionally, we chose (iii) to examine a model with a classic link function for skewness in the GLM literature. The link function of model (iii) is fixed and wrong specified in all setups. We still select this model to study if the loglog link still can capture the essential part of the actual skewness. Models (iv) and (v) model different types of skewness with a single additional parameter. They are correct in setup 2 and 5, respectively. We want to study how they perform in comparison to the wrong models (i)–(iii) and to the more complex models (vi)–(vii) in these cases. Moreover, it is interesting to see if the modeled skewness can compensate the neglected other type of skewness such that still good latent trait estimates are obtained in the setups where they are wrongly specified. Models (vi) and (vii) are the only models that are correctly specified in setup 3 as they take both types of skewness into account. Model (vi) is the more parsimonious model in setup 3 because it has two additional parameters to the probit-normal model while model (vii) has $d + 1$ additional parameters. There we can study the benefit of parsimony of model (vi) in comparison to model (vii). Moreover, we study how well the other models can compensate their differently strong misspecifications by selecting the remaining parameters. Setup 4 extends this objective in the sense that model (vii) is the only correctly specified model. Since all actual skewness parameters are positive, it might be the case that the majority of the item skewnesses can be captured by a single parameter to obtain a sufficiently good approximation. In contrast, model (vii) might fail to perform better because of its large number of parameters.

2.5.2.2 Results

The complete results of the simulation study are given in Tables 2.9, 2.10 and 2.11. In the following, we will summarize the results and point out a few noteworthy observation.

ICC Estimation Generally, the results of all models are, as expected, better for $n = 2000$ than for $n = 1000$. How the models perform relatively to each other, however, does not differ between $n = 1000$ and $n = 2000$ in most cases. Therefore, we will focus only on comparing $d = 10$ and $d = 20$. One thing of note, though, is that the results for model (vii) improve more greatly with the step from $n = 1000$ to $n = 2000$ than for all other models. For example, in setup 5 the mKSD of model 7 for $d = 10$ decreases from 0.310 for $n = 1000$ to 0.282 for $n = 2000$.

Setup 1: Models (ii) and (iv)–(vi) all perform in a very similar manner. Model (i) performs slightly worse than these models for $d = 10$ but, interestingly, slightly better for $d = 20$. Model (iii) has a considerably worse performance in comparison to all other models for $d = 20$, however, for $d = 10$ the mKSD of model (iii) is slightly better than that of model (i) while the aKSD is still worse.

Setup 2: For $d = 10$, model (vi) performs best overall. It is followed closely by model (iv) with very similar aKSD and worse mKSD values and model (iii), which performs similarly but overall a little worse. Model (i) has minimally better mKSD than model (vi) but worse aKSD values than models (vi, iv and iii). The remaining models (ii), (v) and (vii) perform worse overall than the other models but with slight differences in between them. Model (vii) has better aKSD than than the other models and model (ii) has the best mKSD values of the three. For $d = 20$, model (vi) has the best aKSD, followed by models (iv) and (i) with almost identical values and model (iii) with minimal differences. The next best aKSD has model (vii), followed with some distance by models (ii) and (v) whose aKSD values are equal. The best fit in terms of mKSD has model (i) by a large margin. It is followed by models (iii), (vi) and (iv) in that order with relatively similar values. With considerably worse mKSD values than those before them, models (vii), (ii) and (v) follow in that order.

Setup 3: The results for this setup are heterogeneous, therefore the comparison is more complex. Looking at the aKSD, model (vi) has the overall best values for $d = 10$ and $d = 20$. Model (v) has slightly better values for $d = 10$ but worse ones for $d = 20$. Models (ii) and (iii) perform very similarly with aKSD values close to the best models for $d = 10$ but worse performances for $d = 20$. Interestingly, model (i) has the second to worst aKSD value for $d = 10$ but the second best for $d = 20$. Model (vii) behaves similarly but to a lesser extend. For both d considered, model (iv) performs worst. When comparing the mKSD, however, model (iii) has the best value for $d = 10$ and the second best for $d = 20$. It is followed by model (ii) for both $d = 10$ and $d = 20$. Model (iv) follows for $d = 10$ but has the second to worst value for $d = 20$. In contrast, model (i) has the best mKSD of all models for $d = 20$ but follows after model (iv) for $d = 10$. Models (vi) and (vii) follow in that order, performing similarly for both $d = 10$ and $d = 20$. The worst performance in terms of mKSD has model (v).

Setup 4: The results for this setup are, interestingly, almost identical to that for setup 3, both in absolute and relative numbers. We will therefore only point out the differences here. When comparing the aKSD, model (vii) performs better for $d = 10$ in this setup, following after model (iii). For $d = 20$, model (iii) performs worse in comparison to the setup before and has the second to worst aKSD value here. Looking at the mKSD, the only difference for $d = 10$ is that the values of models (i) and (vi) are then closer to each other

than before. Analogously to the aKSD, model (iii) performs worse for $d = 20$ in comparison to setup 3.

Setup 5: Model (v) has the best aKSD values, followed by model (vi). For $d = 10$, model (ii) has the third best value, followed by (i) and (vii) while model (iv) and (iii) have notably worse aKSD values. For $d = 20$, model (i) has the third best values, followed by model (ii) and then model (vii) while model (iv) and (iii) again have the worst aKSD values. The best mKSD values are generally distributed in a similar manner to the aKSD values. However, there is one crucial difference: While model (i) only has the fourth ($n = 1000$) resp. fifth ($n = 2000$) best mKSD value for $d = 10$, it has the best mKSD value of all models for $d = 20$.

Latent Trait Estimation Generally, there is almost no difference between the performance of a model with respect to MAE or AAE. The SdAE values of models with smaller MAE/AAE values are smaller, too. In setup 1, the results for all models are very close. In the setups 2–5, the results are farther apart. The best models in setup 2 are model (iii) and model (iv). The best models in setup 3–5 are model (v)–(vii), where model (vi) is slightly better than the other two in setup 3–4 with model (vii) on the second place. In setup 5, model (v) is the best and slightly better than model (vi) on the second place. In setup 2, n has a large influence while d does not have a notable influence on the order of the models. In setup 3–5, d has an influence on the performance of some models in comparison to other models but d does not have this influence on all models.

Setup 1: Model (ii) and (iv) have the best values, followed closely by model (v) and then model (i) and model (vi) in that order. The worst performance overall has model (vii). Model (iii) has better values for $d = 10$ than model (vii) but performs very similarly for $d = 20$.

Setup 2: Model (iv) has the best performance overall and model (iii) has almost identical values with the exception of slightly worse values for $d = 20$ and $n = 1000$. They are followed by model (vi), which performs almost identically to models (iv) and (iii) for $n = 2000$ but worse for $n = 1000$. Then follows (ii), which performs slightly better for $d = 10$ in both cases of n. Afterwards, the ranking becomes more complex. Model (vii) has the next best overall performances with its best values being for $d = 20$ and $n = 2000$. Second to worst performs model (i) with the worst overall values for $n = 2000$ and then the worst overall performance has model (v) with the worst values for $n = 2000$.

Setup 3: For $d = 10$, the best result have model (v) and (vi), followed by model (vii) with slightly worse AAE and identical MAE. The next best results have model (iii) and (iv) with the same results but a considerable increase in values compared to the three aforementioned models. The next worse result has model (ii) and the worst result has model (i). For $d = 20$, the best result has model (vi) which is followed by model (vii) with slight differences. The next best model is model (v), followed closely by model (ii). With some difference, the next best model is model (iii). Models (i) and (iv) have the worst performance with only minor differences.

Setup 4: Analogously to the ICC results, setups 3 and 4 are almost identical in both absolute and relative numbers here. There are only very minor differences: model (vii) has AAE results on par with models (v) and (vi) in this setup while those models have very slightly worse MAE values for $d = 10$ and $n = 1000$ in comparison to setup 3. The performance of model (i) and model (iv) for $d = 20$ and $n = 2000$ differs slightly here, with model (i) having a better AAE than model (iv) in this setup in contrast to equal values in

setup 3.

Setup 5: For $d = 10$, the best result has model (v), followed first by model (vi) and then by model (vii). Models (i)–(iv) have similar values which are notably larger than those of models (v)–(vii). Of models (i)–(iv), model (i) has the best and model (ii) the worst results. For $d = 20$, the best three models are again models (v)–(vii) in that order. They are followed by model (i) and then (ii) with some difference. Model (iv) has the next best results with further difference and model (iii) has the worst results which are similar to that of model (iv).

To also get a visual impression of the structure in the final estimates of the latent traits, take a look at the scatter plots of the traits (vertical axis) and their estimates (horizontal axis) in Figure 2.9. We see a slight curve in the loglog model (iii), probit-skew normal model (v) and skew-probit-normal model (iv) and a tiny in the Rasch-type model (i) and the probit model (ii) as a small trend in the extreme values. In all cases, the estimation of small traits is less precise and concentrated on a few response pattern.

2.5.2.3 Discussion

We studied five setups to cover various practical situations. In practice, it is not commonly possible to decide whether the supposed model is true or misspecified. In the first setup, we considered a case in which a probit model holds which is the ideal case in practice. In the second setup, the true ICCs had the same positive skewness parameter. In the third setup, the true structural model additionally had a negative skewness. In the fourth setup, the item skewnesses were different (but still all positive) and the true structural model was skew. The fifth and final setup, the structural model had a skewness but the probit link was true for all items. In the first setup, the probit-normal model performed best but remarkably similar to the more complex models that contain the probit model as a special case. In the second to fifth setup, the probit-normal model was always wrong and also had worse fits compared to the other models. In contrast, the models with a single skewness parameter for all items and/or the structural model had good fits in the setups where the items or the structural model had non-zero skewness. Especially the model with a single skewness parameter for all items and a skewness parameter for the structural model had very good fits overall. The model with a skewness parameter for each item and a skewness parameter for the structural model – which is the only true model in fourth setup – suffered heavily from the large number of parameters and even in fourth setup, a model with a single skewness parameter for all items was superior. Fitting a model with one skewness parameter for all items and one skewness parameter for the structural model is hence worth trying as an alternative to a probit-normal model in practical settings, where it is commonly unknown a-priori if there is any skewness in the data. Models with more complex skewness structure, however, should only be considered if a large number of responses is observed because the fit is expected to be bad for a small number of responses due to the complexity of the model. Next, we will discuss all models in more detail.

Parsimony is a very important model attribute, a fact we can observe in this simulation study when focusing on model (i). Model (i), the Rasch-type model, is the most parsimonious of all models examined here because it only has a single parameter for each item. It was misspecified for all setups of this study, yet still produced comparably good estimates of the ICCs for $d = 20$. For cases of much larger d with a comparably large or even smaller n, Rasch-type models can provide good estimates and maintain some robustness.

Model (iii), the loglog-normal, incorporates ICC skewness (albeit in a fixed manner). It

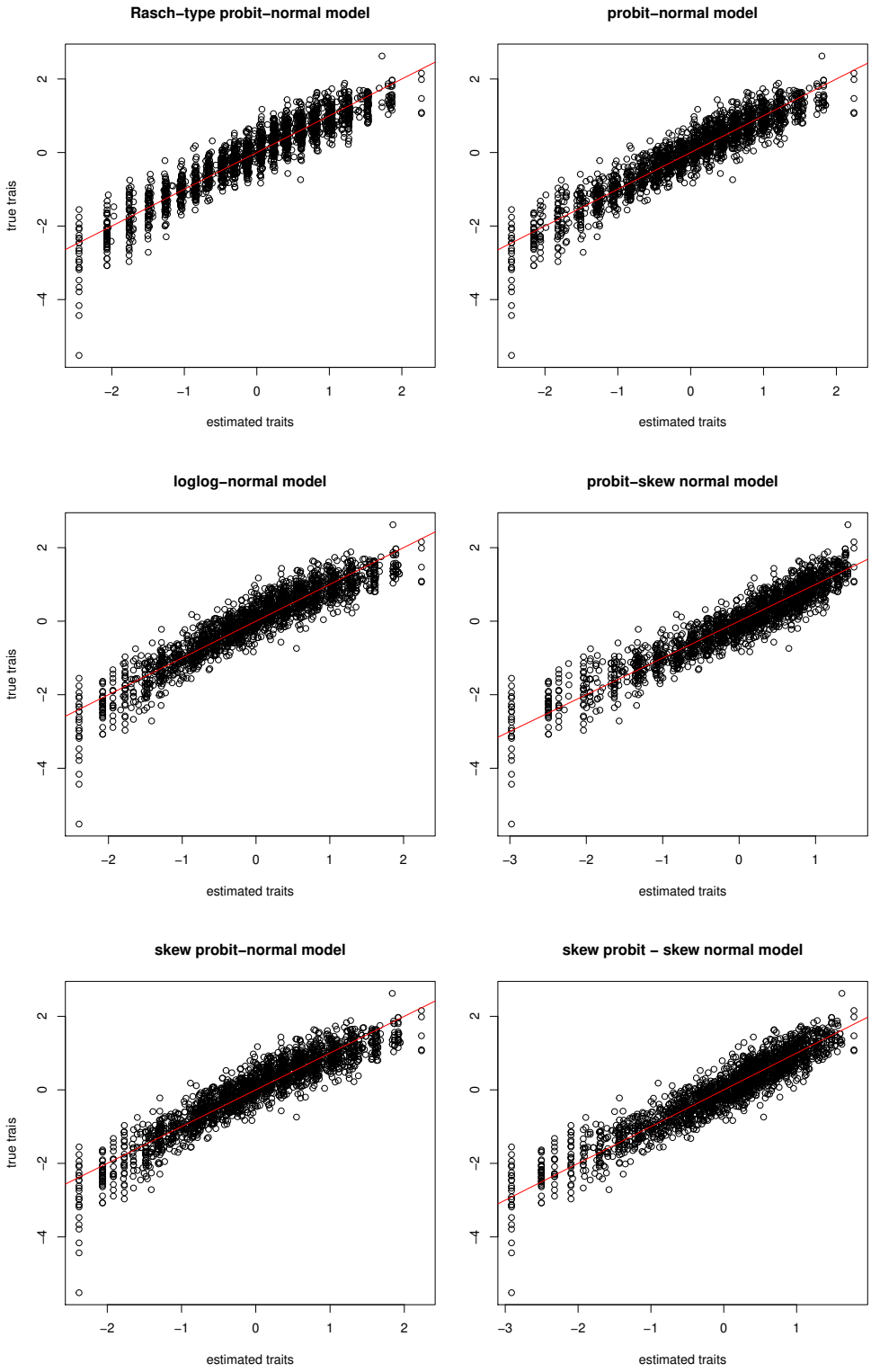

Figure 2.9: *Example scatter plots of the underlying true traits against their EAP estimates for the first six models in setup 4 of the simulation study in Subsection 2.5.2 with $d = 20$ and $n = 2000$. The red line is the identity. The closer the circles are to the red line the better is the estimate.*

was also misspecified in all setups but models ICC skewness in the correct direction in setups 2–4. As a result, it performed especially well in setup 2, where the ICC was skew but the structural model was not. It also yielded good estimates for the ICCs in setup 3 where both ICC and structural model were skew, though. This justifies the use of the loglog and cloglog link functions in IRT althugh they are currently primarily discussed in the GLM literature. Especially in settings with a smaller number of responses, these fixed link functions can be interesting because the skewness parameter can be expected to be difficult to estimate for a few hundred responses.

The loglog-normal struggles, however, in setups where there is skewness but in a different direction because it cannot be adapted for that. Models (iv)–(vi) improve upon this by introducing flexible skewness. Model (iv), the skew-probit-normal model, incorporates flexible ICC skewness, whereas model (v), the probit-skew-normal model, incorporates flexible structural model skewness. Model (vi), the skew-probit-skew-normal model, combines these two approaches and incorporated both flexible ICC as well as flexible structural model skewness. However, these models are still relatively parsimonious because they only have one skewness parameter (models (iv) and (v)) or one for a common skewness of all ICCs and one for the structural model (model (vi)). Thanks to this combination of flexibility and parsimony, these models in general and model (vi) especially have the best overall performance in this simulation study. Of particular interest for evaluating models (iv)-(vi) is setup 1. There, the probit-normal model (model (ii)) is the correct model as well as the most parsimonious one. However, neither its ICC nor its latent trait estimates are notably better than those of models (iv)–(vi). This means that the loss of accuracy when using the more complicated models (iv)–(vi) is not critical. Moreover, in the case of setup 4, where model (vii) is the only correctly specified model, model (vi) still has better ICC and latent trait estimates, which can explained by the greater parsimony of model (vi). Note, however, that in this study, the true underlying ICC skewness parameters are all positive. In a case of positive and negative skewness of ICCs, model (vi) may perform worse than model (vii). In conclusion, we propose that versions of model (vi) might be favorable for many practical data analysis situations for data sets of similar dimension due to its option to capture skewness in the ICCs and the structural model.

Model (vii), the (all) skew-probit-skew-normal model with one skewness parameter for each item, is the most complex of the models examined here. It has three parameters for each item and is thus the least parsimonous of the models by a large margin. In setups 3–5, it yields only slightly worse latent trait estimations than models (iv)–(vi) but considerably worse ICC estimations. However, in setups 1 and 2 it performs among the worst of all models examined. One should note, however, that the improvement of the ICC estimations for the step from $n = 1000$ to $n = 2000$ improve by the largest margin for model (vii). We can conclude from the asymptotic theory that model (vii) will provide good results for $n \to \infty$. One example of a practical application with very large n are large educational surveys, such as the PISA 2018 survey where approximately 710,000 students participated (cf. OECD, 2019, p. 12, Box 1.1). For n that large, the skewness parameters can be expected to be estimated reliably so that the estimates of model (vii) will be better than those produced by more parsimonious models in almost all cases.

2.5.3 Model Selection

Now we discuss which model from the simulation study in Subsection 2.5.2 is selected in which setup by the AIC. Additionally, we test model (i)–(vi) against model (vii) using a

			model (i) Rasch-type normal		model (ii) probit normal		model (iii) loglog normal		model (iv) skew-probit normal		model (v) probit skew-normal		model (vi) skew-probit skew-normal		model (vii) all skew-probit skew-normal	
setup	d	n	aKSD	mKSD	aKSD	mKSD	aKSD	mKSD	aKSD	mKSD	aKSD	mKSD	aKSD	mKSD	aKSD	mKSD
1	10	1000	0.170	0.302	0.166	0.272	0.197	0.294	0.166	0.272	0.164	0.273	0.167	0.266	0.175	0.285
		2000	0.167	0.292	0.163	0.255	0.195	0.276	0.163	0.254	0.163	0.253	0.163	0.250	0.170	0.259
	20	1000	0.172	0.309	0.174	0.338	0.198	0.383	0.174	0.338	0.173	0.341	0.172	0.340	0.179	0.371
		2000	0.173	0.299	0.175	0.326	0.201	0.384	0.175	0.324	0.175	0.325	0.173	0.327	0.176	0.344
2	10	1000	0.188	0.309	0.190	0.346	0.182	0.329	0.180	0.327	0.191	0.380	0.179	0.310	0.187	0.362
		2000	0.185	0.300	0.188	0.333	0.180	0.314	0.176	0.311	0.188	0.338	0.176	0.298	0.182	0.337
	20	1000	0.192	0.335	0.208	0.454	0.194	0.390	0.191	0.396	0.208	0.494	0.190	0.393	0.202	0.464
		2000	0.193	0.328	0.208	0.449	0.197	0.379	0.193	0.385	0.208	0.450	0.192	0.378	0.199	0.435
3	10	1000	0.190	0.310	0.186	0.293	0.186	0.280	0.201	0.294	0.182	0.336	0.186	0.315	0.187	0.326
		2000	0.189	0.307	0.185	0.285	0.185	0.269	0.201	0.281	0.181	0.332	0.180	0.310	0.183	0.318
	20	1000	0.198	0.354	0.203	0.390	0.201	0.373	0.212	0.412	0.203	0.447	0.196	0.391	0.200	0.415
		2000	0.196	0.349	0.201	0.373	0.198	0.363	0.210	0.406	0.199	0.433	0.192	0.371	0.196	0.389
4	10	1000	0.189	0.306	0.185	0.284	0.185	0.273	0.200	0.287	0.181	0.326	0.185	0.306	0.186	0.318
		2000	0.189	0.303	0.185	0.276	0.185	0.262	0.201	0.275	0.180	0.323	0.181	0.304	0.182	0.311
	20	1000	0.196	0.349	0.199	0.376	0.201	0.380	0.211	0.417	0.197	0.428	0.193	0.378	0.198	0.403
		2000	0.193	0.343	0.197	0.360	0.199	0.374	0.209	0.410	0.194	0.417	0.190	0.357	0.193	0.375
5	10	1000	0.178	0.308	0.176	0.289	0.210	0.328	0.204	0.322	0.164	0.270	0.173	0.281	0.186	0.310
		2000	0.177	0.307	0.175	0.278	0.209	0.320	0.201	0.308	0.163	0.258	0.168	0.263	0.180	0.282
	20	1000	0.182	0.320	0.185	0.380	0.217	0.468	0.205	0.439	0.175	0.340	0.178	0.345	0.191	0.404
		2000	0.179	0.307	0.182	0.376	0.214	0.469	0.202	0.438	0.173	0.319	0.175	0.324	0.186	0.391

Table 2.9: *The average values of the average and the maximum Kolmogorov–Smirnov distance of the estimated ICCs to the true ICCs based on 100 generated data sets for each case of the simulation study in Subsection 2.5.2.*

			model (i) Rasch-type normal			model (ii) probit normal			model (iii) loglog normal			model (iv) skew-probit normal		
setup	d	n	MAE	AAE	SdAE	MAE	AAE	SdAE	MAE	AAE	SdAE	MAE	AAE	SdAE
1	10	1000	0.300	0.357	0.274	0.299	0.356	0.272	0.301	0.358	0.274	0.299	0.356	0.272
		2000	0.297	0.356	0.275	0.296	0.354	0.273	0.298	0.356	0.275	0.296	0.354	0.273
	20	1000	0.225	0.270	0.210	0.225	0.269	0.209	0.229	0.274	0.213	0.225	0.269	0.209
		2000	0.224	0.269	0.209	0.223	0.268	0.208	0.227	0.272	0.212	0.223	0.268	0.208
2	10	1000	0.295	0.352	0.272	0.291	0.348	0.269	0.290	0.347	0.268	0.290	0.347	0.268
		2000	0.294	0.352	0.273	0.290	0.347	0.269	0.289	0.345	0.268	0.289	0.345	0.268
	20	1000	0.223	0.269	0.210	0.220	0.265	0.208	0.218	0.263	0.206	0.217	0.262	0.206
		2000	0.223	0.268	0.209	0.219	0.263	0.207	0.217	0.261	0.205	0.217	0.261	0.205
3	10	1000	0.296	0.356	0.285	0.295	0.355	0.284	0.291	0.351	0.282	0.291	0.351	0.282
		2000	0.294	0.354	0.284	0.293	0.353	0.283	0.290	0.349	0.280	0.290	0.349	0.280
	20	1000	0.225	0.272	0.218	0.219	0.264	0.212	0.223	0.270	0.218	0.225	0.273	0.221
		2000	0.224	0.271	0.218	0.217	0.262	0.211	0.221	0.268	0.218	0.223	0.271	0.221
4	10	1000	0.295	0.355	0.284	0.294	0.354	0.284	0.290	0.350	0.282	0.290	0.350	0.282
		2000	0.293	0.354	0.284	0.292	0.352	0.283	0.289	0.349	0.280	0.289	0.349	0.280
	20	1000	0.225	0.271	0.218	0.219	0.265	0.213	0.224	0.272	0.220	0.226	0.275	0.223
		2000	0.223	0.270	0.218	0.218	0.263	0.212	0.222	0.270	0.220	0.224	0.273	0.222
5	10	1000	0.301	0.362	0.291	0.303	0.364	0.292	0.301	0.363	0.291	0.301	0.363	0.291
		2000	0.298	0.361	0.290	0.300	0.363	0.291	0.299	0.362	0.291	0.299	0.362	0.290
	20	1000	0.226	0.273	0.220	0.228	0.276	0.223	0.237	0.288	0.234	0.235	0.285	0.231
		2000	0.226	0.272	0.220	0.227	0.275	0.222	0.237	0.287	0.234	0.234	0.283	0.231

Table 2.10: *The average values of the median and mean absolute error and the average standard deviation of the absolute error for the estimates of latent traits by the EAP based on 100 generated data sets for each case of the simulation study in Subsection 2.5.2 - Part I.*

setup	d	n	model (v) probit skew-normal			model (vi) skew-probit skew-normal			model (vii) all skew-probit skew-normal		
			MAE	AAE	SdAE	MAE	AAE	SdAE	MAE	AAE	SdAE
1	10	1000	0.299	0.357	0.273	0.301	0.359	0.275	0.304	0.362	0.278
		2000	0.296	0.354	0.274	0.297	0.355	0.275	0.300	0.359	0.278
	20	1000	0.225	0.270	0.210	0.227	0.272	0.212	0.229	0.276	0.216
		2000	0.223	0.268	0.208	0.225	0.270	0.210	0.226	0.272	0.213
2	10	1000	0.300	0.358	0.279	0.292	0.350	0.271	0.296	0.354	0.274
		2000	0.290	0.348	0.271	0.289	0.347	0.270	0.293	0.350	0.273
	20	1000	0.232	0.282	0.226	0.219	0.264	0.208	0.224	0.270	0.213
		2000	0.220	0.264	0.208	0.217	0.261	0.205	0.219	0.264	0.208
3	10	1000	0.282	0.339	0.268	0.282	0.339	0.270	0.282	0.340	0.269
		2000	0.281	0.337	0.266	0.281	0.337	0.266	0.281	0.338	0.266
	20	1000	0.219	0.263	0.205	0.215	0.258	0.202	0.216	0.259	0.203
		2000	0.217	0.261	0.204	0.212	0.255	0.201	0.213	0.256	0.201
4	10	1000	0.282	0.338	0.268	0.281	0.339	0.269	0.281	0.339	0.269
		2000	0.281	0.337	0.266	0.281	0.337	0.266	0.281	0.337	0.266
	20	1000	0.219	0.263	0.206	0.215	0.259	0.203	0.217	0.260	0.204
		2000	0.217	0.261	0.205	0.213	0.256	0.202	0.214	0.257	0.202
5	10	1000	0.290	0.350	0.279	0.291	0.353	0.281	0.295	0.356	0.283
		2000	0.289	0.349	0.277	0.289	0.350	0.278	0.292	0.353	0.281
	20	1000	0.220	0.264	0.209	0.221	0.265	0.210	0.224	0.269	0.214
		2000	0.218	0.263	0.208	0.219	0.263	0.209	0.221	0.267	0.212

Table 2.11: *The average values of the median and mean absolute error and the average standard deviation of the absolute error for the estimates of latent traits by the EAP based on 100 generated data sets for each case of the simulation study in Subsection 2.5.2 - Part II.*

LR test (cf. Theorem 2.4.1 (b)). Since model (vii) is correct in each case, this can be considered a test of fit of the parsimonious models. Simultaneously, it is a test of the link function and the structural model for the models (ii)–(vii) because the predictor is the same in these cases – in model (i), all items have a common discrimination parameter which is tested, too. In Table 2.12 and 2.13 we summarize how often each model was selected by the AIC. Additionally, we summarize the empirical distributions of the asymptotic p-values by their 25%, 50% and 75% quantiles in order to not restrict the presentation to arbitrary significance levels.

2.5.3.1 Results

Setup 1: Models (ii) and (iv)–(vii) are correct. This is correctly captured by the p-values of the LR tests. Models (i) and (iii) (correctly) have much smaller p-values corresponding to the sample size and would be rejected by the consideration of the common significance levels. The most parsimonious correct model is model (ii), which is selected by the AIC in most cases. In a few cases, other correct models are selected. In 14 cases, the wrong model (i) is selected. However, 12 of these 14 cases are for $d = 10$ and $n = 1000$, i.e. the smallest data sets.

Setup 2: Models (iv), (vi) and (vii) are correct. Additionally, model (iii) is close in the sense that the fixed skewness has a relatively small departure from the correct model ($\gamma = 3.5$ versus $\gamma \approx 4.57$, recall the decreasing effect of γ for $|\gamma| \to \infty$). This is visible in the p-values of the LR tests for the models (iv) and (vi). For $n = 1000$ the test fails to reject model (iii) and the p-values decrease for $n = 2000$, but the percentage of selection is high for model (iii) and smaller data sets. For $d = 20$ and $n = 2000$, the most parsimonious correct model (iv) is most often selected. Overall, these two models are selected in most cases. The wrong model (v) is not reliably rejected for $n = 1000$ but is correctly rejected for $n = 2000$. Models (i) and (ii) are correctly detected as wrong and model (ii) is only selected in two cases.

Setup 3: Models (vi) and (vii) are the only correct models and model (vi) is the more parsimonious one. This is reflected correctly in the data.

Setup 4: Model (vii) is the only correct model. However, model (vi) is selected in most cases. For $d = 20$ and $n = 2000$, the LR test has a higher rejection rate for this model than under the hypothesis. However, in the other three cases the LR test does not detect that model (vi) is wrong. For $d = 10$ and $n = 1000$, i.e. the smallest data set, model (iv) is selected in 20 cases and has high p-values. Except for a few data sets, all remaining wrong models are reliably rejected and not selected by the AIC.

Setup 5: Models (v)–(vii) are correct and model (v) is the most parsimonious one. For comparably small data sizes, models (iii)–(iv) are rejected in comparably few cases and, especially model (iii) for $d = 10$ and $n = 1000$, selected by the AIC in several cases. Otherwise, the selection and test decisions correspond to the expectations from the theory.

2.5.3.2 Discussion

The LR tests as well as the model selection using the AIC led overall to the correct decisions. One notable exception is the selection of model (i) in some cases in setup 1. This can be explained by the comparably small data sets and the general parsimony of model (i) while it is similar to the correct model (ii). The observation of the previous simulation study that the latent trait estimates using model (i) are only slightly worse than those using the

correct models imply that the wrong decisions using the AIC are not critical in these cases. The situation is similar in setup 2 with the wrong selections of model (iii). The last notable exception is setup 4 with the wrong selections of model (vi). This can be explained by the too large number of parameters of model (vii), resulting in the more parsimonious (but still similar and flexible) model (vi) being preferred by the AIC. The high p-values of model (iv) for $d = 10$ and $n = 1000$ can be explained by the comparably small data size and the advantage of parsimony. We therefore conclude that the nested model comparison using the LR test and the model selection using the AIC are reliable for skew-probit-skew-normal models. We expect that this is generally true for EGLLVM.

2.6 Summary, Discussion and Outlook

In this chapter we introduced an extension of GLLVM for binary data with respect to parametric link functions and parametric structural models (EGLLVM). We have seen that several flexible IRT models which have been discussed in the literature are contained in the EGLLVM class.

By introducing a new model for guessing behavior of test participants which is contained in the EGLLVM model class we illustrated further areas of application of the EGLLVM class. The new guessing model provides additional interesting theoretical properties of their ICCs in comparison to other guessing models. Applications to a simulated and a real data set revealed issues with estimating the link function parameters of the new model which are similar to the issues of fitting the 3PL model.

We proved the asymptotic existence, consistency and asymptotic normality of the MMLE for EGLLVM under suitable conditions on the model components. In addition to the fact that these are generally important properties, the consistency is interesting because it ensures that the link function parameters can eventually be estimated well if only enough data is available. This especially enables the use of flexible EGLLVM for large data sets such as these of the OECD PISA surveys. Moreover, this asymptotic theory contains many classic parametric IRT models for which these properties of the MMLE are commonly stated but no proof is available, such as the commonly used logit and probit IRT models.

Using this asymptotic result of the MMLE for EGLLVM, we also proved that the Wald, score, and likelihood-ratio tests have the same asymptotic χ^2 distribution which they have for many other models. For simple hypotheses, the conditions are the same as the conditions on the asymptotics of the MMLE. Moreover, we provide a set of further conditions to perform conditional tests on nested models of flexible complexity using all three types of tests. Especially using the likelihood-ratio test, we can test the link function of a (E)GLLVM or a structural model against a parametric alternative. An important example for many practical settings is the test of the probit link in a GLLVM for binary data against a skew-probit link with a single skewness parameter for each item. Hence, we can test if for any item a skew alternative allows for a significantly better fit. We can especially test if a loglog or cloglog link – i.e. the two most important fixed skew link functions for binary GLM – would be significantly better for some items because the skew-probit link allows a very precise approximation of the loglog and cloglog link.

The fact that these tests for the link functions and the structural model are conditional tests has the advantage that it provides a higher sensitivity against a specific alternative which is given by the more complex model. The test alternative also can reveal aspects of the departure from the stated model. A disadvantage is that the more complex model has

			model (i) Rasch-type normal				model (ii) probit normal				model (iii) loglog normal			
setup	d	n	PS	Q1	median	Q3	PS	Q1	median	Q3	PS	Q1	median	Q3
1	10	1000	12	1.86e-04	1.84e-03	1.91e-02	73	2.97e-01	5.50e-01	7.71e-01	0	4.28e-05	6.53e-04	6.47e-03
		2000	0	3.59e-09	3.52e-07	4.10e-05	82	1.97e-01	4.76e-01	7.95e-01	0	5.86e-11	1.07e-08	1.18e-06
	20	1000	2	8.36e-07	1.80e-05	3.99e-04	66	2.68e-01	5.63e-01	7.69e-01	0	7.24e-10	5.53e-08	3.74e-06
		2000	0	7.74e-15	4.42e-12	1.78e-10	85	3.27e-01	5.64e-01	7.97e-01	0	0	0	4.44e-15
2	10	1000	0	2.04e-08	1.91e-06	2.00e-05	2	1.22e-03	6.88e-03	4.22e-02	68	2.54e-01	5.68e-01	8.46e-01
		2000	0	0	5.50e-15	1.23e-12	0	9.46e-07	5.05e-05	9.62e-04	45	1.90e-01	4.05e-01	6.47e-01
	20	1000	0	0	0	0	0	4.74e-06	4.18e-05	4.36e-04	40	1.71e-01	5.07e-01	7.28e-01
		2000	0	0	0	0	0	1.14e-12	6.09e-11	7.47e-09	23	1.25e-01	3.59e-01	5.94e-01
3	10	1000	0	1.01e-13	2.89e-12	4.08e-10	0	1.93e-12	6.84e-11	5.01e-09	3	2.54e-02	6.99e-02	2.48e-01
		2000	0	0	0	0	0	0	0	0	0	1.93e-04	3.91e-03	2.07e-02
	20	1000	0	0	0	0	0	0	0	0	0	7.90e-04	1.04e-02	4.03e-02
		2000	0	0	0	0	0	0	0	0	0	3.56e-07	1.21e-05	1.89e-04
4	10	1000	0	5.99e-14	2.11e-12	5.69e-10	0	1.02e-12	1.13e-10	6.58e-09	3	1.49e-02	6.71e-02	2.37e-01
		2000	0	0	0	0	0	0	0	0	0	2.33e-04	2.27e-03	1.36e-02
	20	1000	0	0	0	0	0	0	0	1.11e-16	1	2.24e-03	1.08e-02	7.07e-02
		2000	0	0	0	0	0	0	0	0	0	3.03e-06	4.61e-05	4.43e-04
5	10	1000	0	4.79e-11	1.27e-08	4.75e-07	0	1.42e-04	3.30e-03	1.51e-02	41	1.73e-01	3.73e-01	6.35e-01
		2000	0	0	0	6.94e-16	0	1.46e-07	5.18e-06	2.56e-04	21	9.06e-02	2.54e-01	4.83e-01
	20	1000	0	1.96e-13	6.18e-12	8.37e-10	0	3.97e-06	5.62e-05	5.57e-04	12	3.30e-02	1.44e-01	3.87e-01
		2000	0	0	0	0	0	5.47e-12	1.23e-10	2.41e-09	2	4.77e-03	3.13e-02	1.22e-01

Table 2.12: *The percentage of the selection of each model by the minimal AIC value (PS) and the 25%, 50% and 75% quantiles of the asymptotic p-values of likelihood ratio tests against model (vii) based on 100 generated data sets for each case of the simulation study in Subsection 2.5.3 - Part I.*

			model (iv) skew-probit normal				model (v) probit skew-normal				model (vi) skew-probit skew-normal				model (vii) all skew-probit skew-normal
setup	d	n	PS	Q1	median	Q3	PS	Q1	median	Q3	PS	Q1	median	Q3	PS
1	10	1000	6	2.64e-01	5.51e-01	7.64e-01	8	2.83e-01	5.49e-01	8.13e-01	1	2.31e-01	4.93e-01	7.52e-01	0
		2000	7	1.73e-01	4.29e-01	7.44e-01	11	1.81e-01	4.19e-01	7.50e-01	0	1.55e-01	4.04e-01	6.93e-01	0
	20	1000	13	2.99e-01	5.58e-01	7.75e-01	15	2.59e-01	5.34e-01	7.70e-01	4	2.47e-01	5.27e-01	7.48e-01	0
		2000	8	2.94e-01	5.41e-01	7.87e-01	4	3.06e-01	5.21e-01	7.60e-01	3	2.68e-01	5.57e-01	7.59e-01	0
2	10	1000	9	3.27e-01	6.28e-01	8.46e-01	9	1.51e-01	4.32e-01	7.12e-01	12	3.48e-01	6.05e-01	8.67e-01	0
		2000	41	3.23e-01	5.57e-01	7.41e-01	4	4.07e-07	3.01e-05	1.53e-03	8	2.73e-01	5.17e-01	7.44e-01	2
	20	1000	36	2.83e-01	6.02e-01	7.72e-01	8	8.14e-02	2.31e-01	5.44e-01	16	3.00e-01	6.07e-01	8.11e-01	0
		2000	68	3.39e-01	5.97e-01	8.22e-01	0	4.93e-13	5.46e-11	6.78e-09	8	3.08e-01	5.76e-01	8.13e-01	1
3	10	1000	20	1.29e-01	3.31e-01	6.36e-01	2	9.05e-03	3.43e-02	1.37e-01	75	4.51e-01	7.18e-01	9.16e-01	0
		2000	7	9.84e-03	5.47e-02	1.86e-01	0	1.40e-04	1.53e-03	1.02e-02	92	3.18e-01	5.62e-01	8.02e-01	1
	20	1000	2	8.24e-03	3.69e-02	1.59e-01	0	6.39e-04	3.65e-03	1.66e-02	98	3.74e-01	6.24e-01	8.41e-01	0
		2000	0	6.86e-05	4.79e-04	9.18e-03	0	2.16e-07	2.57e-06	2.06e-05	100	3.59e-01	5.63e-01	7.76e-01	0
4	10	1000	20	9.98e-02	3.45e-01	6.44e-01	2	7.58e-03	3.50e-02	1.40e-01	74	4.70e-01	7.02e-01	9.19e-01	1
		2000	5	1.37e-02	6.95e-02	1.89e-01	0	1.71e-04	8.41e-04	7.35e-03	93	3.61e-01	5.34e-01	7.50e-01	2
	20	1000	2	7.38e-03	4.25e-02	1.44e-01	0	1.76e-03	6.13e-03	3.33e-02	93	2.26e-01	4.59e-01	7.43e-01	4
		2000	0	4.72e-05	7.40e-04	6.01e-03	0	1.70e-06	1.75e-05	1.58e-04	97	1.15e-01	3.16e-01	5.69e-01	3
5	10	1000	11	1.83e-01	4.03e-01	7.18e-01	45	2.48e-01	4.16e-01	6.71e-01	2	2.28e-01	4.48e-01	7.20e-01	1
		2000	8	1.20e-01	3.37e-01	6.53e-01	65	2.54e-01	4.60e-01	7.64e-01	4	2.08e-01	4.90e-01	7.39e-01	2
	20	1000	10	1.09e-01	3.22e-01	5.62e-01	73	2.36e-01	4.69e-01	7.18e-01	5	2.18e-01	4.46e-01	7.06e-01	0
		2000	7	5.97e-02	1.61e-01	4.16e-01	85	2.89e-01	5.28e-01	7.85e-01	5	2.38e-01	4.81e-01	7.72e-01	1

Table 2.13: *The percentage of the selection of each model by the minimal AIC value (PS) and the 25%, 50% and 75% quantiles of the asymptotic p-values of likelihood ratio tests against model (vii) based on 100 generated data sets for each case of the simulation study in Subsection 2.5.3 - Part II.*

to be fitted, too. A goodness of link test that is not only an conditional goodness of link test is then given by a two step testing procedure. First, a goodness of fit of the complex model that is the alternative in the desired test for the link function parameters or the structural model. Thus the condition of the test of a link function or structural model against a complex alternative is tested by nested model comparison. Then, the nested model test is performed. A foundation for these goodness of fit tests is given by the literature regarding goodness of fit tests of IRT models and contingency tables through the consideration of the data as 2^d contingency table.

By performing three simulation studies, we confirmed the theory that the skewness parameters of skew probit models and skew normal structural models can consistently be estimated for $n \to \infty$. However, we also revealed that the skew-probit link functions have estimation issues which are similar to the 3PL model. These are important observations because the skew-probit and similar models are currently only discussed in the literature by applying Bayesian methods, which avoid this issue via the prior distribution. Our results suggest that this is no unique issue of the skew-probit link function but may be a result of the comparably small influence of the link function parameters on the distribution of the response vector in comparison to the factor loadings and the intercept parameters. A second simulation study reveals that skew-probit models with a common skewness parameter for all items provide a good trade-off between flexibility and model parsimony. This study moreover showed that the ICC and latent trait estimates for comparably parsimony skew-probit models are similarly good as the estimates for probit-normal models if the latter is true. The skew-probit-skew-normal normal therefore can be favorable in practice.

The result of the third simulation study is that LR tests for nested model comparisons (and especially for testing the link functions) as well as the AIC for selecting a model work well and are reliable for a few thousand responses. These tools for testing and selecting link functions do not seem to be affected from the estimation issues of models with a skewness parameter for each item in the examined data sizes. If they work well for a few hounded n is currently unknown because we only considered $n = 1000$ and $n = 2000$ in this study. We conclude that these tools for selecting and testing link functions and structural models of EGLLVMs is notable robust against the estimation issues of parametric link functions.

In future research it would be interesting to examine the performance of the comparably parsimonious skew-probit-skew-normal models for other sizes of the data sets. It would also be interesting to study other parametric link functions and multidimensional models.

The current theory contains tests for the skewness using conditional tests. The study of GoF tests for EGLLVM therefore is therefore interesting for future research. This could contain the adaption of GoF tests for classic IRT models to EGLLVM, e.g. the limited-information test discussed by Cai et al. (2006).

Chapter 3

The Asymptotic Posterior Distribution of Latent Vectors for Binary Items

Parts of this chapter are contained in a draft that has been submitted for publication but has yet to be accepted or rejected. The draft contains the literature summary in this introduction and Sections 3.1 to 3.4. The content is adapted to the context of the dissertation and the chapter is not identical to the submitted draft.

We have seen in Remark 1.1.2 that item response theory (IRT) fits naturally in the emprical Bayesian framework. In Subsection 1.3.4 we have then seen that this results practically in the two steps[1] of

(i) fitting the (M)IRT model to the data and then

(ii) estimating the latent traits (LTs) for each observed response pattern.

In Section 2.3 we have then proved that the marginal maximum likelihood estimator for (M)IRT model parameters is consistent and asymptotically efficient for $n \to \infty$ for a large class of IRT models when item number of items d is held fixed, where n denotes the sample size. Now we suppose that the (M)IRT model is given and we focus on the second step of estimating the latent variable/vector. In particular we analyze the asymptotics for $d \to \infty$.

In the context of IRT methodology, statistical inference for the examinee's ability often relies on the assumption that its posterior distribution given the test response is a normal distribution. As this is usually hard to justify and in contradiction to common models of the distribution examinee's abilities in the population, it can assumed to be well approximated by a normal distribution for a long test. This assumption of asymptotic posterior normality (APN) is part of the famous Dutch Identity conjecture of Holland (1990), who mentioned then that he was not aware of a thorough discussion of APN of latent variables and that it would be an interesting area for future research. Shortly afterwards, Chang and Stout (1993) proved the APN for univariate LTs, mentioning that APN for multivariate LTs can be proved, but without providing further details or discussing the associated regularity conditions required. As far as we know, APN of multivariate LTs has not yet been proved for MIRT models in a general context, although posterior normality or APN is assumed quite often under various IRT setups (cf. e.g. Anderson and Vermunt, 2000; Anderson et al., 2007; Li, 2010; Wang and Chang, 2011; Hessen, 2012; Paek, 2016; Anderson and Yu, 2017). Sometimes this assumption is justified by the APN in a Bayesian framework (pointing to the 'Bernstein–von Mises Theorem') without however discussing further details (e.g. the computationally efficient adaptive quadrature methods for high dimensional item factor analysis (Schilling and Bock, 2005) and for generalized linear mixed models (Rabe-Hesketh et al., 2002) are based on the APN assumption).

[1] Except a few alternative approaches like the joint maximum likelihood estimation of the model parameters and the latent traits or a joint Bayesian estimation.

In this chapter, we study the APN for multivariate LT models. We focus on models for dichotomous items. APN of LTs, univariate or multivariate, is related to Bayesian asymptotics. In the light of this connection, we embed APN of LTs in the approach of Ghosal et al. (1995), who discussed asymptotic posterior distributions in a very general setup that includes the regular cases and some non-regular cases as well. They also proved a general result on the asymptotic equivalence of the Bayes and maximum likelihood estimates, a well-known result for the regular cases. In particular, we generalize the approach and results of Chang and Stout (1993), CS hereafter, linking them to the semiproper centering concept of Ghosal et al. (1995), GGS hereafter, and embedding them in their approach. Since the conditions for APN imposed in GGS are not convenient in an IRT framework, we provide alternative conditions that correspond one to one to the conditions of CS. For the case of univariate LTs, this alternative approach has the advantage that it also applies to models with non-monotone item response functions (their monotonicity is an assumption required in the CS setup).

Furthermore we discuss conditions under which the existence of the maximum likelihood estimators (MLEs) for latent vectors is ensured. The consistency of MLEs under mild conditions, an issue which was suggested as a future research topic by Sinharay (2015), follows as a natural consequence from the proof of the APN. Then, we prove the consistency of expected a posteriori (EAP) estimators for multivariate LTs.

Additionally, we discuss the necessary regularity conditions to obtain the proper centering concept of GGS in the IRT framework (PCT for *proper centering theory* hereafter) by adapting the approach by Ghosal (1997, 1999). We simultaneously prove the asymptotic normality of the MLE and the EAP for MIRT models, another issue in MIRT pointed out by Sinharay (2015). In the special case of univariate LTs, we obtain the asymptotic normality of a large class of Bayes estimators for the LT.

Moreover, we prove the locally uniform convergence of the moment generating function (MGF) of the posterior distribution to the MGF of a multivariate normal distribution in probability. This strong theoretical property of the posterior distribution is generally new in the MIRT framework to the best of our knowledge. The convergence of the MGF justifies the approximation of arbitrary moments of the posterior distribution by that of a normal distribution – a property which cannot be concluded from the semiproper centering theory. In particular, the convergence of the MGF of the posterior was an important assumption of Zhang and Stout (1997) in their discussion of the Dutch Identity conjecture. We are optimistic that this result will help to make the conjecture rigorous in future research.

The chapter is organized as follows. Basic notation and the adopted IRT framework is set in Section 3.1 while the CS theory for a univariate LT is briefly reviewed in Subsection 3.2. The approach of GGS is discussed and linked to the APN of LTs and the CS results in Subsection 3.3.

The subject of Section 3.4 is the generalization of the CS theory (SCT for *semiproper centering theory* hereafter) and the subject of Section 3.5 is the PCT. Both sections are organized the same way. First, the necessary conditions are stated and commented (Subsection 3.4.1 and 3.5.1). Next, the main results regarding the APN and additional properties of the posterior and point estimators are formulated (Subsection 3.4.2 and 3.5.2). This is followed by a summary and discussion of the results with a formulation of future research perspectives (Subsection 3.4.3 and 3.5.3). The discussion of the conditions for an example model is subject of Subsection 3.4.4 and 3.5.4. Finally, the theorems and preliminary results are proved in Subsection 3.4.5 and 3.5.5.

3.1 Preliminaries

This chapter deals with the asymptotics for $d \to \infty$ and hence needs an adapted notation. Consider a test consisting of d binary response variables $\boldsymbol{Y}^{(d)} = (Y_1, \dots, Y_d)^\intercal$, with $Y_i \in \{0,1\}$, (1 denotes a correct response, 0 an incorrect one), $i \in [d] := \{1, \dots, d\}$ and defined over a probability space $(\Omega, \mathcal{A}, \mathsf{P})$. Thus, the manifest probability for a specific response pattern $\boldsymbol{y}^{(d)}$ is given by $P(\boldsymbol{y}^{(d)}) := \mathsf{P}(\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)})$. In a MIRT modeling framework, manifest probabilities are derived via conditioning on an absolutely continuous latent vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_q)^\intercal \in \Theta \subseteq \mathbb{R}^q$, defined over the same probability space as the binary items with probability density function (pdf) and cumulative distribution function (cdf) $\mathfrak{h}$ and $\mathcal{H}$, respectively. In particular, the conditional probability mass function (pmf) of $Y_i \,|\, \boldsymbol{\eta}$ is given by

$$\mathsf{P}(Y_i = y \,|\, \boldsymbol{\eta}) = P_i(\boldsymbol{\eta})^y (1 - P_i(\boldsymbol{\eta}))^{1-y}, \quad y \in \{0,1\}, \quad i \in [d], \tag{3.1}$$

with $P_i \colon \Theta \to (0,1)$ being known as the i-th item response function. In MIRT modeling, specific assumptions are usually imposed on the conditional distribution $\mathsf{P}(\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)} \,|\, \boldsymbol{\eta})$; namely the assumption of *local independence*

$$P^{(d)}(\boldsymbol{y}^{(d)} \,|\, \boldsymbol{\eta}) := \mathsf{P}(\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)} \,|\, \boldsymbol{\eta}) = \prod_{i=1}^{d} \mathsf{P}(Y_i = y_i \,|\, \boldsymbol{\eta}), \quad \boldsymbol{y}^{(d)} \in \{0,1\}^d, \tag{3.2}$$

and that of *monotonicity* for the item response functions P_i, i.e. for $i \in [d]$

$$P_i(\boldsymbol{\eta}) \text{ is strictly monotonic in every dimension of } \boldsymbol{\eta} \text{ being measured.} \tag{3.3}$$

Note that assumption (3.3), which is required in the CS approach, is relaxed in our setup. In the following, we write $\{Y_i\}_{i \in \mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta})$ for a sequence of Bernoulli random variables that fulfill (3.1) and (3.2) for all $d \in \mathbb{N}$.

Due to assumption (3.2) and using (3.1), the manifest probabilities are derived through the following integral

$$P(\boldsymbol{y}^{(d)}) = \int_\Theta \prod_{i=1}^{d} P_i(\boldsymbol{\eta})^{y_i} (1 - P_i(\boldsymbol{\eta}))^{1-y_i} \, \mathcal{H}(\mathrm{d}\boldsymbol{\eta}) \tag{3.4}$$

The posterior density of $\boldsymbol{\eta}$, given an observed response $\boldsymbol{y}^{(d)} \in \{0,1\}^d$, is then given by

$$h(\boldsymbol{\eta} \mid \boldsymbol{y}^{(d)}) := \mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)}) = \frac{\mathsf{P}(\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)} \mid \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta})}{\mathsf{P}(\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)})} = \frac{\exp(\ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{y}^{(d)})) \mathfrak{h}(\boldsymbol{\eta})}{P(\boldsymbol{y}^{(d)})}, \tag{3.5}$$

where $\ell^{(d)}(\,\bullet \,|\, \boldsymbol{y}^{(d)}))$ is the log-likelihood coresponding to (3.1), given by

$$\ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{y}^{(d)})) = \sum_{i=1}^{d} \left(y_i \lambda_i(\boldsymbol{\eta}) - \psi(\lambda_i(\boldsymbol{\eta})) \right), \tag{3.6}$$

with $\lambda_i(\boldsymbol{\eta}) := \operatorname{logit}(P_i(\boldsymbol{\eta}))$, $\boldsymbol{\eta} \in \Theta$, denoting the logit of item $i \in \mathbb{N}$ and the function ψ being defined as

$$\psi \colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \log\left(1 + \exp(x)\right). \tag{3.7}$$

Let $\hat{\boldsymbol{\eta}}_d = \hat{\boldsymbol{\eta}}(\boldsymbol{y}^{(d)})$ denote the MLE of the true value of the latent vector $\boldsymbol{\eta}_0$, based on a test realization $\boldsymbol{y}^{(d)}$. Furthermore, the Fisher information matrix of the test at point $\boldsymbol{\eta}$ is given by

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}) \coloneqq \mathsf{E}_{\boldsymbol{\eta}}\left(\nabla \ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})\ \nabla^{\mathsf{T}} \ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})\right) = \sum_{i=1}^{d} \mathcal{I}_i(\boldsymbol{\eta}), \tag{3.8}$$

where $\mathcal{I}_i(\,\bullet\,)$ is the i-th item information matrix

$$\begin{aligned}\mathcal{I}_i(\boldsymbol{\eta}) \coloneqq& \mathsf{E}_{\boldsymbol{\eta}}\left(\nabla \log\left(P_i(\boldsymbol{\eta})^{Y_i}(1-P_i(\boldsymbol{\eta}))^{1-Y_i}\right)\ \nabla^{\mathsf{T}} \log\left(P_i(\boldsymbol{\eta})^{Y_i}(1-P_i(\boldsymbol{\eta}))^{1-Y_i}\right)\right)\\ =& P_i(\boldsymbol{\eta})(1-P_i(\boldsymbol{\eta}))\nabla\lambda_i(\boldsymbol{\eta})\nabla\lambda_i(\boldsymbol{\eta})^{\mathsf{T}}, \quad \boldsymbol{\eta} \in \Theta.\end{aligned}$$

This work studies the APN of $\boldsymbol{\eta}$ for $d \to \infty$ based on a sequence of random variables $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta})$, as defined above. Particularly, we shall prove that under certain conditions, (3.8) is invertible at $\hat{\boldsymbol{\eta}}_d$ and $\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}$ is approximately normal distributed with $\mathcal{N}_q(\hat{\boldsymbol{\eta}}_d, \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1})$. This makes the approximation of probabilities of the type

$$\mathsf{P}\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}\,(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \in B \,\middle|\, \boldsymbol{Y}^{(d)}\right), \quad B \in \mathcal{B}^q, \tag{3.9}$$

possible where $\mathcal{B}^q$ denotes the Borel-σ-algebra of $\mathbb{R}^q$. Moreover, we can consider $\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1}$ an estimator of $\boldsymbol{\Sigma}_d \coloneqq \mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})$ if $\hat{\boldsymbol{\eta}}_d$ exists and $\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)$ is regular. To simplify the notation and get rid of unnecessary case studies, we define $\widehat{\boldsymbol{\Sigma}}_d$ as follows

$$\widehat{\boldsymbol{\Sigma}}_d \coloneqq \begin{cases} \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1}, & \text{if } \hat{\boldsymbol{\eta}}_d \text{ and } \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1} \text{ exist,} \\ \mathrm{I}_q, & \text{otherwise,} \end{cases} \tag{3.10}$$

for all $d \in \mathbb{N}$.

Next, we define some functions that are useful for the derivations in this chapter. For all $d \in \mathbb{N}$, set $Z^{(d)} \colon \Theta \times \Theta \to \mathbb{R}$ with

$$Z^{(d)}(\boldsymbol{\eta}, \boldsymbol{\eta}') \coloneqq \prod_{i=1}^{d} Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}'), \tag{3.11}$$

where $Z_i \colon \Theta \times \Theta \to \mathbb{R}$, $i \in \mathbb{N}$, is defined as

$$Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}') \coloneqq \frac{P_i(\boldsymbol{\eta})^{Y_i}(1-P_i(\boldsymbol{\eta}))^{1-Y_i}}{P_i(\boldsymbol{\eta}')^{Y_i}(1-P_i(\boldsymbol{\eta}'))^{1-Y_i}}\ . \tag{3.12}$$

Note that for given d and $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta$, (3.11) is the likelihood ratio of the likelihoods for $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$. Furthermore

$$\log\left(Z^{(d)}(\boldsymbol{\eta}, \boldsymbol{\eta}')\right) = \sum_{i=1}^{d} \log Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}') = \ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}' \mid \boldsymbol{Y}^{(d)}), \tag{3.13}$$

while $-\mathsf{E}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}_0))$ is the Kullback–Leibler divergence between the conditional distributions of Y_i given $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_0$, respectively. A basic approach for deriving APN results relies on a quadratic approximation of (3.13).

3.2 Review of APN for univariate latent traits

In the case of a single latent variable ($q = 1$, $\boldsymbol{\eta} = \eta$), Chang and Stout (1993) proved the APN of the univariate latent trait, adopting the approach of Walker (1969) for binary Y_i, $i \in [d]$, that are independent but not identically distributed (inid). We briefly review their results, so that we can extend their approach to the multivariate case ($q > 1$) in the SCT.

In addition to the general assumptions (3.2) and (3.3), they also introduced the following regularity conditions.

(CS1) [i] Let $\eta \in \Theta$, where $\Theta \subseteq (-\infty, \infty)$ is a bounded or unbounded interval.

[ii] Let the prior density $\mathfrak{h}$ be continuous and positive at the true value η_0.

(CS2) $P_i(\eta)$ is twice continuously differentiable with the first two derivatives being uniformly bounded in absolute value with respect to both η and i in some closed interval $\Theta_0 \subset \Theta$ around η_0.

(CS3) For every fixed $\eta \neq \eta_0$, $\eta \in \Theta$, there is a $c(\eta) > 0$ so that

$$\limsup_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathsf{E}_{\eta_0} \log Z_i(\eta, \eta_0) \leq -\, c(\eta) \tag{3.14}$$

and $\sup_{i\in\mathbb{N}} |\lambda_i(\eta)| < \infty$.

(CS4) If restricted to Θ_0, the following sets of functions are uniformly bounded:

$$\left\{ \left| \frac{\mathrm{d}\mathcal{I}_i}{\mathrm{d}\eta} \right| \mid i \in \mathbb{N} \right\}, \quad \left\{ \left| \frac{\mathrm{d}^2\lambda_i}{\mathrm{d}\eta^2} \right| \mid i \in \mathbb{N} \right\}, \quad \left\{ \left| \frac{\mathrm{d}^3\lambda_i}{\mathrm{d}\eta^3} \right| \mid i \in \mathbb{N} \right\}.$$

(CS5) Asymptotically, the average information at η_0 is bounded away from 0, i.e.

$$\liminf_{d\to\infty} \frac{\mathcal{I}^{(d)}(\eta_0)}{d} > 0.$$

Remark 3.2.1
With respect to the prior of η, additional to (CS1[ii]), Chang and Stout (1993) implicitly assumed in their proofs its properness, which was stated explicitly in the earlier associated technical report (Chang and Stout, 1991, p. 15). ◇

Chang and Stout (1993) argued convincingly that conditions (CS1)–(CS5) are realistic and non-restrictive in practice for commonly used IRT models of well-designed tests. They particularly commented on condition (CS3) and (3.14), which plays an important role in the proof of their main theorem. CS3 is required when the item responses $\{Y_i\}_{i\in\mathbb{N}}$ are independent but not identically distributed. If they are iid (CS3) is automatically satisfied, which however is not necessarily the case in IRT models.

Remark 3.2.2
Meaningful models for applications do not depend on a specific compact interval in Θ since η_0 usually is unknown. The conditions depending on η_0 should be satisfied for almost all

$\eta_0 \in \Theta$ and for almost each η_0, there should be some (arbitrarily small) interval Θ_0.[2] In the usual models this is satisfied. ◇

Their main results are expressed in the three theorems given below.

Theorem 3.2.3 (Chang and Stout, 1993, Theorem 1)
Suppose that conditions (CS1)–(CS5) hold for a fixed η_0. Let $\hat{\eta}_d$ be the MLE of η_0 and $\hat{\sigma}_d = (\mathcal{I}^{(d)}(\hat{\eta}_d))^{-1/2}$. Then, for $-\infty \leq a < b \leq \infty$, the posterior probability of $\hat{\eta}_d + a\hat{\sigma}_d < \eta < \hat{\eta}_d + b\hat{\sigma}_d$ tends in P_{η_0} to the probability of $Z \in (a, b)$ for $Z \sim \mathcal{N}(0, 1)$.[3] That means

$$A_d := \int_{\hat{\eta}_d + a\hat{\sigma}_d}^{\hat{\eta}_d + b\hat{\sigma}_d} h(\eta \mid \boldsymbol{Y}^{(d)})\, \mathrm{d}\eta \xrightarrow{\mathsf{P}_{\eta_0}} \int_a^b \phi_1(\eta)\, \mathrm{d}\eta =: A$$

for $d \to \infty$. ◇

Theorem 3.2.4 (Chang and Stout, 1993, Theorem 2)
Suppose that conditions (CS1)–(CS5) hold for a fixed η_0 and let $\hat{\eta}_d$ and $\hat{\sigma}_d$ be defined as in Theorem 3.2.3. Then, for $-\infty \leq a < b \leq \infty$, the posterior probability A_d tends to A P_{η_0}–almost surely as $d \to \infty$. ◇

Theorem 3.2.5 (Chang and Stout, 1993, Theorem 3)
Assume $\Theta = \Theta_0$, a finite interval.[4] Suppose that conditions (CS1)–(CS5) hold for all $\eta_0 \in \Theta_0$ and let $\hat{\eta}_d$ and $\hat{\sigma}_d$ be defined as in Theorem 3.2.3. Then, for $-\infty \leq a < b \leq \infty$, the posterior probability A_d tends to A in manifest probability P as $d \to \infty$. ◇

The result of Theorem 3.2.5 does not depend on the true value η_0 and is thus of special practical interest for estimation and prediction purposes. As Chang and Stout (1993) comment, Theorems 1 and 3 cover sampling from a fixed ability sub-population and from the whole population, respectively. An important by-product of the proof of the APN of latent variables distributions is the establishment of the weak and strong consistency of the MLE of η under milder conditions than Lord (1983).

Due to the theorems above, the following approximation for a large d and any observed response pattern $\boldsymbol{y} \in \{0, 1\}^d$ is justified (i.e. the construction of asymptotic credible intervals)

$$\mathsf{P}(a \leq \eta \leq b \mid \boldsymbol{Y}^{(d)} = \boldsymbol{y}) \approx \Phi(b\,;\, \hat{\eta}_d, \hat{\sigma}^2) - \Phi(a\,;\, \hat{\eta}_d, \hat{\sigma}^2), \tag{3.15}$$

for all $a \leq b \in \mathbb{R}$, where $\hat{\eta}_d$ is the MLE of η_0 based on the sample $\boldsymbol{y}$, $\hat{\sigma}^2 = (\mathcal{I}^{(d)}(\hat{\eta}_d))^{-1}$ and $\Phi(\,\bullet\,;\, \hat{\eta}_d, \hat{\sigma}^2)$ denotes the cdf of $\mathcal{N}(\hat{\eta}_d, \hat{\sigma}^2)$. Approximation (3.15) is of special practical importance in the context of long tests where the exact computation of posterior probabilities

[2] Actually, Chang and Stout (1991) proved some properties only for Θ_0 in Lemma 3.1 but used them for arbitrary closed intervals in Θ. To be more specific: (36)–(38) are shown on Θ_0 but they are required to show (39) that supposes a closed interval that is not necessarily contained in Θ_0. Since Chang and Stout (1993) only provide a sketch their proofs, the conditions regarding Θ_0 should be considered for all closed intervals.

[3] The proof of Chang and Stout (1991) for the corresponding Theorem 3.1 only contains the case that $-\infty < a < b < \infty$ because their equation (75) is only true in this case. Since none the other proofs of Chang and Stout (1991, 1993) contain the extension for $a = -\infty$ or $b = \infty$, Theorem 3.2.3, 3.2.4 and 3.2.5 should be used only for $-\infty < a < b < \infty$.

[4] Chang and Stout (1991) proved a version of this theorem without this restriction. As we will also see in the proof of Theorem 3.4.1, this restriction is actually not necessary. Theorem 3.2.5 can therefore be regarded as true for all Θ that satisfy (CS1).

for latent variables is commonly intractable. Furthermore, (3.15) allows the approximation of the posterior if the exact distribution $\mathcal{H}$ of η is unavailable or uncertain.

Finally, Chang and Stout (1993) noted that their theory, under suitable regularity conditions, can be extended to prove the APN for latent vectors of general multidimensional IRT models, without however further commenting on the proving procedure or the regularity conditions required. Next we discuss the current literature regarding APN, which can be seen in the light of the MIRT context.

3.3 A framework for asymptotic posterior distributions of multivariate latent traits

The theory of APN of the latent variables is naturally linked to Bayesian procedures and results on the convergence of posterior distributions. Especially interesting and inspiring is the fundamental contribution by GGS, who consider asymptotic multivariate posterior distributions (not necessarily normal) in a very general and flexible framework discussing different types of convergence, relying on earlier works by Ghosh et al. (1994) and Ibragimov and Has'minskii (1981). In particular, they studied the convergence of suitably centered and normalized posteriors. Their results provide a foundation which can be adopted for the APN in the IRT setup.

Next, we adjust their approach for MIRT models and embed the results of Chang and Stout (1993) in this framework. We consider the APN for multivariate latent vectors ($q \geq 1$) in the Sections 3.4 and 3.5 hereafter.

Following Ghosal et al. (1995, Definition 2), we distinguish two types of limiting behavior of posterior distributions and link them to the statistic used for the centering of the posterior distribution of the latent vector in the following definition.

Definition 3.3.1

Let $\mathcal{Q}$ be a probability measure that is defined on $(\mathbb{R}^q, \mathcal{B}^q)$.[5] Further let $\{\tilde{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ and $\{\boldsymbol{S}_d\}_{d\in\mathbb{N}}$ be two sequences of random elements with values in $\mathbb{R}^q$ and $\mathbb{R}^{q\times q}$, respectively. Suppose that $\tilde{\boldsymbol{\eta}}_d$ and $\boldsymbol{S}_d$ are constant given $\boldsymbol{Y}^{(d)}$ for all $d \in \mathbb{N}$.[6] The statistic $\tilde{\boldsymbol{\eta}}_d$ is called a *proper centering* (with limiting distribution $\mathcal{Q}$) if

$$\sup_{A\in\mathcal{B}^q} \left| \mathsf{P}\left(\boldsymbol{S}_d(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d) \in A \,\middle|\, \boldsymbol{Y}^{(d)}\right) - \mathcal{Q}(A) \right| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0, \text{ as } d \to \infty. \tag{3.16}$$

The statistic $\tilde{\boldsymbol{\eta}}_d$ is called *semiproper centering* (with limiting distribution $\mathcal{Q}$) if, for each $A \in \mathcal{B}^q$,

$$\mathsf{P}\left(\boldsymbol{S}_d(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d) \in A \,\middle|\, \boldsymbol{Y}^{(d)}\right) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathcal{Q}(A), \text{ as } d \to \infty. \tag{3.17}$$

The statistic $\tilde{\boldsymbol{\eta}}_d$ is called *compatible* (with the posterior) if

$$\left(\boldsymbol{S}_d(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0), h^{**}\left(\bullet \,\middle|\, \boldsymbol{Y}^{(d)}\right)\right)$$

[5] Ghosal et al. (1995) define proper and semiproper centering for a general family of numbers $\{\mathcal{Q}(A)\}_{A\in\mathcal{B}^q}$. However, Ghosh et al. (1994) note that the family $\{\mathcal{Q}(A)\}_{A\in\mathcal{B}^q}$ defines a probability measure in the case of proper centering and a countably additive probability content in the case of semiproper centering. Moreover, we are only interested in limiting probability measures and this is consistent with Ghosal et al. (1995, Definition 2) so that we can still apply the required results.

[6] That means that they can be considered by the functional relation $\tilde{\boldsymbol{\eta}}_d = \tilde{\boldsymbol{\eta}}_d(\boldsymbol{Y}^{(d)})$ and $\boldsymbol{S}_d = \boldsymbol{S}_d(\boldsymbol{Y}^{(d)})$. But in some cases $\{\boldsymbol{S}_d\}_{d\in\mathbb{N}}$ is simply a sequence of constant matrices.

converges in distribution for $d \to \infty$, where $h^{**}(\,\bullet \mid \boldsymbol{Y}^{(d)})$ denotes the density of the posterior distribution of $\boldsymbol{S}_d(\boldsymbol{\eta} - \boldsymbol{\eta}_0)$ given $\boldsymbol{Y}^{(d)}$. ◇

Remark 3.3.2

- We primary interested in the case of asymptotic posterior normality. Then $\mathcal{Q}(A) = \mathsf{P}(\boldsymbol{Z} \in A)$ for all $A \in \mathcal{B}^q$ and for a q-variate standard normal distributed random vector $\boldsymbol{Z} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. However, in order to prove the proper centering of Bayes estimators that are not the EAP it is necessary to define a more general limiting behavior of the posterior in order to apply Proposition 2.2 of Ghosh et al. (1994).

- The two important choices for $\boldsymbol{S}_d$ which we consider in this work are $\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}$ and $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}$. The first is generally important because it can be computed in practice. The latter is however important for several proofs of the main results of this chapter and can additionally be interesting and important in theoretical discussions. Another choice which is often considered in classical discussions of APN in Bayesian settings is $\boldsymbol{S}_d = (-\nabla\nabla^\intercal \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \mid \boldsymbol{Y}^{(d)}))^{1/2}$, i.e. the observed information matrix (cf. e.g. Schervish, 1995, Section 7.4.2). However, $-\nabla\nabla^\intercal \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \mid \boldsymbol{Y}^{(d)})$ can be indefinite or negative definite while the expected information has a stronger theoretical justification through the Cramér–Rao bound. ◇

Proper and semiproper centering correspond to uniform and pointwise convergence of the posterior of the standardized latent vector $\boldsymbol{S}_d(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d)$, respectively. Hence, proper centering is a stronger property than semiproper centering and is consequently expected to require stronger assumptions.

In view of this, one can easily recognize that Theorem 3.2.3 of Chang and Stout (1993) is the semiproper centering of the MLE since it can be formulated as

$$\mathsf{P}\left(\mathcal{I}^{(d)}(\hat{\eta}_d)^{1/2}\,(\eta - \hat{\eta}_d) \in [a,b] \,\middle|\, \boldsymbol{Y}^{(d)}\right) \xrightarrow{\mathsf{P}_{\eta_0}} \mathsf{P}\big(Z \in [a,b]\,\big), \qquad d \to \infty,$$

for $Z \sim \mathcal{N}(0,1)$, $A = [a,b] \subset \mathbb{R}$, $\boldsymbol{S}_d = \mathcal{I}^{(d)}(\hat{\eta}_d)^{1/2}$ and $\tilde{\eta}_d = \hat{\eta}_d$ being the MLE of η_0 based on $\boldsymbol{Y}^{(d)}$. Thus for the extension of the CS theory for multivariate LTs, we focus on semiproper centering.

The asymptotic results of GGS adjusted in our setup primarily focus on the convergence of the posterior distribution of the standardized latent vector

$$\boldsymbol{\eta}^* = \boldsymbol{S}_d(\boldsymbol{\eta} - \boldsymbol{\eta}_0). \tag{3.18}$$

We need the likelihood ratio (3.11) expressed in terms of $\boldsymbol{\eta}^*$, which is denoted by

$$Z^{*(d)}(\boldsymbol{\eta}^*) \coloneqq \begin{cases} Z^{(d)}(\boldsymbol{\eta}_0 + \boldsymbol{S}_d^{-1}\boldsymbol{\eta}^*, \boldsymbol{\eta}_0), & \text{if} \quad \boldsymbol{\eta}^* \in \Theta_d, \\ 0, & \text{otherwise}, \end{cases} \tag{3.19}$$

for $d \in \mathbb{N}$ where $\Theta_d \coloneqq \boldsymbol{S}_d(\Theta - \boldsymbol{\eta}_0)$.

The primary conditions (GGS) of Ghosal et al. (1995) are given as follows:

(GGS1) For some $M > 0$, $m_1 \geq 0$ and $\alpha > 0$ holds

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*)^{1/2} - Z^{*(d)}(\boldsymbol{\eta}_2^*)^{1/2}\right|^2\right) \leq M(1 + R^{m_1})\|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|^\alpha$$

for all $\boldsymbol{\eta}_j^* \in \Theta_d$ satisfying $\|\boldsymbol{\eta}_j^*\| \leq R$, $j = 1, 2$.

(GGS2) For all $\boldsymbol{\eta}^* \in \Theta_d$ holds

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}^*)^{1/2}\right) \leq \exp\Big(-g_d\big(\|\boldsymbol{\eta}^*\|\big)\Big),$$

where $\{g_d\}_{d\in\mathbb{N}}$ is a sequence of real-valued functions on $[0,\infty)$ satisfying the following:

(a) for a fixed $d \geq 1$, $\lim_{x\to\infty} g_d(x) = \infty$;

(b) for any $N > 0$,

$$\lim_{x\to\infty}\lim_{d\to\infty} x^N \exp(-g_d(x)) = \lim_{d\to\infty}\lim_{x\to\infty} x^N \exp(-g_d(x)) = 0.$$

(GGS3) There is a family $\{Z(\boldsymbol{\eta}^*)\colon \boldsymbol{\eta}^* \in \mathbb{R}^q\}$ of non identically zero random variables so that for all $n \in \mathbb{N}$ and $\boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_n^* \in \mathbb{R}^q$ holds:

$$(Z^{*(d)}(\boldsymbol{\eta}_1^*), \ldots, Z^{*(d)}(\boldsymbol{\eta}_n^*)) \xrightarrow{\mathcal{D}} (Z(\boldsymbol{\eta}_1^*), \ldots, Z(\boldsymbol{\eta}_n^*)),$$

as $d \to \infty$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

In the binary IRT context, we can write $Z^{*(d)}$ as

$$Z^{*(d)}(\boldsymbol{\eta}^*) = \exp\left(\sum_{i=1}^{d}\left(Y_i\left(\lambda_i\left(\boldsymbol{\eta}_0 + \boldsymbol{S}_d^{-1}\boldsymbol{\eta}^*\right) - \lambda_i\Big(\boldsymbol{\eta}_0\Big)\right)\right.\right.$$
$$\left.\left. - \left(\psi\left(\lambda_i\left(\boldsymbol{\eta}_0 + \boldsymbol{S}_d^{-1}\boldsymbol{\eta}^*\right)\right) - \psi\left(\lambda_i\left(\boldsymbol{\eta}_0\right)\right)\right)\right),$$

and for APN, it must hold

$$Z(\boldsymbol{\eta}^*) = \exp\left(\boldsymbol{\xi}^T\boldsymbol{\eta}^* - \frac{1}{2}\|\boldsymbol{\eta}^*\|^2\right)$$

with $\boldsymbol{\xi} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ in (GGS3). Under the conditions (GGS1)–(GGS3), Ghosal et al. (1995) provided the following general result.

Theorem 3.3.3 (Ghosal et al., 1995, Theorem 1)
Assume that conditions (GGS1)–(GGS3) hold. If either a proper centering or a compatible semiproper centering sequence $\{\tilde{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ exists, then there is a random vector $\boldsymbol{W}$ so that

(a) $\boldsymbol{S}_d(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{D}} \boldsymbol{W}$ for $d \to \infty$ and

(b) $\frac{Z(\boldsymbol{x}-\boldsymbol{W})}{\int_{\mathbb{R}^q} Z(\boldsymbol{x}^*-\boldsymbol{W})\,\mathrm{d}\boldsymbol{x}^*}$ is one point distributed for almost all $\boldsymbol{x} \in \mathbb{R}^q$, where Z is as defined in condition (GGS3).

Conversely, if (b) holds for a random vector $\boldsymbol{W}$ then any Bayes estimator (with respect to a prior and loss considered by Ghosal et al., 1995) is a compatible proper centering. ◇

Applying Theorem 3.3.3 to an appropriate Bayes estimator for $\boldsymbol{\eta}_0$, the APN of an MIRT model under conditions (GGS1) to (GGS3) is derived. Unfortunately, knowledge regarding the posterior distribution is required to compute a Bayes estimator. This reduces the thus obtained APN to limited practical importance. To extend Theorem 3.3.3 to the MLE, i.e. for $\tilde{\boldsymbol{\eta}}_d = \hat{\boldsymbol{\eta}}_d$, the asymptotic equivalence of the MLE and an arbitrary Bayes estimator has to be shown. Ghosal et al. (1995, cf. Corollary 1) gave the following strengthened form of (GGS1) to allow that conclusion:

(GGS1') For some $M > 0$, $m_1 \geq 0$ and $m \geq \alpha > q$ holds

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*)^{1/m} - Z^{*(d)}(\boldsymbol{\eta}_2^*)^{1/m}\right|^m\right) \leq M(1+R^{m_1})\|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|^\alpha \tag{3.20}$$

for all $\boldsymbol{\eta}_j^* \in \Theta_d$ satisfying $\|\boldsymbol{\eta}_j^*\| \leq R$, $j = 1, 2$.

However, comparing conditions (GGS1') to (GGS3) to conditions (CS1) to (CS5), we observe that they are far more abstract. The single relatively abstract condition of Chang and Stout (1993), namely (3.14) in (CS3), is equivalent to the identifiability of the latent trait. Hence, though (CS1) to (CS5) are directly applicable and can be verified for IRT models, (GGS1') to (GGS3) are not easily specified or demonstrated, even for the simple case of a common probit model, i.e.

$$P_i(\boldsymbol{\eta}) = \Phi(\boldsymbol{\alpha}_i^T\boldsymbol{\eta} + \beta_i), \qquad \boldsymbol{\eta} \in \Theta,\, i \in \mathbb{N}.$$

Consider as a further example the IRT model by Lee and Bolt (2018) or its multidimensional version

$$P_i(\boldsymbol{\eta}) = \Phi\left(\frac{\boldsymbol{\alpha}_i^T\boldsymbol{\eta} + \beta_i}{\sqrt{2}(1+\exp(-\boldsymbol{\delta}_i^T\boldsymbol{\eta}))^{-1/2}}\right), \qquad \boldsymbol{\eta} \in \Theta,\, i \in \mathbb{N}. \tag{3.21}$$

For these models, the likelihood ratio $Z^{*(d)}$ becomes quite unhandy. Having recognized the restricted direct applicability of the GGS results in MIRT practice, we next explore APN of latent vectors under alternative more convenient conditions to get semiproper centering, generalizing thus the results of Chang and Stout (1993) for $q > 1$. In the subsequent Section 3.5, we discuss the stronger conditions necessary to receive proper centering in an MIRT framework.

3.4 Semiproper Centering Theory

3.4.1 Regularity Conditions

Throughout this section, we assume that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta})$, i.e. $\{Y_i\}_{i\in\mathbb{N}}$ are Bernoulli random variables fulfilling (3.1) and (3.2) and that the true latent vector $\boldsymbol{\eta}_0$ lies in the interior of the parameter space, i.e. $\boldsymbol{\eta}_0 \in \Theta \setminus \partial\Theta$. The asymptotic results of Subsection 3.4.2 rely on the following regularity conditions.

(CS1′) [i] The set Θ is closed, convex and has a non-empty interior.

[ii] The prior density $\mathfrak{h}$ of $\boldsymbol{\eta}$ is proper and continuous at $\boldsymbol{\eta}_0$ with $\mathfrak{h}(\boldsymbol{\eta}_0) > 0$.

(CS2′) P_i is thrice continuously differentiable, $i \in \mathbb{N}$. If restricted to a compact subset $K \subseteq \Theta$, all $\left|\frac{\partial P_i}{\partial \eta_k}\right|$ and $\left|\frac{\partial^2 P_i}{\partial \eta_k \partial \eta_j}\right|$ are uniformly bounded for all $i \in \mathbb{N}$, $1 \leq j, k \leq q$. Moreover, there are constants $0 < \zeta_0(K) < \zeta_1(K) < 1$ which are independent of $i \in \mathbb{N}$ such that

$$\zeta_0(K) \leq \inf_{(i,\boldsymbol{\eta})\in\mathbb{N}\times K} P_i(\boldsymbol{\eta}) \leq \sup_{(i,\boldsymbol{\eta})\in\mathbb{N}\times K} P_i(\boldsymbol{\eta}) \leq \zeta_1(K). \tag{3.22}$$

(CS3′) For each $\boldsymbol{\eta} \in \Theta$, $\boldsymbol{\eta} \neq \boldsymbol{\eta}_0$, there is a $c(\boldsymbol{\eta}) < 0$ such that

$$\limsup_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathsf{E}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}_0)) = \limsup_{d\to\infty} \frac{1}{d} \mathsf{E}_{\boldsymbol{\eta}_0}(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})) \leq c(\boldsymbol{\eta})$$

and additionally

$$\sup_{\boldsymbol{\eta} \in \Theta \setminus B_\delta(\boldsymbol{\eta}_0)} c(\boldsymbol{\eta}) < 0 \tag{3.23}$$

holds for all $\delta > 0$ if Θ is unbounded.

(CS4′) If restricted to any compact set $K \subseteq \Theta$, the following set of functions is uniformly bounded

$$\left\{ \left| \frac{\partial^3 P_i}{\partial \eta_k \partial \eta_g \partial \eta_u} \right| \,:\, i \in \mathbb{N}, 1 \leq k, g, u \leq q \right\}.$$

(CS5′) For all $\boldsymbol{\eta} \in \Theta$ holds

$$\liminf_{d\to\infty} \nu_{\min} \left(\frac{1}{d} \sum_{i=1}^{d} \nabla \lambda_i(\boldsymbol{\eta}) \nabla \lambda_i(\boldsymbol{\eta})^\intercal \right) > 0, \tag{3.24}$$

where $\nu_{\min}$ denotes the smallest eigenvalue.

These regularity conditions correspond one-to-one to conditions (CS1)–(CS5). For the comparison of these conditions, bear in mind that convexity and connectivity are equivalent properties in $\mathbb{R}$. The convexity condition in (CS1′) is at first glance stronger but it does not impose a real practical restriction, since non-convex Θ are only rarely needed in MIRT. Analogous to the CS theory (cf. Remark 3.2.2), conditions involving $\boldsymbol{\eta}_0$ – like $\mathfrak{h}(\boldsymbol{\eta}_0) > 0$ – should be interpreted as $\mathfrak{h} > 0$ almost surely. Condition (CS1′[ii]) on $\mathfrak{h}$ seems more strict than (CS1[ii]). However, Chang and Stout (1993) additionally require a proper prior (cf. Remark 3.2.1). Thus, under the consideration that $\boldsymbol{\eta}_0$ is still unknown and that we consider $\mathbb{R}^q$ instead of $\mathbb{R}$, the requirements on proper priors in (CS1′[ii]) are analogous to (CS1[ii]). Finally, note that in the generalization of the conditions (CS3) and (CS4), some requirements have been removed because the remaining requirements on λ_i, $i \in \mathbb{N}$, and its derivatives are implied from the conditions (CS2′) and (CS4′).

A common assumption in one dimensional IRT models ($q = 1$) is the strict monotonicity assumption (3.3) of P_i in η for all $i \in \mathbb{N}$. Conceptually, this represents the notion that a more able subject has a higher probability of responding correctly to any item of an educational test. Thus, models fulfilling this strict monotonicity assumption are easier to interpret. However models with non–generalized–linear latent variable effects can be more adequate in practice. For example, Rizopoulos and Moustaki (2008) considered IRT models with possibly non–monotonic latent variable dependencies (like polynomial effects). In order to allow for more flexible modeling options, in our semiproper centering theory we abandon the requirement on strict monotonicity of $\boldsymbol{\eta} \mapsto P_i(\boldsymbol{\eta})$, $i \in \mathbb{N}$, in each component. Since the results of Chang and Stout (1993) rely on this monotonicity assumption, the merit of the current contribution is not only the extension of the CS-results to latent vectors ($q > 1$) but also to univariate latent variables in case of a non-monotonic latent variable effect.

If all P_i, $i \in \mathbb{N}$, are strictly monotonic in each component, then the requirement (3.22) of condition (CS2′) is satisfied as in the univariate case. Otherwise, the requirement in (3.22) is generally not really restrictive; it is the technical formulation of the notion that the response probabilities can (but not necessarily have to) tend to zero or one only if $\|\boldsymbol{\eta}\| \to \infty$. Assumption (3.23) in (CS3′) serves for ensuring the identifiability of the latent vector in case of $\|\boldsymbol{\eta}\| \to \infty$. Hence, this condition is quite natural for a statistical model. In the univariate case, (3.23) of (CS3′) is implied by the strict monotonicity, too. But in contrast to (3.22), (3.23) cannot be concluded directly from the strict monotonicity of all P_i in each component if $q > 1$. Moreover, while a single item can suffice in the univariate case for identifiability, there are always at least q items needed in the q-dimensional one. Similarly, the average test information $\frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta})$ is always singular for $d < q$ since it is a sum of d rank-one matrices. Condition (CS5′) ensures that $\frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta})$ becomes regular for $d \to \infty$ and can be interpreted as a condition to ensure that the asymptotic posterior of $\boldsymbol{\eta}$ is regular q-dimensional distributed and does not have a lower dimensional support.

To get a better impression of the conditions, we exemplary discuss them in Section 3.4.4 for (3.21), the multidimensional version of a model of Lee and Bolt (2018).

3.4.2 Main Results

Our main contribution in this section is the generalization of Theorems 1 and 3 of Chang and Stout (1993) for $q > 1$ under the assumptions (CS1′) to (CS5′) provided next. Together with the new results regarding the consistency of the posterior and that of the EAP, as well as the generalization of the proof of the consistency of the MLE to $q > 1$, this generalization constitutes the semiproper centering theory. The proofs of all results in this section along with some preliminary lemmas are given in Section 3.4.5. First of all, we prove that the MLE $\hat{\boldsymbol{\eta}}_d$ is a semiproper centering (cf. Definition 3.3.1).

Theorem 3.4.1 (Semiproper centering of the MLE)
Let $\boldsymbol{Z} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ be a q-variate standard normal distributed random vector and let $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ be a sequence of binary response variables satisfying (CS1′)–(CS5′) for a fixed $\boldsymbol{\eta}_0 \in \Theta \setminus \partial\Theta$. Then, for all $B \in \mathcal{B}^q$

$$\mathsf{P}\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}\,(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \in B \,\middle|\, \boldsymbol{Y}^{(d)}\right) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathsf{P}(\boldsymbol{Z} \in B). \tag{3.25}$$

If $\boldsymbol{\eta}_0 \sim \mathcal{G}$, where $\mathcal{G}$ is an absolutely continuous proper distribution with $\operatorname{supp}(\mathcal{G}) \subseteq \Theta$, *then additionally*

$$\mathsf{P}\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}\,(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \in B \,\middle|\, \boldsymbol{Y}^{(d)}\right) \xrightarrow{\mathsf{P}} \mathsf{P}(\boldsymbol{Z} \in B) \tag{3.26}$$

for all $B \in \mathcal{B}^q$. ◇

The distribution $\mathcal{G}$ can be different from $\mathcal{H}$ used in the model. Hence, the asymptotic is robust to misspecifications of $\mathcal{H}$ as long as the support is sufficiently large.

As already noted in Section 3.2, this result can be used for the construction of credible regions for $\boldsymbol{\eta}$. Additionally, Theorem 3.4.1 allows the interpretation of the MLE as a Bayesian estimator of $\boldsymbol{\eta}_0$ and thus enables the use of $\hat{\boldsymbol{\eta}}_d$ to derive some kind of *objective posterior*, in the sense that it is prior-free constructed.

Another important concept in the asymptotic analysis of Bayesian procedures is the consistency of the posterior distribution, which forms a basis for the asymptotic validity

of most inferential methods. Roughly speaking, it can be understood as a law of large numbers like $\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)} \to \boldsymbol{\eta}_0$, $d \to \infty$. We formulate it explicitly in the following theorem.

Theorem 3.4.2 (Weak consistency of the posterior distribution)
Let $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$, consider a prior $\mathfrak{h}$ for $\boldsymbol{\eta}$ and assume conditions (CS1′)–(CS5′) for a fixed $\boldsymbol{\eta}_0 \in \Theta$. Then

$$\mathsf{P}(\boldsymbol{\eta} \in B \mid \boldsymbol{Y}^{(d)}) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \delta_{\boldsymbol{\eta}_0}(B) \coloneqq \begin{cases} 1, & \boldsymbol{\eta}_0 \in B \\ 0, & \boldsymbol{\eta}_0 \notin B \end{cases}, \qquad d \to \infty,$$

for all Borel-sets $B \in \mathcal{B}^q$ with $\boldsymbol{\eta}_0 \notin \partial B$. ◇

Furthermore, analogously to Chang and Stout (1993), the consistency of the MLE $\hat{\boldsymbol{\eta}}_d$ under conditions (CS1′) to (CS5′) is received as a by-product of the proof of Theorem 3.4.1, along with an assertion of its existence. Note that although the MLE is commonly known as consistent in IRT and MIRT settings, Sinharay (2015) indicated the lack of corresponding asymptotic results under milder conditions than some of the usual ones (such as test lengthening by *strictly parallel forms*). Thus, the following Theorem is a contribution towards this direction.

Theorem 3.4.3 (Existence and consistency of the MLE)
Consider a sequence $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ such that the conditions (CS1′[i]), (CS2′) and (CS3′) are satisfied for a fixed $\boldsymbol{\eta}_0 \in \Theta$. Then a sequence $\{\hat{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ of measurable mappings exists so that

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\nabla \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\, \boldsymbol{Y}^{(d)}) = \mathbf{0}\right) = 1,$$
$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\ell^{(d)}(\hat{\boldsymbol{\eta}}_d \mid \boldsymbol{Y}^{(d)}) = \max_{\boldsymbol{\eta}\in\Theta} \ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})\right) = 1$$

and $\hat{\boldsymbol{\eta}}_d \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \boldsymbol{\eta}_0$ for $d \to \infty$. ◇

Remark 3.4.4
The consistency of a penalized MLE is implied from the consistency of the MLE. Consider a modified log-likelihood function

$$\tilde{\ell}^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) = \ell^{(d)}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) + \log(\mathcal{W}(\boldsymbol{\eta})), \qquad \boldsymbol{\eta} \in \Theta,\, d \in \mathbb{N},$$

for some continuously differentiable, positive and bounded function $\mathcal{W}$. Then we can replace ℓ with $\tilde{\ell}$ in Theorem 3.4.3. The validity of this replacement is proved after Theorem 3.4.3 is proved. ◇

For $\mathcal{W} = \mathfrak{h}$, the penalized MLE in Remark 3.4.4 becomes the maximum a-posteriori estimator (MAP), which is an important estimator for $\boldsymbol{\eta}$ in IRT, also because it ensures the existence of estimates in cases in which the MLE becomes infinite (for example when $\sum_{i=1}^{d} y_i = 0$ or d). The restriction on $\mathfrak{h}$ in Remark 3.4.4 is stronger than (CS1′[ii]), but still mild.

Finally, we obtain the existence and consistency of the expected a-posteriori estimator (EAP) for estimating $\boldsymbol{\eta}_0$. To the best of our knowledge, a proof of these properties in such a general setup and under comparably mild or milder conditions on the MIRT model does not yet exist in the related literature.

Theorem 3.4.5 (Existence and Consistency of the EAP)
Consider $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$*, a prior* $\mathfrak{h}$ *for* $\boldsymbol{\eta}$ *and assume that conditions (CS1′)–(CS5′) hold for a fixed* $\boldsymbol{\eta}_0 \in \Theta$*. Further consider a continuous mapping* $f\colon \Theta \to \mathbb{R}$ *such that* $\int_\Theta f(\boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}(\boldsymbol{\eta})$ *exists. Then the posterior expected value* $\mathsf{E}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)})$ *exists for all* $d \in \mathbb{N}$ *and is consistent for* $f(\boldsymbol{\eta}_0)$*, i.e.*

$$\mathsf{E}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)}) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} f(\boldsymbol{\eta}_0), \qquad \text{for } d \to \infty.$$

If in particular $\mathsf{E}(\boldsymbol{\eta})$ *exists, then the posterior expected value* $\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})$ *exists for all* $d \in \mathbb{N}$ *and is weakly consistent for* $\boldsymbol{\eta}_0$*.* ◇

3.4.3 Summary, Discussion and Outlook

In this section, we proved the APN of LTs by semiproper centering of the MLE under mild conditions that are fulfilled by a broad class of MIRT models for binary items. Furthermore, we provided results on the weak consistency of posterior distributions, the existence and consistency of the MLE, the MAP and EAP estimators under similar mild conditions.

Our results assume a proper prior $\mathfrak{h}$. This is appropriate in IRT settings where the prior is a model of the population distribution of the latent traits. However, in a Bayesian framework improper priors can also be considered. If this is the case, then the proper prior assumption in (CS1′[ii]) can be replaced by the following condition if the posterior is still proper

$$\frac{1}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta}_0)} \int_{\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}\big(d^{-q/2}\big), \quad \text{for all} \quad \delta > 0. \tag{3.27}$$

Condition (3.27) is sufficient for the derivation of the results stated here and is satisfied by a proper prior.

The APN for a univariate LT for polytomous items by semiproper centering of the MLE was discussed by Chang (1996). The extension of the results to MIRT models with polytomous items is an interesting subject for future reseach. The strengthening of the results to proper centering of estimators of $\boldsymbol{\eta}$ is the subject of Section 3.5.

3.4.4 Example: Regularity Conditions for Model (3.21)

Let us demonstrate the regularity conditions (CS1′) to (CS5′) in the case of model (3.21), the multidimensional version of a model of Lee and Bolt (2018).

If $\mathcal{H}$ is one of the usual structural models or any other regular distribution, for example $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$, a mixture of normals or a uniform distribution on some compact set, then (CS1′) is directly satisfied. If the parameter sequences $\{\boldsymbol{\alpha}_i\}_{i\in\mathbb{N}}$ and $\{\boldsymbol{\delta}_i\}_{i\in\mathbb{N}}$ behave as they were two independent iid sequences drawn from absolutely continuous regular distributions in some bounded region in $\mathbb{R}^q$ and the sequence $\{\beta_i\}_{i\in\mathbb{N}}$ is in an arbitrary bounded subset of $\mathbb{R}$, then the conditions (CS2') and (CS4') are directly satisfied. Furthermore, in almost all cases, the latent vector is identifiable if q arbitrary items are given. Hence, (CS3') is satisfied, too. The gradient of the response probabilities is given by

$$\begin{aligned}\nabla P_i(\boldsymbol{\eta}) = \phi_1\left(\frac{\boldsymbol{\alpha}_i^\intercal\boldsymbol{\eta} + \beta_i}{\sqrt{2}(1+\exp(-\boldsymbol{\delta}_i^\intercal\boldsymbol{\eta}))^{-1/2}}\right) \\ \times\left(\sqrt{\frac{1+\exp(-\boldsymbol{\delta}_i^\intercal\boldsymbol{\eta})}{2}}\boldsymbol{\alpha}_i - \frac{\exp(-\boldsymbol{\delta}_i^\intercal\boldsymbol{\eta})}{2\sqrt{2}}\frac{\boldsymbol{\alpha}_i^\intercal\boldsymbol{\eta} + \beta_i}{\sqrt{1+\exp(-\boldsymbol{\delta}_i^\intercal\boldsymbol{\eta})}}\boldsymbol{\delta}_i\right),\end{aligned} \tag{3.28}$$

for all $i \in \mathbb{N}$ and $\boldsymbol{\eta} \in \Theta$. In particular, we see from (3.28) that $\{\nabla P_i(\boldsymbol{\eta})\}_{i\in\mathbb{N}}$ behaves as an iid sequences drawn from a regular distribution with bounded support in $\mathbb{R}^q$ in almost all cases for all $\boldsymbol{\eta} \in \Theta$. Thus,

$$\frac{1}{d}\sum_{i=1}^{d} \nabla P_i(\boldsymbol{\eta}) \nabla P_i(\boldsymbol{\eta})^\intercal$$

converges to the second moment of this distribution and is, hence, positive definite.

With respect to (CS5′), note that (3.22) in (CS2′) implies that (3.24) is equivalent to

$$\liminf_{d\to\infty} \nu_{\min}\left(\frac{1}{d}\sum_{i=1}^{d} \mathcal{I}_i(\boldsymbol{\eta})^\intercal\right) > 0 \quad \text{and to} \quad \liminf_{d\to\infty} \nu_{\min}\left(\frac{1}{d}\sum_{i=1}^{d} \nabla P_i(\boldsymbol{\eta}) \nabla P_i(\boldsymbol{\eta})^\intercal\right) > 0, \tag{3.29}$$

which in our case ensures that (CS5′) is satisfied.

For illustrative purposes, consider an example with $d = 30$ and $q = 2$ and model parameter values which are independently drawn from the uniform distribution on $(-2, 2)$ for β_i and on $(-0.5, 1)$ for all other parameters (as given in Table 3.1).

i	β_i	α_{i1}	α_{i2}	δ_{i1}	δ_{i2}	i	β_i	α_{i1}	α_{i2}	δ_{i1}	δ_{i2}
1	-0.906	0.946	-0.398	0.230	0.327	16	-0.104	-0.498	0.298	0.359	0.622
2	-1.337	0.893	-0.085	0.696	0.622	17	0.458	0.496	0.267	0.677	0.551
3	-0.913	-0.363	0.703	0.052	0.605	18	-1.882	-0.259	-0.484	0.101	-0.022
4	0.387	0.166	0.313	-0.352	-0.447	19	0.801	0.066	-0.488	0.282	0.697
5	0.676	0.677	0.199	0.701	0.283	20	1.953	-0.169	0.963	-0.467	0.443
6	-0.084	0.865	0.833	0.957	0.724	21	-1.712	0.326	0.979	-0.146	-0.347
7	0.138	0.688	-0.444	0.942	-0.043	22	1.725	0.146	0.875	0.843	0.092
8	1.608	0.296	0.472	-0.096	-0.101	23	1.980	0.826	0.517	0.195	0.825
9	-0.882	0.016	-0.378	0.088	0.458	24	0.171	0.197	0.722	0.069	-0.442
10	1.240	0.320	0.924	-0.250	0.386	25	0.499	-0.101	-0.319	0.943	0.071
11	-0.919	0.867	-0.391	0.051	0.202	26	-1.775	0.530	-0.188	0.819	-0.336
12	0.216	0.093	0.028	0.152	0.339	27	1.523	-0.302	0.646	0.581	0.718
13	1.900	0.726	0.007	0.842	-0.072	28	1.548	0.272	0.469	-0.207	-0.241
14	0.510	-0.068	-0.105	0.132	-0.256	29	-0.823	-0.113	-0.147	0.027	-0.282
15	-1.109	0.072	-0.017	0.370	0.623	30	0.822	0.661	-0.485	-0.277	-0.017

Table 3.1: *Hypothetical parameter values for model* (3.21) *and the first* 30 *items.*

In Figure 3.1 (top) visualizations of $\mathsf{E}_{\boldsymbol{\eta}_0}(\log(Z_5(\boldsymbol{\eta}, \boldsymbol{\eta}_0)))$ are provided for two exemplary values of $\boldsymbol{\eta}_0 \in \Theta$ and the parameters in Table 3.1. In particular, we can recognize lines in Θ for which $\mathsf{E}_{\boldsymbol{\eta}_0}(\log(Z_5(\boldsymbol{\eta}, \boldsymbol{\eta}_0))) = 0$ holds. In Figure 3.1 (bottom), surfaces of $\frac{1}{30}\sum_{i=1}^{30} \mathsf{E}_{\boldsymbol{\eta}_0}(\log(Z_i(\boldsymbol{\eta}, \boldsymbol{\eta}_0)))$ are illustrated for two further exemplary $\boldsymbol{\eta}_0$ values. The surfaces are drawn over $[-1, 1]^2$. There is a nearly parabolic surface which illustrates that there is no reason to doubt for (3.23) in (CS3′).

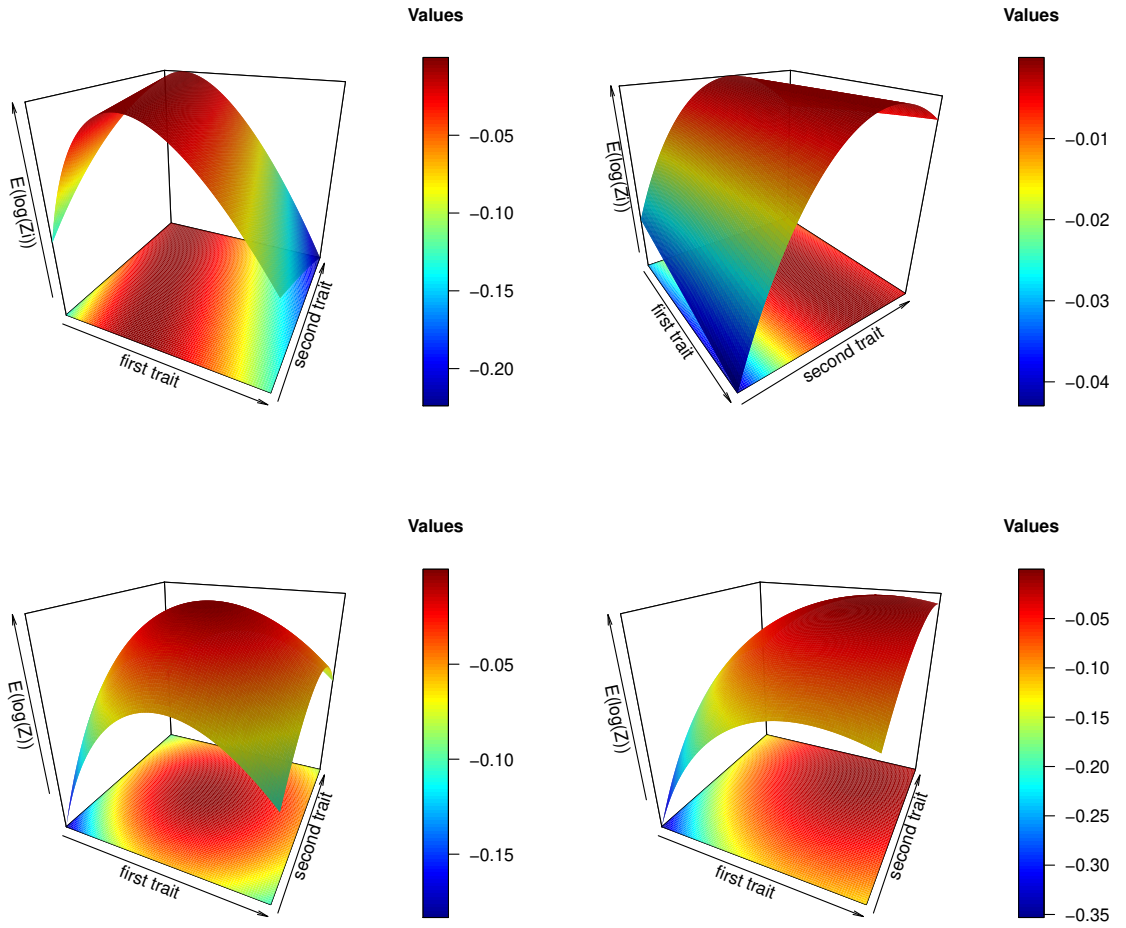

Figure 3.1: *Top left:* $\mathsf{E}_{(-0.5,0.1)}(\log(Z_5(\boldsymbol{\eta}, (-0.5, 0.1))))$,
top right: $\mathsf{E}_{(-0.1,0.3)}(\log(Z_9(\boldsymbol{\eta}, (-0.1, 0.3))))$,
bottom left: $\frac{1}{30}\sum_{i=1}^{30} \mathsf{E}_{(0,0)}(\log(Z_i(\boldsymbol{\eta}, (0, 0))))$,
bottom right: $\frac{1}{30}\sum_{i=1}^{30} \mathsf{E}_{(0.5,0.5)}(\log(Z_i(\boldsymbol{\eta}, (0.5, 0.5))))$.

Figure 3.2 provides the minimal smallest eigenvalue of $\frac{1}{d}\sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta})\nabla\lambda_i(\boldsymbol{\eta})^\intercal$ on $[-1, 1]^2$ for $d \in \{1, \ldots, 30\}$, cf. (CS5′). Overall, in this case, the regularity conditions can be considered as justified to apply the APN for arbitrary response patterns on the illustrated 30 items.

3.4.5 Proofs

Lemma 3.4.6

Let a sequence of mappings $\{P_i\}_{i\in\mathbb{N}}$ *be given so that the condition* (CS2′) *is satisfied. Additionally, let* K *be an arbitrary compact set for which a convex and compact set* K^+ *exists such that* $K \subseteq K^+ \subseteq \Theta$. *Then the family of mappings* $\{P_i|_K\}_{i\in\mathbb{N}}$ *is equicontinuous, where* $P_i|_K$ *denotes the restriction of* P_i *on* K *for* $i \in \mathbb{N}$. ◇

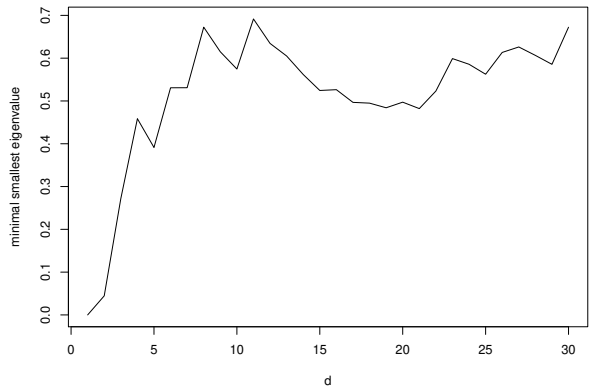

Figure 3.2: *Minimal eigenvalue of* $\frac{1}{d}\sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta})\nabla\lambda_i(\boldsymbol{\eta})^\intercal$ *for* $\boldsymbol{\eta} \in [-1,1]^2$.

Proof.
The proof is based on arguments of Chang and Stout (1991, Lemma 3.1).

Denote by ∇P_i the gradient of P_i. Then we get from the multivariate mean value theorem and the Cauchy–Schwarz inequality

$$|P_i(\boldsymbol{\eta}) - P_i(\boldsymbol{\eta}')| = |\nabla P_i(\tilde{\boldsymbol{\eta}})^\intercal(\boldsymbol{\eta} - \boldsymbol{\eta}')| \leq \|\nabla P_i(\tilde{\boldsymbol{\eta}})\| \cdot \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$$

for a point $\tilde{\boldsymbol{\eta}} \in \{(1-c)\boldsymbol{\eta} + c\boldsymbol{\eta}', c \in [0,1]\}$ on the line between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$. Due to (CS2′), if restricted to K^+, all $\left|\frac{\partial P_i}{\partial \eta_k}\right|$ are uniformly bounded for all $i \in \mathbb{N}$, $1 \leq k \leq q$. Then there is a finite number $\zeta(K^+)$ such that

$$\sup_{(i,\boldsymbol{\eta}) \in \mathbb{N} \times K^+} \|\nabla P_i(\boldsymbol{\eta})\| = \zeta(K^+)$$

and hence for all $\boldsymbol{\eta}, \boldsymbol{\eta}' \in K^+$

$$|P_i(\boldsymbol{\eta}) - P_i(\boldsymbol{\eta}')| \leq \zeta(K^+)\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \quad \text{for all } i \in \mathbb{N}. \tag{3.30}$$

Especially

$$|P_i(\boldsymbol{\eta}) - P_i(\boldsymbol{\eta}')| < \varepsilon \tag{3.31}$$

for all $\boldsymbol{\eta}, \boldsymbol{\eta}' \in K^+$ with $\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| < \delta = \frac{\varepsilon}{\zeta(K^+)}$ for $\varepsilon > 0$ and for all $i \in \mathbb{N}$. Notice, (3.30) and (3.31) are still true, if we take $\boldsymbol{\eta}, \boldsymbol{\eta}' \in K$, where $K \subset K^+$. Hence, the family of maps $\{P_i|_K\}_{i \in \mathbb{N}}$ is equicontinuous for any compact set K, for which a convex and compact set $K^+ \subset \Theta$ exists such that $K \subset K^+$. □

Lemma 3.4.7
Suppose that the conditions (CS2′) *and* (CS4′) *hold. If restricted to any convex and compact set* $K \subseteq \Theta$*, the two families of mappings* $\{\nabla\nabla^\intercal\lambda_i\}_{i \in \mathbb{N}}$ *and* $\{\mathcal{I}_i\}_{i \in \mathbb{N}}$ *are equicontinuous.* ⋄

Proof.
The proofs for $\{\nabla\nabla^\intercal\lambda_i\}_{i \in \mathbb{N}}$ and $\{\mathcal{I}_i\}_{i \in \mathbb{N}}$ can be formulated equivalently.[7] Hence, let $F^{(i)}$ either be $\mathcal{I}_i$ or $\nabla\nabla^\intercal\lambda_i$ for all $i \in \mathbb{N}$ to simplify the notation. Further, we denote by $F^{(i)}_{jk}$

[7] Condition (CS2′) actually suffices for $\mathcal{I}_i$. However, the steps are identical for $\{\nabla\nabla^\intercal\lambda_i\}_{i \in \mathbb{N}}$ and $\{\mathcal{I}_i\}_{i \in \mathbb{N}}$.

the (j,k)th component of $F^{(i)}$ for $j,k=1,\dots,q$ and all $i\in\mathbb{N}$. The conditions (CS2′) and (CS4′) imply that $\nabla F_{jk}^{(i)}(\boldsymbol{\eta})$ exists for all $j,k=1,\dots,q$, $i\in\mathbb{N}$ and $\boldsymbol{\eta}\in K$. Moreover, (CS2′) and (CS4′) imply that there is a $C<\infty$ such that

$$\sup\left\{\left|\frac{\partial F_{jk}^{(i)}(\boldsymbol{\eta})}{\partial\eta_\ell}\right| : j,k,\ell=1,\dots,q,\ i\in\mathbb{N},\ \boldsymbol{\eta}\in K\right\}\leq C. \tag{3.32}$$

Due to the multivariate mean value theorem and the convexity of K, for all $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\in K$, $j,k=1,\dots,q$ and $i\in\mathbb{N}$ there is a $c\in[0,1]$ so that

$$F_{jk}^{(i)}(\boldsymbol{\eta}_1)-F_{jk}^{(i)}(\boldsymbol{\eta}_2)=\nabla F_{jk}^{(i)}(\tilde{\boldsymbol{\eta}})^{\intercal}(\boldsymbol{\eta}_1-\boldsymbol{\eta}_2),$$

where $\tilde{\boldsymbol{\eta}}=c\boldsymbol{\eta}_1+(1-c)\boldsymbol{\eta}_2$. Then, by the Cauchy–Schwarz–inequality and (3.32) we get

$$|F_{jk}^{(i)}(\boldsymbol{\eta}_1)-F_{jk}^{(i)}(\boldsymbol{\eta}_2)|\leq\|\nabla F_{jk}^{(i)}(\tilde{\boldsymbol{\eta}})\|\cdot\|\boldsymbol{\eta}_1-\boldsymbol{\eta}_2\|\leq C\sqrt{q}\|\boldsymbol{\eta}_1-\boldsymbol{\eta}_2\|. \tag{3.33}$$

Notice that the right-hand side of (3.33) is independent of i,j,k. We therefore obtain the following estimate for the maximum-norm of $F^{(i)}(\boldsymbol{\eta}_1)-F^{(i)}(\boldsymbol{\eta}_2)$

$$\|F^{(i)}(\boldsymbol{\eta}_1)-F^{(i)}(\boldsymbol{\eta}_2)\|_{\max}:=\max_{1\leq j,k\leq d}|F_{jk}^{(i)}(\boldsymbol{\eta}_1)-F_{jk}^{(i)}(\boldsymbol{\eta}_2)|\leq C\sqrt{q}\|\boldsymbol{\eta}_1-\boldsymbol{\eta}_2\|,$$

for each $i\in\mathbb{N}$. The equivalence of all norms on $\mathbb{R}^{d\times d}$ implies that there is a $C'>0$ so that

$$\|F^{(i)}(\boldsymbol{\eta}_1)-F^{(i)}(\boldsymbol{\eta}_2)\|\leq C'\|F^{(i)}(\boldsymbol{\eta}_1)-F^{(i)}(\boldsymbol{\eta}_2)\|_{\max}\leq C'C\sqrt{q}\|\boldsymbol{\eta}_1-\boldsymbol{\eta}_2\|$$

for all $i\in\mathbb{N}$ and all $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\in K$. The final step is done analogously to (3.31). □

Lemma 3.4.8
Let a sequence $\{Y_i\}_{i\in\mathbb{N}}\sim\mathcal{P}(\boldsymbol{\eta}_0)$, for a sequence $\{P_i\}_{i\in\mathbb{N}}$ and a fixed $\boldsymbol{\eta}_0\in\Theta$, be given so that the conditions (CS1′[i]), (CS2′) *and* (CS3′) *are satisfied. Then*

$$\limsup_{d\to\infty}\frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)})-\ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right)\leq c(\boldsymbol{\eta})<0,\qquad \mathsf{P}_{\boldsymbol{\eta}_0}\text{–almost surely}, \tag{3.34}$$

holds for every $\boldsymbol{\eta}\in\Theta\setminus\{\boldsymbol{\eta}_0\}$, where $c(\boldsymbol{\eta})$ is the constant of condition (CS3′). ◇

Proof.
The proof is based on arguments of Chang and Stout (1991, Lemma 3.1).

Notice that for each $i\in\mathbb{N}$

$$\begin{aligned}\mathsf{Var}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0))&=\mathsf{Var}_{\boldsymbol{\eta}_0}(Y_i)(\lambda_i(\boldsymbol{\eta})-\lambda_i(\boldsymbol{\eta}_0))^2\\&=P_i(\boldsymbol{\eta}_0)(1-P_i(\boldsymbol{\eta}_0)(\lambda_i(\boldsymbol{\eta})-\lambda_i(\boldsymbol{\eta}_0))^2\end{aligned}$$

holds, where Z_i is defined in equation (3.12). Due to (CS2′), for each $\boldsymbol{\eta}\in\Theta$ there is a constant $0<M(\boldsymbol{\eta},\boldsymbol{\eta}_0)<\infty$ so that

$$\sup_{i\in\mathbb{N}}\mathsf{Var}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0))\leq 2\sup_{i\in\mathbb{N}}|\lambda_i(\boldsymbol{\eta})|^2+2\sup_{i\in\mathbb{N}}|\lambda_i(\boldsymbol{\eta}_0)|^2=M(\boldsymbol{\eta},\boldsymbol{\eta}_0)<\infty.$$

Hence, we get

$$\sum_{i=1}^{\infty}\frac{\mathsf{Var}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0))}{i^2}\leq M(\boldsymbol{\eta},\boldsymbol{\eta}_0)\sum_{i=1}^{\infty}\frac{1}{i^2}<\infty.$$

An application of the strong law of large numbers on the sequence $\{\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0)\}_{i\in\mathbb{N}}$ yields

$$\frac{1}{d}\sum_{i=1}^{d}\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0)-\frac{1}{d}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0))\overset{\mathsf{P}_{\boldsymbol{\eta}_0}-\text{a.s.}}{\longrightarrow}0,\quad d\to\infty.$$

From the last step, equation (3.13) and (CS3′) it follows now that there is a constant $c(\boldsymbol{\eta})<0$ so that (3.34) holds. □

Lemma 3.4.9
Consider $\{Y_i\}_{i\in\mathbb{N}}\sim\mathcal{P}(\boldsymbol{\eta}_0)$ and assume that conditions (CS1′[i]), (CS2′) and (CS3′) are satisfied for a fixed $\boldsymbol{\eta}_0\in\Theta$. Then, for any $\delta>0$ there is a $k(\delta)<0$ such that

$$\lim_{d\to\infty}\mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)}\frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)})-\ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right)<k(\delta)\right)=1,$$

where

$$B_\delta(\boldsymbol{\eta}_0):=\Big\{\boldsymbol{\eta}\in\mathbb{R}^q\colon\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|\le\delta\Big\}.\tag{3.35}$$

◇

Proof.
This proof is based on arguments by Chang and Stout (1991, Lemma 3.1).

For each $i\in\mathbb{N}$, we define the map $H_i:\Theta^2\to\mathbb{R}_{\ge0}$ by

$$(\boldsymbol{\eta},\boldsymbol{\eta}')\mapsto\left|\log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right|+\left|\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right|=:H_i(\boldsymbol{\eta},\boldsymbol{\eta}').$$

Since the image of P_i does not include $\{0,1\}$, H_i is continuous. Moreover, for any $\boldsymbol{\eta}'\in\Theta$ and any $\delta>0$ such that $\bar{B}_\delta(\boldsymbol{\eta}')\subset\Theta$, the map

$$\bar{B}_\delta(\boldsymbol{\eta}')\to\mathbb{R}_{\ge0},\quad\boldsymbol{\eta}\mapsto H_i(\boldsymbol{\eta},\boldsymbol{\eta}')$$

is continuous on the compact set $\bar{B}_\delta(\boldsymbol{\eta}')$ and thus has a maximum value. We denote this maximum value by

$$\hat{H}_i(\delta,\boldsymbol{\eta}'):=\max_{\boldsymbol{\eta}\in\bar{B}_\delta(\boldsymbol{\eta}')}H_i(\boldsymbol{\eta},\boldsymbol{\eta}').$$

In the following we assume, that $\delta>0$ is sufficiently small so that $\bar{B}_\delta(\boldsymbol{\eta}')\subset\Theta$ for the selected value $\boldsymbol{\eta}'\in\Theta$. Letting $\delta\to0$ yields $\bar{B}_\delta(\boldsymbol{\eta}')\to\{\boldsymbol{\eta}'\}$ and therefore, $\lim_{\delta\to0}\hat{H}_i(\delta,\boldsymbol{\eta}')=0$ for each $i\in\mathbb{N}$ and $\boldsymbol{\eta}'\in\Theta$. Since $Y_i\in\{0,1\}$, we get from the triangle inequality

$$\begin{aligned}|\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}')|&=\left|\log\left(P_i(\boldsymbol{\eta})^{Y_i}(1-P_i(\boldsymbol{\eta})^{1-Y_i}\right)-\log\left(P_i(\boldsymbol{\eta}')^{Y_i}(1-P_i(\boldsymbol{\eta}')^{1-Y_i}\right)\right|\\&=\left|Y_i\log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)+(1-Y_i)\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right|\\&\le\left|\log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right|+\left|\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right|=H_i(\boldsymbol{\eta},\boldsymbol{\eta}')\le\hat{H}_i(\delta,\boldsymbol{\eta}'),\end{aligned}\tag{3.36}$$

for $\delta \geq \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$. For any fixed $\boldsymbol{\eta}' \in \Theta$ we get

$$\begin{aligned}\hat{H}_i(\delta, \boldsymbol{\eta}') &= \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left| \log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right| + \left|\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right| \right\} \\ &\leq \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left| \log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right| \right\} + \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left|\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right| \right\}.\end{aligned}$$

Applying the multivariate mean value theorem to $\log(\bullet)$ we get

$$\begin{aligned}\max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left| \log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right| \right\} &= \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left| \log\left(P_i(\boldsymbol{\eta})\right) - \log\left(P_i(\boldsymbol{\eta}')\right)\right| \right\} \\ &= \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \frac{1}{\xi(P_i(\boldsymbol{\eta}), P_i(\boldsymbol{\eta}'))} \left| P_i(\boldsymbol{\eta}) - P_i(\boldsymbol{\eta}')\right| \right\}, \qquad (3.37)\end{aligned}$$

where $\xi(P_i(\boldsymbol{\eta}), P_i(\boldsymbol{\eta}'))$ is a point between $P_i(\boldsymbol{\eta})$ and $P_i(\boldsymbol{\eta}')$. Let $\zeta_0(K)$ and $\zeta_1(K)$ be given for each compact $K \subset \Theta$ as in (CS2′). Additionally, for each compact and convex set $K \subseteq \Theta$ there is a $\zeta_3(K) > 0$ such that

$$\left| P_i(\boldsymbol{\eta}) - P_i(\boldsymbol{\eta}')\right| \leq \zeta_3(K) \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \qquad (3.38)$$

for all $i \in \mathbb{N}$ due to Lemma 3.4.6. By a combination of (3.37), the analogously derived version for $1 - P_i(\boldsymbol{\eta})$ instead of $P_i(\boldsymbol{\eta})$, (3.38) and (3.22) of condition (CS2'), for all compact and convex $K \supseteq \bar{B}_\delta(\boldsymbol{\eta}')$ we obtain

$$\begin{aligned}\max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left| \log\left(\frac{P_i(\boldsymbol{\eta})}{P_i(\boldsymbol{\eta}')}\right)\right| \right\} + \max_{\boldsymbol{\eta} \in \bar{B}_\delta(\boldsymbol{\eta}')} \left\{ \left|\log\left(\frac{1-P_i(\boldsymbol{\eta})}{1-P_i(\boldsymbol{\eta}')}\right)\right| \right\} &\leq \left(\frac{\zeta_3(K)}{\zeta_0(K)} + \frac{\zeta_3(K)}{1-\zeta_1(K)} \right) \delta \\ &=: C(K)\delta\end{aligned}$$

with $C(K) < \infty$. For each $\varepsilon > 0$ there is therefore a $\delta > 0$ so that $\frac{1}{d}\sum_{i=1}^{d} \hat{H}_i(\delta, \boldsymbol{\eta}') < \varepsilon$ for all $d \in \mathbb{N}$. This implies

$$\lim_{\delta \to 0} \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d} \hat{H}_i(\delta, \boldsymbol{\eta}') = 0. \qquad (3.39)$$

Next, we will show that for any $\boldsymbol{\eta}_i \neq \boldsymbol{\eta}_0$ there is a sufficiently small $\delta_i > 0$ and a sufficiently large $c_i < 0$ such that

$$\lim_{d \to \infty} \mathsf{P}_{\boldsymbol{\eta}_0} \left(\sup_{\boldsymbol{\eta} \in \bar{B}_{\delta_i}(\boldsymbol{\eta}_i)} \frac{1}{d} \left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}) \right) < c_i < 0 \right) = 1. \qquad (3.40)$$

First, for any $\boldsymbol{\eta} \in \bar{B}_{\delta_i}(\boldsymbol{\eta}_i)$ equation (3.36) implies

$$\begin{aligned}\frac{1}{d} \left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}) \right) &\leq \frac{1}{d} \left(\ell^{(d)}(\boldsymbol{\eta}_i \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}) \right) \\ &\quad + \frac{1}{d} \left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_i \,|\, \boldsymbol{Y}^{(d)}) \right) \\ &\leq \frac{1}{d} \left(\ell^{(d)}(\boldsymbol{\eta}_i \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}) \right) + \frac{1}{d} \sum_{i=1}^{d} \hat{H}_i(\delta, \boldsymbol{\eta}_i).\end{aligned}$$

Since $\boldsymbol{\eta}_i \in \Theta$, we get for $\boldsymbol{\eta} = \boldsymbol{\eta}_i$ from Lemma 3.4.8

$$\limsup_{d\to\infty} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}_i \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})\right) \leq c(\boldsymbol{\eta}_i) < 0 \quad \mathsf{P}_{\boldsymbol{\eta}_0}\text{– almost surely.}$$

Equation (3.39) implies for all $\boldsymbol{\eta}' = \boldsymbol{\eta}_i$ and for each $\varepsilon > 0$ that there is a sufficiently small $\delta > 0$ such that $\limsup_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \hat{H}_i(\delta, \boldsymbol{\eta}_i) < \varepsilon$. Therefore, there is a $\delta > 0$ sufficiently small and a negative number c_i, for example $c_i = \frac{c(\boldsymbol{\eta}_i)}{2}$ so that (3.40) holds.

Equation (3.40) still holds if we replace $\overline{B}_\delta(\boldsymbol{\eta}_i)$ by an arbitrary subset of $\overline{B}_\delta(\boldsymbol{\eta}_i)$. Especially, for all $\boldsymbol{\eta}_i \neq \boldsymbol{\eta}_0$, there exists a sufficiently small δ_i and a constant $c_i < 0$ such that for all compact sets $K_i \subset \overline{B}_{\delta_i}(\boldsymbol{\eta}_i)$

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in K_i} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})\right) < c_i\right) = 1 \tag{3.41}$$

holds.

Next, we make a case-by-case analysis. First, we assume that Θ is compact. Then for each $\delta > 0$ such that $\Theta\backslash B_\delta(\boldsymbol{\eta}_0) \neq \emptyset$, we define $\Theta_0 := \Theta\backslash B_\delta(\boldsymbol{\eta}_0)$, which is compact, too. For each $\delta' > 0$, $\bigcup_{\boldsymbol{\eta}\in\Theta_0} B_{\delta'}(\boldsymbol{\eta})$ is a covering of Θ_0 and since Θ_0 is compact, there are $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N \in \Theta_0$ for an $N \in \mathbb{N}$ such that $\Theta_0 \subset \bigcup_{j=1}^{N} B_{\delta'}(\boldsymbol{\eta}_j)$. Notice that also $\Theta_0 = \bigcup_{j=1}^{N} K_i$ holds, where $K_i := \bar{B}_{\delta'}(\boldsymbol{\eta}_j) \cap \Theta_0$, $i = 1, \ldots, N$. Particularly, for every K_i there is a compact and convex superset $\tilde{K}_i \supset K_i$ with $\tilde{K}_i \subset \Theta$ due to (CS1′[i]). Then, we can define the events

$$A_i^{(d)} := \left\{\omega \in \Omega\colon \sup_{\boldsymbol{\eta}\in K_i} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}(\omega)) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}(\omega))\right) < c_i\right\}, \tag{3.42}$$

for some $c_1, \ldots, c_N < 0$. Equation (3.41) implies that $\mathsf{P}(A_i^{(d)}) \to 1$ for $d \to \infty$ and

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\bigcap_{i=1}^{N} A_i^{(d)}\right) = 1. \tag{3.43}$$

Equation (3.43) is still true, if we replace in (3.42) the constant c_i by $k = \max\{c_1, \ldots, c_N\} < 0$. We hence get

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\backslash B_\delta(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})\right) < k\right) = 1. \tag{3.44}$$

Now, we assume that Θ is unbounded. Then we can define the sequence $\{\Theta_j\}_{j\in\mathbb{N}_0}$ with $\bigcup_{j\in\mathbb{N}_0} \Theta_j = \Theta \setminus B_\delta(\boldsymbol{\eta}_0)$ by setting

$$\Theta_j = \left\{\boldsymbol{\eta} \in \Theta\colon \delta + j \leq \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \delta + j + 1\right\}, \qquad j \in \mathbb{N}_0.$$

On each Θ_j, $j \in \mathbb{N}_0$, we can apply the analysis of the first case with Θ_j instead of $\Theta \setminus B_\delta(\boldsymbol{\eta}_0)$ and $k_j := \sup_{\boldsymbol{\eta}\in\Theta_j} c(\boldsymbol{\eta})/2 < 0$ instead of $k < 0$. Letting $d \to \infty$ we get with probability tending to one

$$\begin{aligned}
&\sup_{\boldsymbol{\eta}\in\Theta\backslash B_\delta(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})\right) \\
&\quad = \sup_{j\in\mathbb{N}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta_j} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)})\right)\right) \\
&\quad \leq \sup_{j\in\mathbb{N}_0} k_j \leq \sup_{\boldsymbol{\eta}\in\Theta\backslash B_\delta(\boldsymbol{\eta}_0)} c(\boldsymbol{\eta})/2 =: 2 \cdot k(\delta).
\end{aligned}$$

Since equation (3.23) in condition (CS3′) implies $\sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} c(\boldsymbol{\eta})/2 < 0$, we finish the proof by setting $k(\delta) := \sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} c(\boldsymbol{\eta})/3 < 0$. □

Using this lemma, we can show the existence and consistency of the MLE, i.e. Theorem 3.4.3. We do it now because both properties are necessary for the remaining part of the proofs.

Proof of Theorem 3.4.3.
The proof of the consistency is based on the proof of Corollary 3.1 of Chang and Stout (1991). Similar to classic maximum likelihood theory (cf. for example Lehmann and Casella, 1998), we define the set

$$\begin{aligned} C_{d,\delta} :=& \{\boldsymbol{y}\in\{0,1\}^d \,|\, \ell^{(d)}(\boldsymbol{c}\,|\,\boldsymbol{y}) < \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{y}) \text{ for all } \boldsymbol{c}\in\partial B_\delta(\boldsymbol{\eta}_0)\} \\ =& \left\{\boldsymbol{y}\in\{0,1\}^d \,\middle|\, \frac{1}{d}\ell^{(d)}(\boldsymbol{c}\,|\,\boldsymbol{y}) < \frac{1}{d}\ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{y}) \text{ for all } \boldsymbol{c}\in\partial B_\delta(\boldsymbol{\eta}_0)\right\}, \end{aligned}$$

for $d\in\mathbb{N}$ and $\delta>0$ such that $B_\delta(\boldsymbol{\eta}_0)\subset\Theta$. By the definition of $C_{d,\delta}$, at least at the point $\boldsymbol{\eta}_0\in B_\delta(\boldsymbol{\eta}_0)$, the mapping $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})$ takes a larger value in the interior of the open ball $B_\delta(\boldsymbol{\eta}_0)$ than in the boundary $\partial B_\delta(\boldsymbol{\eta}_0)$ for all $\boldsymbol{y}\in C_{d,\delta}$. So, for all $\boldsymbol{y}\in C_{d,\delta}$, $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})$ has at least one local maximum in $B_\delta(\boldsymbol{\eta}_0)$. For all $\boldsymbol{y}\in C_{d,\delta}$, we denote by $M_{\boldsymbol{y}}\subset B_\delta(\boldsymbol{\eta}_0)$ the set of all local maximum points of $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})\big|_{B_\delta(\boldsymbol{\eta}_0)}$. Since $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})\big|_{B_\delta(\boldsymbol{\eta}_0)}\in C^2(B_\delta(\boldsymbol{\eta}_0),\mathbb{R})$ due to (CS2′), $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})\big|_{B_\delta(\boldsymbol{\eta}_0)}$ is continuously differentiable and so for all $\boldsymbol{y}\in C_{d,\delta}$ each point $\boldsymbol{\eta}^*\in M_{\boldsymbol{y}}$ satisfies the likelihood equations, i.e. $\nabla\ell^{(d)}(\boldsymbol{\eta}^*\,|\,\boldsymbol{y}) = \mathbf{0}$. Notice that $\partial B_\delta(\boldsymbol{\eta}_0)\subset\Theta\setminus B_\delta(\boldsymbol{\eta}_0)$ for all $\delta>0$ with $B_\delta(\boldsymbol{\eta}_0)\subset\Theta$. Lemma 3.4.9 now implies $\mathsf{P}_{\boldsymbol{\eta}_0}(\boldsymbol{Y}^{(d)}\in C_{d,\delta})\overset{d\to\infty}{\longrightarrow}1$. In particular, the probability is tending to one such that

$$\frac{1}{d}\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) < \frac{1}{d}\ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)}) \qquad \text{for all } \boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0).$$

Therefore, with probability tending to one for $d\to\infty$, at least one of the local maxima of $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})$ in the interior of $B_\delta(\boldsymbol{\eta}_0)$ has to be a global maximum of $\ell^{(d)}(\bullet\,|\,\boldsymbol{y})$.

Next, we have to select a specific maximum point in a way that it is a measurable mapping $(\{0,1\}^d,\mathrm{Pot}(\{0,1\}^d))\to(\Theta,\mathcal{B}(\Theta))$ to define a sequence of random variables, which are statistics of the response variables.

We write $\Theta_\delta := \overline{B}_\delta(\boldsymbol{\eta}_0)\cap\Theta\subset\Theta$ for a restricted compact parameter space.[8] Notice,

$$\ell^{(d)}(\bullet\,|\,\boldsymbol{y})\colon(\Theta_\delta,\mathcal{B}(\Theta_\delta))\to(\mathbb{R},\mathcal{B}^1)$$

is continuous for each fixed $\boldsymbol{y}\in\{0,1\}^d$ and

$$\ell^{(d)}(\boldsymbol{\eta}\,|\,\bullet)\colon(\{0,1\}^d,\mathrm{Pot}(\{0,1\}^d))\to(\mathbb{R},\mathcal{B}^1)$$

is measurable for each fixed $\boldsymbol{\eta}\in\Theta_\delta$. Then by the remark of Witting and Müller-Funk (1995, page 173)

$$(\{0,1\}^d,\mathrm{Pot}(\{0,1\}^d))\to(\mathbb{R},\mathcal{B}^1),\quad \boldsymbol{y}\mapsto\sup_{\boldsymbol{\eta}\in\Theta_\delta}\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{y})$$

[8] If Θ is compact, we do not have to restrict Θ. Nevertheless, we do it for the simplicity of the formulation. In fact, by Witting and Müller-Funk (1995, page 173), the following argument is in principal true if we do not restrict, even if Θ is unbounded. But then the corresponding random variables can take the "value" infinity, which we want to avoid.

is a measurable mapping. Further, since Θ_δ is compact, for each $\boldsymbol{y} \in \{0,1\}^d$ the mapping $\ell^{(d)}(\bullet \,|\, \boldsymbol{y})$ takes the maximum over Θ_δ and we can select an $\boldsymbol{\eta}^* \in \Theta_\delta$ such that

$$\sup_{\boldsymbol{\eta}\in\Theta_\delta} \ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{y}) = \ell^{(d)}(\boldsymbol{\eta}^* \,|\, \boldsymbol{y}).$$

We may now apply an argument of Witting and Müller-Funk (1995, page 173, "Hilfssatz" 6.7 part b))[9] to get the existence of a measurable mapping

$$\tilde{\boldsymbol{\eta}}_d \colon (\{0,1\}^d, \mathrm{Pot}(\{0,1\}^d)) \to (\Theta, \mathcal{B}(\Theta))$$

such that $\sup_{\boldsymbol{\eta}\in\Theta_\delta} \ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{y}) = \ell^{(d)}(\tilde{\boldsymbol{\eta}}_d(\boldsymbol{y}) \,|\, \boldsymbol{y})$ for all $\boldsymbol{y} \in \{0,1\}^d$. This enables us to well define the sequence $\{\hat{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ by setting $\hat{\boldsymbol{\eta}}_d := \tilde{\boldsymbol{\eta}}_d(\boldsymbol{Y}^{(d)})$ for each $d \in \mathbb{N}$. By the previous part, the probability is tending to one for $d \to \infty$ such that $\hat{\boldsymbol{\eta}}_d$ is a local maximum in Θ_0 and a global maximum of $\ell^{(d)}(\,\bullet \mid \boldsymbol{Y}^{(d)})$ (and thus really the maximum likelihood estimator).

It remains to prove the consistency. We assume that $\hat{\boldsymbol{\eta}}_d$ is the (restricted) MLE for $Y_1, \ldots, Y_d$, $d \in \mathbb{N}$, given in previous part. Since it is the maximum of $\ell^{(d)}(\bullet \,|\, \boldsymbol{Y}^{(d)})\big|_{B_\delta(\boldsymbol{\eta}_0)}$ we have

$$\ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\, \boldsymbol{Y}^{(d)}) = \log\left(\frac{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d)}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \boldsymbol{\eta}_0)}\right) \geq 0. \tag{3.45}$$

Therefore, it is sufficient to proof for all $\varepsilon > 0$ and $\delta > 0$ that there is an $N(\varepsilon,\delta) \in \mathbb{N}$ such that $\mathsf{P}_{\boldsymbol{\eta}_0}(\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| < \delta) > 1 - \varepsilon$, for all $d > N(\varepsilon,\delta)$. Suppose $\hat{\boldsymbol{\eta}}_d$ is not consistent, then there exists $\varepsilon_0 > 0$ and $\delta_0 > 0$ such that for any $N \in \mathbb{N}$ there exists some $d > N$ satisfying $\mathsf{P}_{\boldsymbol{\eta}_0}(\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| > \delta_0) > \varepsilon_0$. Therefore, we can obtain a subsequence $\{\hat{\boldsymbol{\eta}}_{d_i}\}_{i\in\mathbb{N}}$ so that $\mathsf{P}_{\boldsymbol{\eta}_0}(\|\hat{\boldsymbol{\eta}}_{d_i} - \boldsymbol{\eta}_0\| > \delta_0) > \varepsilon_0$, for all $i \in \mathbb{N}$. This especially implies

$$\varepsilon_0 \leq \limsup_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}(\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| > \delta_0) \leq \mathsf{P}_{\boldsymbol{\eta}_0}(A_0), \tag{3.46}$$

with

$$A_0 := \bigcap_{d\in\mathbb{N}} \bigcup_{m\geq d} \left\{ \omega \in \Omega \,\middle|\, \|\hat{\boldsymbol{\eta}}_m(\omega) - \boldsymbol{\eta}_0\| > \delta_0 \right\},$$

since $\limsup_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}(E_d) \leq \mathsf{P}_{\boldsymbol{\eta}_0}(\limsup_{d\to\infty} E_d)$ for all sequences of events $\{E_d\}_{d\in\mathbb{N}}$. For all $\omega \in A_0$ we get $\hat{\boldsymbol{\eta}}_d(\omega) \in \Theta \backslash B_{\delta_0}(\boldsymbol{\eta}_0)$ for an infinite number of $d \in \mathbb{N}$ and this implies

$$\sup_{\boldsymbol{\eta}\in\Theta\backslash B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\, \boldsymbol{Y}^{(d)})\right) \geq 0$$

for infinitely many $d \in \mathbb{N}$. In particular,

$$A_0 \subset \bigcap_{d\in\mathbb{N}} \bigcup_{m\geq d} \left\{ \omega \in \Omega \,\middle|\, \sup_{\boldsymbol{\eta}\in\Theta\backslash B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{m}\left(\ell^{(m)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(m)}(\omega)) - \ell^{(m)}(\hat{\boldsymbol{\eta}}_m(\omega) \,|\, \boldsymbol{Y}^{(m)}(\omega))\right) \geq 0 \right\}.$$

[9] Witting and Müller-Funk (1995) "Hilfssatz" 6.7 part b) of page 173 states in our notation and in English: Consider a function $g(\bullet,\bullet)$ so that for each fixed $\boldsymbol{\eta} \in \Theta \subset \mathbb{R}^q$ the mapping $g(\bullet,\boldsymbol{\eta})\colon (\mathbb{R}^d, \mathcal{B}^d) \to (\mathbb{R}^d, \mathcal{B}^d)$ is measurable and for each fixed $\boldsymbol{x} \in \mathbb{R}^d$ the mapping $g(\boldsymbol{x},\bullet)\colon \Theta \to \mathbb{R}^d$ is continuous. Assume that Θ is compact or there is a sequence of compact sets $\{U_i\}_{i\in\mathbb{N}}$ with $\Theta = \bigcup_{i\in\mathbb{N}} U_i$. Further, assume that there is a mapping $\vartheta\colon \mathbb{R}^d \to \Theta$ such that $\boldsymbol{x} \mapsto g(\boldsymbol{x},\vartheta(\boldsymbol{x})) =: h(\boldsymbol{x})$ is measurable, then there is a measurable mapping $\tilde{\vartheta}\colon (\mathbb{R}^d, \mathcal{B}^d) \to (\Theta, \mathcal{B}(\Theta))$ such that $h(\boldsymbol{x}) = g(\boldsymbol{x},\tilde{\vartheta}(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Now, (3.45) and (3.46) imply

$$\varepsilon_0 \leq \mathsf{P}_{\boldsymbol{\eta}_0}(A_1) \tag{3.47}$$

for the event

$$A_1 := \bigcap_{d\in\mathbb{N}} \bigcup_{m\geq d} \left\{ \omega \in \Omega \,\middle|\, \sup_{\boldsymbol{\eta}\in\Theta\setminus B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{m}\left(\ell^{(m)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(m)}(\omega)) - \ell^{(m)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(m)}(\omega))\right) \geq 0 \right\}.$$

But, according to Lemma 3.4.9, there is a $c = c(\delta_0) < 0$ such that for the event

$$B := \bigcap_{d\in\mathbb{N}} \bigcup_{m\geq d} \left\{ \omega \in \Omega \,\middle|\, \sup_{\boldsymbol{\eta}\in\Theta\setminus B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{m}\left(\ell^{(m)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(m)}(\omega)) - \ell^{(m)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(m)}(\omega))\right) < c \right\}$$

we get

$$\begin{aligned} 1 &= \lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\setminus B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right) < c < 0 \right) \\ &= \limsup_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\setminus B_{\delta_0}(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right) < c < 0 \right) \\ &\leq \mathsf{P}_{\boldsymbol{\eta}_0}(B), \end{aligned} \tag{3.48}$$

where we can replace "$\leq$" in the last step by "$=$". Notice that $A_1 \cap B = \emptyset$, since

$$0 \leq \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right) \leq c < 0$$

is impossible. Combining (3.47) and (3.48) results in

$$1 \geq \mathsf{P}_{\boldsymbol{\eta}_0}(A_1 \cup B) = \mathsf{P}_{\boldsymbol{\eta}_0}(A_1) + \mathsf{P}_{\boldsymbol{\eta}_0}(B) \geq \varepsilon_0 + 1 > 1 \qquad \square$$

The existence and consistency of penelized MLEs is directly obtained as we see in the following.

Proof of Remark 3.4.4.
Due to the requirement that $\mathcal{W}$ is positive and bounded, there is a positive constant C_1 so that $\sup_{\boldsymbol{\eta}\in\Theta} \mathcal{W}(\boldsymbol{\eta}) < C_1 < \infty$. Let $d \in \mathbb{N}$ and $\delta > 0$ be selected arbitrarily, then

$$\begin{aligned} &\sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\tilde{\ell}^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \tilde{\ell}^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right) \\ &\quad\leq \sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \left(\frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right)\right) \\ &\qquad + \sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \left(\frac{1}{d}\left(\log \mathcal{W}(\boldsymbol{\eta}) - \log \mathcal{W}(\boldsymbol{\eta}_0)\right)\right) \\ &\quad\leq \sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \left(\frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right)\right) + \frac{1}{d}\left(\log(C_1) - \log \mathcal{W}(\boldsymbol{\eta}_0)\right). \end{aligned} \tag{3.49}$$

The term $d^{-1}(\log(C_1) - \log \mathcal{W}(\boldsymbol{\eta}_0))$ approaches zero as $d \to \infty$. By applying Lemma 3.4.9 on the first term on the right-hand side of (3.49) results in

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\tilde{\ell}^{(d)}(\boldsymbol{\eta}\,|\,\boldsymbol{Y}^{(d)}) - \tilde{\ell}^{(d)}(\boldsymbol{\eta}_0\,|\,\boldsymbol{Y}^{(d)})\right) < \tilde{k}(\delta)\right) = 1$$

for each $\tilde{k}(\delta) \in (k(\delta), 0)$, e.g. $\tilde{k}(\delta) := k(\delta)/2$, where $k(\delta)$ is given from Lemma 3.4.9. The remaining steps of the proof go along the lines of the proof of Theorem 3.4.3 via simply replacing ℓ by $\tilde{\ell}$. □

To prepare the proof of the next lemma, we need a further technical tool.

Lemma 3.4.10
Let $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ be symmetric and positive definite, $\boldsymbol{x} \in \mathbb{R}^q$ and $\boldsymbol{B} \in \mathbb{R}^{q\times q}$ for $q \in \mathbb{N}$. Then

$$|\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}| \leq \sqrt{\frac{\nu_{max}(\boldsymbol{A})}{\nu_{min}(\boldsymbol{A})}}\|\boldsymbol{B}\|\, \boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x}. \qquad \diamond$$

Proof.
We first observe that

$$\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} \geq \nu_{\min}(\boldsymbol{A})\|\boldsymbol{x}\|^2. \tag{3.50}$$

Since $\boldsymbol{A}$ is symmetric and positive definite, we can define a scalar product and a norm by setting

$$\langle \boldsymbol{v}, \boldsymbol{w}\rangle_{\boldsymbol{A}} := \boldsymbol{v}^\intercal \boldsymbol{A}\boldsymbol{w}, \quad \text{and} \quad \|\boldsymbol{v}\|_{\boldsymbol{A}} := \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle_{\boldsymbol{A}}}, \qquad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^q.$$

Additionally, we define the matrix norm

$$\|\boldsymbol{V}\|_{\boldsymbol{A}} := \sup_{\boldsymbol{v}\in\mathbb{R}^q\setminus\{\boldsymbol{0}\}} \frac{\|\boldsymbol{V}\boldsymbol{v}\|_{\boldsymbol{A}}}{\|\boldsymbol{v}\|_{\boldsymbol{A}}}, \qquad \boldsymbol{V} \in \mathbb{R}^{q\times q}. \tag{3.51}$$

Since each symmetric and positive definite matrix has a unique symmetric and positive definite square root matrix, we get $\|\boldsymbol{V}\boldsymbol{v}\|_{\boldsymbol{A}} = \|\boldsymbol{A}^{1/2}\boldsymbol{V}\boldsymbol{v}\|$. The induced matrix norm is always compatible to the corresponding vector norm and hence we get from the sub-multiplicativity

$$\|\boldsymbol{V}\boldsymbol{v}\|_{\boldsymbol{A}} \leq \|\boldsymbol{A}^{1/2}\| \cdot \|\boldsymbol{V}\boldsymbol{v}\| = \nu_{\max}(\boldsymbol{A}^{1/2})\|\boldsymbol{V}\boldsymbol{v}\| = \nu^{1/2}_{\max}(\boldsymbol{A})\|\boldsymbol{V}\boldsymbol{v}\|. \tag{3.52}$$

Applying (3.50) and (3.52) to (3.51), we get an estimate for an upper bound

$$\|\boldsymbol{V}\|_{\boldsymbol{A}} \leq \sup_{\boldsymbol{v}\in\mathbb{R}^q\setminus\{\boldsymbol{0}\}} \frac{\nu^{1/2}_{\max}(\boldsymbol{A})\|\boldsymbol{V}\boldsymbol{v}\|}{\nu^{1/2}_{\min}(\boldsymbol{A})\|\boldsymbol{v}\|} = \sqrt{\frac{\nu_{\max}(\boldsymbol{A})}{\nu_{\min}(\boldsymbol{A})}}\|\boldsymbol{V}\|.$$

Finally, by Cauchy-Schwarz's inequality and sub-multiplicativity of compatible matrix norms we obtain

$$\langle \boldsymbol{x}, \boldsymbol{B}\boldsymbol{x}\rangle^2_{\boldsymbol{A}} \leq \|\boldsymbol{x}\|^2_{\boldsymbol{A}}\, \|\boldsymbol{B}\boldsymbol{x}\|^2_{\boldsymbol{A}} \leq \|\boldsymbol{B}\|^2_{\boldsymbol{A}}\|\boldsymbol{x}\|^4_{\boldsymbol{A}} \leq \frac{\nu_{\max}(\boldsymbol{A})}{\nu_{\min}(\boldsymbol{A})}\|\boldsymbol{B}\|^2\, \langle \boldsymbol{x}, \boldsymbol{x}\rangle^2_{\boldsymbol{A}}.$$

This is equivalent to

$$\Leftrightarrow |\langle \boldsymbol{x}, \boldsymbol{B}\boldsymbol{x}\rangle_{\boldsymbol{A}}| \leq \sqrt{\frac{\nu_{\max}(\boldsymbol{A})}{\nu_{\min}(\boldsymbol{A})}}\|\boldsymbol{B}\|\, \langle \boldsymbol{x}, \boldsymbol{x}\rangle_{\boldsymbol{A}},$$

which completes the proof. □

Lemma 3.4.11
Let a sequence $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ be given and suppose that conditions (CS1′)–(CS5′) hold for a fixed $\boldsymbol{\eta}_0 \in \Theta$. Let I_q be the $q \times q$ identity matrix. Denote by $\nabla\nabla^\intercal$ the Hessian-operator and let $\widehat{\boldsymbol{\Sigma}}_d$ be given by (3.10). Then a sequence $\{a_d\}_{d\in\mathbb{N}}$, $a_i \in [0,1]$, exists such that for

$$R_d(\boldsymbol{\eta}) \coloneqq \widehat{\boldsymbol{\Sigma}}_d(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)})),$$

with $\boldsymbol{\eta}_d^* \coloneqq a_d\hat{\boldsymbol{\eta}}_d + (1-a_d)\boldsymbol{\eta}$, holds with probability tending to 1 for $d \to \infty$:

$$\ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) - \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\, \boldsymbol{Y}^{(d)}) = \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \tag{3.53}$$

$$= -\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)(\mathrm{I}_q - R_d(\boldsymbol{\eta}))(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d), \tag{3.54}$$

$\boldsymbol{\eta} \in \Theta$. Moreover, for any $\varepsilon > 0$, a $\delta > 0$ and a $\delta' > 0$ exist such that

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in B_\delta(\boldsymbol{\eta}_0)} \|R_d(\boldsymbol{\eta})\| < \varepsilon\right) = 1 \tag{3.55}$$

and

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\Big((1+\varepsilon)Q_d(\boldsymbol{\eta}) \le -\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)(\mathrm{I}_q - R_d(\boldsymbol{\eta}))(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \le (1-\varepsilon)Q_d(\boldsymbol{\eta})\Big) = 1, \tag{3.56}$$

for all $\boldsymbol{\eta} \in B_{\delta'}(\boldsymbol{\eta}_0)$, where

$$Q_d(\boldsymbol{\eta}) \coloneqq -\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d), \quad \boldsymbol{\eta} \in B_\delta(\boldsymbol{\eta}_0),\ d \in \mathbb{N}. \qquad \diamond$$

Proof.
This proof of (3.53)–(3.55) is based on arguments by Chang and Stout (1991, Lemma 3.2).

First, we get (3.53) directly by using a second order Taylor expansion of $\ell^{(d)}(\bullet \,|\, \boldsymbol{Y}^{(d)})$ at $\hat{\boldsymbol{\eta}}_d$, since $\nabla\ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\, \boldsymbol{Y}^{(d)}) = \mathbf{0}$. Theorem 3.4.3 implies that for each $\varepsilon > 0$ and $\delta > 0$ there is an $N(\varepsilon,\delta) \in \mathbb{N}$ such that $\mathsf{P}_{\boldsymbol{\eta}_0}(\hat{\boldsymbol{\eta}}_d \in B_\delta(\boldsymbol{\eta}_0)) > 1-\varepsilon$ for all $d > N(\varepsilon,\delta)$. Therefore, we assume without loss of generality that $\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| \le \delta$ for the discussed d and δ. Further, due to (CS5′) there is an $N \in \mathbb{N}$ so that

$$\inf_{\boldsymbol{\eta}\in B_\delta(\boldsymbol{\eta}_0)} \nu_{\min}\left(\frac{1}{d}\mathcal{I}_d(\boldsymbol{\eta})\right) > 0$$

for all $d > N$. Notice that $\|A^{-1}\| = 1/\nu_{\min}(A)$ holds for the selected matrix norm for all $A \in \mathbb{R}^{q\times q}$ with $\det(A) \neq 0$. Therefore, we assume without restriction that $\frac{1}{d}\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)$ is regular and that there is a constant $C_0 > 0$ (which is independent of d) such that

$$\left\|\left(\frac{1}{d}\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)\right)^{-1}\right\| \le \frac{1}{C_0}, \tag{3.57}$$

for $d > N(\delta)$.

Using the sub-multiplicative property of matrix norms and (3.57), we get

$$\begin{aligned}\|R_d(\boldsymbol{\eta})\| &= \left\| \left(\frac{1}{d} \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) \right)^{-1} \frac{1}{d} (\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)})) \right\| \\ &\leq \left\| \left(\frac{1}{d} \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) \right)^{-1} \right\| \cdot \left\| \frac{1}{d} (\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)})) \right\| \\ &\leq \frac{1}{C_0} \left\| \frac{1}{d} (\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)})) \right\|.\end{aligned}$$

We now study a decomposition of $\|\frac{1}{d}(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + H_d(\boldsymbol{\eta}_d^*))\|$ in order to prove (3.55). Notice first that

$$\nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) = \sum_{i=1}^{d} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta})(Y_i - P_i(\boldsymbol{\eta})) - \mathcal{I}^{(d)}(\boldsymbol{\eta})$$

for all $\boldsymbol{\eta} \in \Theta$. The triangle inequality next implies

$$\begin{aligned}C_0\|R_d(\boldsymbol{\eta})\| \leq& \left\| \frac{1}{d} \Big(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) + \nabla\nabla^\intercal \ell^{(d)}(\boldsymbol{\eta}_d^* \,|\, \boldsymbol{Y}^{(d)}) \Big) \right\| \\ =& \left\| \frac{1}{d} \sum_{i=1}^{d} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*)(Y_i - P_i(\boldsymbol{\eta}_d^*)) - \frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta}_d^*) + \frac{1}{d}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d) \right\| \\ =& \Bigg\| \frac{1}{d} \sum_{i=1}^{d} \Big(\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*) - \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) + \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \Big) \\ &\qquad \times \Big(Y_i - P_i(\boldsymbol{\eta}_d^*) - P_i(\boldsymbol{\eta}_0) + P_i(\boldsymbol{\eta}_0) \Big) - \frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta}_d^*) + \frac{1}{d}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d) \Bigg\| \\ =& \Bigg\| \frac{1}{d} \sum_{i=1}^{d} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)(Y_i - P_i(\boldsymbol{\eta}_0)) + \frac{1}{d} \sum_{i=1}^{d} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)(P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*)) \\ &+ \frac{1}{d} \sum_{i=1}^{d} \big(\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*) - \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \big)(Y_i - P_i(\boldsymbol{\eta}_d^*)) + \frac{1}{d} \Big(- \mathcal{I}^{(d)}(\boldsymbol{\eta}_d^*) + \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d) \Big) \Bigg\| \\ \leq& \left\| \frac{1}{d} \sum_{i=1}^{d} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)(Y_i - P_i(\boldsymbol{\eta}_0)) \right\| + \frac{1}{d} \sum_{i=1}^{d} \left\| \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)(P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*)) \right\| \\ &+ \frac{1}{d} \sum_{i=1}^{d} \left\| \big(\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*) - \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \big)(Y_i - P_i(\boldsymbol{\eta}_d^*)) \right\| \qquad (3.58) \\ &+ \frac{1}{d} \sum_{i=1}^{d} \left\| - \mathcal{I}_i(\boldsymbol{\eta}_d^*) + \mathcal{I}_i(\hat{\boldsymbol{\eta}}_d) \right\|.\end{aligned}$$

Since $Y_i \in \{0,1\}$, $P_i \in (0,1)$, $i \in \mathbb{N}$, and due to Corollary 3.4.7, there are constants $C_1, C_2 > 0$ so that

$$\| - \mathcal{I}_i(\boldsymbol{\eta}_d^*) + \mathcal{I}_i(\hat{\boldsymbol{\eta}}_d)\| \leq C_1 \|\boldsymbol{\eta}_d^* - \hat{\boldsymbol{\eta}}_d\|, \qquad (3.59)$$

and

$$\begin{aligned}\|(\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*) - \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0))(Y_i - P_i(\boldsymbol{\eta}_d^*))\| &\leq \|\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_d^*) - \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)\| \\ &\leq C_2 \|\boldsymbol{\eta}_d^* - \boldsymbol{\eta}_0\|, \qquad (3.60)\end{aligned}$$

for $i \in \mathbb{N}$. Further, due to (CS2′), the norm-equivalence and Lemma 3.4.6, there are constants $C_4, C_5 > 0$ so that

$$\begin{aligned}\|\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)(P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*))\| &\leq C_4\|\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)(P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*))\|_{\max} \\ &\leq C_4\|C_3\mathbf{1}_{q\times q}(P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*))\|_{\max} \\ &\leq C_3C_4|P_i(\boldsymbol{\eta}_0) - P_i(\boldsymbol{\eta}_d^*)| \\ &\leq C_3C_4C_5\|\boldsymbol{\eta}_0 - \boldsymbol{\eta}_d^*\| \end{aligned} \tag{3.61}$$

for all $i \in \mathbb{N}$ with the maximum-norm $\| \bullet \|_{\max}$, the $q \times q$ matrix matrix of ones $\mathbf{1}_{q\times q}$

$$C_3 := \sup_{(i,j,k,\boldsymbol{\eta})\in\mathbb{N}\times\{1,\dots,q\}^2\times B_\delta(\boldsymbol{\eta}_0)} \left|\frac{\partial^2\lambda_i(\boldsymbol{\eta})}{\partial\eta_j\partial\eta_k}\right|.$$

Next, for each $(j,k) \in \{1,\dots,q\}^2$ and $i \in \mathbb{N}$ we get

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\frac{\partial^2\lambda_i(\boldsymbol{\eta}_0)}{\partial\eta_j\partial\eta_k}Y_i\right) = \frac{\partial^2\lambda_i(\boldsymbol{\eta}_0)}{\partial\eta_j\partial\eta_k}P_i(\boldsymbol{\eta}_0)$$

and

$$\mathsf{Var}_{\boldsymbol{\eta}_0}\left(\frac{\partial^2\lambda_i(\boldsymbol{\eta}_0)}{\partial\eta_j\partial\eta_k}Y_i\right) \leq C_3^2\mathsf{Var}_{\boldsymbol{\eta}_0}(Y_i) = C_3^2P_i(\boldsymbol{\eta}_0)(1 - P_i(\boldsymbol{\eta}_0)) \leq \frac{C_3^2}{4}.$$

Therefore,

$$\sum_{i=1}^{\infty}\frac{\mathsf{Var}\left(\frac{\partial^2\lambda_i(\boldsymbol{\eta}_0)}{\partial\eta_j\partial\eta_k}Y_i\right)}{i^2} \leq \frac{C_3^2\pi^2}{24} < \infty.$$

By applying the strong law of large numbers, we obtain

$$\frac{1}{d}\sum_{i=1}^{d}\left(\frac{\partial^2\lambda_i(\boldsymbol{\eta}_0)}{\partial\eta_j\partial\eta_k}(Y_i - P_i(\boldsymbol{\eta}_0))\right) \overset{\mathsf{P}_{\boldsymbol{\eta}_0}\text{-a.s.}}{\longrightarrow} 0.$$

Especially, each component of $\frac{1}{d}\sum_{i=1}^{d}\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)(Y_i - P_i(\boldsymbol{\eta}_0))$ is almost surely converging to zero and hence

$$\left\|\frac{1}{d}\sum_{i=1}^{d}\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)(Y_i - P_i(\boldsymbol{\eta}_0))\right\| \overset{\mathsf{P}_{\boldsymbol{\eta}_0}\text{-a.s.}}{\longrightarrow} 0. \tag{3.62}$$

Combining (3.58), (3.59), (3.60), (3.61) and (3.62), it follows that

$$\begin{aligned}\|R_d(\boldsymbol{\eta})\| &\leq \frac{o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1) + \frac{1}{d}\sum_{i=1}^{d}(C_2 + C_3C_4C_5)\|\boldsymbol{\eta}_0 - \boldsymbol{\eta}_d^*\| + \frac{1}{d}\sum_{i=1}^{d}C_1\|\boldsymbol{\eta}_d^* - \hat{\boldsymbol{\eta}}_d\|}{C_0} \\ &= \frac{C_2 + C_3C_4C_5}{C_0}\|\boldsymbol{\eta}_0 - \boldsymbol{\eta}_d^*\| + \frac{C_1}{C_0}\|\boldsymbol{\eta}_d^* - \boldsymbol{\eta}_0 + \boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}_d\| + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1) \\ &\leq \frac{C_2 + C_3C_4C_5}{C_0}\|\boldsymbol{\eta}_0 - \boldsymbol{\eta}_d^*\| + \frac{C_1}{C_0}(\|\boldsymbol{\eta}_d^* - \boldsymbol{\eta}_0\| + \|\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}_d\|) + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1),\end{aligned}$$

where

$$\|\boldsymbol{\eta}_d^* - \boldsymbol{\eta}_0\| \leq \max(\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\|, \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|) \leq \|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| + \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|,$$

since $\boldsymbol{\eta}_d^* \in \{a\hat{\boldsymbol{\eta}}_d + (1-a)\boldsymbol{\eta} : a \in [0,1]\}$. For any $\varepsilon > 0$ and d sufficiently large, we can set

$$\delta = \frac{C_0}{2C_1 + C_2 + C_3C_4C_5}\frac{\varepsilon}{2}$$

to obtain (3.55).

It now remains to show (3.56). By assumption, $\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)$ is always symmetric and for $d > N$ positive definite. Now, let $d > N$ be fixed, then

$$-\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)(\mathrm{I}_q - R_d(\boldsymbol{\eta}))(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) = Q_d(\boldsymbol{\eta}) + \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) R_d(\boldsymbol{\eta})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d).$$

Therefore, it is sufficient to show that for each $\varepsilon > 0$, it exists a $\delta > 0$ such that

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\left|(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) R_d(\boldsymbol{\eta})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)\right| \leq -2\varepsilon Q_d(\boldsymbol{\eta})\right) = 1$$

For all $\boldsymbol{\eta} \in B_\delta(\boldsymbol{\eta}_0)$. Using Lemma 3.4.10 with $\boldsymbol{A} = \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)$, $\boldsymbol{B} = R_d(\boldsymbol{\eta})$ and $\boldsymbol{x} = \boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d$, we get

$$\left|\frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d) R_d(\boldsymbol{\eta})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)\right| \leq -\sqrt{\frac{\nu_{\max}(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}{\nu_{\min}(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}}\, \|R_d(\boldsymbol{\eta})\|\, Q_d(\boldsymbol{\eta}).$$

By assumption (CS5') and the consistency of the MLE $\hat{\boldsymbol{\eta}}_d$, there is a constant $C_1' > 0$ such that

$$\mathsf{P}_{\boldsymbol{\eta}_0}\left(\limsup_{d\to\infty} \sqrt{\frac{\nu_{\max}(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}{\nu_{\min}(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}} \leq C_1'\right) = \mathsf{P}_{\boldsymbol{\eta}_0}\left(\limsup_{d\to\infty} \sqrt{\frac{\nu_{\max}(1/d \cdot \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}{\nu_{\min}(1/d \cdot \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d))}} \leq C_1'\right) = 1.$$

We finish the proof by applying (3.55) with $\varepsilon = \varepsilon_1/C_1'$ for an arbitrary $\varepsilon_1 > 0$. □

Lemma 3.4.12

Let $\tilde{\Phi}_q(B) := \mathsf{P}(\boldsymbol{Z} \in B)$ for all $B \in \mathcal{B}^q$ with $\boldsymbol{Z} \sim \mathcal{N}_q(\boldsymbol{0}, \mathrm{I}_q)$. Consider a sequence $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ for a fixed $\boldsymbol{\eta}_0 \in \Theta$, for which conditions (CS1′ [i]), (CS2′)–(CS5′), and either (CS1′ [ii]) or (3.27) are satisfied. Then, the following holds.

1. *For every function f that is either absolutely bounded by a constant $c > 0$ or for which the integral $\int_{\mathbb{R}^q} f(\boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}$ exists and for every $\delta > 0$,*

$$\frac{\int_{\mathbb{R}^q \setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta}) P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det(\widehat{\boldsymbol{\Sigma}}_d^{1/2})} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0, \qquad d \to \infty. \tag{3.63}$$

2. *Consider a sequence $\{G_d\}_{d\in\mathbb{N}}$ with $G_d\colon (\Theta, \mathcal{B}(\Theta)) \to (\Theta, \mathcal{B}(\Theta))$ satisfying either*

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(G_d(B) \subset B_\delta(\boldsymbol{\eta}_0)\right) = 1, \quad \text{for all} \quad \delta > 0, \tag{3.64}$$

or

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(G_d(B) \supset B_\delta(\boldsymbol{\eta}_0)\right) = 1, \quad \textit{for one} \quad \delta > 0, \tag{3.65}$$

for all bounded $B \in \mathcal{B}^q$*. Then*

$$\frac{\int_{G_d(B)} P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det(\hat{\boldsymbol{\Sigma}}_d^{1/2})} - \tilde{\Phi}_q\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}(G_d(B) - \hat{\boldsymbol{\eta}}_d)\right)\mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$$

for $d \to \infty$*. In particular, in case of* (3.65)

$$\tilde{\Phi}_q\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}(G_d(B) - \hat{\boldsymbol{\eta}}_d)\right) = \mathsf{P}\left(\boldsymbol{Z} \in \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}(G_d(B) - \hat{\boldsymbol{\eta}}_d)\right) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 1.$$

holds. ◇

Proof.
This proof is partially based on arguments by Chang and Stout (1991, Theorem 3.1).

1. With regard to the left hand-side of (3.63), note that it can be written as

$$\begin{aligned}&\frac{\int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det(\widehat{\boldsymbol{\Sigma}}_d^{1/2})} \\ &\quad = \exp\left(\ell^{(d)}(\boldsymbol{\eta}_0 \,|\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\boldsymbol{Y}^{(d)})\right)\frac{H_d}{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)},\end{aligned} \tag{3.66}$$

in terms of the log-likelihood function, where

$$H_d \coloneqq \int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\exp\left(\ell^{(d)}(\boldsymbol{\eta} \,|\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\boldsymbol{Y}^{(d)})\right)\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}.$$

Since $\hat{\boldsymbol{\eta}}_d$ is a maximum of $\ell^{(d)}(\,\bullet\,|\boldsymbol{Y}^{(d)})$,

$$\exp\left(\ell^{(d)}(\boldsymbol{\eta}_0 \,|\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\hat{\boldsymbol{\eta}}_d \,|\boldsymbol{Y}^{(d)})\right) \le 1$$

always holds and therefore

$$\left|\frac{\int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)}\right| \le \left|\frac{H_d}{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)}\right|.$$

If $\mathcal{H}$ is improper and f is bounded by a constant, (3.27) directly implies

$$\left|\frac{H_d}{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)}\right| = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1).$$

In any other case the conditions imply that f is $\mathcal{H}$-integrable and hence $\mathsf{E}(|f \circ \boldsymbol{\eta}|) < \infty$. Lemma 3.4.9 implies that there is a $c(\delta) < 0$ such that

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}\in\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} \frac{1}{d}\left(\ell^{(d)}(\boldsymbol{\eta} \,|\boldsymbol{Y}^{(d)}) - \ell^{(d)}(\boldsymbol{\eta}_0 \,|\boldsymbol{Y}^{(d)})\right) < c(\delta)\right) = 1. \tag{3.67}$$

Further,

$$\begin{aligned}\left|\int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\exp(dc(\delta))\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\right| &= \exp(dc(\delta))\left|\int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\right| \\ &\le \exp(dc(\delta))\int_{\mathbb{R}^q\setminus B_\delta(\boldsymbol{\eta}_0)} \big|f(\boldsymbol{\eta})\big|\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} \\ &\le \exp(dc(\delta))\mathsf{E}(|f\circ\boldsymbol{\eta}|). \end{aligned} \tag{3.68}$$

Combining (3.67) and (3.68) results in

$$\lim_{d\to\infty}\mathsf{P}_{\boldsymbol{\eta}_0}\Big(|H_d|\le\exp(dc(\delta))\mathsf{E}(|f\circ\boldsymbol{\eta}|)\Big)=1.$$

Condition (CS2') implies that all eigenvalues $\nu_i(d^{-1}\mathcal{I}^{(d)}(\boldsymbol{\eta}))$, $i=1,\dots,q$, of $d^{-1}\mathcal{I}^{(d)}(\boldsymbol{\eta})$ are bounded for all $\boldsymbol{\eta}$ in an arbitrary compact subset of Θ and all $d\in\mathbb{N}$. By utilizing this fact and the consistency of $\hat{\boldsymbol{\eta}}_d$ for $d\to\infty$ we then get

$$\begin{aligned}\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)^{-1} &= \sqrt{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{-1}\right)} = \sqrt{d^q\det\left(\frac{1}{d}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)\right)} \\ &= d^{q/2}\sqrt{\prod_{i=1}^q \nu_i\left(\frac{1}{d}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)\right)} = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(d^{q/2}\right).\end{aligned}$$

The fact that $d\mapsto\exp(dc(\delta))$ is decreasing faster than any polynomial implies

$$\frac{\mathsf{E}(|f\circ\boldsymbol{\eta}|)\exp(dc(\delta))}{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)}\xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}}0$$

and this finishes the proof of the first part.

2. Let $B\in\mathcal{B}^q$ be an arbitrary bounded Borel set and define for $\delta>0$ and $d\in\mathbb{N}$ the set

$$M_{\delta,d}:=B_\delta(\boldsymbol{\eta}_0)\cap G_d(B)$$

and the integral

$$H_d':=\int_{M_{\delta,d}} P^{(d)}(\boldsymbol{Y}^{(d)}\,|\,\boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}.$$

By the definition of $R_d(\boldsymbol{\eta})$ in Lemma 3.4.11, we get

$$\begin{aligned}&\frac{H_d'}{P^{(d)}(\boldsymbol{Y}^{(d)}\,|\,\hat{\boldsymbol{\eta}}_d)\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \\ &= \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{\det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)}\int_{M_{\delta,d}}\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\exp\left(-\frac{1}{2}(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}_d)^\intercal\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)(\mathrm{I}-R_d(\boldsymbol{\eta}))(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}_d)\right)\mathrm{d}\boldsymbol{\eta}.\end{aligned} \tag{3.69}$$

From (CS1′), i.e. the continuity of $\mathfrak{h}$ and $\mathfrak{h}(\boldsymbol{\eta}_0)>0$, follows that for every $\varepsilon_1>0$, there is a $\delta_1>0$ such that

$$1-\varepsilon_1\le\inf_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\le\inf_{\boldsymbol{\eta}\in M_{\delta_1,d}}\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\le\sup_{\boldsymbol{\eta}\in M_{\delta_1,d}}\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\le\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\le1+\varepsilon_1. \tag{3.70}$$

Furthermore, from (3.56) in Lemma 3.4.11 we get for any $\varepsilon_2 > 0$ and appropriate $\delta_2 = \delta_2(\varepsilon_2) > 0$:

$$\begin{aligned}
&(1 - o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)) \int_{M_{\delta_2,d}} \exp\left(-\frac{1+\varepsilon_2}{2} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \right) \mathrm{d}\boldsymbol{\eta} \\
&\leq \int_{M_{\delta_2,d}} \exp\left(-\frac{1}{2} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)(\mathrm{I} - R_d(\boldsymbol{\eta}_d^*))(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \right) \mathrm{d}\boldsymbol{\eta} \qquad (3.71) \\
&\leq (1 + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)) \int_{M_{\delta_2,d}} \exp\left(-\frac{1-\varepsilon_2}{2} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d)^\intercal \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \right) \mathrm{d}\boldsymbol{\eta}.
\end{aligned}$$

Next, the transformation theorem, (3.70) and (3.71) imply

$$\begin{aligned}
&\tilde{\Phi}_q\Big(\sqrt{1+\varepsilon_2}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2} \big(M_{\delta_2,d} - \hat{\boldsymbol{\eta}}_d\big) \Big) \mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} \frac{1-\varepsilon_1}{(1+\varepsilon_2)^{q/2}} \big(1 - o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\big) \\
&\leq \frac{H_d'}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\big(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\big)} \qquad (3.72) \\
&\leq \tilde{\Phi}_q\Big(\sqrt{1-\varepsilon_2}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2} \big(M_{\delta_2,d} - \hat{\boldsymbol{\eta}}_d\big) \Big) \mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} \frac{1+\varepsilon_1}{(1-\varepsilon_2)^{q/2}} \big(1 + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\big).
\end{aligned}$$

In the case of (3.64), $\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\big(G_d(B) = M_{\delta_2,d}\big) = 1$ holds. Selecting ε_1 and ε_2 arbitrarily small leads to

$$\frac{\int_{G_d(B)} P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d) \det(\hat{\boldsymbol{\Sigma}}_d^{1/2})} = \tilde{\Phi}_q(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2}(G_d(B) - \hat{\boldsymbol{\eta}}_d)) \mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1).$$

In the case of equation (3.65), we get for each $\delta_2 < \delta$: $\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\big(B_{\delta_2}(\boldsymbol{\eta}_0) = M_{\delta_2,d}\big) = 1$. Condition (CS5′) implies that[10]

$$\sqrt{1+\varepsilon_2}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2} \big(B_{\delta_2}(\boldsymbol{\eta}_0) - \hat{\boldsymbol{\eta}}_d\big) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathbb{R}^d,$$

which further implies

$$\tilde{\Phi}_q\Big(\sqrt{1+\varepsilon_2}\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2} \big(B_{\delta_2}(\boldsymbol{\eta}_0) - \hat{\boldsymbol{\eta}}_d\big) \Big) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 1.$$

Finally, the further valid selection of arbitrarily small $\varepsilon_1, \varepsilon_2 > 0$ in (3.72) and the application of Lemma 3.4.12(1.) to $f = \mathbb{1}_{G_d(B) \setminus B_{\delta_2}(\boldsymbol{\eta}_0)}$ completes the proof. □

Lemma 3.4.12(2.) and a utilization of the continuous mapping theorem with $\mathbb{R} \setminus \{0\} \to \mathbb{R}$, $x \mapsto x^{-1}$ directly imply to the following corollary by setting $G_d(B) := \mathbb{R}^q$, $d \in \mathbb{N}$.

Corollary 3.4.13

Suppose a sequence $\{Y_i\}_{i\in\mathbb{N}}$ for a fixed $\boldsymbol{\eta}_0 \in \Theta$, for which conditions (CS1′)–(CS5′) hold. Then

$$\left(\frac{P(\boldsymbol{Y}^{(d)})}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\big(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\big)} \right)^{-1} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \frac{1}{\mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2}}$$

for $d \to \infty$. ◇

[10] This convergence is defined as follows: Let $\{A_d\}_{d\in\mathbb{N}}$ be a sequence of random sets with $A_d \subset \mathbb{R}^q$. Then we say $A_d \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathbb{R}^q$, if for each $\varepsilon > 0$ and each compact $K \subset \mathbb{R}^q$, there exists an $N \in \mathbb{N}$ so that $\mathsf{P}_{\boldsymbol{\eta}_0}\big(K \subset A_d\big) > 1 - \varepsilon$, for all $d > N$.

Proof of Theorem 3.4.1.
This proof is partially based on arguments by Chang and Stout (1991, Theorem 3.1 and 3.3).[11]

Analogously to the proof of Lemma 3.4.11, we can assume without loss of generality that $\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1/2}$ exists. Set

$$G_d(B) \coloneqq \left\{ \widehat{\boldsymbol{\Sigma}}_d^{1/2} \boldsymbol{x} + \hat{\boldsymbol{\eta}}_d \,:\, \boldsymbol{x} \in B \right\} \equiv \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1/2} B + \hat{\boldsymbol{\eta}}_d, \quad B \in \mathcal{B}^q,\ d \in \mathbb{N}. \tag{3.73}$$

To show (3.25) for bounded B we utilize the reformulation

$$\begin{aligned} &\mathsf{P}\left(\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{1/2} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_d) \in B \,\middle|\, \boldsymbol{Y}^{(d)} \right) \\ &\quad = \frac{\int_{G_d(B)} P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta}) \,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \left(\frac{P(\boldsymbol{Y}^{(d)})}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \right)^{-1} \end{aligned} \tag{3.74}$$

for each $d \in \mathbb{N}$. Since $\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0$, we get from (CS5′) and the fact that $\|\boldsymbol{A}^{-1}\| = 1/\nu_{\min}(\boldsymbol{A})$ for any regular symmetric matrix $\boldsymbol{A}$:

$$\left\|\widehat{\boldsymbol{\Sigma}}_d\right\| = \left\|\mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d)^{-1}\right\| = \frac{1}{d} \left\| \left(\frac{1}{d} \mathcal{I}^{(d)}(\hat{\boldsymbol{\eta}}_d) \right)^{-1} \right\| = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(\frac{1}{d}\right).$$

This implies $\left\|\widehat{\boldsymbol{\Sigma}}_d\right\| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0$ and especially $\widehat{\boldsymbol{\Sigma}}_d \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathbf{0}$. Hence, for each $\delta_2 > 0$ and $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $\mathsf{P}_{\boldsymbol{\eta}_0}(G_d(B) \subset B_{\delta_2}(\boldsymbol{\eta}_0)) > 1 - \varepsilon$ for all $d > N$, i.e. condition (3.64) of Lemma 3.4.12(2.) is satisfied, where G_d is defined in (3.73). We can therefore apply Lemma 3.4.12(2.) and get

$$\frac{\int_{G_d(B)} P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta}) \,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \tilde{\Phi}_q(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)^{1/2}(G_d(B) - \hat{\boldsymbol{\eta}}_d)) \mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} \tag{3.75}$$

for $d \to \infty$. Further, we get

$$\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)^{1/2}\big(G_d(B) - \hat{\boldsymbol{\eta}}_d\big) = \mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)^{1/2}\big(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)^{-1/2} B + \hat{\boldsymbol{\eta}}_d - \hat{\boldsymbol{\eta}}_d\big) = B, \tag{3.76}$$

from the definition of G_d. Combining (3.74), (3.75), (3.76) and Corollary 3.4.13 results in (3.25) for bounded B.

For the first part it remains to show the case of unbounded $B \in \mathcal{B}^q$. In order to show this, we define the sequence of random probability measures $\{\Psi_d\}_{d\in\mathbb{N}}$ on $(\mathbb{R}^q, \mathcal{B}^q)$ by setting

$$\Psi_d(A) \coloneqq \int_A |\det(\mathcal{I}_d(\hat{\boldsymbol{\eta}}_d)^{1/2})|\, h(G_d^{-1}(\boldsymbol{\eta}) \,|\, \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}, \quad A \in \mathcal{B}^q, d \in \mathbb{N}.$$

Due to the transformation theorem,

$$\Psi_d(A) = \int_{G_d(A)} h(\boldsymbol{\eta} \,|\, \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}$$

[11] In the corresponding proof by Chang and Stout (1991) for the one dimensional case, they proposed the result for unbounded intervals. But in their proof, they did not consider this case. Our proof shows that their results are still valid.

holds for all $d \in \mathbb{N}$ and $A \in \mathcal{B}^q$. Notice that Ψ_d is the posterior probability distribution of the affine transformation $G_d^{-1}(\boldsymbol{\eta})$ of $\boldsymbol{\eta}$ given $\boldsymbol{Y}^{(d)}$. So, $\Psi_d(A)$ is well-defined and finite for each Borel set $A \in \mathcal{B}^q$ and $d \in \mathbb{N}$.

Now, let B be an unbounded Borel set, then it is always possible to decompose it into a sequence $\{B_m\}_{m\in\mathbb{N}}$ with B_m bounded, $\bigcup_{m\in\mathbb{N}} B_m = B$ and $B_m \cap B_M = \emptyset$ for all $m, M \in \mathbb{N}$ with $m \neq M$. We get from the definition of convergence in probability and our just given proof of (3.25) for bounded B

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}(|\Psi_d(B_m) - \tilde{\Phi}_q(B_m)| < \varepsilon) = 1, \quad \text{for all } m \in \mathbb{N}, \tag{3.77}$$

for any $\varepsilon > 0$. Especially, (3.77) also holds for $\varepsilon = \frac{6\varepsilon'}{\pi^2 m^2}$ with $\varepsilon' > 0$. Since the probability of each of these events is tending to one, this is also true for the (countable) intersection of these events. Thus, we get

$$\begin{aligned} |\Psi_d(B) - \tilde{\Phi}_q(B)| = \left| \sum_{m\in\mathbb{N}} \left(\Psi_d(B_m) - \tilde{\Phi}_q(B_m) \right) \right| &\leq \sum_{m\in\mathbb{N}} \left| \Psi_d(B_m) - \tilde{\Phi}_q(B_m) \right| \\ &< \sum_{m\in\mathbb{N}} \frac{6\varepsilon'}{\pi^2 m^2} = \varepsilon', \end{aligned}$$

with probability tending to one for $d \to \infty$ for any $\varepsilon' > 0$.

For the second part, let $B \in \mathcal{B}^q$ be chosen arbitrarily. We define

$$H_{d,\varepsilon}(\boldsymbol{\eta}') \coloneqq \mathsf{P}\left(\left| \Psi_d(B) - \tilde{\Phi}_q(B) \right| > \varepsilon \,\middle|\, \boldsymbol{\eta}_0 = \boldsymbol{\eta}' \right)$$

for all $d \in \mathbb{N}$, $\varepsilon > 0$ and $\boldsymbol{\eta}' \in \Theta$, where ϕ_q is the pdf of $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. Notice that $0 \leq H_{d,\varepsilon} \leq 1$ is true for all $d \in \mathbb{N}$ and $\varepsilon > 0$. Hence, for each $\varepsilon > 0$, $\boldsymbol{\eta} \mapsto 1$ dominates the sequence $\{H_{d,\varepsilon}\}_{d\in\mathbb{N}}$, while $\boldsymbol{\eta} \mapsto 1$ is always $\mathcal{G}$-integrable for a proper $\mathcal{G}$. An application of Lebesgue's theorem of dominated convergence to $\{H_{d,\varepsilon}\}_{d\in\mathbb{N}}$ results in

$$\lim_{d\to\infty} \mathsf{P}\left(\left| \Psi_d(B) - \tilde{\Phi}_q(B) \right| > \varepsilon \right) = \lim_{d\to\infty} \int_\Theta H_{d,\varepsilon} \,\mathrm{d}\mathcal{G} = \int_\Theta \lim_{d\to\infty} H_{d,\varepsilon} \,\mathrm{d}\mathcal{G} = 0, \tag{3.78}$$

because $\lim_{d\to\infty} H_{d,\varepsilon}(\boldsymbol{\eta}_0, \varepsilon) = 0$ for all $\varepsilon > 0$ and $\boldsymbol{\eta}_0 \in \mathbb{R}^q$ due to (3.25). Equation (3.78) directly implies (3.26) and finishes the proof. □

Proof of Theorem 3.4.2.
Similar to (3.74), we utilize the reformulation

$$\mathcal{H}(B|\boldsymbol{Y}^{(d)}) = \frac{\int_B P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta}) \,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \left(\frac{P(\boldsymbol{Y}^{(d)})}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \right)^{-1} \tag{3.79}$$

for an arbitrary $B \in \mathcal{B}^q$ with $\boldsymbol{\eta}_0 \notin \partial B$ and each $d \in \mathbb{N}$. If $\boldsymbol{\eta}_0 \in B$, then condition (3.65) of Lemma 3.4.12(2.) is satisfied with $G_d(B) \coloneqq B$ for all $d \in \mathbb{N}$. However, if $\boldsymbol{\eta}_0 \notin B$, then the condition of Lemma 3.4.12(1.) is satisfied with $f \coloneqq \mathbb{1}_B$. We get therefore

$$\frac{\int_B P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta}) \mathfrak{h}(\boldsymbol{\eta}) \,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \,|\, \hat{\boldsymbol{\eta}}_d) \det\left(\widehat{\boldsymbol{\Sigma}}_d^{1/2}\right)} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} \mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2} \mathbb{1}_B(\boldsymbol{\eta}_0) \tag{3.80}$$

for $d \to \infty$. The claim follows from a combination of (3.79), (3.80) and Corollary 3.4.13.□

Proof of Theorem 3.4.5.
In a first step, the existence of $\mathsf{E}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)})$ for all functions $f\colon \Theta \to \mathbb{R}$, which are continuous and for which the integral $\int_\Theta f(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}$ exists, will be proved. In a second step its consistency for $f(\boldsymbol{\eta}_0)$ will be discussed.

We get

$$\int_\Theta |f(\boldsymbol{\eta})|\mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{y}^{(d)}) = \frac{\int_\Theta |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{y}^{(d)})} \leq \frac{\int_\Theta |f(\boldsymbol{\eta})|\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{y}^{(d)})} < \infty$$

for each $d \in \mathbb{N}$ and for all $\boldsymbol{y}^{(d)} \in \{0,1\}^d$, because $P^{(d)}(\boldsymbol{y}^{(d)} \mid \boldsymbol{\eta}) \in (0,1)$, $P^{(d)}(\boldsymbol{y}^{(d)})$ is positive and independent of $\boldsymbol{\eta} \in \Theta$, and $\int_\Theta |f(\boldsymbol{\eta})|\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}$ exists if and only if $\int_\Theta f(\boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}$ exists. Hence, $\mathsf{E}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)})$ exists. Furthermore, it remains integrable for $d \to \infty$, as shown next. Notice that the last statement does not follow directly, because $P^{(d)}(\boldsymbol{y}^{(d)}) \to 0$ for any sequence $\{y_i\}_{i\in\mathbb{N}}$ and $d \to \infty$.

Similar to (3.74), we start with the representation

$$\begin{aligned}&\int_\Theta |f(\boldsymbol{\eta})|\,\mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})\\ &\quad= \frac{\int_\Theta |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det(\widehat{\boldsymbol{\Sigma}}_d)^{1/2}} \left(\frac{P^{(d)}(\boldsymbol{Y}^{(d)})}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det\left(\widehat{\boldsymbol{\Sigma}}_d\right)^{1/2}}\right)^{-1},\end{aligned} \tag{3.81}$$

for each $d \in \mathbb{N}$. We decompose for an arbitrary $\delta > 0$

$$\begin{aligned}\int_\Theta |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} &= \int_{B_\delta(\boldsymbol{\eta}_0)} |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\\ &\quad+ \int_{\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}.\end{aligned}$$

Due to the integrability of f with respect to $\mathcal{H}$, Lemma 3.4.12(1.) implies

$$\frac{\int_{\Theta\setminus B_\delta(\boldsymbol{\eta}_0)} |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det(\hat{\boldsymbol{\Sigma}})^{1/2}} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0, \qquad d \to \infty. \tag{3.82}$$

Since f is continuous, we get

$$\sup_{\boldsymbol{\eta}\in B_\delta(\boldsymbol{\eta}_0)} |f(\boldsymbol{\eta})| =: C_1 < \infty. \tag{3.83}$$

Hence,

$$\int_{B_\delta(\boldsymbol{\eta}_0)} |f(\boldsymbol{\eta})|P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} \leq C_1 \int_{B_\delta(\boldsymbol{\eta}_0)} P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}. \tag{3.84}$$

An application of Lemma 3.4.12(2.) on the right-hand side results in

$$C_1 \frac{\int_{B_\delta(\boldsymbol{\eta}_0)} P^{(d)}(\boldsymbol{Y}^{(d)} \mid \boldsymbol{\eta})\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}}{P^{(d)}(\boldsymbol{Y}^{(d)} \mid \hat{\boldsymbol{\eta}}_d)\det\left(\widehat{\boldsymbol{\Sigma}}_d\right)^{1/2}} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} C_1\mathfrak{h}(\boldsymbol{\eta}_0)(2\pi)^{q/2}, \tag{3.85}$$

because $G_d(B) \coloneqq B_\delta(\boldsymbol{\eta}_0)$, $d \in \mathbb{N}$, satisfies equation (3.65). By summing up (3.81)–(3.85) and applying Corollary 3.4.13 to the second factor on the right-hand side of (3.81) we obtain

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_\Theta |f(\boldsymbol{\eta})|\,\mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) < C_1\right) = 1.$$

Next, we compute the value of $\mathsf{E}(f \circ \boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})$ for $d \to \infty$. Since $\mathcal{H}(\,\bullet \mid \boldsymbol{Y}^{(d)})$ is always a probability measure,

$$\begin{aligned}
\left| \int_\Theta f(\boldsymbol{\eta})\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0) \right| &= \left| \int_\Theta f(\boldsymbol{\eta})\, \mathrm{d}\mathcal{H}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0) \mathcal{H}(\,\Theta \mid \boldsymbol{Y}^{(d)}) \right| \\
&= \Bigg| \int_{B_\delta(\boldsymbol{\eta}_0)} \big(f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0)\big)\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \\
&\quad + \int_{\Theta \setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0) \mathcal{H}(\,\Theta \setminus B_\delta(\boldsymbol{\eta}_0) \mid \boldsymbol{Y}^{(d)}) \Bigg| \\
&\leq \left| \int_{B_\delta(\boldsymbol{\eta}_0)} \big(f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0)\big)\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \right| + \left| \int_{\Theta \setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \right| \\
&\quad + |f(\boldsymbol{\eta}_0)| \cdot \mathcal{H}(\,\Theta \setminus B_\delta(\boldsymbol{\eta}_0) \mid \boldsymbol{Y}^{(d)})
\end{aligned}$$

holds for each $\delta > 0$. Using a similar representation to (3.81) and applying (3.82) and Corollary 3.4.13 to it, leads to

$$\left| \int_{\Theta \setminus B_\delta(\boldsymbol{\eta}_0)} f(\boldsymbol{\eta})\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \right| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0 \quad \text{for } d \to \infty.$$

Further, Theorem 3.4.2 implies

$$|f(\boldsymbol{\eta}_0)| \cdot \mathcal{H}(\,\Theta \setminus B_\delta(\boldsymbol{\eta}_0) \mid \boldsymbol{Y}^{(d)}) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0 \quad \text{for } d \to \infty.$$

Finally, since f is continuous at $\boldsymbol{\eta}_0$, for each $\varepsilon > 0$ there is a $\delta' > 0$ such that $|f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0)| < \varepsilon$ for all $\boldsymbol{\eta} \in B_{\delta'}(\boldsymbol{\eta})$. Therefore,

$$\begin{aligned}
\left| \int_{B_{\delta'}(\boldsymbol{\eta}_0)} \big(f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0)\big)\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \right| &\leq \int_{B_{\delta'}(\boldsymbol{\eta}_0)} \big|f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0)\big|\, \mathcal{H}(\mathrm{d}\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) \\
&\leq \varepsilon \cdot \mathcal{H}(B_{\delta'}(\boldsymbol{\eta}_0) \mid \boldsymbol{Y}^{(d)}) \leq \varepsilon
\end{aligned}$$

for each $\varepsilon > 0$. Since $\delta > 0$ was chosen arbitrarily in the decomposition of Θ, we get for each $\varepsilon > 0$

$$\begin{aligned}
\lim_{d \to \infty} \mathsf{P}_{\boldsymbol{\eta}_0}\Big(\big|\mathsf{E}(f \circ \boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0)\big| > \varepsilon\Big) &= 1 - \lim_{d \to \infty} \mathsf{P}_{\boldsymbol{\eta}_0}\Big(\big|\mathsf{E}(f \circ \boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0)\big| \leq \varepsilon\Big) \\
&= 0.
\end{aligned}$$

which is what we had to show for the first part. The second part follows directly by considering the mappings $\boldsymbol{\eta} \mapsto \eta_j$, $j \in \{1, \ldots, q\}$, in the first part, which are continuous and by assumption integrable. □

3.5 Proper Centering Theory

3.5.1 Regularity Conditions

Throughout this section, we assume that $\{Y_i\}_{i \in \mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta})$, i.e. $\{Y_i\}_{i \in \mathbb{N}}$ are Bernoulli random variables fulfilling (3.1) and (3.2), and that the true latent vector $\boldsymbol{\eta}_0$ lies in the interior of the parameter space, i.e. $\boldsymbol{\eta}_0 \in \Theta \setminus \partial\Theta$. The asymptotic results of Subsection 3.5.2 rely on the following regularity conditions.

(PCT1) [i] Θ is compact, convex and has a non-empty interior.

[ii] $\mathcal{H}$ is proper and its pdf $\mathfrak{h}$ is Lipschitz continuous and positive on Θ.

(PCT2) Each $\lambda_i \colon \Theta \to \mathbb{R}$, $i \in \mathbb{N}$, is thrice continuously differentiable with

$$\sup\left\{ |\lambda_i(\boldsymbol{\eta})| \,:\, i \in \mathbb{N}, \boldsymbol{\eta} \in \Theta \right\} < \infty,$$

$$\sup\left\{ \left| \frac{\partial \lambda_i(\boldsymbol{\eta})}{\partial \eta_k} \right| \,:\, i \in \mathbb{N}, 1 \leq k \leq q, \boldsymbol{\eta} \in \Theta \right\} < \infty,$$

$$\sup\left\{ \left| \frac{\partial^2 \lambda_i(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_g} \right| \,:\, i \in \mathbb{N}, 1 \leq k, g \leq q, \boldsymbol{\eta} \in \Theta \right\} < \infty,$$

$$\sup\left\{ \left| \frac{\partial^3 \lambda_i(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_g \partial \eta_u} \right| \,:\, i \in \mathbb{N}, 1 \leq k, g, u \leq q, \boldsymbol{\eta} \in \Theta \right\} < \infty.$$

(PCT3) It holds

$$\inf_{\{\boldsymbol{\eta}_i\}_{i\in\mathbb{N}} \in \Theta_{\mathcal{L}}} \liminf_{d\to\infty} \nu_{\min}\left(\frac{1}{d} \sum_{i=1}^{d} \nabla \lambda_i(\boldsymbol{\eta}_i) \nabla \lambda_i(\boldsymbol{\eta}_i)^\intercal \right) > 0,$$

where $\nu_{\min}$ denotes the smallest eigenvalue and

$$\Theta_{\mathcal{L}} \coloneqq \left\{ \{\boldsymbol{\eta}_i\}_{i\in\mathbb{N}} \,:\, \boldsymbol{\eta}_i = c_i \boldsymbol{\eta}_\partial + (1-c_i)\boldsymbol{\eta}_\circ,\, c_i \in [0,1],\, i \in \mathbb{N},\, \boldsymbol{\eta}_\partial \in \partial\Theta,\, \boldsymbol{\eta}_\circ \in \Theta \setminus \partial\Theta \right\}.$$

When comparing these conditions with conditions (CS1′)–(CS5′) of the semiproper centering theory, we notice in particular the restriction to compact Θ in (PCT1) in comparison to (CS1′). But, as Chang and Stout (1993, p. 51) noted while discussing the regularity conditions for their posterior asymptotic theory of unidimensional latent traits:[12]

> "Although Θ could be $(-\infty, \infty)$, we always assume without loss of generality that θ_0 is contained in a finite interval (e.g., $[-a, a]$ for some fixed $a > 0$). This is not restrictive from the psychometric viewpoint, taking [V]ar$(\theta) = 1$ to fix the scale unit indeterminacy, because the same educational decision is made about people with $\theta = 5$ and people with $\theta = 25$; moreover the proportion of people with $|\theta| \geq 5$ is negligible. Further, the amount of test information available to distinguish $\theta = 5$ from $\theta = 25$ is also negligible. Thus, assuming $-5 \leq \theta \leq 5$, say, does no practical damage in all realistic applications."

Thus, from a Bayesian point of view, it is consequent to set Θ to be compact. Choosing a prior distribution like a multivariate standard normal distribution with restriction to an arbitrary compact set (e.g. $\overline{B}_5(\mathbf{0})$) is valid and it is practically identical to the more common distributions with unbounded support.

The condition (PCT1[ii]) is stricter than the corresponding condition (CS1'[ii]) but the Lipschitz continuity is satisfied by all commonly applied models. Under the restriction to compact Θ, (PCT2) is equivalent to (CS2') and (CS4').

[12] Notice that Chang and Stout (1993) denote the latent variable by θ, which we denote by $\boldsymbol{\eta}$. To increase the readability in this quote, we do not change the notation.

While (PCT3) is generally stricter than (CS5'), it ensures the identifiability of $\boldsymbol{\eta}_0 \in \Theta$ (i.e. (CS3')), as we will see in the proof. Under the restriction to compact Θ and given (PCT2), the conditions (PCT3) and (CS5') are equivalent for all EGLLVM for binary data. In Subsection 3.5.4, we will illustrate that (PCT3) is realistic for other models, too.

Overall, the proper centering theory can be applied to many MIRT models under the slight restriction of a compact support of $\mathcal{H}$. Moreover, the PCT is nested to the semiproper centering theory in the following sense.

Proposition 3.5.1
The set of conditions (PCT1)–(PCT3) *implies the set of conditions* (CS1′)–(CS5′). ◇

3.5.2 Main Results

Our main contribution in this section is the strengthening of results of the semiproper centering theory in Subsection 3.4.2 under the conditions (PCT1)–(PCT3) with respect to the following:

1. Strengthening of Theorem 3.4.1 to proper centeringing and inclusion of Bayes estimators for the centering of the posterior (cf. Definition 3.3.1).

2. Strengthening of Theorem 3.4.2 to strong consistency of the posterior distribution.

3. Strengthening of Theorem 3.4.3 to the asymptotic normality and efficiency of the MLE.

4. Strengthening of Theorem 3.4.5 to
   a) the asymptotic normality and efficiency of $\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})$,
   b) the convergence rate of $\mathsf{E}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)})$ for appropriate f with the same order as $\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)})$.

As a by-product, we obtain the asymptotic normality and efficiency of all Bayes estimators of $\boldsymbol{\eta}_0$ in the special case of $q = 1$ with respect to the loss functions considered by GGS. These new results constitute the proper centering theory together with a proof of the fact that the PCT is nested to the SCT in the sense of Proposition 3.5.1.

The asymptotic normality of estimators of $\boldsymbol{\eta}_0$ was indicated by Sinharay (2015) to be an open issue to prove in MIRT settings without the condition of test lengthening by *strictly parallel forms*. With the proper centering theory, we provide such a result. The proofs of all results in this section along with some preliminary results are given in Subsection 3.5.5.

First of all, we prove that the MLE and appropriate Bayes estimators are proper centering (cf. Definition 3.3.1). In the proper centering theory, this is closely related to the asymptotic efficiency of these estimators and, therefore, formulated jointly in the following theorem.

Theorem 3.5.2 (Proper centering and efficiency of the MLE and Bayes estimators)
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for a* $\boldsymbol{\eta}_0 \in \Theta \setminus \partial\Theta$, *that conditions* (PCT1)–(PCT3) *are satisfied and let* $\boldsymbol{Z} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. *Then the maximum likelihood estimator and each Bayes estimator with respect to a loss function satisfying the conditions stated by GGS exists for all* $d \in \mathbb{N}$. *Suppose* $\{\tilde{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$, *with* $\tilde{\boldsymbol{\eta}}_d \colon (\Omega, \mathcal{A}, \mathsf{P}) \to (\Theta, \mathcal{B}(\Theta))$ *for all* $d \in \mathbb{N}$, *is such a sequence of maximum likelihood or Bayes estimators. Then there is a constant* $\boldsymbol{c} \in \mathbb{R}^q$ *such*

that

$$\sup_{A\in\mathcal{B}^q}\left|\mathsf{P}\left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}_d)\in A\,\middle|\,\boldsymbol{Y}^{(d)}\right)-\mathsf{P}\Big(\boldsymbol{Z}\in(A-\boldsymbol{c})\Big)\right|\xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}}0,$$
$$\sup_{A\in\mathcal{B}^q}\left|\mathsf{P}\left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}_d)\in A\,\middle|\,\boldsymbol{Y}^{(d)}\right)-\mathsf{P}\Big(\boldsymbol{Z}\in(A-\boldsymbol{c})\Big)\right|\xrightarrow{\mathsf{P}}0,$$
$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\tilde{\boldsymbol{\eta}}_d-\boldsymbol{\eta}_0)-\boldsymbol{c}\xrightarrow{\mathcal{D}}\boldsymbol{Z},$$

for $d\to\infty$. In particular, if $\{\tilde{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ is the sequence of MLEs or EAPs, then $\boldsymbol{c}=\boldsymbol{0}$. If moreover $q=1$, then $\boldsymbol{c}=\boldsymbol{0}=0$ for all considered estimators. If $\boldsymbol{c}=\boldsymbol{0}$, then $\tilde{\boldsymbol{\eta}}_d-\boldsymbol{\eta}_0=o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$ and the normalizing factor $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}$ can be replaced by $\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}$. In particular, if $\boldsymbol{c}=\boldsymbol{0}$, then the considered sequence is a proper centering (cf. Definition 3.3.1) and a consistent and asymptotic efficient sequence of estimators for $\boldsymbol{\eta}_0$. The sequence of MLEs or EAPs and, if $q=1$, the sequences of all considered estimators are asymptotically stochastic equivalent in the sense that they converge to the same random variable/vector. ◇

Note that the theorem does not state that $\boldsymbol{c}\neq\boldsymbol{0}$ for any Bayes estimator for $q>1$, rather it is left open. If a specific MIRT model additionally satisfies condition (GGS1'), i.e. equation (3.20), then $\boldsymbol{c}=\boldsymbol{0}$ in all cases.

We can also consider the estimation of $f(\boldsymbol{\eta}_0)$ using the posterior expected value similar to Theorem 3.4.5. An important corollary of this theorem is the following, indicating the same convergence rate as for estimating just $\boldsymbol{\eta}_0$.

Corollary 3.5.3
Suppose that $\{Y_i\}_{i\in\mathbb{N}}\sim\mathcal{P}(\boldsymbol{\eta}_0)$ for $\boldsymbol{\eta}_0\in\Theta\setminus\partial\Theta$ and that conditions (PCT1)–(PCT3) *are satisfied. Further assume that $f\colon\Theta\to\mathbb{R}$ is twice continuously differentiable with bounded second order partial derivatives on Θ. Then*

$$\mathsf{E}_{\boldsymbol{\eta}_0}(f(\boldsymbol{\eta})\mid\boldsymbol{Y}^{(d)})-f(\boldsymbol{\eta}_0)=\mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(\frac{1}{\sqrt{d}}\right)\tag{3.86}$$

for $d\to\infty$. If $q=1$, then with $\eta:=\boldsymbol{\eta}$ and $\eta_0:=\boldsymbol{\eta}_0$

$$\begin{aligned}&\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(f(\eta)\mid\boldsymbol{Y}^{(d)})-f(\eta_0)\Big)\\&\quad=f'(\eta_0)\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(\eta\mid\boldsymbol{Y}^{(d)})-\eta_0\Big)+\mathcal{O}_{\mathsf{P}_{\eta_0}}\left(\frac{1}{\sqrt{d}}\right)\end{aligned}\tag{3.87}$$

holds for $d\to\infty$, where f' denotes the first derivative of f. If for $q=1$ moreover $f'(\eta_0)\neq0$, then

$$\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(f(\eta)\mid\boldsymbol{Y}^{(d)})-f(\eta_0)\Big)\xrightarrow{\mathcal{D}}Z\sim\mathcal{N}(0,f'(\eta_0)^2)\tag{3.88}$$

for $d\to\infty$. ◇

The existence of all expectations in the last corollary is always ensured by the fact that Θ must be compact in the proper centering theory.

The next theorem formulates the strengthening of Theorem 3.4.2.

Theorem 3.5.4 (Strong consistency of the posterior distribution)
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Then*

$$\mathsf{P}(\boldsymbol{\eta} \in B \mid \boldsymbol{Y}^{(d)}) \to \delta_{\boldsymbol{\eta}_0}(B) := \begin{cases} 1, & \boldsymbol{\eta}_0 \in B \\ 0, & \boldsymbol{\eta}_0 \notin B \end{cases}, \qquad \mathsf{P}_{\boldsymbol{\eta}_0}\text{-a.s. for } d \to \infty$$

for all Borel-sets $B \in \mathcal{B}^q$ *with* $\boldsymbol{\eta}_0 \notin \partial B$. ◇

In some theoretical discussions it is necessary to consider the convergence of the moment generating function (MGF) of the posterior distribution. One important example in the context of IRT is the discussion of the Dutch Identity Conjecture by Zhang and Stout (1997, Proposition 1), who assumed the uniform convergence of the MGF of the posterior distribution to the MGF of a normal distribution. It is therefore important to note that the proper centering theory implies the convergence of the MGF of the conditional distribution of

$$\boldsymbol{\eta}^* := \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \qquad \text{conditional on } \boldsymbol{Y}^{(d)} \tag{3.89}$$

to the MGF of a multivariate normal distribution locally uniformly in probability. To the best of our knowledge, a result such as ours is generally new in the IRT literature and we expect that it will be helpful in future research.

Theorem 3.5.5
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Then the moment generating function of the distribution of* $\boldsymbol{\eta}^*$ *conditional on* $\boldsymbol{Y}^{(d)}$ *exists for all* $d \in \mathbb{N}$ *on* $\mathbb{R}^q$. *Additionally*

$$\sup_{\boldsymbol{t}\in K} \left| \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\mathsf{T}\boldsymbol{\eta}^*)\, h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\, \mathrm{d}\boldsymbol{\eta}^* - \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\mathsf{T}\boldsymbol{\eta}^*)\, \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q)\, \mathrm{d}\boldsymbol{\eta}^* \right| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0$$

for each bounded set $K \subset \mathbb{R}^q$, *where* $h^*(\,\bullet \mid \boldsymbol{Y}^{(d)})$ *is the pdf of* $\boldsymbol{\eta}^* \mid \boldsymbol{Y}$ *of equation* (3.89) *and*

$$\Delta_d := \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \sum_{i=1}^{d} (Y_i - P_i(\boldsymbol{\eta}_0)) \nabla \lambda_i(\boldsymbol{\eta}_0) \tag{3.90}$$

for all $d > D$, *where* $D \in \mathbb{N}$ *is the index so that* $\mathcal{I}^{(d')}(\boldsymbol{\eta}_0)$ *is regular for all* $d' > D$. ◇

Note that the limiting normal distribution in the last theorem is not constant because the distribution of Δ_d converges to $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. This might be surprising at first glance but it is because the posterior distribution of $\boldsymbol{\eta}$ is here centered at $\boldsymbol{\eta}_0$ and not at any estimator which is asymptotically normal distributed. Hence, this is a purely distributional result and absolutely independent of the properties of any point estimator.

3.5.3 Summary, Discussion and Outlook

In this section, we proved the APN of LTs by proper centering of the MLE and various Bayes estimators under appropriate conditions that are stronger than the conditions (CS1')–(CS5') but are still fulfilled by a broad class of MIRT models for binary items. Furthermore, we provided results on the strong consistency of posterior distributions and the existence,

asymptotic normality and asymptotic efficiency of the MLE and the EAP estimators under the same conditions. For univariate LT, we also proved the existence, asymptotic normality and asymptotic efficiency of a large class of Bayes estimators for the LT. Additionally, we proved the locally uniform convergence of the MGF of the standardized posterior to the MGF of a multivariate normal distribution.

In contrast to the results in Subsection 3.4.2, the results of PCT cannot simply be generalized to improper priors.

An important open issue in (M)IRT is the formally clear formulation of the Dutch Identity Conjecture and its proof. The PCT provides a strong toolkit which might allow the resolution of this issue in future research. Consider for example a univariate logit model and the following small proposition, which we derived using the PCT.

Proposition 3.5.6
Let $\{Y_i\}_{i\in\mathbb{N}_0} \sim \mathcal{P}(\eta_0)$ *with* $\lambda_i(\eta) = \alpha_{i0} + \alpha_{i1}\eta$, $\eta \in \Theta = [-c, c]$, $i \in \mathbb{N}_0$, *for some* $\alpha_{i0} \in (-c, c), \alpha_{i1} \in (0, c)$, $i \in \mathbb{N}_0$, *and arbitrary* $\eta_0 \in (-c, c)$, $c > 0$ *and a prior distribution* $\mathcal{H}$ *satisfying (PCT1[ii]). Then*

$$\frac{\operatorname{logit}\left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)})\right) - \lambda_0(\eta_0)}{\sigma_d} = \alpha_{01}\sigma_d \sum_{i=1}^{d} \alpha_{i1}(Y_i - P_i(\eta_0)) + o_{\mathsf{P}_{\eta_0}}(1) \tag{3.91}$$

for $\sigma_d = \mathcal{I}^{(d)}(\eta_0)^{-1/2}$ *and* $d \to \infty$. ◇

Equation (3.91) can be interpreted as the asymptotic linearity of $\operatorname{logit}\left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)})\right)$ in $\boldsymbol{Y}^{(d)}$ in the form

$$\operatorname{logit}\left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)})\right) \approx \left(\lambda_i(\eta_0) - \alpha_{01}\sigma_d^2 \sum_{i=1}^{d} \alpha_{i1} P_i(\eta_0)\right) + \alpha_{01}\sigma_d^2 \sum_{i=1}^{d} \alpha_{i1} Y_i.$$

This is not just an asymptotic justification of the heuristically formulated conditionally defined model approach of Subsection 1.2.3 but also an approach to tackle the Dutch Identity Conjecture in future research.

In Chapter 4 we will discuss the application of the PCT to derive a new MCEM algorithm for MMLE of logit models, which further illustrates the broad range of applicability of the PCT.

3.5.4 Example: Regularity Conditions for Model (3.21)

Consider again the multidimensional version of the model of Lee and Bolt (2018) with the specification of the example in Subsection 3.4.4. By considering any compact Θ, e.g. $\Theta = [-1, 1]^2$, the conditions (PCT1[i]) and (PCT2) follow from the same arguments as in Subsection 3.4.4. If $\mathcal{H}$ is the uniform distribution or a truncated normal distribution, (PCT1[ii]) is clear. Under the assumption of a regular further parameterization analogous to the first d items, the abstract condition (PCT3) is reasonable (compare the visualization of $\min_{(\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_d)\in\Theta^d} \nu_{\min} \frac{1}{d} \sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta}_i)\nabla\lambda_i(\boldsymbol{\eta}_i)^\intercal$ in Figure 3.3). The selection of the square $\Theta = [-1, 1]^2$ is made for convenience of creating the figure but the fact that the boundary points are -1 and 1 has no influence on the fact that the eigenvalues are positive. Overall, there is no reason to doubt the conditions and so the results of the PCT can be applied.

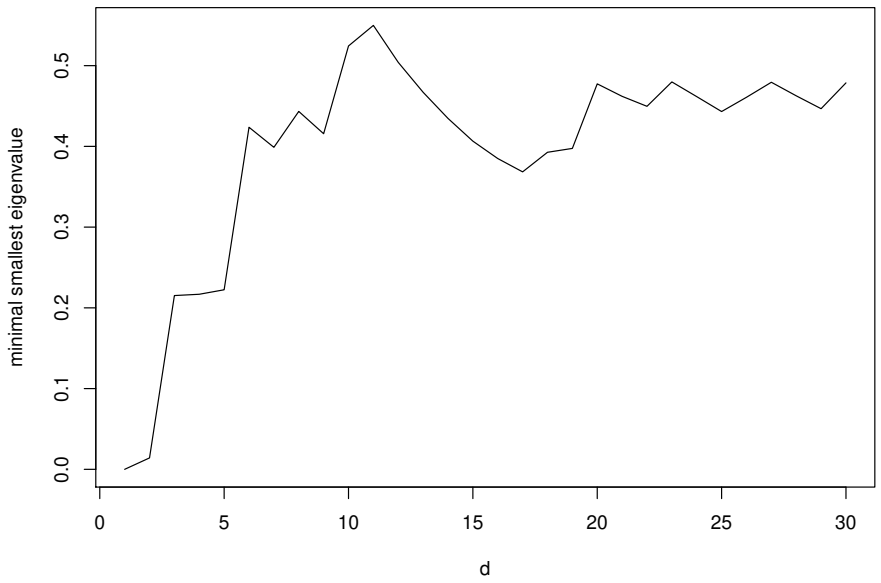

Figure 3.3: *Values of* $\min_{(\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_d)\in\Theta^d} \nu_{\min}\left(\frac{1}{d}\sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta}_i)\nabla\lambda_i(\boldsymbol{\eta}_i)^\intercal\right)$ *for* $\Theta = [-1,1]^2$.

3.5.5 Proofs

For the proofs of the proper centering theory, we have to start with the centering and scaling of the posterior using $\boldsymbol{\eta}_0$, i.e. estimator free. The necessary results to switch to quantities which are not dependent on $\boldsymbol{\eta}_0$ will be considered later. Hence, in the following we set $\boldsymbol{S}_d \coloneqq \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}$ in (3.19), i.e. $Z^{*(d)}$ is given by

$$Z^{*(d)}(\boldsymbol{\eta}^*) = \begin{cases} Z^{(d)}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*, \boldsymbol{\eta}_0), & \text{if} \quad \boldsymbol{\eta}^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0), \\ 0, & \text{otherwise.} \end{cases} \tag{3.92}$$

Lemma 3.5.7
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Then there is a* $B > 0$ *such that*

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*)^{1/2} - Z^{*(d)}(\boldsymbol{\eta}_2^*)^{1/2}\right|^2\right) \leq B\|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|^2, \qquad \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0),$$
$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*) - Z^{*(d)}(\boldsymbol{\eta}_2^*)\right|\right) \leq 2\sqrt{B}\|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|, \qquad \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0).$$

Further, there is a $D \in \mathbb{N}$ *and a* $\varepsilon > 0$ *such that*

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}^*)^{1/2}\right) \leq \exp(-\varepsilon\|\boldsymbol{\eta}^*\|^2), \qquad \boldsymbol{\eta}^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0) \tag{3.93}$$

for all $d > D$. *Additionally, there is a* $0 < \delta_0 < 1$ *such that for each* $0 < \delta < \delta_0$

$$\mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{\mathbb{R}^q} Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* \leq \mathfrak{h}(\boldsymbol{\eta}_0)\frac{(2\delta)^q}{4}\right) \leq 4(qB)^{1/2}\delta$$

holds with the same B *as in the first part.* ◇

To prove this lemma, we need a small technical device.

Lemma 3.5.8
For $x_1, \dots, x_d \in [0,1]$

$$1 - \prod_{i=1}^{d}(1 - x_i) \leq \sum_{i=1}^{d} x_i \tag{3.94}$$

holds. ◇

Proof.
Notice that the left hand side is always smaller or equal than one such that the result is trivial for $\sum_{i=1}^{d} x_i \geq 1$.

We prove the statement by induction. Inequality (3.94) is equivalent to

$$1 - \sum_{i=1}^{d} x_i \leq \prod_{i=1}^{d}(1 - x_i),$$

which is satisfied for $d = 1$. Hence, assume that (3.94) is satisfied for $d \in \mathbb{N}$, then we get

$$\begin{aligned}\prod_{i=1}^{d+1}(1 - x_i) = \left(\prod_{i=1}^{d}(1 - x_i)\right)(1 - x_{d+1}) &\geq \left(1 - \sum_{i=1}^{d} x_i\right)(1 - x_{d+1}) \\ &= 1 - \sum_{i=1}^{d+1} x_i + \left(\sum_{i=1}^{d} x_i\right) x_{d+1} \\ &\geq 1 - \sum_{i=1}^{d+1} x_i.\end{aligned}$$

□

Proof of Lemma 3.5.7.
The proof is based on arguments by Ghosal (1999, Lemma 2.3 and 2.4) and Ibragimov and Has'minskii (1981, Lemma I.5.1).

We fix $d \in \mathbb{N}$ sufficiently large such that $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)$ is regular, which is always possible due to condition (PCT2). Notice that

$$H^{(d)}(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*) \coloneqq \mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*)^{1/2} - Z^{*(d)}(\boldsymbol{\eta}_2^*)^{1/2}\right|^2\right) \tag{3.95}$$

equals the squared Hellinger distance[13] between the counting densities

$$P^{(d)}(\,\bullet \mid \boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_1^*)$$

and

$$P^{(d)}(\,\bullet \mid \boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_2^*)$$

[13] Let $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$, be an open and convex subset of $\mathbb{R}^k$. Consider a σ-finite measurable space $(S, \mathcal{S}, \mu)$ and a family of probability densities $\{f_\theta\}_{\theta\in\Theta}$ with respect to the measure μ. Then the Hellinger distance between f_{θ_1} and f_{θ_2} is defined as $H\colon \Theta^2 \to [0,1]$, $(\theta_1, \theta_2) \mapsto \sqrt{\frac{1}{2}\int_S \left|f_{\theta_1}^{1/2} - f_{\theta_2}^{1/2}\right|^2 \mathrm{d}\mu}$.

for all $\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$. Similar to $H^{(d)}$, we denote by $H_{i,d}(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)$ for all $\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$ the Hellinger distance between $\mathsf{P}(Y_i = \bullet \mid \boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_1^*)$ and $\mathsf{P}(Y_i = \bullet \mid \boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_2^*)$. Notice that due to the conditional independence of $Y_1, \ldots, Y_d$, $H^{(d)}$ denotes the Hellinger distance between the product measures of the distributions of $Y_1, \ldots, Y_d$ for $\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*$. In constrast is $H_{i,d}$ the Hellinger distance of the distributions for a single Y_i, $i = 1, \ldots, d$, $d \in \mathbb{N}$.

First, we get directly from Ibragimov and Has'minskii (1981, Section 7.3, page 81–82) for the general family of Bernoulli counting densities,

$$f\colon \{0,1\} \times \mathbb{R} \to (0,1), \qquad (y, \theta) \mapsto \exp\left(y\theta - \log(1 + \exp(\theta))\right) =: f_\theta(y),$$

the estimate

$$\frac{ah^2}{1+h^2} \leq H(\theta, \theta + h)^2 \leq Ah^2, \tag{3.96}$$

for all $\theta \in K$ on any compact set $K \subset \mathbb{R}$ and some constants $a = a(K)$, $A = A(K) > 0$, where $H(\theta, \theta + h)$ denotes the Hellinger distance between f_θ and $f_{\theta+h}$.

By condition (PCT2), there is always a $k \in \mathbb{N}$ such that

$$\overline{\bigcup_{i \in \mathbb{N}} \lambda_i(\Theta)} \subset [-k, k].$$

The compactness of $[-k, k]$ implies that (3.96) holds for all $\theta \in \overline{\bigcup_{i\in\mathbb{N}} \lambda_i(\Theta)}$. Hence, there are two constants $a = a([-k,k])$, $A = A([-k,k])$ so that

$$\frac{a}{1+(2k)^2}h^2 \leq \frac{ah^2}{1+h^2} \leq H_{i,d}(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)^2 \leq Ah^2,$$

for all $\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$ and each $i \in \{1, \ldots, d\}$ with

$$h = \left|\lambda_i\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_1^*\right) - \lambda_i\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}_2^*\right)\right|.$$

The multivariate mean value theorem implies that

$$h = \left|\nabla\lambda_i(\tilde{\boldsymbol{\eta}})^\intercal \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\left(\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\right)\right|,$$

for a

$$\tilde{\boldsymbol{\eta}} \in \left\{\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\left(c\boldsymbol{\eta}_1^* + (1-c)\boldsymbol{\eta}_2^*\right)\colon c \in [0,1]\right\},$$

where $\nabla\lambda_i(\tilde{\boldsymbol{\eta}})$ is well defined because $\tilde{\boldsymbol{\eta}} \in \Theta$ due to the convexity of Θ, compare (PCT1).

Next, the common properties of the Hellinger distance for product measures and an application of Lemma 3.5.8 imply that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{\eta}_1^*)^{1/2} - Z^{*(d)}(\boldsymbol{\eta}_2^*)^{1/2}\right|^2\right) &= H^{(d)}(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)^2 \\
&= 1 - \prod_{i=1}^{d}\left(1 - H_i(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_1^*)^2\right) \\
&\leq \sum_{i=1}^{d} H_i(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_1^*)^2 \\
&\leq Ad \sup_{(i,\boldsymbol{\eta}) \in \mathbb{N}\times\Theta}\left\{\|\nabla\lambda_i(\boldsymbol{\eta})\|^2\right\} \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\|^2 \\
&\quad \times \|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|^2.
\end{aligned}$$

Since $\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\|^2 = \mathcal{O}(d^{-1})$ due to condition (PCT3), condition (PCT2) ensures that

$$B \coloneqq \sup_{d\in\mathbb{N}} \left\{ Ad \sup_{(i,\boldsymbol{\eta})\in\mathbb{N}\times\Theta} \left\{ \|\nabla\lambda_i(\boldsymbol{\eta})\|^2 \right\} \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\|^2 \right\} < \infty,$$

which implies the first estimate. Set $X = Z^{*(d)}(\boldsymbol{w})$ and $Y = Z^{*(d)}(\boldsymbol{w}')$ for some $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$. As $X, Y \geq 0$ P–almost surely, Hölder's inequality leads to

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\eta}_0}|X - Y| &= \mathsf{E}_{\boldsymbol{\eta}_0}\left|(\sqrt{X} - \sqrt{Y})(\sqrt{X} + \sqrt{Y})\right| \\ &\leq \left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{X} - \sqrt{Y}\right|^2\right)^{1/2} \left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{X} + \sqrt{Y}\right|^2\right)^{1/2}.\end{aligned}$$

Next, Minkowski's inequality implies

$$\begin{aligned}\left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{X} + \sqrt{Y}\right|^2\right)^{1/2} &\leq \left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{X}\right|^2\right)^{1/2} + \left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{Y}\right|^2\right)^{1/2} \\ &= \left(\mathsf{E}_{\boldsymbol{\eta}_0}(X)\right)^{1/2} + \left(\mathsf{E}_{\boldsymbol{\eta}_0}(Y)\right)^{1/2} = 2,\end{aligned}$$

as $\mathsf{E}_{\boldsymbol{\eta}_0}(Z^{*(d)}(\boldsymbol{w})) = 1$ for all $\boldsymbol{w} \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$. Additionally, the first part of the lemma implies

$$\left(\mathsf{E}_{\boldsymbol{\eta}_0}\left|\sqrt{X} - \sqrt{Y}\right|^2\right)^{1/2} \leq \sqrt{B}\|\boldsymbol{w} - \boldsymbol{w}'\|.$$

Further, for all $d \in \mathbb{N}$ with $\varepsilon_1 = a(1 + (2k))^{-2}$ and an $\boldsymbol{\eta}^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$ for some $\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_d$ between $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}^*$ we get

$$\begin{aligned}\sum_{i=1}^{d} H_i(\boldsymbol{\eta}^*, \mathbf{0})^2 &\geq \varepsilon_1 \sum_{i=1}^{d} \boldsymbol{\eta}^{*\intercal}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\nabla\lambda_i(\boldsymbol{\eta}_i)\nabla\lambda_i(\boldsymbol{\eta}_i)^\intercal\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^* \\ &= \varepsilon_1\boldsymbol{\eta}^{*\intercal}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \sum_{i=1}^{d} \left(\nabla\lambda_i(\tilde{\boldsymbol{\eta}}_i)\nabla\lambda_i(\tilde{\boldsymbol{\eta}}_i)^\intercal\right)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*.\end{aligned}$$

We define the set of all sequences on a line between $\boldsymbol{\eta}_0$ and an arbitrary boundary point by

$$\Theta_{\mathcal{L}_0} \coloneqq \left\{ \{\boldsymbol{\eta}_i\}_{i\in\mathbb{N}} \colon \boldsymbol{\eta}_i = c_i\boldsymbol{\eta}_\partial + (1 - c_i)\boldsymbol{\eta}_0,\ c_i \in [0,1],\ i \in \mathbb{N},\ \boldsymbol{\eta}_\partial \in \partial\Theta \right\}$$

Recall, if $\boldsymbol{A}$, $\boldsymbol{B}$ are symmetric positive definite matrices, then there is a $\boldsymbol{B}^{1/2}$ so that $\boldsymbol{ABA} = (\boldsymbol{AB}^{1/2})(\boldsymbol{AB}^{1/2})^\intercal$, which is itself symmetric and positive definite. Condition (PCT3) implies

$$\begin{aligned}&\liminf_{d\to\infty} |\nu|_{\min}\left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \sum_{i=1}^{d} \nabla\lambda_i(\tilde{\boldsymbol{\eta}}_i)\nabla\lambda_i(\tilde{\boldsymbol{\eta}}_i)^\intercal\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right) \\ &\quad> \inf_{\{\boldsymbol{\eta}_i\}_{i\in\mathbb{N}}\in\Theta_{\mathcal{L}_0}} \liminf_{d\to\infty} |\nu|_{\min}\left(d^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}d^{-1}\sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta}_i)\nabla\lambda_i(\boldsymbol{\eta}_i)^\intercal d^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right) \\ &\quad\geq \inf_{\{\boldsymbol{\eta}_i\}_{i\in\mathbb{N}}\in\Theta_{\mathcal{L}}} \liminf_{d\to\infty} |\nu|_{\min}\left(d^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}d^{-1}\sum_{i=1}^{d} \nabla\lambda_i(\boldsymbol{\eta}_i)\nabla\lambda_i(\boldsymbol{\eta}_i)^\intercal d^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right) \\ &\quad= \frac{c}{\varepsilon_1} > 0,\end{aligned}$$

for some $c > 0$, where $\Theta_{\mathcal{L}}$ is the sequence space on all lines between an interior and a boundary point that is defined in condition (PCT3). Hence, there is an $\varepsilon_2 > 0$ and a $D > 0$ such that

$$\sum_{i=1}^{d} H_i(\boldsymbol{\eta}^*, \mathbf{0})^2 \geq \varepsilon_2 \|\boldsymbol{\eta}^*\|^2$$

for all $d > D$. Using again the properties of the Hellinger distance leads to the second estimate with

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}^*)^{1/2}\right) = 1 - H^{(d)}(\boldsymbol{\eta}^*, \mathbf{0})^2 &= \prod_{i=1}^{d} \left(1 - H_i(\boldsymbol{\eta}^*, \mathbf{0})^2\right) \\
&= \exp\left(-\sum_{i=1}^{d} \log\left(\frac{1}{1 - H_i(\boldsymbol{\eta}^*, \mathbf{0})^2}\right)\right) \\
&\leq \exp\left(-\sum_{i=1}^{d} H_i(\boldsymbol{\eta}^*, \mathbf{0})^2\right) \\
&\leq \exp\left(-\varepsilon_2 \|\boldsymbol{\eta}^*\|^2\right).
\end{aligned}$$

It remains to show the last estimate. Condition (PCT1) implies

$$\log\left(\frac{\mathfrak{h}(\boldsymbol{\eta})}{\mathfrak{h}(\boldsymbol{\eta}_0)}\right) \leq K \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|$$

for all $\boldsymbol{\eta} \in \Theta$ and for some $K > 0$. In particular,

$$\begin{aligned}
\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*\right) &\geq \mathfrak{h}(\boldsymbol{\eta}_0) \exp\left(-K \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*\|\right) \\
&\geq \mathfrak{h}(\boldsymbol{\eta}_0) \exp\left(-K \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\| \, \|\boldsymbol{\eta}^*\|\right)
\end{aligned}$$

with $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*$ for any $\boldsymbol{\eta}^* \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$. Notice that $\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\| \to 0$ for $d \to \infty$ due to (PCT3). If we now restrict $\boldsymbol{\eta}^*$ to be an element of $[-\delta, \delta]^q$ for an $\delta \in (0, 1)$, then we obtain $K\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\| \leq \log(2)$ and particularly

$$\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*\right) \geq \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{2}.$$

This implies that there is a $\delta_0 > 0$ and a $D > 0$ such that

$$\begin{aligned}
&\int_{\mathbb{R}^q} Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* \\
&\quad \geq \int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w}\right)\,\mathrm{d}\boldsymbol{w} > \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{2}\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}
\end{aligned}$$

for all $\delta < \delta_0$ and $d > D$. Since $Z^{*(d)}(\mathbf{0}) = 1$, Markov's inequality and Tonelli's Theorem

imply that

$$\begin{aligned}
\mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} < 2^{q-1}\delta^q\right) &= \mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} - 2^q\delta^q < -2^{q-1}\delta^q\right) \\
&= \mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w}) - Z^{*(d)}(\mathbf{0})\,\mathrm{d}\boldsymbol{w} < -2^{q-1}\delta^q\right) \\
&\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} \left|Z^{*(d)}(\boldsymbol{w}) - Z^{*(d)}(\mathbf{0})\right|\,\mathrm{d}\boldsymbol{w} > 2^{q-1}\delta^q\right) \\
&\leq 2^{-q+1}\delta^{-q}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} \left|Z^{*(d)}(\boldsymbol{w}) - Z^{*(d)}(\mathbf{0})\right|\,\mathrm{d}\boldsymbol{w}\right) \\
&= 2^{-q+1}\delta^{-q}\int_{[-\delta,\delta]^q} \mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{w}) - Z^{*(d)}(\mathbf{0})\right|\right)\,\mathrm{d}\boldsymbol{w}.
\end{aligned}$$

Finally, applying the second part to the integrand yields

$$2^{-k+1}\delta^{-k}\int_{[-\delta,\delta]^q} \mathsf{E}_{\boldsymbol{\eta}_0}\left(\left|Z^{*(d)}(\boldsymbol{w}) - Z^{*(d)}(\mathbf{0})\right|\right)\,\mathrm{d}\boldsymbol{w} \leq 2^{-k+1}\delta^{-k}\cdot 2\sqrt{Bq}\delta(2\delta)^q = 4\sqrt{Bq}\delta$$

and hence

$$\begin{aligned}
&\mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{\mathbb{R}^q} Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* < \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{4}(2\delta)^q\right) \\
&\quad\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w}\right)\,\mathrm{d}\boldsymbol{w} < \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{4}(2\delta)^q\right) \\
&\quad\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(\frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{2}\int_{[-\delta,\delta]^q} Z^{*(d)}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} < \frac{\mathfrak{h}(\boldsymbol{\eta}_0)}{4}(2\delta)^q\right) \leq 4\sqrt{Bq}\delta. \qquad \square
\end{aligned}$$

This lemma enables us to prove the following connection of (PCT1)–(PCT3) and (CS3').

Corollary 3.5.9
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Then there are* $C > 0$ *and* $D \in \mathbb{N}$ *such that for all* $\boldsymbol{\eta} \in \Theta$ *and all* $d > D$

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{(d)}(\boldsymbol{\eta},\boldsymbol{\eta}_0)^{1/2}\right) \leq \exp\left(-Cd\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2\right). \tag{3.97}$$

Especially,

$$\limsup_{d\to\infty}\frac{1}{d}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0)\right) \leq -2C\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2 =: c(\boldsymbol{\eta}) < 0 \tag{3.98}$$

holds for all $\boldsymbol{\eta} \in \Theta \setminus \{\boldsymbol{\eta}_0\}$. $\diamond$

Proof.
First of all, we note that

$$\boldsymbol{\eta}^* := \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\boldsymbol{\eta}-\boldsymbol{\eta}_0) \in \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta-\boldsymbol{\eta}_0)$$

holds for all $d \in \mathbb{N}$ and $\boldsymbol{\eta} \in \Theta$. We therefore get from inequality (3.93) in Lemma 3.5.7

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{(d)}(\boldsymbol{\eta},\boldsymbol{\eta}_0)^{1/2}\right) = \mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}^*)^{1/2}\right) &\leq \exp(-\varepsilon\|\boldsymbol{\eta}^*\|^2)\\ &= \exp\left(-\varepsilon(\boldsymbol{\eta}-\boldsymbol{\eta}_0)^\intercal \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)(\boldsymbol{\eta}-\boldsymbol{\eta}_0)\right)\\ &\leq \exp\left(-\varepsilon d\nu_{\min}\left(\frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)\right)\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2\right),\end{aligned}$$

for all $d > D$ with D as in Lemma 3.5.7. Next, condition (PCT3) implies (3.97). The second part follows from an application of Jensen's inequality to the first part with

$$\begin{aligned}\frac{1}{d}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\log Z_i(\boldsymbol{\eta},\boldsymbol{\eta}_0)\right) = \frac{1}{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\frac{2}{2}\log\left(Z^{(d)}(\boldsymbol{\eta},\boldsymbol{\eta}_0)\right)\right) &= \frac{2}{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\log\left(Z^{(d)}(\boldsymbol{\eta},\boldsymbol{\eta}_0)^{1/2}\right)\right)\\ \leq \frac{2}{d}\log\left(\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{(d)}(\boldsymbol{\eta},\boldsymbol{\eta}_0)^{1/2}\right)\right) &\leq \frac{2}{d}\log\left(\exp\left(-cd\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2\right)\right)\\ &= -2c\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2,\end{aligned}$$

for all $d > D$, as the logarithm is concave and strictly increasing. □

The inequation (3.98) directly implies condition (CS3') of the semiproper centering theory. Hence, we can directly conclude that the conditions of the PCT imply that of the SCT, i.e. Proposition 3.5.1 is proved.

Proposition 3.5.10

Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT2) *and* (PCT3) *are satisfied and that* Δ_d *is defined as in* (3.90) *for all* $d > D$. *Then* Δ_d *is almost surely not one-point distributed at* $\mathbf{0}$ *for any finite* $d > D$, $\Delta_d \xrightarrow{\mathcal{D}} \boldsymbol{\xi} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ *and* $\Delta_d = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$. ⋄

Proof.

First, notice that the conditional independence of $Y_1, \ldots, Y_d$ and $Y_i \neq 0$, $\mathsf{P}_{\boldsymbol{\eta}_0}$–almost surely directly imply that $\Delta_d \neq 0$, $\mathsf{P}_{\boldsymbol{\eta}_0}$–almost surely. Additionally, $\Delta_d \xrightarrow{\mathcal{D}} \boldsymbol{\xi}$ directly implies $\Delta_d = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$. Hence, it remains to prove $\Delta_d \xrightarrow{\mathcal{D}} \boldsymbol{\xi}$.

First, condition (PCT2) implies that the sequence $\{Y_i\}_{i\in\mathbb{N}}$ satisfies the Lindeberg condition. We define $\boldsymbol{X}_i := \nabla\lambda_i(\boldsymbol{\eta}_0)Y_i$ for all $i \in \mathbb{N}$ and

$$\theta_d := \nu_{\min}\left(\frac{1}{d}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)\right), \qquad \phi_d := \sum_{i=1}^{d} P_i(\boldsymbol{\eta}_0)(1-P_i(\boldsymbol{\eta}_0)) \leq \frac{d}{4},$$

for all $d \in \mathbb{N}$. Due to condition (PCT2), there is a $c < \infty$ such that $\sup_{i\in\mathbb{N}}\|\nabla\lambda_i(\boldsymbol{\eta}_0)\| \leq c$. Additionally, condition (PCT3) ensures that there is a $D \in \mathbb{N}$ and an $\delta > 0$ such that $\theta_d > \delta > 0$ for all $d > D$. We get

$$\begin{aligned}&\frac{1}{d\theta_d}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\|\boldsymbol{X}_i - \mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{X}_i)\|^2 \cdot \mathbb{1}\left(\|\boldsymbol{X}_i - \mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{X}_i)\|^2 \geq \varepsilon d\theta_d\right)\right)\\ &\quad\leq c'\frac{1}{\phi_d}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}\left((Y_i - P_i(\boldsymbol{\eta}_0))^2 \cdot \mathbb{1}\left((Y_i - P_i(\boldsymbol{\eta}_0))^2 \geq \varepsilon'\phi_d\right)\right)\end{aligned}$$

for each $\varepsilon > 0$ and all $d > D$ with $c' := c^2(4\delta)^{-1}$ and $\varepsilon' := c^{-2}\varepsilon\delta$, i.e. the Lindeberg condition of $\{Y_i\}_{i\in\mathbb{N}}$ implies the multivariate Lindeberg condition of $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ and this finishes the proof. □

Lemma 3.5.11
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT2) *and* (PCT3) *are satisfied. Then for each* $\delta > 0$ *there is a sequence* $\{\beta_d\}_{d\in\mathbb{N}}$ *with* $\beta_d = \mathcal{O}(d^{-1/2})$ *such that*

$$\lim_{d\to\infty} \mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{\eta}^*\in B_\delta(\mathbf{0})}\left\{\left|\log Z^{*(d)}(\boldsymbol{\eta}^*) - \left(\Delta_d^\intercal\boldsymbol{\eta}^* - \frac{1}{2}\|\boldsymbol{\eta}^*\|^2\right)\right| - \|\boldsymbol{\eta}^*\|^2\beta_d\right\} \le 0\right) = 1$$

for $Z^{*(d)}$ *given in* (3.92) *and* Δ_d *given in* (3.90). ◇

Proof.
The proof is based on the arguments of Ghosal (1997, 1999, Lemma 2.1 in both publications).

We develop a second order Taylor expansion of $\log Z^{*(d)}(\,\bullet\,)$ at point $\boldsymbol{u}_0 = \mathbf{0}$ with evaluation at an arbitrary $\boldsymbol{u} \in B_\delta(\mathbf{0})$. So holds

$$\log Z^{*(d)}(\boldsymbol{u}) = \Delta_d^\intercal\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^\intercal\nabla\nabla^\intercal\log Z^{*(d)}(\mathbf{0})\boldsymbol{u} + \frac{1}{6}\boldsymbol{u}^\intercal\nabla\nabla^\intercal_{\boldsymbol{u}}\log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u},$$

with immediate point $\boldsymbol{u}^* \in \{c\boldsymbol{u}\colon c \in [0,1]\}$ between $\boldsymbol{u}$ and $\mathbf{0}$ where $\nabla\nabla^\intercal$ denotes the Hessian–operator and the operator $\nabla\nabla^\intercal_{\boldsymbol{u}}$ denotes the element–wise directional derivatives of the Hessian in direction $\boldsymbol{u}$. We denote the first three derivatives of ψ, which is defined in (3.7), by ψ', ψ'', ψ'''. We get

$$\nabla\log Z^{*(d)}(\mathbf{0}) = \Delta_d,$$

due to $\psi'(\lambda_i(\boldsymbol{\eta}_0)) = \mathsf{E}_{\boldsymbol{\eta}_0}(Y_i) = P_i(\boldsymbol{\eta}_0)$. Direct computation of $\nabla\nabla^\intercal\log Z_d$ and evaluation at $\boldsymbol{\eta}_0$ results in

$$\begin{aligned}\nabla\nabla^\intercal\log Z_d(\mathbf{0}) = &\sum_{i=1}^d\left(\Big(Y_i - \psi'(\lambda_i(\boldsymbol{\eta}_0))\Big)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right)\\ &-\sum_{i=1}^d\left(\psi''(\lambda_i(\boldsymbol{\eta}_0))\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\nabla\lambda_i(\boldsymbol{\eta}_0)\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right).\end{aligned} \tag{3.99}$$

By plugging-in the definition of $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)$ in (3.99), we get

$$-\boldsymbol{u}^\intercal\sum_{i=1}^d\psi''(\lambda_i(\boldsymbol{\eta}_0))\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\nabla\lambda_i(\boldsymbol{\eta}_0)\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u} = -\|\boldsymbol{u}\|^2$$

for the quadratic form of the second sum in (3.99). Next, we discuss the quadratic form of the first sum in (3.99). To simplify the notation, we set $\boldsymbol{H}_i := \nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)$, $V_i := (Y_i - \psi'(\lambda_i(\boldsymbol{\eta}_0)))$ for $i = 1,\ldots,d$, $\boldsymbol{B}_d := \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}$, unit vectors $\boldsymbol{e}_j$, $j = 1,\ldots,q$, and $\boldsymbol{u} = (u_1,\ldots,u_q)^\intercal$ and obtain

$$\begin{aligned}&\left|\boldsymbol{u}^\intercal\sum_{i=1}^d\Big(Y_i - \psi'(\lambda_i(\boldsymbol{\eta}_0))\Big)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0)\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}\right|^2 \qquad (3.100)\\ &= \left|\boldsymbol{u}^\intercal\sum_{i=1}^d V_i\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{u}\right|^2 \le \left(\sum_{j=1}^q\sum_{k=1}^q u_ju_k\right)^2\left|\sum_{j=1}^q\sum_{k=1}^q\sum_{i=1}^d V_i\boldsymbol{e}_j^\intercal\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{e}_k\right|^2.\end{aligned}$$

Using (3.100) and Markov's inequality and supposing that $B > 0$ is arbitrarily selected, we get

$$\begin{aligned}
&\mathsf{P}_{\boldsymbol{\eta}_0}\left(\sup_{\boldsymbol{u}\colon \|\boldsymbol{u}\|=1}\left\{\left|\boldsymbol{u}^\intercal\sum_{i=1}^{d}V_i\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{u}\right|^2\right\}\geq B\|\boldsymbol{B}_d^{-1/2}\|\right)\\
&\quad\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(\left|\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{i=1}^{d}V_i\boldsymbol{e}_j^\intercal\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{e}_k\right|^2\geq B\|\boldsymbol{B}_d^{-1/2}\|\right)\\
&\quad\leq \sum_{j=1}^{q}\sum_{k=1}^{q}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\sum_{i=1}^{d}V_i\boldsymbol{e}_j^\intercal\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{e}_k\right)^2\left(B\|\boldsymbol{B}_d^{-1/2}\|\right)^{-2}\\
&\quad\leq \sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{i=1}^{d}\mathsf{E}_{\boldsymbol{\eta}_0}(V_i^2)\left(\boldsymbol{e}_j^\intercal\boldsymbol{B}_d^{-1/2}\boldsymbol{H}_i\boldsymbol{B}_d^{-1/2}\boldsymbol{e}_k\right)^2\left(B\|\boldsymbol{B}_d^{-1/2}\|\right)^{-2}\\
&\quad\leq \frac{1}{4}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{i=1}^{d}\|\boldsymbol{e}_j\|^4\cdot\|\boldsymbol{B}_d^{-1/2}\|^4\cdot\|\boldsymbol{H}_i\|^2\cdot\left(B\|\boldsymbol{B}_d^{-1/2}\|\right)^{-2}\\
&\quad= \frac{\|\boldsymbol{B}_d^{-1/2}\|^2}{4B^2}q^2d\sup_{i=1,\ldots,d}\|\boldsymbol{H}_i\|^2\leq CB^{-2},
\end{aligned}$$

for a constant $C > 0$, where the last step followed from (PCT2) with $\sup_{i\in\mathbb{N}}\|\boldsymbol{H}_i\| < \infty$ and (PCT3) with $\|\boldsymbol{B}_d^{-1/2}\| = \mathcal{O}(d^{-1/2})$. Hence, by selecting B sufficiently large, the probability will be arbitrary small for $d \to \infty$. Additionally, it is still converging to zero, as the right hand sight in the inequality in the first term is converging to zero in order $\|\boldsymbol{B}_d^{-1/2}\| = \mathcal{O}(d^{-1/2})$.

Finally, we have to show that the remainder, i.e. $6^{-1}\boldsymbol{u}^\intercal\nabla\nabla_{\boldsymbol{u}}^\intercal\log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u}$, converges to zero. We define $G\colon \mathbb{R}^q \mapsto \mathbb{R}^{q\times q}$,

$$\begin{aligned}
\boldsymbol{u}\mapsto\sum_{i=1}^{d}\Big(&\Big(Y_i-\psi'(\lambda_i(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}))\Big)\nabla\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u})\\
&-\psi''(\lambda_i(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u})))\nabla\lambda_i(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u})\\
&\times\nabla^\intercal\lambda_i(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u})\Big).
\end{aligned}\tag{3.101}$$

Using G, we get

$$\nabla\nabla^\intercal\log Z^{*(d)}(\boldsymbol{u}) = \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}G(\boldsymbol{u})\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}$$

for the Hessian so that we can use this representation to compute $\nabla\nabla_{\boldsymbol{u}}^\intercal\log Z^{*(d)}(\boldsymbol{u}^*)$ in the remainder. In order to do this, we have to compute the gradient of each element to compute the element-wise directional derivative, as stated above. To formulate the next step, we define some differential operators. First, let us define the operator of the element-wise kth partial derivative of a matrix valued mapping $\boldsymbol{F}\colon \mathbb{R}^q \to \mathbb{R}^{q\times q}$ as

$$\partial_k\boldsymbol{F}(\boldsymbol{x}) := \begin{pmatrix}\frac{\partial F_{11}(\boldsymbol{x})}{\partial x_k} & \cdots & \frac{\partial F_{1q}(\boldsymbol{x})}{\partial x_k}\\ \vdots & \ddots & \vdots\\ \frac{\partial F_{q1}(\boldsymbol{x})}{\partial x_k} & \cdots & \frac{\partial F_{qq}(\boldsymbol{x})}{\partial x_k}\end{pmatrix}.$$

Further, we define the operator of the element-wise directional derivative in direction $\boldsymbol{h}$ of $\boldsymbol{F}\colon \mathbb{R}^q \to \mathbb{R}^{q\times q}$ as

$$\mathcal{D}_{\boldsymbol{h}}\boldsymbol{F}(\boldsymbol{x}) := \begin{pmatrix} \nabla_{\boldsymbol{h}} F_{11}(\boldsymbol{x}) & \dots & \nabla_{\boldsymbol{h}} F_{1q}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \nabla_{\boldsymbol{h}} F_{q1}(\boldsymbol{x}) & \dots & \nabla_{\boldsymbol{h}} F_{qq}(\boldsymbol{x}) \end{pmatrix},$$

where $\nabla_{\boldsymbol{h}}$ denotes the directional derivative of a mapping in direction h. We can arrange all kth elements of the element-wise gradients in a matrix and obtain

$$\begin{aligned}\partial_k \nabla\nabla^\intercal \log Z^{*(d)}(\boldsymbol{u}) &= \partial_k \left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} G(\boldsymbol{u}) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right) \\ &= \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \partial_k \Big(G(\boldsymbol{u})\Big) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\end{aligned}$$

because derivation is a linear operator. Therefore, setting $\boldsymbol{v} := \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}$ leads to

$$\boldsymbol{u}^\intercal \nabla\nabla^\intercal_{\boldsymbol{u}} \log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u} = \boldsymbol{u}^\intercal \mathcal{D}_{\boldsymbol{u}} \nabla\nabla^\intercal \log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u} = \boldsymbol{v}^\intercal \mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\boldsymbol{v}.$$

Notice that the elements of $\mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)$ are given by

$$\Big(\mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\Big)_{jk} = \nabla G(\boldsymbol{u}^*)^\intercal_{jk}\boldsymbol{u}, \qquad j,k = 1,\dots,q,$$

due to the definition of directional derivatives, where $G(\boldsymbol{u}^*)_{jk}$ denotes the (j,k)th element of $G(\boldsymbol{u}^*)$. By the definition of G in (3.101) for each $\nabla G(\boldsymbol{u}^*)_{jk}$, $j,k = 1,\dots,q$, there is a function $K_{jk}\colon \mathbb{R}^q \to \mathbb{R}^q$ such that

$$\nabla G(\boldsymbol{u}^*)^\intercal_{jk}\boldsymbol{u} = K_{ik}(\boldsymbol{u}^*)^\intercal \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}.$$

This implies that

$$\left|\frac{1}{6}\boldsymbol{u}^\intercal \nabla\nabla^\intercal_{\boldsymbol{u}} \log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u}\right| = \frac{1}{6}\left|\boldsymbol{v}^\intercal \mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\boldsymbol{v}\right| \le \frac{1}{6}|\nu|_{\max}\Big(\mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\Big)\|\boldsymbol{v}\|^2$$

with

$$\begin{aligned}|\nu|_{\max}\Big(\mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\Big) \le \sqrt{\sum_{j=1}^q \sum_{k=1}^q (\mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)_{jk})^2} &= \sqrt{\sum_{j=1}^q \sum_{k=1}^q \Big(K_{ik}(\boldsymbol{u}^*)^\intercal \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}\Big)^2} \\ &\le \sqrt{\sum_{j=1}^q \sum_{k=1}^q \|K_{ik}(\boldsymbol{u}^*)\|^2 \cdot \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}\|^2} \\ &= \sqrt{\sum_{j=1}^q \sum_{k=1}^q \|K_{ik}(\boldsymbol{u}^*)\|^2} \cdot \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}\|,\end{aligned}$$

where the first step is implied by the fact that the Frobenius norm provides an upper bound for the largest eigenvalue and the second inequality results from an application of the Cauchy–Scharz–inequality.

As $Y_i \in \{0,1\}$, we obtain a sure bound (not only almost sure). Each component of each K_{jk} is a sum of order d of finite products of derivatives of ψ and λ_i of up to the third order,

evaluated over the relatively compact area $B_\delta(\mathbf{0})$. Due to the structure of $\psi, \dots, \psi'''$, we get from condition (PCT2) that

$$\sqrt{\sum_{j=1}^{q}\sum_{k=1}^{q} \|K_{ik}(\boldsymbol{u}^*)\|^2} = \mathcal{O}(d)$$

holds. Hence,

$$\frac{1}{6}\left|\boldsymbol{u}^\intercal \nabla\nabla_{\boldsymbol{u}}^\intercal \log Z^{*(d)}(\boldsymbol{u}^*)\boldsymbol{u}\right| = \frac{1}{6}|\boldsymbol{v}^\intercal \mathcal{D}_{\boldsymbol{u}} G(\boldsymbol{u}^*)\boldsymbol{v}| \leq \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}\|^3 \mathcal{O}(d) = \mathcal{O}(d^{-1/2}),$$

what we had to show. □

This leads to the following definition of an approximate likelihood ratio:

$$\tilde{Z}^{(d)} \colon \mathbb{R}^q \to \mathbb{R}, \qquad \boldsymbol{\eta}^* \mapsto \exp\left(\Delta_d^\intercal \boldsymbol{\eta}^* - \frac{1}{2}\|\boldsymbol{\eta}^*\|^2\right). \tag{3.102}$$

Proposition 3.5.10 and Lemma 3.5.11 immediately imply the following important corollary, which relates the PCT conditions to the GGS condition, in particular (GGS3), cf. Section 3.3.

Corollary 3.5.12
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT2) *and* (PCT3) *are satisfied. Then there is a random vector* $\boldsymbol{\xi} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ *such that all finite dimensional marginal distributions of* $Z^{*(d)}$ *converge to the finite dimensional marginal distributions of* $\tilde{Z}$ *with*

$$\tilde{Z} \colon \mathbb{R}^q \to \mathbb{R}, \qquad \boldsymbol{\eta}^* \mapsto \exp\left(\boldsymbol{\xi}^\intercal \boldsymbol{\eta}^* - \frac{1}{2}\|\boldsymbol{\eta}^*\|^2\right),$$

for $d \to \infty$. ◇

Lemma 3.5.13
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT2) *and* (PCT3) *are satisfied. Then, with* $\tilde{Z}^{(d)}$ *of* (3.102), $Z^{*(d)}$ *of* (3.92) *and for each* $\delta > 0$

$$\left(\int_{\mathbb{R}^q} \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*\right)^{-1} \int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^* = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(d^{-1/2}). \qquad \diamond$$

Proof.
The proof is based on the arguments by Ghosal (1997, Lemma 2.2).

For each $c > 0$ there is a $C > 0$ so that $\left|1 - x^{-1}\right| \leq C|\log(x)|$ for all $x \in (c, \infty)$. Denote by $\{\beta_d\}_{d\in\mathbb{N}}$ the sequence from Lemma 3.5.11 with $\beta_d = \mathcal{O}(d^{-1/2})$. As $\beta_d\|\boldsymbol{\eta}^*\|^2 \to 0$ for all $\boldsymbol{\eta}^* \in B_\delta(\mathbf{0})$ and all $\delta > 0$, Lemma 3.5.11 implies that

$$\left|\log\left(\frac{Z^{*(d)}(\boldsymbol{\eta}^*)}{\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}\right)\right| \leq \beta_d\|\boldsymbol{\eta}^*\|^2$$

with probability tending to one for $d \to \infty$. Additionally, for any $c \in (0,1)$ the probability is tending to one such that

$$0 < c < \frac{Z^{*(d)}(\boldsymbol{\eta}^*)}{\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}.$$

Hence, there is a $B > 0$ such that

$$\left| 1 - \frac{\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}{Z^{*(d)}(\boldsymbol{\eta}^*)} \right| \leq B\beta_d \|\boldsymbol{\eta}^*\|^2$$

with probability probability tending to one for $d \to \infty$. Multiplying the last equation by $Z^{*(d)}(\boldsymbol{\eta}^*)$ and using Lemma 3.5.11 again, we obtain

$$|Z^{*(d)}(\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)| \leq B\beta_d \|\boldsymbol{\eta}^*\|^2 \tilde{Z}^{(d)}(\boldsymbol{\eta}^*) \exp(\beta_d \|\boldsymbol{\eta}^*\|^2).$$

Setting $\gamma_d = C\beta_d$ with $C = \max(B, 1)$, we get directly

$$\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)| \, \mathrm{d}\boldsymbol{\eta}^* \leq \gamma_d \int_{B_\delta(\mathbf{0})} \|\boldsymbol{\eta}^*\|^2 \exp\left(\Delta_d^\intercal \boldsymbol{\eta}^* - \frac{(1 - 2\gamma_d)\|\boldsymbol{\eta}^*\|^2}{2} \right) \mathrm{d}\boldsymbol{\eta}^*.$$

Defining $c_d \coloneqq 1 - 2\gamma_d$, $d \in \mathbb{N}$ and using

$$\left\| c_d^{1/2} \boldsymbol{\eta}^* - c_d^{-1/2} \Delta_d \right\|^2 = c_d \|\boldsymbol{\eta}^*\|^2 - \Delta_d^\intercal \boldsymbol{\eta}^* + c_d^{-1} \|\Delta_d\|^2,$$

we conclude that

$$\begin{aligned} &\int_{B_\delta(\mathbf{0})} \|\boldsymbol{\eta}^*\|^2 \exp\left(\Delta_d^\intercal \boldsymbol{\eta}^* - c_d \frac{\|\boldsymbol{\eta}^*\|^2}{2} \right) \mathrm{d}\boldsymbol{\eta}^* \\ &\qquad = \exp\left(\frac{\|\Delta_d\|^2}{2c_d} \right) \int_{B_\delta(\mathbf{0})} \|\boldsymbol{\eta}^*\|^2 \exp\left(-\frac{\|c_d^{1/2} \boldsymbol{\eta}^* - c_d^{-1/2} \Delta_d\|^2}{2} \right) \mathrm{d}\boldsymbol{\eta}^*. \end{aligned}$$

As the integrand is positive, we get an estimate from above by replacing the integration area $B_\delta(\mathbf{0})$ by $\mathbb{R}^q$ on the right-hand side. Using the transformation $\boldsymbol{u}(\boldsymbol{x}) \coloneqq c_d^{-1/2} \boldsymbol{x} + c_d^{-1} \Delta_d$, $\boldsymbol{x} \in \mathbb{R}^q$, with the functional determinant $c_d^{-q/2}$ results in

$$\begin{aligned} &\int_{\mathbb{R}^q} \|\boldsymbol{\eta}^*\|^2 \exp\left(-\frac{\|c_d^{1/2} \boldsymbol{\eta}^* - c_d^{-1/2} \Delta_d\|^2}{2} \right) \mathrm{d}\boldsymbol{\eta}^* \\ &\qquad = c_d^{-q/2} \int_{\mathbb{R}^q} \|c_d^{-1/2} \boldsymbol{x} + c_d^{-1} \Delta_d\|^2 \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2} \right) \mathrm{d}\boldsymbol{x} \\ &\qquad = c_d^{-1-q/2} \int_{\mathbb{R}^q} \|\boldsymbol{x} + c_d^{-1/2} \Delta_d\|^2 \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2} \right) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

The fact that

$$\int_{\mathbb{R}^q} \|\boldsymbol{x} + \boldsymbol{a}\|^2 \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2} \right) \mathrm{d}\boldsymbol{x} = (2\pi)^{q/2} (\|\boldsymbol{a}\|^2 + q)$$

for all $\boldsymbol{a} \in \mathbb{R}^q$ implies that

$$\int_{\mathbb{R}^q} \|\boldsymbol{x} + c_d^{-1/2}\Delta_d\|^2 \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right) \mathrm{d}\boldsymbol{x} = (2\pi)^{q/2}(c_d^{-1}\|\Delta_d\|^2 + q).$$

Putting the last steps together results in

$$\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)| \,\mathrm{d}\boldsymbol{\eta}^* \leq \gamma_d c_d^{-1-q/2} \exp\left(\frac{\|\Delta_d\|^2}{2c_d}\right)(2\pi)^{q/2}(c_d^{-1}\|\Delta_d\|^2 + q).$$

Finally, the result follows from the fact that $c_d \to 1$, $\gamma_d \to 0$, $\|\Delta_d\|^2 = \mathcal{O}_\mathsf{P}(1)$ for $d \to \infty$ and

$$\int_{\mathbb{R}^q} \tilde{Z}^{(d)}(\boldsymbol{\eta}^*) \,\mathrm{d}\boldsymbol{\eta}^* = \exp\left(\frac{\|\Delta_d\|^2}{2}\right)(2\pi)^{q/2}$$

for every $d \in \mathbb{N}$. □

Until now, we only discussed the behavior inside some neighborhood of $\boldsymbol{\eta}_0$. It is also important that the distribution tail is decreasing sufficiently fast. Under the given assumptions, we can show, that this is true for the normalized posterior.

Lemma 3.5.14
Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Consider a mapping* $f\colon \mathbb{R}^q \to (0,\infty)$ *such that a further mapping* $F\colon (0,\infty) \to (0,\infty)$ *exists with* $f\big|_{B_\delta(\mathbf{0})} \leq F(\delta)$ *for all* $\delta > 0$ *and* $F(\delta) = \mathcal{O}(\exp(C\delta))$ *for* $\delta \to \infty$ *and some* $C > 0$*. Then, for any* $C \geq 0$ *and for all* $k \in \mathbb{N}$ *there is is a constant* $\delta > 0$ *such that*

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\mathbb{R}^q\setminus\bar{B}_\delta(\mathbf{0})} f(\boldsymbol{\eta}^*)h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \leq C. \qquad \diamond$$

Proof.
The proof is based on the arguments by Ghosal (1999, Lemma 2.5) and Ibragimov and Has'minskii (1981, Lemma I.5.2).

The strategy is now to cut the integration area into small pieces so that we get for each $\delta \in \mathbb{N}$:

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\mathbb{R}^q\setminus\overline{B}_\delta(\mathbf{0})} f(\boldsymbol{\eta}^*)h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \\
&\quad= \mathsf{E}_{\boldsymbol{\eta}_0}\left(\sum_{H=\delta}^{\infty}\int_{\overline{B}_{H+1}(\mathbf{0})\setminus\overline{B}_H(\mathbf{0})} f(\boldsymbol{\eta}^*)h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \\
&\quad= \sum_{H=\delta}^{\infty}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\overline{B}_{H+1}(\mathbf{0})\setminus\overline{B}_H(\mathbf{0})} f(\boldsymbol{\eta}^*)h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \\
&\quad\leq \sum_{H=\delta}^{\infty}\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\overline{B}_{H+1}(\mathbf{0})\setminus\overline{B}_H(\mathbf{0})} F(\|\boldsymbol{\eta}^*\|)h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \\
&\quad\leq \sum_{H=\delta}^{\infty} F(H+1)\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\overline{B}_{H+1}(\mathbf{0})\setminus\overline{B}_H(\mathbf{0})} h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \,\mathrm{d}\boldsymbol{\eta}^*\right) \\
&\quad\leq \sum_{H=\delta}^{\infty} F(H+1)g(H) \leq C < \infty
\end{aligned}$$

for an appropriate function $g\colon \mathbb{N} \to (0,\infty)$. The change of series and expectation in the second step is valid by Tonelli's theorem, as $f(\boldsymbol{\eta}^*)h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \geq 0$ for all $d, k \in \mathbb{N}$ and all $\boldsymbol{\eta}^*$. In particular, as all summands of the series are positive, too, we can fix $C > 0$ and find a corresponding $\delta > 0$ to get a proper bound for the tail-moment. Hence, we only have to determine an appropriate mapping g so that

$$\mathsf{E}_{\boldsymbol{\eta}_0}\left(\int_{\overline{B}_{H+1}(\mathbf{0})\setminus\overline{B}_H(\mathbf{0})} h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^*\right) \leq g(H) \tag{3.103}$$

and the series $\sum_{H=1}^{\infty} F(H+1)g(H)$ converges.

First, we fix $d, H > 0$ and set

$$\begin{aligned} M &:= \overline{B}_{H+1}(\mathbf{0}) \setminus \overline{B}_H(\mathbf{0}), \\ I &:= \int_M Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{\eta}^*. && (3.104) \\ W &:= \int_{\mathbb{R}^q} Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*, && (3.105) \\ Q &:= \frac{I}{W} = \int_M h(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^*. && (3.106) \end{aligned}$$

This means that (3.103) is reformulated as $\mathsf{E}_{\boldsymbol{\eta}_0}(Q) \leq g(H)$.

Let $L > 0$ be some further constant such that $L^q \in \mathbb{N}$ which will be specified later. We decompose the set $\Gamma \cap \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta - \boldsymbol{\eta}_0)$ into L^q identical disjoint sets $\Delta_1, \ldots, \Delta_{L^q}$ (identical in the sense that $\Delta_i = \boldsymbol{Q}\Delta_j$ for some rotation matrix $\boldsymbol{Q}$ and all $i, j = 1, \ldots, L^q$). We select $\boldsymbol{\eta}_i \in \Delta_i$, $i = 1, \ldots, L^q$, and approximate I by setting

$$\begin{aligned} S :=& \sum_{i=1}^{L^q} \int_{\Delta_i} Z^{*(d)}(\boldsymbol{\eta}_i)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} && (3.107) \\ =& \sum_{i=1}^{L^q} Z^{*(d)}(\boldsymbol{\eta}_i) \int_{\Delta_i} \mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \\ \leq& \max_{i=1,\ldots,L^q}\left\{Z^{*(d)}(\boldsymbol{\eta}_i)\right\} \int_{M\cap\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\Theta-\boldsymbol{\eta}_0)} \mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \\ \leq& \max_{i=1,\ldots,L^q}\left\{Z^{*(d)}(\boldsymbol{\eta}_i)\right\} \sup_{\boldsymbol{\eta}\in\Theta}\{\mathfrak{h}(\boldsymbol{\eta})\}\mathscr{L}^q(M) \\ =& \max_{i=1,\ldots,L^q}\left\{Z^{*(d)}(\boldsymbol{\eta}_i)\right\} \sup_{\boldsymbol{\eta}\in\Theta}\{\mathfrak{h}(\boldsymbol{\eta})\}\frac{\pi^{q/2}\left((H+1)^q - H^q\right)}{\Gamma\left(\frac{q}{2}+1\right)}, \end{aligned}$$

where $\mathscr{L}^q$ denotes the q-dimensional Lebesgue measure, Γ denotes the Gamma-function. Notice that

$$\sup_{\boldsymbol{\eta}\in\Theta}\{\mathfrak{h}(\boldsymbol{\eta})\} < \infty$$

because of (PCT1). We define

$$\gamma\colon \mathbb{N} \to (0,1), \qquad k \mapsto \sup_{\boldsymbol{\eta}\in\Theta}\{\mathfrak{h}(\boldsymbol{\eta})\}\frac{\pi^{q/2}\left((k+1)^q - k^q\right)}{\Gamma\left(\frac{q}{2}+1\right)}.$$

It follows for any $b \in (0, \frac{1}{2}]$ and sufficiently large d with ε from Lemma 3.5.7 using Markov's inequality that

$$\begin{aligned}
\mathsf{P}_{\boldsymbol{\eta}_0}\left(S > \frac{1}{4}\exp(-\varepsilon b H^2)\right) &\le \mathsf{P}_{\boldsymbol{\eta}_0}\left(\max_{i=1,\dots,L^q}\left\{Z^{*(d)}(\boldsymbol{\eta}_i)\right\}\gamma(H) > \frac{1}{4}\exp(-\varepsilon b H^2)\right) \\
&= \mathsf{P}_{\boldsymbol{\eta}_0}\left(\max_{i=1,\dots,L^q}\left\{Z^{*(d)}(\boldsymbol{\eta}_i)^{1/2}\right\}\gamma(H)^{1/2} > \frac{1}{2}\exp\left(-\frac{1}{2}\varepsilon b H^2\right)\right) \\
&\le \sum_{i=1}^{L^q}\mathsf{P}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}_i)^{1/2}\gamma(H)^{1/2} > \frac{1}{2}\exp\left(-\frac{1}{2}\varepsilon b H^2\right)\right) \\
&\le \sum_{i=1}^{L^q}\mathsf{E}_{\boldsymbol{\eta}_0}\left(Z^{*(d)}(\boldsymbol{\eta}_i)^{1/2}\right)\gamma(H)^{1/2}\cdot 2\exp\left(\frac{1}{2}\varepsilon b H^2\right) \\
&\le 2\exp\left(\frac{1}{2}\varepsilon b H^2\right)\gamma(H)^{1/2}\sum_{i=1}^{L^q}\exp\left(-\varepsilon\|\boldsymbol{\eta}_i\|^2\right) \\
&\le 2L^q\exp\left(\frac{1}{4}\varepsilon H^2\right)\gamma(H)^{1/2}\exp\left(-\varepsilon H^2\right) \\
&= 2L^q\exp\left(-\frac{3}{4}\varepsilon H^2\right)\gamma(H)^{1/2}. \qquad (3.108)
\end{aligned}$$

Further, we get from Lemma 3.5.7 the following upper bound for the expected difference of S and I that are defined in (3.104) and (3.104), respectively:

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\eta}_0}|S-I| &= \mathsf{E}_{\boldsymbol{\eta}_0}\left|\sum_{i=1}^{L^q}\int_{\Delta_i}\left(Z^{*(d)}(\boldsymbol{\eta}_i) - Z^{*(d)}(\boldsymbol{w})\right)\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\right| \\
&\le \sum_{i=1}^{L^q}\int_{\Delta_i}\mathsf{E}_{\boldsymbol{\eta}_0}\left|Z^{*(d)}(\boldsymbol{\eta}_i) - Z^{*(d)}(\boldsymbol{w})\right|\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \\
&\le \sum_{i=1}^{L^q}\int_{\Delta_i}2\sqrt{B}\|\boldsymbol{\eta}_i - \boldsymbol{w}\|\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}.
\end{aligned}$$

By the construction of $\Delta_1, \dots, \Delta_{L^q}$ there is a constant $0 < \beta = \mathcal{O}(H)$ such that

$$\|\boldsymbol{\eta}_i - \boldsymbol{w}\| \le \operatorname{diam}(\Delta_i) := \sup_{\boldsymbol{v},\boldsymbol{w}\in\Delta_i}\|\boldsymbol{v} - \boldsymbol{w}\| \le \frac{\beta}{L}$$

for all $i = 1, \dots, L^q$. Then with $B_1 := 2\sqrt{B}\beta$

$$\mathsf{E}_{\boldsymbol{\eta}_0}|S-I| \le \frac{B_1}{L}\int_\Gamma \mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \le \frac{B_1}{L}\gamma(H) \qquad (3.109)$$

holds. By the finite additivity of measures we get for any two random variables X, Y and constants $c > 0$, $d \in (0,1)$

$$\mathsf{P}(X + Y > c) \le \mathsf{P}\left(\{X > cd\} \cup \{Y > c(1-d)\}\right) \le \mathsf{P}(X > cd) + \mathsf{P}(Y > c(1-d)),$$

as necessarily $X > cd$ or $Y > c(1-d)$ if $X + Y > c$. Hence, for each $b \in (0, \frac{1}{2})$ we obtain

$$\begin{aligned}
\mathsf{P}_{\boldsymbol{\eta}_0}\left(I \geq \exp\left(-\frac{1}{2}b\varepsilon H^2\right)\right) &\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(S + |I-S| \geq \exp\left(-\frac{1}{2}b\varepsilon H^2\right)\right) \\
&\leq \mathsf{P}_{\boldsymbol{\eta}_0}\left(S \geq \frac{1}{2}\exp\left(-\frac{1}{2}b\varepsilon H^2\right)\right) \\
&\quad + \mathsf{P}_{\boldsymbol{\eta}_0}\left(|I-S| \geq \frac{1}{2}\exp\left(-\frac{1}{2}b\varepsilon H^2\right)\right) \\
&\leq 2L^q \exp\left(-\frac{3}{4}\varepsilon H^2\right)\gamma(H)^{1/2} \\
&\quad + \frac{B_1}{L}\gamma(H) \cdot 2\exp\left(\frac{1}{2}b\varepsilon H^2\right).
\end{aligned}$$

The change of the first term in the last step follows from (3.108) with a valid replacement of $\frac{1}{2}\exp\left(-\frac{1}{2}b\varepsilon H^2\right)$ by $\frac{1}{4}\exp\left(-\frac{1}{2}b\varepsilon H^2\right)$as it only increases the probability of the event. The second part of the last step results from an application of Markov's inequality using (3.109). Selecting L such that

$$1 \leq L^q \exp\left(-\frac{1}{4}\varepsilon H^2\right) \leq 2^q$$

ensures that

$$\frac{1}{L}\exp\left(\frac{1}{2}b\varepsilon H^2\right) \leq \exp\left(\left(\frac{b}{2} - \frac{1}{4q}\right)\varepsilon H^2\right).$$

Selecting $b = (4q)^{-1}$ implies

$$\mathsf{P}_{\boldsymbol{\eta}_0}\left(I \geq \exp\left(-\frac{1}{2}b\varepsilon H^2\right)\right) \leq 2^{q+1}\exp\left(-\frac{1}{2}\varepsilon H^2\right)\gamma(H)^{1/2} + 2B_1\gamma(H)\exp\left(-\frac{1}{8q}\varepsilon H^2\right).$$

Now we are able to show the desired bound for Q that is defined in (3.106). Since $Q \leq 1$, we get with $\alpha_1 \coloneqq C^{-1}\delta^q$ and $\alpha_2 \coloneqq \exp(-b\varepsilon H^2)$ for some $C, \delta, b, \varepsilon > 0$

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\eta}_0}(Q) \leq \frac{\alpha_2}{\alpha_1}\mathsf{P}_{\boldsymbol{\eta}_0}\left(Q \leq \frac{\alpha_2}{\alpha_1}\right) + 1 \cdot \mathsf{P}_{\boldsymbol{\eta}_0}\left(Q \geq \frac{\alpha_2}{\alpha_1}\right) &\leq \frac{\alpha_2}{\alpha_1} + \mathsf{P}_{\boldsymbol{\eta}_0}\left(\{I \geq \alpha_2\} \cup \{W \leq \alpha_1\}\right) \\
&\leq \frac{\alpha_2}{\alpha_1} + \mathsf{P}_{\boldsymbol{\eta}_0}(I \geq \alpha_2) + \mathsf{P}_{\boldsymbol{\eta}_0}(W \leq \alpha_1),
\end{aligned}$$

where I and W are given in (3.104) and (3.105), respectively. Lemma 3.5.7 implies that $\mathsf{P}_{\boldsymbol{\eta}_0}(W \leq \alpha_1) \leq C'\delta$ for some $C' > 0$. Now, selecting $\delta = \exp(-b\varepsilon(2q)^{-1}H^2)$ leads to

$$\mathsf{E}_{\boldsymbol{\eta}_0}(Q) = \mathcal{O}\left(H^q \exp\left(-C'' \cdot H^2\right)\right)$$

for some $C'' > 0$, which finishes the proof. □

Proposition 3.5.15

Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta \setminus \partial\Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Let* Δ_d *be given by* (3.90) *for all* $d \in \mathbb{N}$ *and* $Z^{*(d)}$ *by* (3.92) *for all* $d > D$*, where* $D \in \mathbb{N}$ *is the index so that* $\mathcal{I}^{(d')}(\boldsymbol{\eta}_0)$ *is regular for all* $d' > D$*. Further consider a mapping* f *satisfying the conditions of Lemma 3.5.14. Then*

$$\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*)h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^* \tag{3.110}$$

exists for all $d \in \mathbb{N}$ *and*

$$\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \overset{\mathsf{P}_{\boldsymbol{\eta}_0}}{\to} 0,$$

for $d \to \infty$. *In particular, this is valid for* $f = \| \bullet \|^k$ *for all* $k \in \mathbb{N}_0$, $f = \exp(\boldsymbol{t}^\intercal \bullet)$ *and* $f = \exp(\|\boldsymbol{t}\| \cdot \| \bullet \|)$ *for all* $\boldsymbol{t} \in \mathbb{R}^q$. ◇

Proof.
The proof is based on arguments by Ghosal (1997, 1999, in each Theorem 2.1). That it is possible to show a noticeably general moment convergence is inspired by Remark 2.1 of Ghosal (1997), where he stated that it is possible to prove the convergence of the first moment.

First, we discuss the existence of (3.110). Notice that

$$\begin{aligned} 0 &\leq \int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \, \mathrm{d}\boldsymbol{\eta}^* \\ &\leq \int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* + \int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \, \mathrm{d}\boldsymbol{\eta}^*. \end{aligned}$$

Since

$$\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) \left| \phi_q(\boldsymbol{\eta}^*; \boldsymbol{\mu}, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \leq \int_{\mathbb{R}^q} F(\boldsymbol{\eta}^*) \phi_q(\boldsymbol{\eta}^*; \boldsymbol{\mu}, \mathrm{I}_q) \, \mathrm{d}\boldsymbol{\eta}^* < \infty$$

for all $\boldsymbol{\mu} \in \mathbb{R}^q$, the existence of $\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \, \mathrm{d}\boldsymbol{\eta}^*$ follows from analyzing the term

$$\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^*.$$

The normal pdf with mean vector Δ_d can be represented using the approximate likelihood ratio $\tilde{Z}^{(d)}$ by writing

$$\phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) = \frac{\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}{\int_{\mathbb{R}^q} \tilde{Z}^{(d)}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}} = \frac{\mathfrak{h}(\boldsymbol{\eta}_0) \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}{\int_{\mathbb{R}^q} \mathfrak{h}(\boldsymbol{\eta}_0) \tilde{Z}^{(d)}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}}$$

for $\boldsymbol{\eta}^* \in \mathbb{R}^q$. Using the triangle inequality for an arbitrary $\delta > 0$, we get

$$\begin{aligned} &\int_{\mathbb{R}^q} f(\boldsymbol{\eta}^*) |h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q)| \, \mathrm{d}\boldsymbol{\eta}^* \\ &\leq \int_{B_\delta(\mathbf{0})} f(\boldsymbol{\eta}^*) \left| \frac{Z^{*(d)}(\boldsymbol{\eta}^*) \mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \boldsymbol{\eta}^*)}{\int_{\mathbb{R}^q} Z^{*(d)}(\boldsymbol{w}) \mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}} \right. \\ &\qquad \left. - \frac{\mathfrak{h}(\boldsymbol{\eta}_0) \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)}{\int_{\mathbb{R}^q} \mathfrak{h}(\boldsymbol{\eta}_0) \tilde{Z}^{(d)}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}} \right| \mathrm{d}\boldsymbol{\eta}^* \\ &\quad + \int_{\mathbb{R}^q \setminus B_\delta(\mathbf{0})} f(\boldsymbol{\eta}^*) h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) \, \mathrm{d}\boldsymbol{\eta}^* + \int_{\mathbb{R}^q \setminus B_\delta(\mathbf{0})} f(\boldsymbol{\eta}^*) \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \, \mathrm{d}\boldsymbol{\eta}^*. \end{aligned}$$

The last two terms on the right-hand side of the last inequation can be made arbitrarily small for sufficiently large δ with arbitrarily high probability by Markov's inequality, Lemma

3.5.14 and a similar property of the normal density. With an argument by Ghosal (1999, Lemma A.1)[14], the first term on the right-hand side is bounded by

$$F(\delta)\int_{\mathbb{R}^q\setminus B_\delta(\mathbf{0})} h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^* + F(\delta)\int_{\mathbb{R}^q\setminus B_\delta(\mathbf{0})} \phi_q(\boldsymbol{\eta}^*;\Delta_d,\mathrm{I}_q)\,\mathrm{d}\boldsymbol{\eta}^* + 3F(\delta)\frac{\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*) - \mathfrak{h}(\boldsymbol{\eta}_0)\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\mathfrak{h}(\boldsymbol{\eta}_0)\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*},$$

where the first two terms can again be made arbitrarily small because $F(\delta)=\mathcal{O}(\exp(\delta))$ and the integrals are $\mathcal{O}(\exp(-\delta^2))$ for $\delta\to\infty$. The triangle inequality implies that numerator of the last term on the right-hand side is bounded in the following way:

$$\begin{aligned}
&\int_{B_\delta(\mathbf{0})}\left|Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0)\right|\mathrm{d}\boldsymbol{\eta}^*\\
&\quad\le \int_{B_\delta(\mathbf{0})}\left|Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*) - Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0)\right|\mathrm{d}\boldsymbol{\eta}^*\\
&\qquad+\int_{B_\delta(\mathbf{0})}\left|Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0)\right|\mathrm{d}\boldsymbol{\eta}^*\\
&\quad\le \sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})}\left|\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w}) - \mathfrak{h}(\boldsymbol{\eta}_0)\right|\int_{B_\delta(\mathbf{0})} Z^{*(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*\\
&\qquad+\int_{B_\delta(\mathbf{0})}\left|Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0)\right|\mathrm{d}\boldsymbol{\eta}^*.
\end{aligned}$$

Hence we get

$$\begin{aligned}
&\frac{\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*)\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*) - \mathfrak{h}(\boldsymbol{\eta}_0)\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\mathfrak{h}(\boldsymbol{\eta}_0)\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*}\\
&\quad\le \sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})}\left|\frac{\mathfrak{h}(\boldsymbol{\eta}_0+\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})}{\mathfrak{h}(\boldsymbol{\eta}_0)} - 1\right|\cdot\frac{\int_{B_\delta(\mathbf{0})} Z^{*(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*}\\
&\qquad+\frac{\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*}.
\end{aligned}\tag{3.111}$$

Lemma 3.5.13 implies that

$$\frac{\int_{B_\delta(\mathbf{0})} |Z^{*(d)}(\boldsymbol{\eta}^*) - \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0.\tag{3.112}$$

[14] Lemma A.1 of Ghosal (1999) states: Suppose that $(S,\mathcal{S},\mu)$ is a measurable space and let $f,g\colon S\to[0,\infty)$ be μ–integrable mappings that satisfy $F\coloneqq\int_S f\mathrm{d}\mu>0$ and $G\coloneqq\int_S g\mathrm{d}\mu>0$. Then $\int_A |F^{-1}f - G^{-1}g|\,\mathrm{d}\mu \le F^{-1}\int_{S\setminus A} f\mathrm{d}\mu + G^{-1}\int_{S\setminus A} g\mathrm{d}\mu + 3G^{-1}\int_A |f-g|\mathrm{d}\mu$ holds for all $A\in\mathcal{S}$.

By Lemma 3.5.11 there is a a sequence $\{\beta_d\}_{d\in\mathbb{N}}$ with $\beta_d = \mathcal{O}(d^{-1/2})$ such that

$$\begin{aligned}\int_{B_\delta(\mathbf{0})} Z^{*(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* &\leq \int_{B_\delta(\mathbf{0})} \tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\exp(\beta_d\|\boldsymbol{\eta}^*\|^2)\,\mathrm{d}\boldsymbol{\eta}^* \\ &\leq \sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})}\{\exp(\beta_d\|\boldsymbol{w}\|^2)\}\int_{B_\delta(\mathbf{0})}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* \\ &\leq \exp(\beta_d(q\delta)^2)\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* \\ &= \exp\left(\beta_d(q\delta)^2 + \frac{\|\Delta_d\|^2}{2}\right)(2\pi)^{q/2}.\end{aligned}$$

Combining this with $\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^* = \exp\left(2^{-1}\|\Delta_d\|^2\right)(2\pi)^{q/2}$ results in

$$\frac{\int_{B_\delta(\mathbf{0})} Z^{*(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*}{\int_{\mathbb{R}^q}\tilde{Z}^{(d)}(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*} \leq \exp(\beta_d(q\delta)^2) \overset{d\to\infty}{\longrightarrow} 1. \tag{3.113}$$

Condition (PCT3) ensures that $\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\| \to 0$ for $d\to\infty$. This implies that

$$\sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})}\frac{\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})}{\mathfrak{h}(\boldsymbol{\beta}_0)} \leq 3.5 \tag{3.114}$$

for $d\to\infty$ and any $\delta > 0$. Inequality (3.114) enables us to utilize the fact that $|x-1| \leq 2|\log(x)|$ for all $x \in (0, 3.5]$. Additionally, the Lipschitz-continuity of $\mathfrak{h}$ in (PCT1) ensures that there is a $K > 0$ such that $\|\log\mathfrak{h}(\boldsymbol{\eta}_0+\boldsymbol{x}) - \log\mathfrak{h}(\boldsymbol{\eta}_0)\| \leq K\|\boldsymbol{x}\|$ for all $\boldsymbol{x}\in\mathbb{R}^q$ such that $\boldsymbol{\eta}_0 + \boldsymbol{x} \in \Theta$. Therefore, we get for $d\to\infty$:

$$\begin{aligned}\sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})}\left|\frac{\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w})}{\mathfrak{h}(\boldsymbol{\eta}_0)} - 1\right| &\leq \sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})} 2|\log\mathfrak{h}(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w}) - \log\mathfrak{h}(\boldsymbol{\eta}_0)| \\ &\leq \sup_{\boldsymbol{w}\in B_\delta(\mathbf{0})} 2K\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{w}\| \\ &\leq 2K(q\delta)^2\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\| \overset{d\to\infty}{\longrightarrow} 0,\end{aligned} \tag{3.115}$$

where the third step follows from the operator norm compatibility.

Due to (3.112), (3.113) and (3.115), the right-hand side of (3.111) converges to zero in probability and that finishes the proof. □

Proposition 3.5.16

Suppose that $\{Y_i\}_{i\in\mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta}_0)$ *for* $\boldsymbol{\eta}_0 \in \Theta$ *and that conditions* (PCT1)–(PCT3) *are satisfied. Consider a sequence* $\{\tilde{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ *of estimators for* $\boldsymbol{\eta}_0$ *such that*

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) = \Delta_d + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1), \tag{3.116}$$

for Δ_d *given in* (3.90). *Then*

$$\sup_{A\in\mathcal{B}^q}\left|\mathsf{P}\left(\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d)\in A \,\middle|\, \boldsymbol{Y}^{(d)}\right) - \tilde{\Phi}_q(A)\right| \overset{\mathsf{P}_{\boldsymbol{\eta}_0}}{\longrightarrow} 0, \tag{3.117}$$

$$\tilde{\boldsymbol{\eta}}_d \overset{\mathsf{P}_{\boldsymbol{\eta}_0}}{\longrightarrow} \boldsymbol{\eta}_0, \tag{3.118}$$

$$\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) \overset{\mathcal{D}}{\longrightarrow} \boldsymbol{\xi} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q), \tag{3.119}$$

as $d\to\infty$. ◇

Proof.
The proof is based on the first part of the proof of Theorem 2.2 by Ghosal (1997).

First, we get

$$\begin{aligned}\|\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\| &= \|\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0)\| \\ &\leq \|\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\| \cdot \|\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0)\| \\ &= \mathcal{O}(d^{-1/2})\mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(\|\Delta_d\|) = \mathcal{O}(d^{-1/2})\mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1) = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(d^{-1/2}) = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\end{aligned}$$

for $d \to \infty$. Lemma 3.4.7 implies that $\{d^{-1/2}\mathcal{I}^{(d)}(\,\bullet\,)^{1/2}\}_{d\in\mathbb{N}}$ is eqicontinuous. A combination of this with Proposition 3.5.10, Slutsky's theorem, the recently shown consistency of $\tilde{\boldsymbol{\eta}}_d$ and the regularity of $\mathcal{I}^{(d)}(\,\bullet\,)^{1/2}$ for sufficiently large d imply that

$$\begin{aligned}\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) &= \mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\tilde{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) \\ &= \mathrm{I}_q(1 + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1))\left(\Delta_d + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\right) \\ &= \Delta_d + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1) \xrightarrow{\mathcal{D}} \boldsymbol{\xi}.\end{aligned}$$

It remains to show the proper centering property. First, we obtain from the general properties of L^1 and total variation distance of Lebesgue-continuous probability measures that

$$\sup_{A\in\mathcal{B}^q}\left|\mathsf{P}_{\boldsymbol{\eta}_0}\left(\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d) \in A \,\middle|\, \boldsymbol{Y}^{(d)}\right) - \Phi_q(A)\right| = \frac{1}{2}\int_{\mathbb{R}^q}\left|h^{**}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*)\right| \mathrm{d}\boldsymbol{\eta}^*,$$

where $h^{**}(\,\bullet \mid \boldsymbol{Y}^{(d)})$ denotes the posterior density of $\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d)$ given $\boldsymbol{Y}^{(d)}$ which is given by

$$h^{**}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) = \frac{\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\boldsymbol{\eta}^*\right) Z^{*(d)}(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\boldsymbol{\eta}^*)}{\int_{\mathbb{R}^q}\mathfrak{h}\left(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\boldsymbol{x}\right) Z^{*(d)}(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{-1/2}\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}.$$

The triangle inequality, the transformation theorem and Proposition 3.5.15 imply that

$$\begin{aligned}&\int_{\mathbb{R}^q}\left|h^{**}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*)\right| \mathrm{d}\boldsymbol{\eta}^* \\ &\quad\leq \int_{\mathbb{R}^q}|\phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d) - \phi_q(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^* + \int_{\mathbb{R}^q}\left|h^{**}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)\right| \mathrm{d}\boldsymbol{\eta}^* \\ &\quad= \int_{\mathbb{R}^q}|\phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d) - \phi_q(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^* + \int_{\mathbb{R}^q}\left|h^{*}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*;\Delta_d,\mathrm{I}_q)\right| \mathrm{d}\boldsymbol{\eta}^* \\ &\quad= \int_{\mathbb{R}^q}|\phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d) - \phi_q(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^* + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\end{aligned}$$

for $d \to \infty$, where

$$\begin{aligned}\boldsymbol{\mu}_d &:= \mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}\left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\Delta_d - \tilde{\boldsymbol{\eta}}_d + \boldsymbol{\eta}_0\right), \\ \boldsymbol{\Sigma}_d &:= \mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1}\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}.\end{aligned}$$

Utilizing Pinsker's inequality and the closed form expression of the Kullback-Leibler divergence for normal distributions and again the relation of L^1 and total variation distances results in

$$\begin{aligned}\int_{\mathbb{R}^q}|\phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d) - \phi_q(\boldsymbol{\eta}^*)|\,\mathrm{d}\boldsymbol{\eta}^* &\leq \sqrt{2\int_{\mathbb{R}^q}\log\left(\frac{\phi_q(\boldsymbol{\eta}^*)}{\phi_q(\boldsymbol{\eta}^*;\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)}\right)\phi_q(\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{\eta}^*} \\ &= \sqrt{\mathrm{tr}\left(\boldsymbol{\Sigma}_d^{-1}\right) - q - \boldsymbol{\mu}_d^{\mathrm{T}}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\mu} + \log\left(\det\left(\boldsymbol{\Sigma}_d\right)\right)}.\end{aligned}$$

The consistency of $\tilde{\boldsymbol{\eta}}_d$ and the continuity and regularity of $\mathcal{I}^{(d)}(\,\bullet\,)$ now imply that

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_d^{-1}\right) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} q, \qquad \det\left(\boldsymbol{\Sigma}_d\right) \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 1$$

for $d \to \infty$. Completely analogous to the computation of the asymptotic normality of $\tilde{\boldsymbol{\eta}}_d$ with estimated asymptotic variance we get $\boldsymbol{\mu}_d^{\mathsf{T}}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\mu} = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$. This finally implies

$$\int_{\mathbb{R}^q} \left|\phi_q(\boldsymbol{\eta}^*; \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) - \phi_q(\boldsymbol{\eta}^*)\right| \,\mathrm{d}\boldsymbol{\eta}^* = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$$

and this is what we had to show. □

Proof of Theorem 3.5.2.
The Proof of the EAP part follows the lines that are sketched in Remark 2.1 of Ghosal (1997), the MLE part is partially based on the second part of Theorem 2.2 of Ghosal (1997).

First we discuss the existence. Due to Proposition 3.5.1, we can apply Theorem 3.4.3 which implies that a sequence of measurable MLEs exists.[15] For the Bayes estimators, we define the posterior-risk as mapping

$$R_d\colon \Theta \times \{0,1\}^d \to (0,\infty), \qquad (\boldsymbol{e},\boldsymbol{y}) \mapsto \int_\Theta l(\boldsymbol{e}-\boldsymbol{\eta}) h(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)} = \boldsymbol{y}) \,\mathrm{d}\boldsymbol{\eta},$$

where l denotes the loss-function for a $d \in \mathbb{N}$. By definition, a Bayes estimator with respect to l is a measurable mapping $\boldsymbol{T}_d\colon (\{0,1\}^d, \mathrm{Pot}(\{0,1\}^d)) \to (\Theta, \mathcal{B}(\Theta))$ satisfying

$$R_d(\boldsymbol{T}_d(\boldsymbol{y}), \boldsymbol{y}) = \inf_{\boldsymbol{e}\in\Theta} R_d(\boldsymbol{e}, \boldsymbol{y}), \qquad \text{for all } \boldsymbol{y} \in \{0,1\}^d. \tag{3.120}$$

Since l is by assumption continuous, existence and measurability of the Bayes estimators is completely analogous to the MLE, just replacing $\ell^{(d)}$ by $-R_d$.

The strategy is now as follows: We will prove that the EAP satisfies the condition of Proposition 3.5.16 and, therefore, is a proper centering with limiting distribution $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ of the normalized posterior. Since the conditions of Theorem 3.3.3 are satisfied due to Lemma 3.5.7 and Corollary 3.5.12 (in particular, the corresponding random variable is $\boldsymbol{\xi}+\boldsymbol{c}$ for some $\boldsymbol{c} \in \mathbb{R}^q$ and especially $\boldsymbol{c} = \mathbf{0}$ for the EAP), we directly get that all Bayes estimators are proper centering with a possibly different limiting measure $\mathcal{Q}$. By Ghosh et al. (1994, Proposition 2.2)[16], the limit of the normalized posterior is given by $\mathcal{Q}(\,\bullet\,) = \tilde{\Phi}_q(\,\bullet - \boldsymbol{c})$ for all Bayes estimators. Generally,

$$\mathbb{R}^q \ni \boldsymbol{x} \mapsto \exp\left(\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}\right)$$

is uniquely maximized at $\boldsymbol{x} = \boldsymbol{\xi}$. Therefore, if (GGS1') is satisfied, Ghosal et al. (1995, Corollary 1) implies $\boldsymbol{c} = \mathbf{0}$ for all Bayes estimators (cf. Section 3.3). In the particular case of $q = 1$, we directly get this from Lemma 3.5.7. The change of the normalizing factor follows from the consistency (due to $\boldsymbol{c} = \mathbf{0}$) and Slutsky's theorem (cf. the proof of Proposition 3.5.16). The version of proper centering with convergence in P instead of $\mathsf{P}_{\boldsymbol{\eta}_0}$ follows from

[15] The required compactness of Θ and the continuity of the log-likelihood, i.e. (PCT1) and (PCT2), also imply the existence of the MLE for all $d \in \mathbb{N}$. However, the measurability still remains to show.

[16] This proposition states that limiting distributions can only differ by the location if there is a difference, i.e. the offset $\boldsymbol{c}$ because the EAP ensures that the general limiting distribution is a normal distribution.

Lebegue's theorem of dominated convergence, as $\mathcal{H}$ is a proper probability distribution and any (conditional) probability has upper bound 1. The argument is completely analogous to the proof of Chang and Stout (1991, Theorem 3) and the last part of the proof of Theorem 3.4.1. Finally, we will show explicitly that the MLE satisfies the condition of Proposition 3.5.16, too.

Since $h^*(\,\bullet \mid \boldsymbol{Y}^{(d)})$ is the posterior density of $\boldsymbol{\eta}^*$, we get

$$\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) = \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2} \left(\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \boldsymbol{\eta}_0\right)$$

from the linearity of the standardization. Generally, $\int_{\mathbb{R}^q} \boldsymbol{\eta}^* \phi_q(\boldsymbol{\eta}^*\,;\,\Delta_d, \mathrm{I}_q)\,\mathrm{d}\boldsymbol{\eta}^* = \Delta_d$ holds. So we get from Theorem 3.5.15

$$\begin{aligned}
&\left\| \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2} \left(\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \boldsymbol{\eta}_0\right) - \Delta_d \right\| \\
&\quad = \left\| \int_{\mathbb{R}^q} \boldsymbol{\eta}^* \left(h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*\,;\,\Delta_d, \mathrm{I}_q)\right) \mathrm{d}\boldsymbol{\eta}^* \right\| \\
&\quad \leq \int_{\mathbb{R}^q} \|\boldsymbol{\eta}^*\| \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*\,;\,\Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1), \qquad (3.121)
\end{aligned}$$

which proves the proper centering of the EAP with $\boldsymbol{c} = \boldsymbol{0}$.

It remains to show that property for the MLE. Further applying Theorem 3.4.3 implies the local maximization of the likelihood at the MLE with probability tending to one for $d \to \infty$ and the consistency of the MLE. Therefore, we assume now that $\{\hat{\boldsymbol{\eta}}_d\}_{d\in\mathbb{N}}$ is the MLE as local optimizer of the likelihood for all d and $\hat{\boldsymbol{\eta}}_d = \boldsymbol{\eta}_0 + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$.

We make a second order Taylor expansion of $f(\,\bullet\,) := \boldsymbol{a}^\intercal \nabla \log Z^{*(d)}(\,\bullet\,)$ for an arbitrary $\boldsymbol{a} \in \mathbb{R}^q$ at development point $\boldsymbol{a}_0 = \boldsymbol{0}$ and evaluate at $\boldsymbol{T} = \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0)$ for $d \in \mathbb{N}$ sufficiently large such that $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)$ is regular. The proof follows along the same lines as the one of Lemma 3.5.11 and hence the arguments are shortened. We get

$$f(\boldsymbol{T}) = f(\boldsymbol{0}) + \nabla f(\boldsymbol{0})^\intercal \boldsymbol{T} + \frac{1}{2} \boldsymbol{T}^\intercal \nabla\nabla^\intercal f(\boldsymbol{a}^*) \boldsymbol{T}$$

with $\boldsymbol{a}^* \in \{c\boldsymbol{T} \colon c \in [0,1]\}$. Since $\hat{\boldsymbol{\eta}}_d$ satisfies as local optimizer the likelihood equations

$$\sum_{i=1}^{d} (Y_i - P_i(\hat{\boldsymbol{\eta}}_d)) \nabla \lambda_i(\hat{\boldsymbol{\eta}}_d) = \boldsymbol{0},$$

it follos that

$$f(\boldsymbol{T}) = \boldsymbol{a}^\intercal \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \sum_{i=1}^{d} (Y_i - P_i(\hat{\boldsymbol{\eta}}_d)) \nabla \lambda_i(\hat{\boldsymbol{\eta}}_d) = 0.$$

By definition, we further get $f(\boldsymbol{0}) = \boldsymbol{a}^\intercal \nabla \log Z^{*(d)}(\boldsymbol{0}) = \boldsymbol{a}^\intercal \Delta_d$. Next, we get

$$\begin{aligned}
\nabla f(\boldsymbol{x}) &= \nabla \left(\boldsymbol{a}^\intercal \nabla \log Z^{*(d)}(\boldsymbol{x})\right)^\intercal = \nabla\nabla^\intercal \log Z^{*(d)}(\boldsymbol{x}) \boldsymbol{a} \\
&= \sum_{i=1}^{d} \Bigg(\left(Y_i - \psi'(\lambda_i(\eta_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}))\right) \\
&\qquad \times \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \nabla\nabla^\intercal \lambda_i(\eta_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \\
&\quad - \psi''(\lambda_i(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u})) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \nabla \lambda_i(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}) \\
&\qquad \times \nabla^\intercal \lambda_i(\boldsymbol{\eta}_0 + \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{u}) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \Bigg) \boldsymbol{a}.
\end{aligned}$$

This implies

$$\nabla f(\mathbf{0})^\intercal \boldsymbol{T} = -\boldsymbol{a}^\intercal \boldsymbol{T} + \boldsymbol{a}^\intercal \sum_{i=1}^{d} (Y_i - P_i(\boldsymbol{\eta}_0)) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \, (\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) \, .$$

We suppose $\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \neq \mathbf{0}$ for infinitely many $i \in \mathbb{N}$ (else, the second term of the right-hand side would be zero and is at least bounded by a variant with $\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \neq \mathbf{0}$ for all $i \in \mathbb{N}$). Denote by $\boldsymbol{V}_{ik}$ the kth column of $\nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0)$. Then, as $\nu_{\max}\left(\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right) = \mathcal{O}(d^{-1/2})$, the multivariate central limit theorem implies (compare the proof of Proposition 3.5.10 for a discussion of the conditions, the arguments here are analogues)

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \sum_{i=1}^{d} (Y_i - P_i(\boldsymbol{\eta}_0)) \boldsymbol{V}_{ik} \twoheadrightarrow \boldsymbol{\xi}_k \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

as $d \to \infty$ for some $\boldsymbol{\mu}_k \in \mathbb{R}^q$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{q \times q}$ and $k = 1, \ldots, q$. Hence,

$$\boldsymbol{a}^\intercal \sum_{i=1}^{d} (Y_i - P_i(\boldsymbol{\eta}_0)) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1).$$

Since $\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0 = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$, we get

$$\boldsymbol{a}^\intercal \sum_{i=1}^{d} (Y_i - P_i(\boldsymbol{\eta}_0)) \mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2} \nabla\nabla^\intercal \lambda_i(\boldsymbol{\eta}_0) \, (\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1).$$

Analogously to the discussion of the reminder in the proof of Lemma 3.5.11, it follows that

$$\frac{1}{2} \nabla\nabla^\intercal f(\boldsymbol{a}^*) \boldsymbol{T} = \frac{1}{2} \boldsymbol{a}^\intercal \nabla\nabla^\intercal_{\boldsymbol{T}} \log Z^{*(d)}(\boldsymbol{a}^*) = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(\|\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0\|\right) = o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1),$$

where $\nabla\nabla^\intercal_{\boldsymbol{T}}$ denotes the component-wise derivative of the Hessian in the direction of $\boldsymbol{T}$. Hence, we obtain by rearranging the terms

$$\left(\boldsymbol{a} + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\right)^\intercal \boldsymbol{T} = \left(\boldsymbol{a} + \frac{1}{2} \nabla\nabla^\intercal f(\boldsymbol{a}^*) \boldsymbol{T}\right)^\intercal \boldsymbol{T} = \boldsymbol{a}^\intercal \Delta_d + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1).$$

Since $\boldsymbol{a} \in \mathbb{R}^q$ was arbitrary, we get from an argument of Lehmann and Casella (1998, Lemma 5.2, p. 465) by selecting q linearly independent vectors in $\mathbb{R}^q$ for $\boldsymbol{a}$

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2} (\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) = \Delta_d + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)$$

and in particular

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2} (\hat{\boldsymbol{\eta}}_d - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{D}} \boldsymbol{\xi}, \tag{3.122}$$

for $d \to \infty$. Finally note that the limiting random vector $\boldsymbol{\xi}$ in equation (3.122) is the same limiting vector as that from the EAP since both apply Proposition 3.5.16 and, thus, prove the asymptotic equivalence to Δ_d. □

Proof of Corollary 3.5.3.
The continuity of f and compactness of Θ imply that $\max_{\boldsymbol{\eta}\in\Theta}|f(\boldsymbol{\eta})|$ exists. Hence, the expectation $\mathsf{E}_{\boldsymbol{\eta}_0}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)})$ exists for all $d \in \mathbb{N}$. Developing a Taylor expansion of f at $\boldsymbol{\eta}_0$ with an $\boldsymbol{\eta}'$ between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_0$ results in

$$f(\boldsymbol{\eta}) - f(\boldsymbol{\eta}_0) = \nabla f(\boldsymbol{\eta}_0)^\intercal(\boldsymbol{\eta} - \boldsymbol{\eta}_0) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^\intercal \nabla\nabla^\intercal f(\boldsymbol{\eta}')(\boldsymbol{\eta} - \boldsymbol{\eta}_0).$$

Therefore, we get

$$\begin{aligned}
&\left|\mathsf{E}_{\boldsymbol{\eta}_0}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0)\right| \\
&\quad= \left|\mathsf{E}_{\boldsymbol{\eta}_0}(\nabla f(\boldsymbol{\eta}_0)^\intercal(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \mid \boldsymbol{Y}^{(d)}) + \frac{1}{2}\mathsf{E}_{\boldsymbol{\eta}_0}\Big((\boldsymbol{\eta} - \boldsymbol{\eta}_0)^\intercal \nabla\nabla^\intercal f(\boldsymbol{\eta}')(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \,\Big|\, \boldsymbol{Y}^{(d)}\Big)\right| \\
&\quad\leq \|\nabla f(\boldsymbol{\eta}_0)\| \cdot \|\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \boldsymbol{\eta}_0\| + \frac{1}{2}\max_{\boldsymbol{\eta}\in\Theta}\|\nabla\nabla^\intercal f(\boldsymbol{\eta})\| \cdot \mathsf{E}_{\boldsymbol{\eta}_0}\Big(\|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|^2 \,\Big|\, \boldsymbol{Y}^{(d)}\Big),
\end{aligned}$$

where $\max_{\boldsymbol{\eta}\in\Theta}\|\nabla\nabla^\intercal f(\boldsymbol{\eta})\| < \infty$ by the assumptions regarding f. Theorem 3.5.2 implies that

$$\|\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \boldsymbol{\eta}_0\| \leq \left\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\right\| \left\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}\Big(\mathsf{E}_{\boldsymbol{\eta}_0}(\boldsymbol{\eta} \mid \boldsymbol{Y}^{(d)}) - \boldsymbol{\eta}_0\Big)\right\| = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(\frac{1}{\sqrt{d}}\right).$$

Applications of the transformation theorem and Proposition 3.5.15 imply that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\eta}_0}\Big(\|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|^2 \,\Big|\, \boldsymbol{Y}^{(d)}\Big) &= \int_{\mathbb{R}^q}\left\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\boldsymbol{\eta}^*\right\|^2 h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^* \\
&\leq \left\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)\right\|^{-1}\int_{\mathbb{R}^q}\left\|\boldsymbol{\eta}^*\right\|^2 h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\,\mathrm{d}\boldsymbol{\eta}^* \\
&= \left\|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)\right\|^{-1}\left(\int_{\mathbb{R}^q}\left\|\boldsymbol{\eta}^*\right\|^2 \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q)\,\mathrm{d}\boldsymbol{\eta}^* + o_{\mathsf{P}_{\boldsymbol{\eta}_0}}(1)\right) \\
&= \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}\left(\frac{1}{d}\right) \qquad (3.123)
\end{aligned}$$

and this implies (3.86).[17]

Next we prove (3.87) and therefore have $\eta = \boldsymbol{\eta}$, $\eta_0 = \boldsymbol{\eta}_0$, $f' = \nabla f$ and set $f'' \coloneqq \nabla\nabla^\intercal f$. Using again a Taylor expansion of f at η_0 with η' between η and η_0 results in

$$\begin{aligned}
&\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(f(\eta) \mid \boldsymbol{Y}^{(d)}) - f(\eta_0)\Big) \\
&\quad= \sqrt{\mathcal{I}^{(d)}(\eta_0)}\left(\mathsf{E}_{\eta_0}\left(f'(\eta_0)(\eta - \eta_0) + \frac{f''(\eta')}{2}(\eta - \eta_0)^2 \mid \boldsymbol{Y}^{(d)}\right)\right) \\
&\quad= f'(\eta_0)\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(\eta \mid \boldsymbol{Y}^{(d)}) - \eta_0\Big) + \sqrt{\mathcal{I}^{(d)}(\eta_0)}\mathsf{E}_{\eta_0}\left(\frac{f''(\eta')}{2}(\eta - \eta_0)^2 \mid \boldsymbol{Y}^{(d)}\right).
\end{aligned}$$

The boundedness of f'' and (3.123) imply that

$$\sqrt{\mathcal{I}^{(d)}(\eta_0)}\mathsf{E}_{\eta_0}\left(\frac{f''(\eta')}{2}(\eta - \eta_0)^2 \mid \boldsymbol{Y}^{(d)}\right) = \mathcal{O}_{\mathsf{P}_{\eta_0}}\left(\frac{1}{\sqrt{d}}\right)$$

and, in consequence, (3.87) is true. An application of Theorem 3.5.2 to the right hand side of (3.87) directly implies (3.88) if $f'(\eta_0) \neq 0$. □

[17] We can recognize that even $\mathsf{E}_{\boldsymbol{\eta}_0}(f(\boldsymbol{\eta}) \mid \boldsymbol{Y}^{(d)}) - f(\boldsymbol{\eta}_0) = \mathcal{O}_{\mathsf{P}_{\boldsymbol{\eta}_0}}(d^{-1})$ if $\nabla f(\boldsymbol{\eta}_0) = \mathbf{0}$.

Proof of Theorem 3.5.4.
Lemma 3.5.7 implies that condition (GGS1) is satisfied with $M = B/2$, $m_1 = 0$, $\alpha = 2$ for all $R > 0$. Further, Lemma 3.5.7 implies that condition (GGS2) is satisfied with $g_d(x) = \varepsilon x^2$ for $x \in [0, \infty)$. Condition (PCT1[ii]) implies the conditions stated by GGS on the prior. The claim follows directly from

$$\sum_{d=D}^{\infty} \|\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{-1/2}\|^4 = \sum_{d=D}^{\infty} \mathcal{O}(d^{-2}) = \mathcal{O}(1),$$

and Proposition 1 of Ghosal et al. (1995)[18], where D is the first index such that $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)$ is regular for all $d \geq D$. □

Proof of Theorem 3.5.5.
For each bounded set $K \subset \mathbb{R}^q$ there is a $\delta > 0$ so that $K \subset B_\delta(\mathbf{0})$. Then Proposition 3.5.15 implies that

$$\begin{aligned}
&\sup_{\boldsymbol{t} \in K} \left| \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\, \mathrm{d}\boldsymbol{\eta}^* - \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q)\, \mathrm{d}\boldsymbol{\eta}^* \right| \\
&\quad \leq \sup_{\boldsymbol{t} \in B_\delta(\mathbf{0})} \left| \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)})\, \mathrm{d}\boldsymbol{\eta}^* - \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q)\, \mathrm{d}\boldsymbol{\eta}^* \right| \\
&\quad \leq \sup_{\boldsymbol{t} \in B_\delta(\mathbf{0})} \int_{\mathbb{R}^q} \left| \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*)\, \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \\
&\quad = \sup_{\boldsymbol{t} \in B_\delta(\mathbf{0})} \int_{\mathbb{R}^q} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \\
&\quad \leq \int_{\mathbb{R}^q} \sup_{\boldsymbol{t} \in B_\delta(\mathbf{0})} \exp(\boldsymbol{t}^\intercal \boldsymbol{\eta}^*) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \\
&\quad \leq \int_{\mathbb{R}^q} \exp(\delta \|\boldsymbol{\eta}^*\|) \left| h^*(\boldsymbol{\eta}^* \mid \boldsymbol{Y}^{(d)}) - \phi_q(\boldsymbol{\eta}^*; \Delta_d, \mathrm{I}_q) \right| \mathrm{d}\boldsymbol{\eta}^* \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0,
\end{aligned}$$

for $d \to \infty$. □

Proof of Proposition 3.5.6.
We recognize that the conditions imply (PCT1)–(PCT3). Next, a linearization of the logit with

$$\operatorname{logit}(x) = \operatorname{logit}(x_0) + \frac{x - x_0}{x_0(1 - x_0)} + \mathcal{O}(|x - x_0|^2)$$

for $x, x_0 \in \mathbb{R}$ and $x \to x_0$ leads to

$$\begin{aligned}
&\sqrt{\mathcal{I}^{(d)}(\eta_0)} \left(\operatorname{logit}\left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)}) \right) - \lambda_0(\eta_0) \right) \\
&\quad = \frac{\sqrt{\mathcal{I}^{(d)}(\eta_0)}}{\mathsf{Var}(Y_0 \mid \eta_0)} \left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)}) - P_0(\eta_0) \right) \\
&\qquad + \sqrt{\mathcal{I}^{(d)}(\eta_0)} \mathcal{O}_{\mathsf{P}_{\eta_0}} \left(\left| \mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)}) - P_0(\eta_0) \right|^2 \right)
\end{aligned} \tag{3.124}$$

[18] Adopted to the notation and framework of this work, Proposition 1 of Ghosal et al. (1995) states: If (GGS1) and (GGS2) are satisfied and if the prior satisfies the conditions of GGS, then the posterior is weakly consistent. If there is moreover some $s > 0$ such that $\sum_{d=1}^{\infty} \|\boldsymbol{S}_d\|^s < \infty$, where $\boldsymbol{S}_d$ is the normalizing factor of the posterior (cf. Definition 3.3.1), then the posterior is strongly consistent. Note that Ghosal et al. (1995) define the consistency of the posterior by the convergence of the posterior probability of complements of neighborhoods of $\boldsymbol{\eta}_0$ (i.e. almost sure or in probability for strong or weak consistency, respectively). However, this implies the statement of Theorem 3.5.4.

with $P_0 \colon \Theta \to (0,1)$, $P_0(\eta) := (1+\exp(-\lambda_0(\eta)))^{-1}$, $x_0 = P_0(\eta_0)$ and $x = \mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)})$. Note that

$$\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)}) = \mathsf{E}(Y_0 \mid \boldsymbol{Y}^{(d)}) = \mathsf{E}\left(\mathsf{E}(Y_0 \mid \eta) \,\middle|\, \boldsymbol{Y}^{(d)}\right) = \mathsf{E}(P_0(\eta) \mid \boldsymbol{Y}^{(d)}).$$

With $f = P_0$ and $f'(\eta_0) = \mathsf{Var}(Y_0 \mid \eta_0)\alpha_{01}$, equation (3.87) in Corollary 3.5.3 implies that

$$\begin{aligned}&\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)}) - P_0(\eta_0)\Big) \\ &\quad = \mathsf{Var}(Y_0 \mid \eta_0)\alpha_{01}\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(\eta \mid \boldsymbol{Y}^{(d)}) - \eta_0\Big) + \mathcal{O}_{\mathsf{P}_{\eta_0}}\left(\frac{1}{\sqrt{d}}\right).\end{aligned}$$

Additionally, the second term on the right hand side of (3.124) also is $\mathcal{O}_{\mathsf{P}_{\eta_0}}(d^{-1/2})$ by (3.86) in Corollary 3.5.3. We therefore get that

$$\begin{aligned}&\sqrt{\mathcal{I}^{(d)}(\eta_0)}\left(\operatorname{logit}\left(\mathsf{P}(Y_0 = 1 \mid \boldsymbol{Y}^{(d)})\right) - \lambda_0(\eta_0)\right) \\ &\quad = \alpha_{01}\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(\eta \mid \boldsymbol{Y}^{(d)}) - \eta_0\Big) + \mathcal{O}_{\mathsf{P}_{\eta_0}}\left(\frac{1}{\sqrt{d}}\right).\end{aligned}$$

From (3.121) in the proof of Theorem 3.5.2 we know that

$$\sqrt{\mathcal{I}^{(d)}(\eta_0)}\Big(\mathsf{E}_{\eta_0}(\eta \mid \boldsymbol{Y}^{(d)}) - \eta_0\Big) = \frac{1}{\sqrt{\mathcal{I}^{(d)}(\eta_0)}}\sum_{i=1}^{d}\alpha_{i1}(Y_i - P_i(\eta_0)) + o_{\mathsf{P}_{\eta_0}}(1)$$

for $d \to \infty$. We finish the proof by replacing $\sqrt{\mathcal{I}^{(d)}(\eta_0)}$ with σ_d^{-1}. □

Chapter 4

Estimation for Multidimensional Logit IRT Models for Many Binary Items

We have seen in Section 1.3 that marginal maximum likelihood estimation (MMLE) is an important approach to fit IRT models and that the expected a-posteriori estimator (EAP) is an important estimator for the latent vectors after the measurement model is fitted (e.g. using the MMLE). In Section 2.3 we have then proved that the MMLE is consistent and asymptotically efficient for $n \to \infty$ for a large class of IRT models, especially the important logit models. This has been assumed to be true in the literature so far, however, we were the first to prove it. Important new results of Chapter 3 were the consistency and asymptotic efficiency of the EAP for $d \to \infty$ under appropriate nonparametric conditions on the IRT model.

For the practical application of the EAP and MMLE, the q-dimensional integrals have to be approximated numerically because the integrals are only explicitly evaluable in very special cases. An approximation that yields good results but remains computationally feasible is especially interesting for MMLE because of the number of integrals that have to be evaluated during an iterative procedure, often the EM-algorithm. As we reviewed in Section 1.3, to find such an approximation can be a quite difficult task and especially for models with many factors it requires advanced numerical techniques.

One important special case of a model with explicitly given $\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y})$ and $\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{y})$, $\boldsymbol{y} \in \{0,1\}^d$, is the logit model with the assumption of a normal distribution of the latent vector given one response pattern which we discussed in detail in Section 1.2 (cf. Assumption (A1) on page 19). For this model, however, MMLE (called MPMLE for manifest probabilities MLE in this case) has to take into account all possible 2^d response patterns, which makes the approach unfeasible for a large number of items.[1] The posterior normality assumption (or its asymptotic justification) hence only solves the integration problems for logit models in theory but not in practice.

By focusing on the important logit models, we derive a new EM-algorithm for MMLE in this chapter called the *posterior normal approximated Monte Carlo EM (aMCEM)-algorithm*. It is feasible for arbitrary latent dimension q and provides a good approximation for large d. The aMCEM-algorithm is based on the proper centering theory (PCT) from the last chapter (cf. Section 3.5). The overall procedure bears similarities to the efficient normal importance sampling MCEM-algorithm (cf. e.g. Schilling and Bock, 2005). However, it increases computational efficiency by eliminating of the application of an iterative optimization procedure for each unique response pattern in each E-step. This is done using an approximation of the EAP that can easily be evaluated. Additionally, the approximation of posterior by a normal distribution simplifies the M-step in such a way that the

[1] The implementation of the MPMLE from the R-package **plRasch** which we used to fit the Rasch model could not be applied for $d > 20$ on the computer we used for this work.

overall procedure can easily be implemented using binary logistic regression software. This makes the aMCEM-algorithm particularly interesting for practitioners without access to an already implemented procedure to fit high dimensional logit models and without much experience in software development.

Current EM-algorithms for high dimensional latent variable models locate the quadrature points for the posterior expected values for each unique response pattern around the MAP in order to reduce the number of quadrature points needed (the so called adaptive Gauss–Hermite quadrature methods, cf. e.g. Schilling and Bock, 2005). This means the E-step requires an iterative optimization for each unique response pattern. The major problem when accelerating the E-step and simplifying its implementation is therefore to eliminate this task. Rizopoulos (2012) hastes the E-step by only updating the quadrature points in the first step. While this still requires the implementation of the MAP, the quadrature points might be less optimal if the parameters are far away from the optimal ones – this necessitates a larger number of quadrature points. In order to allow the adaption of the quadrature points in each E-step but still accelerate the full procedure we instead suggest using a linear approximation of the EAP which is based on the component statistics and the PCT.

With using this approach we derive an estimator for $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \in \mathbb{R}^q$ which underlie the observed response vectors $\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n$, $\boldsymbol{y}_k \in \{0,1\}^d$, $k = 1, \ldots, n$. This estimator can be considered an approximation of the EAP (aEAP) and is, therefore, suitable for centering the posterior distribution. The aEAP can very easily be computed if assumptions regarding the first two central moments of the structural model are given – assumptions which are required to solve the location and scale indeterminacy of the exploratory MIRT models on which we focus in this work. A simulation study we conduct further supports using the aEAP to estimate $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \in \mathbb{R}^q$.

When comparing the aMCEM-algorithm to current specifications of EM-algorithms to compute the MMLE it is expected to observe a competitive performance of the aMCEM-algorithm for large q and a sufficiently large d to justify the asymptotic approximations. This is especially expected for $q \geq 8$ which is the current area of computationally difficult problems. A detailed comparison to the efficient normal importance sampling MCEM-algorithms reveals the similarities and differences between the two approaches. We see that the aMCEM-algorithm generally has some interesting properties in comparison to the current alternative approaches. The aMCEM overall is an interesting approach for high dimensional IRT settings if a logit model is assumed.

This chapter is organized as follows: First, we briefly review results and observations regarding the component statistic and relate the statistic to the EAP for logit models assuming (A1). Then we derive the new and simple linear approximation of the EAP (aEAP) for more general logit models in Section 4.2. Also in this section we discuss the simulation study which supports the applicability of the aEAP.

In Section 4.3 we discuss the aMCEM-algorithm for fitting (multidimensional) logit IRT models. In Subsection 4.3.1 we formulate the required assumptions, introduce the necessary notation and briefly recall important equations and properties of general EM-algorithms to compute the MMLE of logit IRT models. In Subsection 4.3.2 we derive the aMCEM-algorithm using the PCT and the aEAP. Example applications of the aMCEM to two simulated data sets in Subsection 4.3.3 illustrate the algorithm and that it is in general suitable for high dimensional settings. In Subsection 4.3.4 we review current approaches to EM-algorithms for computing the MMLE and compare them to the aMCEM-algorithm.

Finally, we summarize and discuss the results of this chapter in Section 4.4 and point

out interesting future research questions.

4.1 Review of the Component Statistics and Linear Expected A-Posteriori Estimators

We suppose that the following model assumption holds.

(M) Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be an independent sample with $\boldsymbol{Y}_k \sim \mathcal{P}(\boldsymbol{\eta}_k)$ for vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \in \Theta \subseteq \mathbb{R}^q$, $n, d, q \in \mathbb{N}$. Further suppose that a logit model holds with

$$\lambda_i(\boldsymbol{\eta}) = \log\left(\frac{P_i(\boldsymbol{\eta})}{1 - P_i(\boldsymbol{\eta})}\right) = \alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}, \qquad \boldsymbol{\eta} \in \Theta, \quad i = 1, \ldots, d, \tag{4.1}$$

for some parameter values $\alpha_{i0} \in \mathbb{R}$ and $\boldsymbol{\alpha}_i \in \mathbb{R}^q$, $i = 1, \ldots, d$.

We also suppose that the values of $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ are known. Now we focus on the estimation of the unknown but constant vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \in \Theta$.[2]

We set

$$\boldsymbol{A} \coloneqq (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d)^\intercal \in \mathbb{R}^{d \times q},$$

such that row i of the matrix $\boldsymbol{A}$ is given by $\boldsymbol{\alpha}_i^\intercal$. With this notation, the *component statistics* (*components* for short) are given by setting

$$\boldsymbol{X}_k \coloneqq \boldsymbol{A}^\intercal \boldsymbol{Y}_k, \qquad k = 1, \ldots, n. \tag{4.2}$$

One important property of the components is the fact that $\boldsymbol{X}_k$ is the (vector valued) minimal sufficient statistic for $\boldsymbol{\eta}_k$ given that the logit model holds, $k = 1, \ldots, n$ (cf. e.g. Bartholomew et al., 2011, Section 2.5 and equation (4.6)).

Additionally, it can often be observed that the posterior expected value of the latent variables has a nearly linear dependence from the components (cf. e.g. Knott and Albanese, 1993; Bartholomew, 1998, and page 42 of Bartholomew et al., 2011).[3] This is closely related to the following property of the components.

If we suppose that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are drawn from a distribution $\mathcal{H}$ on Θ such that $\mathcal{H}(\bullet \mid \boldsymbol{Y}_k = \boldsymbol{y}_0) = \mathcal{N}_q(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ for an arbitrary $\boldsymbol{y} \in \{0,1\}^d$ and $k \in \{1, \ldots, n\}$,[4] then the expected a-posteriori estimators (EAP, cf. (1.54)) of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are given by

$$\tilde{\boldsymbol{\eta}}_k = \boldsymbol{x}_0 + \boldsymbol{\Sigma} \boldsymbol{X}_k, \qquad k = 1, \ldots, n, \tag{4.3}$$

for some vector $\boldsymbol{x}_0 \in \mathbb{R}^q$ (cf. Proposition 1.2.3 for further details).[5] This means that under posterior normality the EAP is an affine transformation of the components and, due to the linearity of the components, also an affine transformation of the response vectors.

[2] In many practical settings, α_{i0} and/or $\boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ might be estimated instead of being specified a-priori. However, similarly to the common approach to estimate $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ which we discussed in Subsection 1.3.4, we consider $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ to be known and simplify the notation by still writing $\alpha_{i0}, \boldsymbol{\alpha}_i$ instead of $\hat{\alpha}_{i0}, \hat{\boldsymbol{\alpha}}_i$ if it is estimated.

[3] Knott and Albanese (1993) illustrate an example where it cannot be observed.

[4] This means that we assume that condition (A1) on page 19 holds.

[5] Due to the properties of the normal distribution, the right-hand side of (4.3) is also the MAP estimator in this case and, for $q = 1$, the median a-posteriori estimator.

Bartholomew et al. (2011, Section 2.15) proved that the EAP is the gradient of a convex function of the component vector regardless of the distribution of the latent vector, thereby concluding that the components are useful for ranking the individuals $1, \ldots, n$. They further emphasized the importance of the components by linking (4.3) with the conjecture of asymptotic posterior normality (APN) of latent vectors. The validity of this cenjecture is shown by our proof of the APN of latent vectors under suitable conditions in the last chapter. If we apply the model under the common identification constraints, i.e. $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$ for $\boldsymbol{\eta} \sim \mathcal{H}$, then $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ can be reliable measures of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ for relative comparisons but are inappropriate point estimates of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ due to neglecting $\boldsymbol{x}_0$ and $\boldsymbol{\Sigma}$ in (4.3), which are commonly unknown. Moreover, the posterior covariance matrix can depend on the observed response pattern and is then not homoscedastic over the different patterns. Bartholomew (1998) observed nearly constant posterior variances. This suggests that the posterior inhomogeneity can potentially be unproblematic but it is not obvious how to determine a good asymptotic approximation of $\boldsymbol{\Sigma}$.

In logit models which assume the posterior normality of latent vectors like the models in Section 1.2, the homogeneity of the posterior covariance of latent vectors is part of the model (cf. again Proposition 1.2.3). Hence, assuming that one of these models holds, (4.3) can be used to estimate $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ as also suggested by Li (2010) and Hessen (2012). Both examined models where $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d$ are known vectors in $\{0, 1\}^d$. Because $\boldsymbol{x}_0$ is unidentifiable in their models, it has to be set arbitrarily. Hessen (2012) set $\boldsymbol{x}_0 = \mathbf{0}$ and estimated $\boldsymbol{\Sigma}$ jointly with $\alpha_{10}, \ldots, \alpha_{d0}$ during the model fitting. Hence, he obtained the estimates

$$\tilde{\boldsymbol{\eta}}_k = \widehat{\boldsymbol{\Sigma}} \boldsymbol{X}_k, \qquad k = 1, \ldots, n,$$

where $\widehat{\boldsymbol{\Sigma}}$ is the MLE of $\boldsymbol{\Sigma}$. In contrast, Li (2010, p. 91–92) suggested using $\boldsymbol{x}_0 = -\widehat{\boldsymbol{\Sigma}} \overline{\boldsymbol{X}}$, where $\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{X}_k$ such that the average of $\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n$ is zero. Her estimator for $\boldsymbol{\eta}_k$ then is

$$\tilde{\boldsymbol{\eta}}_k = \widehat{\boldsymbol{\Sigma}} \Big(\boldsymbol{X}_k - \overline{\boldsymbol{X}} \Big), \qquad k = 1, \ldots, n, \tag{4.4}$$

where $\hat{\boldsymbol{\Sigma}}$ again is an estimate of $\boldsymbol{\Sigma}$.[6] Using the property

$$\mathsf{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Sigma} \mathsf{Cov}(\boldsymbol{X}) \boldsymbol{\Sigma} + \boldsymbol{\Sigma},$$

where $\boldsymbol{X} = \boldsymbol{A}^{\intercal} \boldsymbol{Y}$ with $\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} \sim \mathcal{H}$, Li (2010) obtained an estimate of $\mathsf{Cov}(\boldsymbol{\eta})$ by replacing $\boldsymbol{\Sigma}$ with $\widehat{\boldsymbol{\Sigma}}$ and estimating $\mathsf{Cov}(\boldsymbol{X})$ with the sample covariance matrix of the observed components. During her simulation studies on analyzing data with covariates, Li (2010) also analyzed (4.4) and one extension of it which includes covariates for $q = 1$. In these simulation studies, the estimator had a good recovery of true values in $[-2, 2]$ and she concluded that it seems to be practically unbiased. In the corresponding figures, there is a visible underestimation of larger values (true values between ≈ -4 and ≈ 4 existed, the great majority were between ≈ -2 and ≈ 2), hence, relatively extreme values are estimated less well. However, the more extreme values are often problematic to estimate: They lead to a higher probability of response vectors in which all responses are either correct or wrong – in these cases we get $\|\hat{\boldsymbol{\eta}}_k\| = \infty$ for the MLE $\hat{\boldsymbol{\eta}}_k$ of $\boldsymbol{\eta}_k$.

Summing up, we see that the components provide a good foundation for the creation of estimators for $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$. Moreover, we can see the potential for approximating the EAP

[6] She used MLE and a pseudo likelihood estimation for $\boldsymbol{\Sigma}$, cf. the next chapter.

linearly using the components for logit IRT models with many items. In the next section, we discuss a simple approach for such a linear approximation and analyze its performance in a small simulation study.

4.2 Approximate Linear Expected A-Posteriori Estimators

4.2.1 Derivation of the Estimator

As in the last section we suppose that (M) holds and that the values of $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ are known. We now suppose that the unobservable vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are drawn from a distribution $\mathcal{H}$ with

$$\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0} \qquad \text{and} \qquad \mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q \tag{4.5}$$

for $\boldsymbol{\eta} \sim \mathcal{H}$, i.e. we consider the most common identification constraints on the location and scale indeterminacy of the prior distribution $\mathcal{H}$.[7]

Remark 4.2.1
Note that if $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ are actually specified a-priori, then there is no location and scale indeterminacy of the prior distribution $\mathcal{H}$ such that $\mathsf{E}(\boldsymbol{\eta})$ and $\mathsf{Cov}(\boldsymbol{\eta})$ do not have to be specified. If the first two central moments of $\mathcal{H}$ are known, then we can replace them in equation (4.5). If they are not known, then the following approach cannot be used. In many common cases, the estimation of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ is the second step after estimating $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$; then identification constraints for the location and scale have to be assumed and (4.5) is no restriction. In this case, we consider the estimates $\hat{\alpha}_{i0}, \hat{\boldsymbol{\alpha}}_i$ of $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ the known values following the two-step approach discussed in Subsection 1.3.4. To simplify the notation and to incorporate the assumption that they are the true values, we still write $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$. This assumption is important for applying the theory of Chapter 3. ◇

Additionally, we suppose that (PCT1) of Section 3.5.1 is satisfied and we make to following further assumption:

(M') The model parameters $\alpha_{10}, \ldots, \alpha_{d0}$ and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d$ are bounded for $d \to \infty$. Additionally, $\|\boldsymbol{\alpha}_i\|$ does not converge to 0 for $d \to \infty$ and $\liminf_{d\to\infty} \nu_{\min}\left(\frac{1}{d}\sum_{i=1}^{d} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^{\mathsf{T}}\right) > 0$ where $\nu_{\min}(\boldsymbol{B})$ denotes the smallest eigenvalue of a symmetric matrix $\boldsymbol{B}$.

Assumption (M') is realistic because it formulates the requirements that the items do not tend to become arbitrarily difficult or easy and that the latent vectors also have an influence on the response probability for long tests – without these requirements, a precise measurement of the latent vectors is impossible. Additionally, it requires that the probabilities of successful response neither are either zero or one for any item (i.e. infinite discrimination if $q = 1$) nor approach this case for any latent vector in long tests. Even if arbitrarily high discrimination might be desired, it is without doubt nearly impossible to create in many practical cases and it is a usual assumption.

[7] Notice that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are still constant. They correspond to $\boldsymbol{\eta}_0$ in Chapter 3 where we discussed the estimation for a single response vector. Analogously to Chapter 3 $\mathcal{H}$ is the prior distribution for estimating $\boldsymbol{\eta}_k$ from the response vector $\boldsymbol{Y}_k$, $k = 1, \ldots, n$. Here, we additionally assume that $\mathcal{H}$ is "correct" in the sense that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ were drawn prior to the generation of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from $\mathcal{H}$. After this generation of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, the vector $\boldsymbol{\eta}_k$ determined the conditional distribution of $\boldsymbol{Y}_k$, $k = 1, \ldots, n$.

The conditions (M) and (M') directly imply the conditions (PCT2) and (PCT3) of Section 3.5.1 so that we can apply the complete theory of Chapter 3. As we discussed there, the assumption of Lipschitz-continuity in (PCT1) is realistic and the compactness assumption of Θ does not make practical damage. However, many common models violate the compactness assumption such as all models that assume that $\mathcal{H} = \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$. At the end of this section we conduct some simulations which take violations of the compactness assumption into account. Next we derive the approximate estimator.

First, we know from the proof of Theorem 3.5.2 that

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_k)^{1/2}\Big(\int_\Theta \boldsymbol{\eta}\, \mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y}_k)\,\mathrm{d}\boldsymbol{\eta} - \boldsymbol{\eta}_k\Big) = \mathcal{I}^{(d)}(\boldsymbol{\eta}_k)^{-1/2}\sum_{i=1}^d \boldsymbol{\alpha}_i\Big(Y_i - P_i(\boldsymbol{\eta}_k)\Big) + o_{\mathsf{P}_{\boldsymbol{\eta}_k}}(1)$$

for $d \to \infty$ and $k = 1, \ldots, n$, where

$$\mathcal{I}^{(d)}(\boldsymbol{\eta}_k) = \sum_{i=1}^d P_i(\boldsymbol{\eta}_k)\big(1 - P_i(\boldsymbol{\eta}_k)\big)\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\intercal.$$

This directly implies

$$\int_\Theta \boldsymbol{\eta}\, \mathfrak{h}(\boldsymbol{\eta} \mid \boldsymbol{Y}_k)\,\mathrm{d}\boldsymbol{\eta} \approx \boldsymbol{x}_k + \boldsymbol{\Sigma}_k\boldsymbol{X}_k$$

for large d where $\boldsymbol{X}_k$ is the component vector and $\boldsymbol{x}_k$ and $\boldsymbol{\Sigma}_k$ are given by

$$\boldsymbol{\Sigma}_k \coloneqq \mathcal{I}^{(d)}(\boldsymbol{\eta}_k)^{-1} \qquad \text{and} \qquad \boldsymbol{x}_k \coloneqq \boldsymbol{\eta}_k - \mathcal{I}^{(d)}(\boldsymbol{\eta}_k)^{-1}\sum_{i=1}^d \boldsymbol{\alpha}_i P_i(\boldsymbol{\eta}_k),$$

$k = 1, \ldots, n$. The theory neither provides direct results on the variation between $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n$ nor on that between $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. However, the independent analysis of $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is not required because it is more important that

$$\boldsymbol{x}_k + \boldsymbol{\Sigma}_k\boldsymbol{X}_k \approx \boldsymbol{x}_m + \boldsymbol{\Sigma}_m\boldsymbol{X}_m \qquad \text{if} \qquad \boldsymbol{X}_k \approx \boldsymbol{X}_m \tag{4.6}$$

for $k, m \in \{1, \ldots, n\}$. The empirical observations by Knott and Albanese (1993) and Bartholomew (1998) mentioned in the last section and the success of the logit models assuming posterior normality with homogeneous posterior covariance suggest that (4.6) can be assumed for moderately large d (the approximate linear dependence of the EAP from the components is even documented for data sets with $d = 5$).

This motivates the attempt to estimate the vectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ using the approach

$$\tilde{\boldsymbol{\eta}}_k = \boldsymbol{m} + \boldsymbol{S}\boldsymbol{X}_k, \qquad k = 1, \ldots, n, \tag{4.7}$$

where $\boldsymbol{m} \in \mathbb{R}^q$, $\boldsymbol{S} \in \mathbb{R}^{q\times q}$ are currently unknown additional parameters. Next we discuss how appropriate values for $\boldsymbol{m}$ and $\boldsymbol{S}$ can be obtained.

Assume that $\hat{\boldsymbol{\eta}}_k^{(d)}$ is a consistent estimator for $\boldsymbol{\eta}_k$ for $d \to \infty$, $k = 1, \ldots, n$. Then by definition $\hat{\boldsymbol{\eta}}_k^{(d)} \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_k}} \boldsymbol{\eta}_k$ for $d \to \infty$, $k = 1, \ldots, n$. According to the identification condition (4.5) and the weak law of large numbers

$$\frac{1}{n}\sum_{k=1}^n \boldsymbol{\eta}_k \xrightarrow{\mathsf{P}} \mathsf{E}(\boldsymbol{\eta}) = \mathbf{0} \qquad \text{holds for } n \to \infty.$$

These two properties imply

$$\frac{1}{n}\sum_{k=1}^{n}\hat{\boldsymbol{\eta}}_k^{(d)} =: \overline{\boldsymbol{\eta}}_{d,n} \xrightarrow{\mathsf{P}} \mathsf{E}(\boldsymbol{\eta}) = \mathbf{0} \tag{4.8}$$

for $d, n \to \infty$. Analogously, we get

$$\frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{\boldsymbol{\eta}}_k^{(d)} - \overline{\boldsymbol{\eta}}_{d,n}\right)\cdot\left(\hat{\boldsymbol{\eta}}_k^{(d)} - \overline{\boldsymbol{\eta}}_{d,n}\right)^{\mathsf{T}} =: \boldsymbol{S}_{d,n} \xrightarrow{\mathsf{P}} \mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q \tag{4.9}$$

for $d, n \to \infty$. Especially the EAP is consistent with the identification constraints in this sense. Its linear approximation (4.7) has to fulfill (4.8) and (4.9), too. Therefore, an easy way to define $\boldsymbol{m}$ and $\boldsymbol{S}$ is given by requiring that (4.8) and (4.9) hold. This is satisfied if $\boldsymbol{m}$, $\boldsymbol{S}$ are obtained by solving the estimating equations

$$\frac{1}{n}\sum_{k=1}^{n}\Big(\boldsymbol{m} + \boldsymbol{S}\boldsymbol{X}_k\Big) = \mathbf{0}, \tag{4.10}$$

$$\frac{1}{n-1}\sum_{k=1}^{n}\Big(\boldsymbol{m} + \boldsymbol{S}\boldsymbol{X}_k\Big)\cdot\Big(\boldsymbol{m} + \boldsymbol{S}\boldsymbol{X}_k\Big)^{\mathsf{T}} = \mathrm{I}_q, \tag{4.11}$$

for the transformation-parameters $\boldsymbol{m} \in \mathbb{R}^q$ and $\boldsymbol{S} \in \mathbb{R}^{q\times q}$, where $\boldsymbol{S}$ is positive definite. Clearly, a solution is given by a standardization of the components $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with

$$\boldsymbol{m} \coloneqq -\widehat{\boldsymbol{\Gamma}}^{-1/2}\overline{\boldsymbol{X}} \qquad \text{and} \qquad \boldsymbol{S} \coloneqq \widehat{\boldsymbol{\Gamma}}^{-1/2},$$

where $\overline{\boldsymbol{X}} \coloneqq \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{X}_k$ and

$$\widehat{\boldsymbol{\Gamma}} \coloneqq \frac{1}{n-1}\sum_{k=1}^{n}\Big(\boldsymbol{X}_k - \overline{\boldsymbol{X}}\Big)\cdot\Big(\boldsymbol{X}_k - \overline{\boldsymbol{X}}\Big)^{\mathsf{T}}.$$

Finally, the *linear approximated expected a-posteriori estimator* (aEAP) of $\boldsymbol{\eta}_k$ in a logit model is given by

$$\tilde{\boldsymbol{\eta}}_k = \widehat{\boldsymbol{\Gamma}}^{-1/2}\Big(\boldsymbol{X}_k - \overline{\boldsymbol{X}}\Big), \tag{4.12}$$

$k = 1, \ldots, n$.

Remark 4.2.2

1. The estimator (4.12) does not depend on the intercept parameters α_{i0}, $i = 1, \ldots, d$.

2. If $\mathcal{H}$ has any first two central moments other than those given in (4.5), then we can simply adapt (4.10) and (4.11) to these moments of $\mathcal{H}$. The resulting aEAP then is a transformation of the components such that the transformed components have the formulated first two central moments of $\mathcal{H}$. Hence, it is not restrictive to assume (4.5). ◇

4.2.2 Simulation Study

The aEAP uses several approximations for which cannot always be said with certainty that they are reliable. In particular, the approximate linearity is not generally valid as Knott and Albanese (1993) showed. Hence, the aEAP only has a heuristic character in its current state. We therefore conduct a simulation study to verify estimator (4.12). Two specific questions which we want to answer with this study are:

1. How does the precision of the estimator evolve for an increasing test length?
2. How strongly do this evolution and the precision of the estimates depend on the distribution of the latent vectors?

4.2.2.1 Setups, Methods and Quality Measures

In order to define several distributions from which we generate the latent vectors in the simulations, we specify some random variables:

$$\begin{aligned} \xi_1 \sim& \mathcal{N}(0,1), \\ \xi_2 + 1 \sim& \text{Exp}(1), \\ (\xi_3 \mid P = p) \sim& \mathcal{N}(p \cdot 0.9, 0.19), \quad P \sim \mathcal{U}(\{-1,1\}), \\ \xi_4 :=& \frac{T - \mathsf{E}(T)}{\sqrt{\mathsf{Var}(T)}}, \end{aligned}$$

where T is a random variable with an Exp(1) distribution with truncation to the interval $(0,3]$. We assume that these random variables are defined over the probability space $(\Omega, \mathcal{A}, \mathsf{P})$. Then we define the following prior distributions

$$\begin{aligned} \mathcal{H}_v :=& \mathsf{P}^{\xi_v}, \quad v = 1,\dots,4, \\ \mathcal{H}_5 :=& \mathcal{N}_3(\mathbf{0}, \mathrm{I}_3), \\ \mathcal{H}_6 :=& \bigotimes_{v=2}^{4} \mathsf{P}^{\xi_v}, \\ \mathcal{H}_7 :=& \mathsf{P}^{\xi_1} \otimes \mathsf{P}^{\xi_3} \otimes \mathsf{P}^{\xi_4}, \end{aligned} \tag{4.13}$$

where P^{ξ_v} is the distribution of ξ_v, $v = 1,\dots,4$. In the following we illustrate some thoughts on how the distributions where selected to study an appropriate set of prior distributions. First, $\mathcal{H}_1,\dots,\mathcal{H}_4$ are priors for the one dimensional case with $q = 1$ and $\mathcal{H}_5,\dots,\mathcal{H}_7$ for the three dimensional case with $q = 3$. Hence, we can get inside to the one and a multidimensional case. The priors $\mathcal{H}_1$ and $\mathcal{H}_5$ are standard normal distributions and the commonly selected priors for $\boldsymbol{\eta}$ in IRT. $\mathcal{H}_2$ is strongly skew and large latent variables occur with a higher probability than for a normal distribution. Especially, $\mathcal{H}_2$ implies that the majority of the latent variables are close to zero and several latent variables are notably large while there are no latent traits which are smaller than -1. $\mathcal{H}_3$ is symmetric but bi-modal such that it is an example for a distribution with more than one mode. $\mathcal{H}_4$ is skew such as $\mathcal{H}_2$ but it has a bounded support such that no extremely large latent variables can be generated by $\mathcal{H}_4$. Moreover, $\mathcal{H}_4$ is the only examined prior with compact support such that we can study the importance of this restriction of (PCT1) by comparing the results to $\mathcal{H}_1,\dots,\mathcal{H}_3$. $\mathcal{H}_6$ and $\mathcal{H}_7$ are products of $\mathcal{H}_1,\dots,\mathcal{H}_4$ such that the comparison of the results for there priors

to the results for $\mathcal{H}_5$ allow us to estimate the influence of deviations from the (commonly supposed) multivariate standard normal distribution for estimating latent vectors with the aEAP. Besides the general differences of this setting to that of Li (2010), we study more different prior distributions (Li (2010) only examined models with $q = 1$ and normal distributions as well as normal linear models for $\mathcal{H}$, partially depending on covariates, which we do not consider). This will give us some insight into situations in which a model like $\mathcal{H} = \mathcal{N}(0, 1)$ or $\mathcal{H} = \mathcal{N}_3(\mathbf{0}, \mathrm{I}_3)$ is inappropriate.

The parameters of the logit measurement model to generate the response data are specified as follows. For a given number d of items, the intercept parameters $\alpha_{10}, \ldots, \alpha_{n0}$ are equally spaced between -2 and 2. We further generate the elements of $\boldsymbol{A}$ as follows:

- for $q = 1$, $\alpha_{11}, \ldots, \alpha_{d1} \overset{\text{iid}}{\sim} \mathcal{U}(\,(0.8,\, 1.2)\,)$,
- for $q = 3$, first column: $\alpha_{11}, \ldots, \alpha_{d1} \overset{\text{iid}}{\sim} \mathcal{U}(\,(-0.5, 1)\,)$,
  second and third column: random permutations of the first column.

To study the quality of the aEAP estimator (4.12), we examine data sets with $n = 100$ responses to d items for $d = 10, \ldots, 100$ in steps of ten for all seven prior distributions. For each d and each prior, we generate $M = 1000$ data sets from the specified distribution and average the values of the *average absolute error* (AAE) and the root mean squared error (RMSE) over the M data sets to measure the quality of the aEAP, where the AAE and RMSE for one data set are defined as

$$\text{AAE} := \frac{1}{qn} \sum_{j=1}^{q} \sum_{k=1}^{n} \left| \tilde{\eta}_{jk} - \eta_{jk} \right|, \qquad \text{RMSE} := \sqrt{\frac{1}{qn} \sum_{j=1}^{q} \sum_{k=1}^{n} \left(\tilde{\eta}_{jk} - \eta_{jk} \right)^2}$$

η_{jk} denotes the *true* value of the latent variable j of subject k (which are generated by the prior distribution) and $\tilde{\eta}_{jk}$ denotes its estimate by the aEAP, $j = 1, \ldots, q$, $k = 1, \ldots, n$.

4.2.2.2 Results

The resulted AAE and RMSE valued are visualized in Figure 4.1. Further, in Figure 4.2 we see four scatter-plots, which compare the true values to their estimates for a single sample of $n = 100$ responses to $d = 60$ items for the one dimensional prior distributions.

The AAE values and the RMSE values are remarkably similar. The AAE and RMSE values decrease in all cases and the evolution is qualitatively similar. Especially, we see that the trend of all curves is $\mathcal{O}(d^{-1/2})$ and this is exactly what we expected for the aEAP. For $q = 1$ we always get lower AAE and RMSE values than for $q = 3$. This difference between the cases $q = 1$ and $q = 3$ also is what we expected because we estimate three times as many parameters for $q = 3$ than for $q = 1$ from equally sized binary data sets such that more information is required to obtain estimates of the same precision.

In the one dimensional case, we observe that we received the most precise estimates in terms of the AAE and RMSE for $\mathcal{H}_3$. This is followed by $\mathcal{H}_1$ and $\mathcal{H}_4$ whose latent variables are nearly equally well estimated with slightly worse results for $\mathcal{H}_4$. We obtained the worst estimates in the case $q = 1$ for $\mathcal{H}_2$. In the three dimensional case, we observe that the AAE and RMSE values for all priors are remarkably close. The best values has $\mathcal{H}_7$, closely followed by $\mathcal{H}_5$. The worst values has $\mathcal{H}_6$ with a slightly larger difference to $\mathcal{H}_5$ than that between $\mathcal{H}_7$ and $\mathcal{H}_5$.

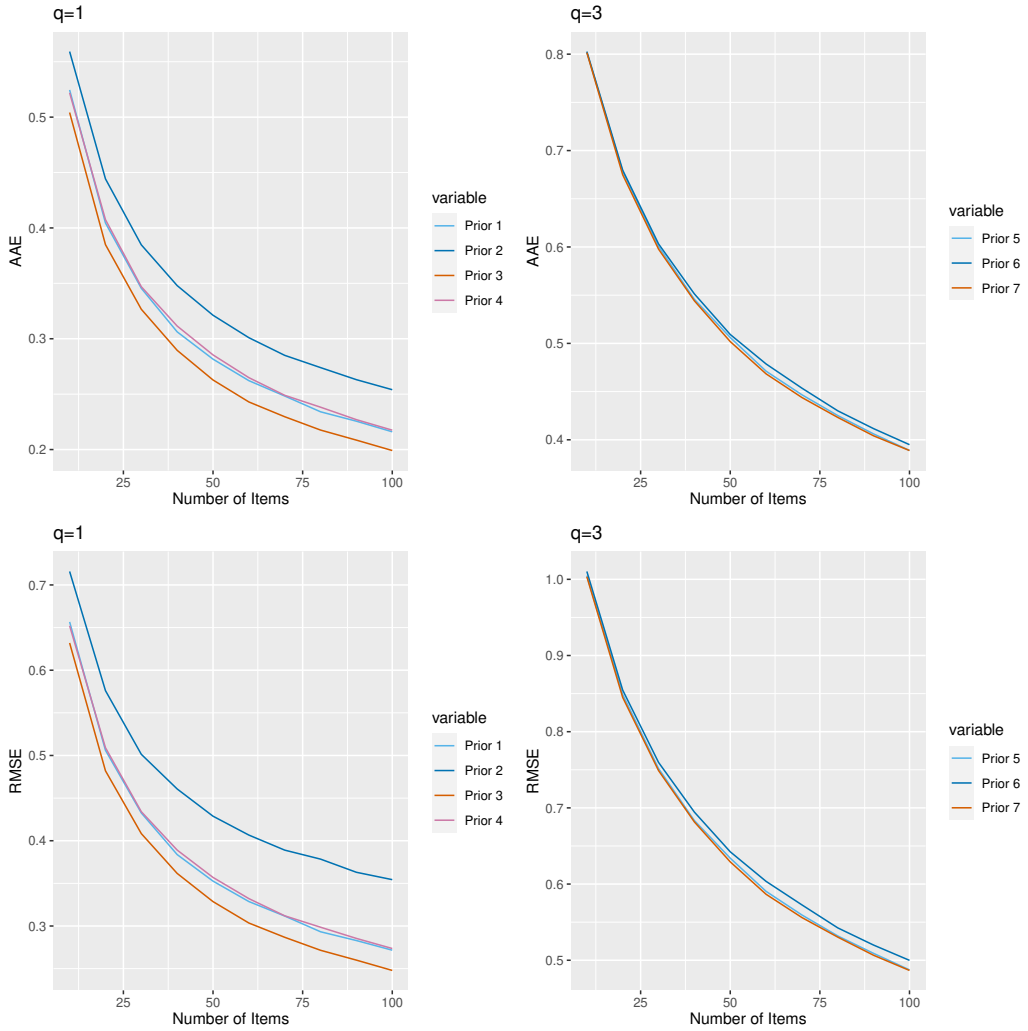

Figure 4.1: *Average absolute error (AAE) and root mean squared error (RMSE) values for the aEAP estimates which are based on* $n = 100$ *responses, averaged over* 1000 *randomly generated data sets for each* d *and each prior for the prior distributions in equation* (4.13).

In Figure 4.2 we see that the estimates for symmetric distributions are relatively uniformly distributed around the red line, which represents a perfect estimate. In the scatter-plot for $\mathcal{H}_2$, we see that the latent variables in $(-1, 0)$ are slightly underestimated and the latent variables in $(0, 2)$ are slightly overestimated. Similar to the single point in the top-right of the scatter-plot, we observed that latent variables with larger values than two again tend to be underestimated. Sometimes, we observe a similar trend but with a less strong difference to the identity for $\mathcal{H}_4$. In cases where several latent variables with comparably large absolute values were generated from $\mathcal{H}_1 = \mathcal{N}(0, 1)$, we also observed a notable under-/overestimation for the larger/smaller latent variables.

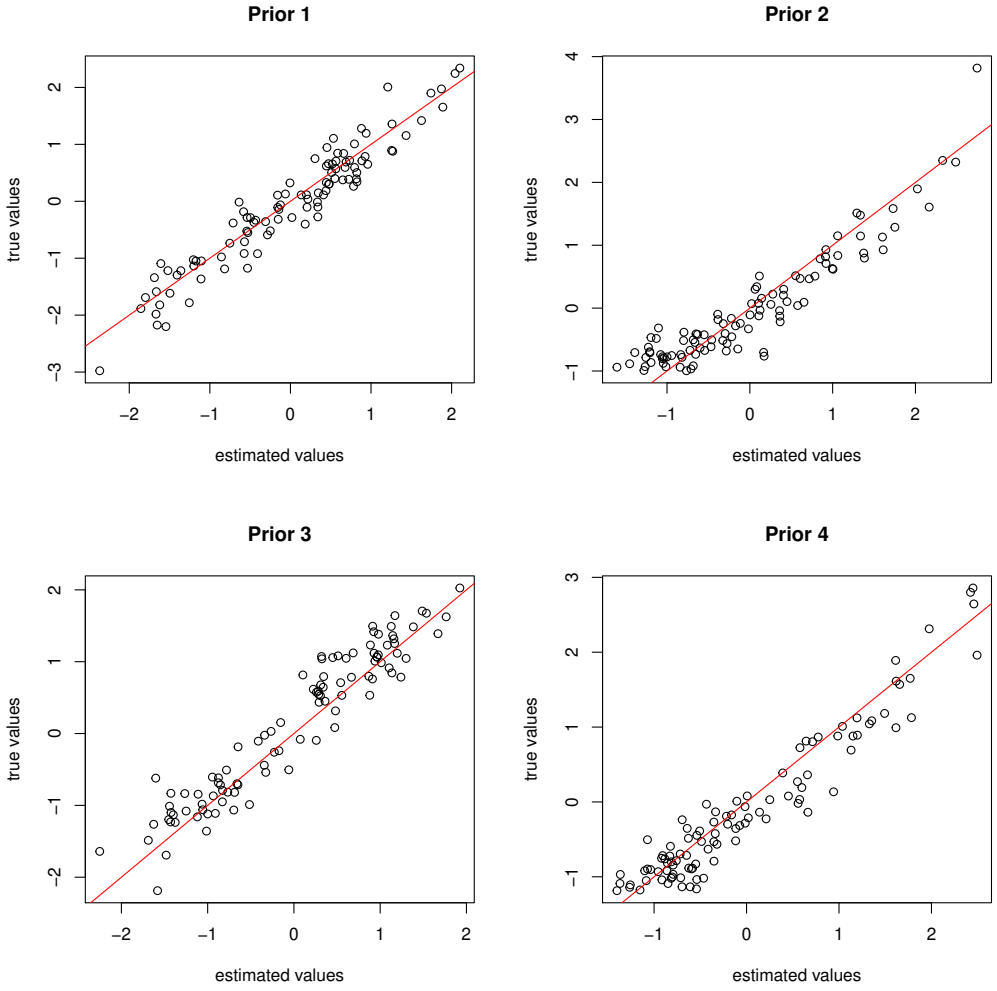

Figure 4.2: *Single sample examples of the aEAP estimator for latent variables for cases with $q = 1$, $d = 60$, $n = 100$. The red line is the identity. The closer the circles are to the red line the better is the estimate.*

4.2.2.3 Discussion

Generally, we see that the aEAP performed well in the studied one dimensional cases of (i) the common standard normal prior, (ii) skew, (iii) multimodal and (iv) bounded-support priors as well as in the multidimensional cases of (v) multivatiate standard normal and (vi)–(vii) skew and multi-modal. There is no indication that the aEAP is crucially sensitive to specific aspects of the prior distribution. We therefore conclude its practical applicability. The application is especially interesting due to its simplicity (neither quadrature nor iterative optimization is required) and its usability in the case of prior uncertainty. Due to the current decrease of all AAE curves to zero with $\mathcal{O}(d^{-1/2})$, we expect the consistency of the aEAP for $d \to \infty$ and the asymptotic normality of the estimates. A theoretical analysis of the aEAP with regard to these properties would be interesting for future research.

The number and absolute value/norm of extremely large latent vectors seem to be the primary factor for the precision of the aEAP: The worst results have $\mathcal{H}_2$ for $q = 1$ and $\mathcal{H}_6$

for $q = 3$ – $\mathcal{H}_2$ equals the standardized exponential distribution and one component of $\mathcal{H}_6$ is one while the standardized exponential distribution has the highest mass on arbitrary large values from all examined one dimensional distributions. This is consistent to the common observation that comparably extreme values of latent variables are difficult to estimate and, therefore, not surprising.[8]

In contrast, the skewness does not seem to have a notable influence on the efficiency even if Figure 4.2 seems to indicate otherwise because the AAE curves for $\mathcal{H}_1$ and $\mathcal{H}_4$ are remarkably close. Similarly, the estimates for $\mathcal{H}_7$ are slightly more precise than for $\mathcal{H}_5$ and $\mathcal{H}_6$ while $\mathcal{H}_7$ changed the standardized exponential distribution from $\mathcal{H}_6$ to the (similarly skew) standardized truncated exponential distribution.

Multi-modality of the distribution does not seem to have any negative influence on the estimation quality of the aEAP. Actually, the latent variables of $\mathcal{H}_3$ are the most precise estimates. Two positive aspects of the standardized bi-modal distribution could be that

1. there is only little mass on extreme values through the distribution of the mass on two distinct bells and

2. these bells produce a larger heterogeneity of the resulting response pattern which allow a more precise estimation.

Our observation that the aEAP for latent variables from symmetric distributions is practically unbiased corresponds to the observation of Li (2010) that the linear EAP of the manifest probabilities model to estimate latent variables are practically unbiased if the latent variables were actually drawn from a normal distribution.

Another important conclusion is that the compact support of $\mathcal{H}$ in (PCT1) does not seem to be a crucial assumption. The simulations we performed support using of the standardized components to estimate $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ in large data sets.

4.3 A new EM-Algorithm for Marginal Maximum Likelihood Estimation in High Dimensional Settings

4.3.1 Preliminaries

Throughout this section, we assume that (M) and (M') hold. Additionally, we suppose that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n \sim \mathcal{H}$ for a $\mathcal{H}$ that satisfies (PCT1). We write $\mathfrak{h}$ for the pdf of $\mathcal{H}$. Now, we suppose that the model parameters $\alpha_{i0}, \boldsymbol{\alpha}_i$, $i = 1, \ldots, d$ are unknown and we want to estimate them from an observed sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ of the response vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. To overcome the location and scale indeterminacy, we suppose that (4.5) is satisfied.

Remark 4.3.1
The method which we will derive is constructed in such a way that it does not use the prior density $\mathfrak{h}$ but approximates $\mathcal{H}(\bullet \mid \boldsymbol{Y}_k = \boldsymbol{y}_k)$ by a multivariate normal distribution using the PCT of Section 3.5.2. Hence, for the application of the following EM-algorithm no explicit assumption regarding $\mathcal{H}$ is required. We can alternatively suppose $(\boldsymbol{\eta}_1 \mid \boldsymbol{Y}_1 = \mathbf{0}) \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

[8] Binary response vectors have a finite number of possible response patterns. IRT allows to discriminate well between values of latent vectors in a region around zero (due to $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$) using a majority of the possible response patterns. But this implies that latent vectors with (comparably) large absolute value/norm only can be estimated on a basis of a few possible response pattern unless the number of items is correspondingly large. The extreme latent vectors are therefore generally more difficult to estimate than the latent vectors in the closer region around $\mathsf{E}(\boldsymbol{\eta})$.

for some $\boldsymbol{\mu} \in \mathbb{R}^q$ and symmetric positive definite $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$, i.e. we suppose that the structural model $\mathcal{H}$ implied by the posterior normality holds (cf. equation (1.34)). Then we suppose that the manifest probabilities model of Theorem/Definition 1.2.4 holds. So, with an adaption to the posterior covariance homogeneity of the manifest probabilities model, the method which we discuss in the following can also be considered a method for the manifest probabilities model. As we discussed in Subsection 1.3.2 and later supported with simulations in the more simple Rasch model case in Subsection 1.3.3.6: Even if the manifest probabilities MMLE (MPMLE) uses the closed form expression of the marginal distribution in (1.38) and overcomes the burden of the integrals, it still requires an evaluation of all 2^d possible response patterns (and not only the observed ones). The computational effort therefore grows exponentially with d and the MPMLE already reaches its computational limits for a quite small number of items.[9] So, if one considers the finite d posterior normality assumption to be valid, then further methods for the manifest probabilities models are required for model fitting for data sets with comparably large d. ◇

We collect all parameters of the measurement model in the vector

$$\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{d(q+1)})^\intercal \coloneqq (\alpha_{i0}, \ldots, \alpha_{dq})^\intercal \in \mathbb{R}^{d(q+1)} =: \Theta_1$$

and denote by $\nabla_{\boldsymbol{\vartheta}}$ the vector of partial derivatives with respect to $\vartheta_1, \ldots, \vartheta_{d(q+1)}$. To indicate the dependence of P_i and λ_i on $\boldsymbol{\vartheta}$, we set $P_i(\boldsymbol{\eta}) = P_i(\boldsymbol{\eta}; \boldsymbol{\vartheta})$ and $\lambda_i(\boldsymbol{\eta}) = \lambda_i(\boldsymbol{\eta}; \boldsymbol{\vartheta})$, $\boldsymbol{\eta} \in \Theta$, $\boldsymbol{\vartheta} \in \Theta_1$. Here, the *complete data* likelihood function which is defined in equation (2.11) on page 71 – considering $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ as unobserved data – becomes

$$\begin{aligned}\mathcal{L}_{\text{complete}}\colon \Theta_1 \times \Theta^n \times \{0,1\}^{dn} &\to (0,\infty),\\ (\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) &\mapsto \prod_{k=1}^{n} \left(\prod_{i=1}^{d} P_i(\boldsymbol{\eta}_k; \boldsymbol{\vartheta})^{y_{ik}} \Big(1 - P_i(\boldsymbol{\eta}_k; \boldsymbol{\vartheta})\Big)^{1-y_{ik}} \right) \mathfrak{h}(\boldsymbol{\eta}_k).\end{aligned}$$

We write for short

$$\mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) =: \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n).$$

As discussed at equation (2.14) in Section 2.3.1, we get the *marginal likelihood function* – or in this notion the *observed data likelihood* – as

$$\mathcal{L}_{\text{obs}}\colon \Theta_1 \to (0,\infty), \qquad \boldsymbol{\vartheta} \mapsto \int_\Theta \ldots \int_\Theta \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)\, \mathrm{d}\boldsymbol{\eta}_1 \ldots \mathrm{d}\boldsymbol{\eta}_n.$$

The general EM-algorithm for obtaining marginal maximum likelihood estimates defines a sequence $\{\boldsymbol{\vartheta}_t\}_{t\in\mathbb{N}_0}$ in Θ_1 as the following iterative procedure:

1. Select a starting value $\boldsymbol{\vartheta}_0 \in \Theta_1$.

2. For each $t \in \mathbb{N}_0$, set $\boldsymbol{\vartheta}_{t+1}$ to an arbitrary point in Θ_1 satisfying

$$\begin{aligned}&\mathsf{E}\left(\log \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}_{t+1}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta}_t\right)\\ &\quad = \max_{\boldsymbol{\vartheta}\in\Theta_1} \mathsf{E}\left(\log \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta}_t\right).\end{aligned} \tag{4.14}$$

[9] Cf. footnote 1 of this chapter on page 171.

Hereby the parameter $\boldsymbol{\vartheta}_t$ in the condition of the conditional expectation denotes that $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_t$ in the posterior distribution of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, while $\boldsymbol{\vartheta}$ in the argument of $\mathcal{L}_{\text{complete}}$ remains variable. Starting values can be selected arbitrarily, generated using GLMs with randomly generated values of the latent vectors (cf. e.g. Rizopoulos, 2006) or even by generating random starting values. In all of our studies we have observed that the algorithm converges to equivalent solutions (i.e. identical solutions after rotating the matrix of factor loadings) even if the starting values were generated completely at random. The selection of good starting value therefore does not seem to be crucial for the convergence of the EM-algorithm for fitting logit IRT models by MMLE. In common practical cases, we only consider local maximization in each iteration to find the next $\boldsymbol{\vartheta}_{t+1}$. This can be achieved using iterative local maximization algorithms. In the following we examine the computation of the EM-iterations via local maximization.

If for any $t \in \mathbb{N}_0$ no $\boldsymbol{\vartheta}_{t+1} \in \Theta_1$ exists such that (4.14) is satisfied, then the EM-sequence is not well defined. This can happen if no marginal maximum likelihood estimate exists for the current data set and a parameter value has to be selected to be equal to infinity. This can then be seen during the iterative procedure when some parameter updates approach infinity in absolute value or are set to $\pm\infty$ in an implementation using floating point arithmetic. However, Theorem 2.3.2 guarantees the existence of the MMLE for $n \to \infty$ under the conditions of this section. The EM-algorithm guarantees that $\mathbb{N} \ni t \mapsto \mathcal{L}_{\text{obs}}(\boldsymbol{\vartheta}_t)$ is non-decreasing. Additionally, we will see in the following that each limit point of the EM-algorithm satisfies the marginal likelihood equations. Hence, we can find local maximal points of $\mathcal{L}_{\text{obs}}$, but as usual, proving that they are global maximal points is often hardly possible. However, the convergence to the same point from different starting points allows the model fitting with reasonable doubt. In a practical application, we terminate the iterative procedure if a certain convergence criterion is satisfied.

In our special case of logit models, we get from the independence of the n random tuples $(\boldsymbol{Y}_1, \boldsymbol{\eta}_1), \ldots, (\boldsymbol{Y}_n, \boldsymbol{\eta}_n)$ that[10]

$$\begin{aligned}
&\mathsf{E}(\log \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta}_t) \\
&\quad = \sum_{i=1}^{d} \left(\sum_{k=1}^{n} \mathsf{E}\left(y_{ik}(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}_k) - \log\left(1 + \exp\left(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta}_k\right)\right) \,\middle|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \right) \right) \\
&\qquad + \sum_{k=1}^{n} \mathsf{E}(\log(\mathfrak{h}(\boldsymbol{\eta}_k)) \mid \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t).
\end{aligned}$$

This has an important consequence: In a single step of the EM-algorithm, the measurement model parameters can be updated item-wise. Additionally, the term $\sum_{k=1}^{n} \mathsf{E}(\log(\mathfrak{h}(\boldsymbol{\eta}_k)) \mid \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t)$ on the right-hand side can be ignored in the following because it does not depend on $\boldsymbol{\vartheta}$. Notice that we get analogously to Fisher's identity in equation (2.23) for an arbitrary $\boldsymbol{\vartheta}_t \in \Theta$:

$$\nabla_{\boldsymbol{\vartheta}} \mathsf{E}(\log \mathcal{L}_{\text{complete}}(\boldsymbol{\vartheta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n, \boldsymbol{\vartheta}_t)$$

[10] For a more detailed discussion of the EM-algorithm under the restriction to one dimensional models compare for example Baker and Kim (2004) or Bartholomew et al. (2011, Section 4.5).

$$= \begin{pmatrix} \sum_{k=1}^n \mathsf{E}\Big(y_{1k} - P_1(\boldsymbol{\eta}_k;\boldsymbol{\vartheta}) \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \\ \sum_{k=1}^n \mathsf{E}\Big(\big(y_{1k} - P_1(\boldsymbol{\eta}_k;\boldsymbol{\vartheta})\big)\eta_{1k} \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \\ \vdots \\ \sum_{k=1}^n \mathsf{E}\Big(\big(y_{1k} - P_1(\boldsymbol{\eta}_k;\boldsymbol{\vartheta})\big)\eta_{qk} \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \\ \sum_{k=1}^n \mathsf{E}\Big(y_{2k} - P_2(\boldsymbol{\eta}_k;\boldsymbol{\vartheta}) \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \\ \vdots \\ \sum_{k=1}^n \mathsf{E}\Big(\big(y_{dk} - P_d(\boldsymbol{\eta}_k;\boldsymbol{\vartheta})\big)\eta_{qk} \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \end{pmatrix}, \tag{4.15}$$

for all $\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_t \in \Theta_1$. So, in each step of the EM-algorithm, we have to solve some kind of conditional expected likelihood equations of a binomial distributed response with unobserved explanatory variable and logit-link. In particular by comparing it to Fisher's identity in equation (2.23) we see that all limit point of the EM-algorithm satisfy the marginal likelihood equations.

An important step for completing the algorithm is to find a good approximation of the posterior expected values in (4.15), which can be feasibly computed because their exact evaluation is not possible. Next we discuss a new solution for the task of approximating (4.15) based on the PCT in Section 3.5.2. This results in a new EM-algorithm for fitting multidimensional logit models. After the discussion of the application of this algorithm to two simulated example data sets we compare it with other approaches to specify EM-algorithms by focusing on high dimensional settings in the sense that d, q are comparably large.

4.3.2 A Posterior Normal Approximated MCEM-Algorithm

Our approach is now to approximate the posterior distributions for the evaluation of the conditional expectations in (4.15) by a normal distribution centered in the EAP. This is justified for sufficiently large d under the assumption of "regular" unknown true item parameters in the sense of condition (M')[11] by Theorem 3.5.2. In this case, the aEAP of equation (4.12) can be considered a valid approximation of the EAP. Denote by $\tilde{\boldsymbol{\eta}}_k$ the aEAP for $\boldsymbol{\eta}_k$, $k \in \mathbb{N}$, then we get

$$\begin{aligned} &\mathsf{E}\Big(\big(y_{ik} - P_i(\boldsymbol{\eta}_k;\boldsymbol{\vartheta})\big)\eta_{jk} \,\Big|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \Big) \\ &\qquad \approx \int_{\mathbb{R}^q} \big(y_{ik} - P_i(\boldsymbol{\eta};\boldsymbol{\vartheta})\big)\eta_j \phi_q(\boldsymbol{\eta}; \tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k;\boldsymbol{\vartheta}_t)^{-1})\, \mathrm{d}\boldsymbol{\eta}, \end{aligned} \tag{4.16}$$

$i = 1, \ldots, d$, $j = 1, \ldots, q$, $k = 1, \ldots, n$, where

$$\mathcal{I}(\tilde{\boldsymbol{\eta}}_k;\boldsymbol{\vartheta}_t) \coloneqq \sum_{i=1}^d P_i(\tilde{\boldsymbol{\eta}}_k;\boldsymbol{\vartheta}_t)\big(1 - P_i(\tilde{\boldsymbol{\eta}}_k;\boldsymbol{\vartheta}_t)\big)\boldsymbol{\alpha}_i^{(t)}\boldsymbol{\alpha}_i^{(t)\mathsf{T}}, \tag{4.17}$$

[11] Thus, the response vectors are supposed to be the first d responses of an theoretical test with $d \to \infty$, so that (PCT2) and (PCT3) are satisfied.

and $\boldsymbol{\alpha}_i^{(t)}$ denotes the values of $\boldsymbol{\alpha}_i$ in $\boldsymbol{\vartheta}_t$, i.e. the tth iteration. The right-hand side of (4.16) can now be approximated by MC-integration. Suppose that $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M \overset{\text{iid}}{\sim} \mathcal{N}_q(\tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k; \boldsymbol{\vartheta}_t)^{-1})$, $M \in \mathbb{N}$, then we get

$$\int_{\mathbb{R}^q} \big(y_{ik} - P_i(\boldsymbol{\eta}; \boldsymbol{\vartheta})\big)\eta_{jk}\phi_q(\boldsymbol{\eta}; \tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k; \boldsymbol{\vartheta}_t)^{-1})\,\mathrm{d}\boldsymbol{\eta} \approx \frac{1}{M}\sum_{m=1}^{M} \big(y_{1k} - P_1(\boldsymbol{\xi}_m; \boldsymbol{\vartheta})\big)\xi_{jm} \tag{4.18}$$

with a standard error of the order $\mathcal{O}_\mathsf{P}\left(\frac{1}{\sqrt{M}}\right)$. As an improvement to (4.18), we can use the common variance reduction techniques to require less quadrature points for the same approximation quality. We use the antithetic variable approach: We generate $\boldsymbol{\xi}'_1, \ldots, \boldsymbol{\xi}'_{M_\text{half}} \overset{\text{iid}}{\sim} \mathcal{N}_q(\tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k; \boldsymbol{\vartheta}_t)^{-1})$, $M_\text{half} \in \mathbb{N}$ and then use the sample

$$(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M) \coloneqq (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{M_\text{half}}, -\boldsymbol{\xi}_1, \ldots, -\boldsymbol{\xi}_{M_\text{half}}) \tag{4.19}$$

of length $M := 2M_\text{half}$.

For the use of (4.16), we have to take a further aspect into account. The assumption (4.5) principally solves the location and scale indeterminacy of the model. In the classic approach using Gauss–Hermite quadrature, this is incorporated in the fixed quadrature points which are determined at the start of the algorithm. In approaches like the adaptive Gauss-Hermite quadrature or the adaptive normal importance sampling by Schilling and Bock (2005), in which the quadrature points are adapted in each step, ensuring these constraints in each step is substantial for the convergence and efficient application of the method. Similarly, we have to incorporate this explicitly in each step of the algorithm. Via the samples $\boldsymbol{\xi}_1^{(k)}, \ldots, \boldsymbol{\xi}_M^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}_q(\tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k; \boldsymbol{\vartheta}_t)^{-1})$, $k = 1, \ldots, n$, we get a sample from the whole population distribution $\mathcal{H}$. Our approach therefore is the standardization of $\boldsymbol{\xi}_1^{(1)}, \ldots, \boldsymbol{\xi}_M^{(1)}, \ldots, \boldsymbol{\xi}_1^{(n)}, \ldots, \boldsymbol{\xi}_M^{(n)}$. Due to the commonly high n, this is valid for ensuring (4.5) while the relation between the sub-samples remains the same. Overall, this is a "Monte–Carlo implementation" of the identification constraints using sampling from the posterior.

For simplicity, we suggest solving the rotation invariance of the solution after the final estimation (e.g. by the varimax solution of the estimates of $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d$). For stability as well as for computational efficiency, it is absolutely reasonable to use the same base sample $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$ and just transform them according to the various normal distributions. The previous standardization of the aEAP is still important to evaluate the information matrix at appropriate points.

The final EM-algorithm for fitting a logit-model using (4.18) with the antithetic variable approach in (4.19) is given by the following pseudo-code.

Algorithm 1 Posterior normal **a**pproximated **M**onte **C**arlo **EM** for MMLE (**aMCEM**)

Require: response vectors $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \{0,1\}^d$, latent dimension $q \in \mathbb{N}$, termination $\varepsilon > 0$, MC sample size $M_{\text{half}} \in \mathbb{N}$, max iterations $T \in \mathbb{N}$, stating value $\boldsymbol{\vartheta}_0 \coloneqq (\alpha_{i0}^{(0)}, \dots, \alpha_{dq}^{(0)})^\intercal \in \mathbb{R}^{(q+1)d}$

Generate $\boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_{M_{\text{half}}} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$
Define $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M$ for $M = 2M_{\text{half}}$ by (4.19)
for $t = 1, \dots, T$ **do**
  Compute the aEAP $\tilde{\boldsymbol{\eta}}_1^{(t)}, \dots, \tilde{\boldsymbol{\eta}}_n^{(t)}$ using $\boldsymbol{\alpha}_1^{(t-1)}, \dots, \boldsymbol{\alpha}_d^{(t-1)}$ via (4.12)
  for all $\tilde{\boldsymbol{\eta}}_1^{(t)}, \dots, \tilde{\boldsymbol{\eta}}_n^{(t)}$ **do**
    Compute $\mathcal{I}(\tilde{\boldsymbol{\eta}}_k^{(t)}; \boldsymbol{\vartheta}_t)$ via (4.17)
  end for
  Set $\boldsymbol{x}_{11}^{(t)}, \dots, \boldsymbol{x}_{nM}^{(t)}$ to the standardization of

$$\mathcal{I}(\tilde{\boldsymbol{\eta}}_k^{(t)}; \boldsymbol{\vartheta}_t)^{-1/2}\Big(\boldsymbol{\xi}_m + \tilde{\boldsymbol{\eta}}_k^{(t)}\Big), \qquad k = 1, \dots, n,\ m = 1, \dots, M. \tag{4.20}$$

  for $i = 1, \dots, d$ **do**
    Compute $\alpha_{i0}^{(t)}$, $\boldsymbol{\alpha}_i^{(t)}$ of $\boldsymbol{\vartheta}_t$ by maximizing $\ell_i^{(t)} \colon \mathbb{R}^{1+q} \to \mathbb{R}$ with

$$\ell_i^{(t)}(\alpha_{i0}, \boldsymbol{\alpha}_i) \coloneqq \sum_{m=1}^{M} \sum_{k=1}^{n} \left(y_{ik}(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{x}_{km}^{(t)}) - \log\Big(1 + \exp(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{x}_{km}^{(t)})\Big) \right) \tag{4.21}$$

  end for
  if $\|\boldsymbol{\vartheta}_t - \boldsymbol{\vartheta}_{t-1}\| \le \varepsilon d(q+1)\big(1 + \|\boldsymbol{\vartheta}_{t-1}\|\big)$ **then**
    break
  end if
end for
return $\boldsymbol{\vartheta}_t$, $\tilde{\boldsymbol{\eta}}_n^{(t)}$, $\mathcal{I}(\tilde{\boldsymbol{\eta}}_k^{(t)}; \boldsymbol{\vartheta}_t)$, $k = 1, \dots, n$

Remark 4.3.2

(i) The gradient of $\ell_i^{(t)}$ in (4.21) is given by

$$\nabla \ell_i^{(t)}(\alpha_{i0}, \boldsymbol{\alpha}_i) = \begin{pmatrix} \sum_{m=1}^{M} \sum_{k=1}^{n} \left(y_{ik} - \frac{1}{1+\exp(-\alpha_{i0} - \boldsymbol{\alpha}_i^\intercal \boldsymbol{x}_{km}^{(t)})} \right) \\ \sum_{m=1}^{M} \sum_{k=1}^{n} \left(y_{ik} - \frac{1}{1+\exp(-\alpha_{i0} - \boldsymbol{\alpha}_i^\intercal \boldsymbol{x}_{km}^{(t)})} \right) x_{km1}^{(t)} \\ \vdots \\ \sum_{m=1}^{M} \sum_{k=1}^{n} \left(y_{ik} - \frac{1}{1+\exp(-\alpha_{i0} - \boldsymbol{\alpha}_i^\intercal \boldsymbol{x}_{km}^{(t)})} \right) x_{kmq}^{(t)} \end{pmatrix}, \tag{4.22}$$

with $\boldsymbol{x}_{km}^{(t)} = (x_{km1}^{(t)}, \dots, x_{kmq}^{(t)})^\intercal$. The maximization of $\ell_i^{(t)}$ hence can be performed easily and quickly by gradient based algorithms. Particularly, the maximization of (4.21) is equivalent to fitting a binary logit GLM with intercept and q explanatory variables, which are the quadrature points. This equivalence allows the efficient use of standard GLM software (e.g. the R-function `glm`) for the implementation of the aMCEM algorithm. Moreover, the algorithm can be easily implemented for general $q \in \mathbb{N}$ to obtain a general procedure for fitting multidimensional logit IRT models for

binary responses. Further aspects of the implementation in R are discussed in Section C.2 of the appendix.

(ii) The restriction to the logit link is not completely necessary. In the aEAP, the logit link is only actually required but the standardized components are at least a rough estimate for the latent vectors when other link functions are supposed. Alternatively, the update of the latent vectors by the MLE is completely supported by the proper centering theory. Alternatively, it seems plausible that the theories of the asymptotic posterior normality can be generalized to the MAP, which is then also applicable. ⋄

4.3.3 Application to two Simulated Example Data Sets

To get an impression of the method, we will now discuss the application to two simulated data sets.

In the first example, we examine a case of $d = 60$ items, $q = 2$ factors and $n = 5000$ responding subjects. For the aMCEM routine, we use 80 quadrature points (i.e. we generate 40 and also use their negative values as antithetic variables). We compare the estimates to estimates which are achieved by using the Gauss–Hermite EM-BFGS-hybrid of the R-package **ltm** with its standard settings. The true factor values are generated from $\mathcal{N}_2(\mathbf{0}, \mathrm{I}_2)$, so the structural model of the ltm method is correctly specified, while the aMCEM method has a lack in justification because the normal distribution is not supported by the proper centering theory. For the comparison of the estimated factor loadings by aMCEM and ltm with the true values, we use the varimax solution. The MMLE estimates of the intercept parameters obtained from ltm have a root mean squared error (RMSE) of ≈ 0.03178 and a RMSE of the factor loadings of ≈ 0.0457. The estimated intercepts of aMCEM have a RMSE of ≈ 0.031 and the factor loadings of aMCEM have a RMSE of ≈ 0.0473. This means the estimates are similarly good. For visual comparison see Figure 4.3. There, we see that the estimates are similar. The largest visual difference is that the aMCEM has slighly worse estimates for the loadings of the first factor. When using the same starting value, aMCEM needed about 3.43 times as long as the ltm method in this example. For the comparison of the required time one has to take into account the single use of the EM-algorithm in aMCEM, which is known to be slow, while ltm switches to a fast quasi-Newton algorithm after a while.

However, these computational aspects reverse for a larger number of factors. As the second example, we study a simulated data set with $d = 120$ items, $q = 8$ factors and $n = 10000$ test participants. For these settings, deterministic numerical quadrature methods cannot be applied feasibly without risking a questionable approximation quality or requiring a large computation time. For aMCEM, we still use 80 (or 2×40, respectively) quadrature points. In this work, we will not conduct comparisons to other Monte Carlo approaches, so at the current it is state only an illustrative example. In this case, the RMSE of the intercepts is ≈ 0.041 and of the factor loadings ≈ 0.061. A visualization is given in Figure 4.4. We see that there generally is a good recovery of the parameters.[12] With random starting values in $(-1, 1)$ for the factor loadings and roughly estimated values of the intercepts, the actual fitting took 324734.09 seconds ($\approx$ 3 d 18 h 12.2 min) and 292151.1 seconds ($\approx$ 3 d 9 h 9.2

[12] As a rough comparison, the error measures are approximately 1.24 to 1.3 times larger than in the smaller data example where we had the relation to ltm. Bear in mind that this depends on the data dimensions, $\mathcal{H}$ and the values of the parameters themself besides of the integral approximation, so it is not possible to estimate the individual influences from the current data set.

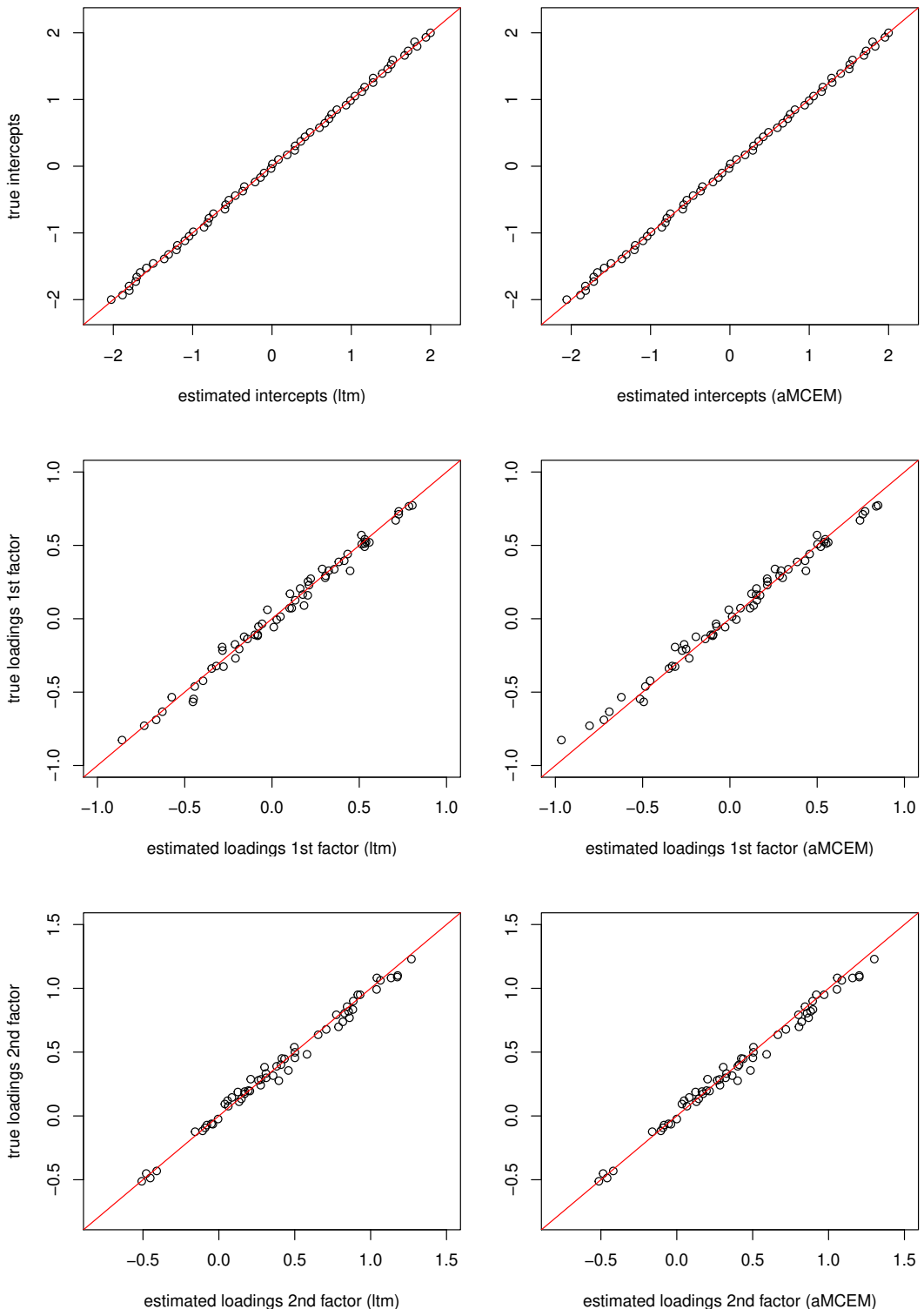

Figure 4.3: *Comparison of MMLE estimates for a simulated data set with $d = 60$ items, $q = 2$ factors and $n = 5000$ observed responses. Left: Estimated parameter values using ltm against the true values. Right: estimated values using aMCEM against the true values. The red line is the identity. The closer the circles are to the red line the better is the estimate.*

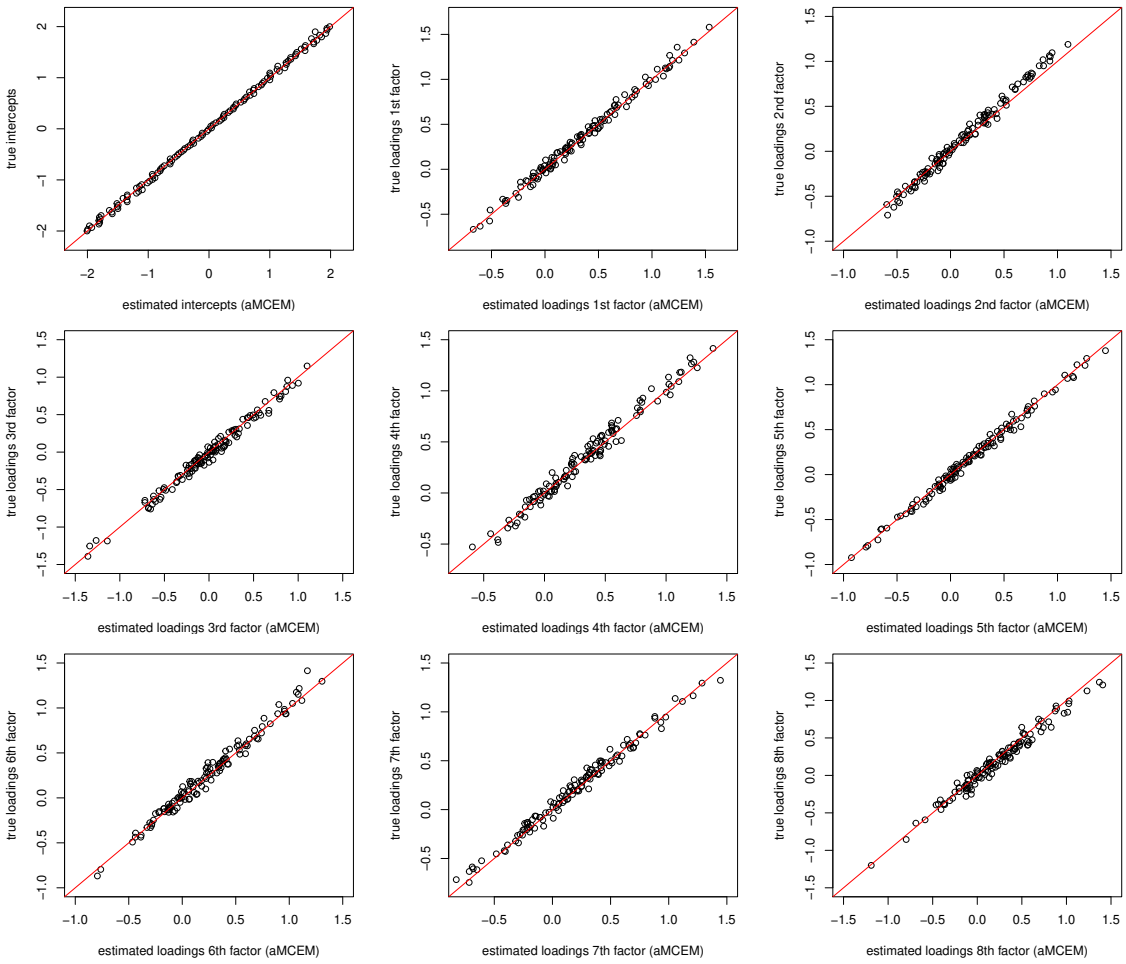

Figure 4.4: *Comparison of MMLE estimates for a simulated data set with $d = 120$ items, $q = 8$ factors and $n = 10000$ observed responses. Estimated values using aMCEM against the true values. The red line is the identity. The closer the circles are to the red line the better is the estimate.*

min) in two trials. We thus estimate that an arbitrary home computer takes between 3 and 4.5 days to complete this operation.

An important observation when looking at these exameplary data sets is that the assumption of the compact support of $\mathfrak{h}$ does not seem to be critical. Additionally we recognize the sufficiently good approximation of the posterior expected values using the approximate posterior normality and the aEAP. Particularly, $d \approx 60$ seems to be sufficient for an application of aMCEM and $M \approx 80$ seems to be sufficient for an eight-factor model. Since a deterministic quadrature method with only two quadrature points per dimension already requires $M = 256$ quadrature points, this is a substantial reduction. In the next subsection, we will briefly review important aspects of different algorithmic approaches for EM-algorithms for fitting logit IRT models with a focus on data sets with large d and q and compare them to the aMCEM-algorithm.

4.3.4 Comparison of the aMCEM-Algorithm to other approaches for EM-Algorithms for MMLE in the case of large q and large d

Bear in mind that many classical approaches for approximating (4.15) are based on

$$\begin{aligned} &\mathsf{E}\left(\big(y_{ik} - P_i(\boldsymbol{\eta}_k; \boldsymbol{\vartheta})\big)\eta_{jk} \,\middle|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \right) \\ &\qquad = \frac{\int_\Theta \big(y_{ik} - P_i(\boldsymbol{\eta}; \boldsymbol{\vartheta})\big)\eta_j \, \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta}}{\int_\Theta \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta}}, \end{aligned} \tag{4.23}$$

$i = 1, \ldots, d$, $j = 1, \ldots, q$, $k = 1, \ldots, n$. Here the numerator and denominator have to be approximated separately. Additionally, notice that $\mathfrak{h} = \phi_q$ is commonly assumed in the literature but the concepts of many algorithms do not depend on this assumption. The derivation and theoretical justification of the aMCEM-algorithm are based on the PCT but the final algorithm is independent of $\mathfrak{h}$ and the simulations suggest that (PCT1) is not critical for the application of the algorithm. We therefore discuss and compare the different approaches in the following without restricting $\mathfrak{h}$.

Deterministic quadrature rules require $M = z^q$ quadrature points for the approximation of q-dimensional integrals when using z quadrature points for each single dimension. This means that we already get $M = 256$ for $q = 8$ and $z = 2$. It is often useful to calculate another fit using $z = 3$ for comparison, possibly recovering a bad approximation. However, this increases M to $3^8 = 6561$, which causes a substantial increase in computation time. Clearly, further increments of z or larger q make this situation even more drastic. Another burden on the integral approximation comes from the fact that the required integrals can be considered posterior expected values (cf. (4.23)). Selecting quadrature points corresponding to the prior (as we do by using classic Gauss–Hermite quadrature and $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$, $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$) requires a large number of quadrature points because the posterior variance decreases with order $\frac{1}{\sqrt{d}}$ and many points therefore only have a small contribution to the integral value. Often selected values for classic Gauss–Hermite quadrature are $z \in \{15, \ldots, 30\}$ such that the integral approximation is infeasible for a large number of factors. The adaptive Gauss–Hermite quadrature EM-algorithms overcome this problem by scaling and shifting the quadrature points so that they correspond to integrals with respect to the $\mathcal{N}_q(\tilde{\boldsymbol{\eta}}_k^*, \tilde{\boldsymbol{\Sigma}}_k^*)$ distributions for all unique response patterns. There $\tilde{\boldsymbol{\eta}}_k^*$ denotes the MAP of $\boldsymbol{\eta}_k$ based on the response pattern $\boldsymbol{y}_k$ and $\tilde{\boldsymbol{\Sigma}}_k^*$ either is $\mathcal{I}(\tilde{\boldsymbol{\eta}}_k^*; \boldsymbol{\vartheta}_t)^{-1}$ or the inverse of the negative Hessian of $\boldsymbol{\eta} \mapsto \log\big(\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{\eta})\big)$ at $\tilde{\boldsymbol{\eta}}_k^*$ (cf. e.g. Rabe-Hesketh et al., 2002; Schilling and Bock, 2005). Supposing asymptotic posterior normality of latent vectors for centering at the MAP[13], this allows a drastic reduction of the required quadrature points for the same approximation quality if d is large. It is realistic that the required number of quadrature points can be reduced to $z \in \{2, \ldots, 5\}$ is many practical applications (Schilling and Bock, 2005). As we noted before, z values in this range are still critical for $q \approx 8$ and even small increments of q or z can result in an infeasible procedure. An additional new drawback of this adaptive approach is that in each E-step, the MAP has to be computed for each unique response pattern which requires the application of iterative procedures in a large number of cases. An approach for accelerating this type of adaptive Gauss–Hermite quadrature was discussed by Rizopoulos (2012), who suggested adapting the quadrature points only in the first iteration, thereby eliminating the repreated application of iterative procedures

[13] Cf. Chapter 3 for a general discussion of asymptotic posterior normality of latent vectors. We conjecture that the theory discussed there can be extended to centering in the MAP.

to compute the MAP. The resulting quadrature points can however be expected to be less efficient in comparison to those resulting for an update in each E-step. Hence, z has to be at least as large for this approach as for the adaptive Gauss–Hermite method, so the general problem remains the same for a large number of factors.

Our approach for updating the quadrature points using the aEAP has the clear advantage that it can be easily and very quickly be done in each E-step. Especially, the aEAP is easier to implement because it only requires the standardization of the component statistics using the current factor loadings.

The problem of the large number of quadrature points can be solved by using MC-integration – as we do in the aMCEM-algorithm – because we can select the total number of quadrature points $M \in \mathbb{N}$ arbitrarily. This has two significant advantages over deterministic quadrature approaches: We can increase the number of quadrature points slightly to check the need for more quadrature points, but it is also possible to use a smaller number if it is appropriate. We successfully fitted models using only 80 quadrature points for the aMCEM-algorithm instead of using 256 quadrature points for an eight-factor model on $d = 120$ items as the two-point adaptive quadrature does.

The direct approach of sampling from $\mathcal{H}$ to approximate numerator and denominator in (4.23) requires instead a large sample size M because the posterior has the majority of its mass in a much smaller area than the prior $\mathcal{H}$. Moreover, the majority of the mass of the posterior is not necessarily in the center of $\mathcal{H}$. Hence, many quadrature points drawn from $\mathcal{H}$ contribute little to the integral approximation similar to classic Gauss–Hermite quadrature. For example Sammel et al. (1997) used $M = 10000$ for a single factor model. An approach which approximates the posterior in (4.15) directly is given by drawing a sample from the posterior using an MCMC-algorithm. Such an approach using a Gibbs-sampler was presented by Meng and Schilling (1996). This approach requires the draw of a complete new sample in each iteration however which causes a substantial increase in computation time compared to other approaches (cf. Schilling and Bock, 2005). Schilling and Bock (2005) improved MCEM-algorithms by using

$$C_k^* \coloneqq \int_\Theta \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = \int_\Theta \frac{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{\eta}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{\eta})}{\phi_q(\boldsymbol{\eta}; \tilde{\boldsymbol{\eta}}_k, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k; \boldsymbol{\vartheta}_t)^{-1})} \phi_q(\boldsymbol{\eta}; \tilde{\boldsymbol{\eta}}_k^*, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k^*; \boldsymbol{\vartheta}_t)^{-1})\,\mathrm{d}\boldsymbol{\eta}, \tag{4.24}$$

where $\tilde{\boldsymbol{\eta}}_k^*$ denotes the MAP of $\boldsymbol{\eta}_k$ based on the response pattern $\boldsymbol{y}_k$ and considering the current measurement model parameters $\boldsymbol{\vartheta}_t$ as known. By drawing the sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \overset{\text{iid}}{\sim} \mathcal{N}_q(\tilde{\boldsymbol{\eta}}_k^*, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k^*; \boldsymbol{\vartheta}_t)^{-1})$ this leads to the MC-approximation

$$C_k^* \approx \frac{1}{M} \sum_{m=1}^{M} w_k(\boldsymbol{x}_m) =: C_k,$$

where

$$w_k(\boldsymbol{x}) \coloneqq \frac{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}_k \mid \boldsymbol{x}; \boldsymbol{\vartheta}_t)\mathfrak{h}(\boldsymbol{x})}{\phi_q(\boldsymbol{x}; \tilde{\boldsymbol{\eta}}_k^*, \mathcal{I}(\tilde{\boldsymbol{\eta}}_k^*; \boldsymbol{\vartheta}_t)^{-1})}, \qquad \boldsymbol{x} \in \mathbb{R}^q.$$

The numerator of (4.23) is then approximated analogously and we overall get

$$\mathsf{E}\left(\big(y_{ik} - P_i(\boldsymbol{\eta}_k; \boldsymbol{\vartheta})\big)\eta_{jk} \,\middle|\, \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{\vartheta}_t \right) \approx \frac{1}{M} \sum_{m=1}^{M} \big(y_{ik} - P_i(\boldsymbol{x}_m; \boldsymbol{\vartheta})\big) x_{mj} \frac{w_k(\boldsymbol{x}_m)}{C_k}. \tag{4.25}$$

This approach is an importance sampling method where the importance distribution is a multivariate normal distribution which is centered at the posterior mode and uses a estimate of the posterior covariance matrix as covariance matrix. Schilling and Bock (2005) argued that this provides a good approximation of the integrals for comparably few quadrature points and a large d under the assumption that asymptotic posterior normality of latent vectors holds for $d \to \infty$.[14] Even if our theory of asymptotic posterior normality in Chapter 3 does not contain the MAP for centering of the posterior, this assumption is strongly supported by our theory and we conjecture that SCT and PCT can be extended to the MAP. Using a single sample of $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$, which is then transformed according to the individual normal distributions, leads to a new algorithm. It is comparably fast and precise as the adaptive Gauss–Hermite quadrature methods but provides a much larger flexibility in selecting M. We call MCEM-algorithms which use this normal importance sampling (NIS) technique NIS-MCEM-algorithms (NIS-MCEM for short). NIS-MCEM-algorithms therefore are advantageous over

1. adaptive Gauss–Hermite quadrature EM-algorithms,
2. classic MCEM-algorithms using direct sampling from the prior and
3. Gibbs-MCEM-algorithms

when the number of factors is large (i.e. $q \geq 8$) and the number of items is large.

Next, we compare NIS-MCEM and aMCEM in more detail. First of all, we recognize that the two major differences between NIS-MCEM and aMCEM are

1. NIS-MCEM centers the posterior at the MAP and aMCEM at the aEAP,
2. $\frac{w_k(\boldsymbol{x}_m)}{C_k} = 1$ holds in the aMCEM-version of the approximation in equation (4.25).

With the notation and assumptions of Theorem 3.5.2, we get $\left|\frac{h'(\boldsymbol{x}|\boldsymbol{Y}^{(d)})}{\phi_q(\boldsymbol{x})} - 1\right| \xrightarrow{\mathsf{P}_{\boldsymbol{\eta}_0}} 0$ for $d \to \infty$ and almost all $\boldsymbol{x} \in \mathbb{R}^q$ where $h'(\bullet \mid \boldsymbol{Y}^{(d)})$ denotes the pdf of $\mathcal{I}^{(d)}(\tilde{\boldsymbol{\eta}}_d)^{1/2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_d)$ given $\boldsymbol{Y}^{(d)}$. The estimator $\tilde{\boldsymbol{\eta}}_d$ can be the MLE, the EAP and for $q = 1$ any Bayes estimator with respect to the loss functions considered by GGS. Assuming Theorem 3.5.2 can be extended to the centering of the posterior at the MAP, we get $\frac{w_k(\boldsymbol{x})}{C_k^*} \approx 1$ for large d, all $\boldsymbol{x} \in \Theta$ and each $\mathfrak{h}$ that satisfies (PCT1). Because both methods justify their approximation of the integrals with the asymptotic posterior normality of latent vectors for $d \to \infty$, the aMCEM method essentially eliminates the unnecessary computation of $C_1, \ldots, C_n$ in each iteration of the EM-algorithm and the unnecessary floating-point operations in the right-hand side of (4.25). This has another advantage: The M-step of the aMCEM-algorithm can directly be implemented using GLM-fitting-procedures, which simplifies the implementation greatly. Additionally, the computation of the MAP requires the application of another iterative optimization procedure for each unique response pattern in each E-step of the EM-algorithm while the aEAP can be computed in a single step by the computation and standardization of the components for all unique response patterns. These differences lead to smaller computation time for the E-step of an aMCEM iteration in comparison to an iteration of the NIS-MCEM. Additionally, it implies a faster computation of the gradient of the item-log-likelihood of the aMCEM in comparison to the NIS-MCEM. A disadvantage

[14] This is essentially the same argument as that for the advantage of the adaptive Gauss-Hermite methods in comparison to the classic Gauss-Hermite methods.

of the aMCEM is that the integral approximation error of NIS-MCEM approaches zero for $M \to \infty$, while aMCEM then still has the error of the approximation of the posterior by a normal distribution if no actual posterior normality is assumed. Hence, the NIS-MCEM is more robust than the aMCEM for small ds. Our simulations in the last subsection with application of the aMCEM suggest that the aEAP as well as the direct approximation of the posterior density are sufficiently good for data sets with $d \geq 60$. How fast the overall M-step of the aMCEM and NIS-MCEM are in comparison and how many iterations they need depend on situational factors such as $\boldsymbol{\vartheta}_t$ and the studied data set. Further simulation based research regarding the objectives of the quality of the approximations which are used for the aMCEM and the NIS-MCEM in comparison is recommended.

Another minor difference between aMCEM and NIS-MCEM is the presence of the supposed prior density $\mathfrak{h}$. NIS-MCEM involves $\mathfrak{h}$ in the computation of the MAP and the w_ks. In contrast, the aMCEM algorithm is completely independent of $\mathfrak{h}$ and only incorporates the assumption $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$. An advantage of the aMCEM is that a specification of $\mathfrak{h}$ is not necessary and an "objective" posterior is obtained. A disadvantage of the aMCEM is that, if there is an assumption regarding $\mathfrak{h}$, $\mathfrak{h}$ can only be included by switching the centering of the posterior densities from the aEAP to the MAP. However, the APN of latent vectors implies that the influence of $\mathfrak{h}$ vanishes for $d \to \infty$ so that this aspect has only limited practical relevance.

This discussion already suggests that the aMCEM is an interesting alternative to the NIS-MCEM and worth a further look in the advanced development of fitting procedures for IRT models in high dimensional settings. Especially, the simplifications of the aMCEM-algorithm ease the implementation of the method in comparison to the NIS-MCEM which can be very interesting for practitioners without access to a software package with an efficient implementation of an advanced NIS-MCEM algorithm.

A fast deterministic method for the integral approximation in EM-algorithms is the Laplace-approximation. For the first order Laplace approximation, the posterior density of the latent vectors is directly approximated by a normal density using the hessian to estimate the covariance matrix. The computational burden of the integral approximations reduces to an evaluation of the integrand only at the mode while the approximation quality increases for $d \to \infty$, too. As Rabe-Hesketh et al. (2002) argued, the first order Laplace approximation is equivalent to a specific type of adaptive Gauss–Hermite quadrature where only one quadrature point is used. Hence, if we want to control the approximation quality straight-forwardly, we face the same problem as with the adaptive Gauss–Hermite quadrature methods. Another approach to control the approximation quality is given by using higher-order Laplace-approximations, but as Bianconcini et al. (2017) argued the Laplace-approximation is very difficult to implement for an higher order than second. As an alternative, Bianconcini et al. (2017) suggested an approximation of the marginal pmf of $\boldsymbol{Y}$ by a specific sum of several s-dimensional integrals where $0 \leq s \leq q$. These s-dimensional integrals can then be approximated using classic Gauss–Hermite quadrature. By selecting s appropriately small, the curse of dimension of classic deterministic quadrature approaches is avoided. This approach is based on a Taylor-expansion of the logarithm of the integrand at the MAP like the Laplace-approximation. Particularly, the approach is identical to the first order Laplace-approximation for $s = 0$ and selecting $s = q$ results in the adaptive Gauss–Hermite quadrature – in this sense it is a generalization of both approaches. Consequently, this integral approximation method can be used to specify another EM-algorithm for high dimensional settings. A deeper comparison of the overall computation time and precision of the estimates to that of the aMCEM and NIS-MCEM via simulation studies

would be revealing.[15] Similar to the comparison of aMCEM with NIS-MCEM, the aMCEM is an interesting alternative to EM-algorithms using this "intermediate-Laplace-adaptive-Gauss–Hermitte-approximation" due to the simple E- and M-steps of aMCEM while it still uses the adaptive concept to reduce the number of quadrature points.

4.4 Summary and Outlook

In this chapter we justified using of the standardized component statistics of a logit model to approximate the EAP of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ given a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ if $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$ for $\boldsymbol{\eta} \sim \mathcal{H}$ are assumed.

By approximating the posterior expected values, which have to be solved for the M-step of the EM-algorithm using the asymptotic posterior normality of latent vectors, we derived a new alternative approach for a feasible EM-algorithm of logit IRT models in high dimensional settings (aMCEM-algorithm). This new approach has similarities with the normal importance sampling MCEM-algorithm (NIS-MCEM-algorithm). By using the standardized components to approximate the EAP, we simplify the E-step in comparison to NIS-MCEM and eliminate the application of an iterative optimization procedure for each unique response pattern in each E-step. However, we can still update the quadrature points in each E-step to get good integral approximations using a small number of quadrature points. The M-step of the new EM-algorithm simplifies to fitting logistic regression models which reduces the full procedure essentially to:

E-step: Compute the standardized components with the current factor loadings and adapt the quadrature points.

M-step: Fit d binary logistic GLMs to update the model parameters.

This procedure can easily be implemented for arbitrary q while only requiring a comparably small number of quadrature points for high q. This is especially interesting for practitioners because not every statistical software package provides an implementation of a NIS-MCEM-algorithm to fit high dimensional logit IRT models.

While our new approach provides a good approximation of the corresponding integrals for $d \to \infty$ due to the asymptotic posterior normality of latent vectors (cf. the last chapter; at least in our simulation based experiments $d \geq 60$ was sufficient), the approximation of the posterior by a normal distribution can be of limited quality for a small d.

In future research, a comparison of the aMCEM-algorithm, the NIS-MCEM-algorithm and the dimension reduction approach by Bianconcini et al. (2017) – another feasible approach which is based on the Laplace-approximation and adaptive Gauss–Hermite quadrature – with respect to the overall fitting time and the estimation quality for different d, n, q, $\mathcal{H}$ and number of quadrature points M in a comprehensive simulation study would allow to make recommendations which algorithm is ideal for which case in practical data analyzes.

For the aMCEM we used the antithetic variable variance reduction technique to decrease the number of quadrature points in the MC-integration. A simulation study regarding its effect would be interesting. Using heavy-tailed importance distributions for variance reduction with importance sampling is often efficient in MC-integration. This approach

[15] Bianconcini et al. (2017) compared the approach for all $s \in \{0, \ldots, 4\}$ for a four-factor model with each other. A comparison to NIS-MCEM has not jet been published.

could be implemented in another aMCEM-algorithm by using for example a multivariate t-distribution (that is located and scaled corresponding to the posterior distribution) as importance distributon. A simulation study that compares the current antithetic variable aMCEM to importance sampling aMCEM-algorithms could be helpful for selecting the optimal variance reduction technique and, potentially, leads to an improved aMCEM-algorithm. Using another importance distribution instead of the normal distribution in (4.24) or using the aEAP instead of the MAP in the E-Step could also lead to other computationally efficient MCEM-algorithms which would be interesting for future comparisons.

By discussing that classic adaptive quadrature (i.e. the local refinement of the quadrature-grid) cannot be used inside of iterative optimization procedures, Lyness (1983) suggested instead first optimizing using a small number of quadrature points M and comparably large termination $\varepsilon > 0$. After convergence of the procedure, it restarts with a reduced ε and increased M. If the optimal values found differ strongly, another increment of M and restart of the procedure can reveal if M is still to small. The comparison of the results for different M is an established approach to check if the integral approximation was sufficiently good. An automatic procedure which tests for the need to increase M and successively reduces ε could be an improvement and could possibly accelerate the overall procedure because the increased M is used only where it is needed.

Chapter 5

Pseudo Likelihood Estimation for Binary Logit IRT Models

Marginal maximum likelihood estimation (MMLE) is a common and efficient method for fitting IRT models. For large data sets and high dimensional models, however, it is computationally challenging. In Chapter 4, we have contributed an improvement for fitting logit models using MMLE but the derived aMCEM-algorithm still is computationally intensive in large, high dimensional settings. Anderson et al. (2007), Li (2010), Paek (2016) and Paek and Anderson (2017) developed computationally efficient pseudo likelihood (PL) methods for fitting specific generalizations of IRT models, i.e. models such as confirmatory versions of the manifest probabilities model (1.21) and extensions but without the requirement that an equivalent latent variable model exists (compare the next section for more details). Their reported success for fast model fitting indicates that their approach also is a foundation for a solution for similar models. For example exploratory models and the subclass of models for which there is always an equivalent latent variable model in order to utilize the estimation procedure for latent variable models.

The models fitted using PL estimation (PLE) cannot generally be interpreted as latent variable models. This prohibits fitting latent variable models using the current PLE algorithms. Additionally, for non-Rasch-type models, the implemented algorithm available is substantially memory consuming and comparably time intensive for data sets with a larger number of items – especially for multidimensional models.

To overcome these problems, we develop a PL method which results in parameter estimates that can be used with certainty for estimating the parameters of exploratory logit IRT models. We then derive two algorithms for computing the PL estimates that can be implemented efficiently.

Our experiments and simulation studies show that our implementations of the two algorithms in R solve the noted problems of the currently available implementation for PLE. Particularly, we achieve time and memory efficient estimates for high dimensional models for large data sets. Our approach therefore proves to be a promising foundation for fitting latent variable models in high dimensional settings. A remaining problem is a bias of the PL estimator for the latent variable model which leads to inconsistent estimates. Our simulation study shows that the bias is principally correctable. Future research could focus on a simple transformation of the PL estimates that either leads to a consistent estimator or at least to a reduction of the bias to a practically negligible order.

This chapter is organized as follows. First, we revisit the literature regarding the PL method with focus on the discussed models, algorithmic approaches and results in Section 5.1. In Section 5.2, we first discuss the considered manifest probabilities model which is the base for our further developments of the PL approach, followed by discussions of two algorithmic approaches for computing the PL estimates. Then, we present and compare

three possible transformations of the PL estimates obtained to achieve estimates for the IRT in which we are primarily interested. In Section 5.3, we analyze the bias and the efficiency of the PL estimator (PLE) for the IRT model in comparison to an optimal linear corrected PLE. To get an impression of the usability of the PLE as an alternative to MMLE for the IRT model of interest, we compare the two methods in Section 5.4. Finally, we summarize the results and point out future research directions regarding PLE for logit IRT models in Section 5.5.

5.1 Revisit of the Approach for Rasch and Confirmatory Logit IRT Models in a Generalized Sense

Anderson et al. (2007) presented the pseudo likelihood approach as a composite likelihood method of fully conditional densities for a special type of Rasch models. They studied the model

$$\log(\mathsf{P}(\boldsymbol{Y}=\boldsymbol{y})) = a + \boldsymbol{b}^\intercal\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\intercal\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal\boldsymbol{y}, \qquad \boldsymbol{y}\in\{0,1\}^d, \tag{5.1}$$

for a known matrix $\boldsymbol{A}\in\{0,1\}^{d\times q}$, unknown parameters $\boldsymbol{b}\in\mathbb{R}^d$ and $\boldsymbol{\Sigma}=(\sigma_{ij})_{i,j=1}^{q,q}\in\mathbb{R}^{q\times q}$ where $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^\intercal$ and $a\in\mathbb{R}$ is the normalization constant. For this model the conditional pmf of Y_i given $Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},Y_{i+1}=y_{i+1},\ldots,Y_d=y_d$ is given by

$$\begin{aligned}&\mathsf{P}(Y_i=y_i \mid Y_v=y_v,\, v=1,\ldots,d, v\neq i)\\ &\quad=\frac{\exp\left(y_i\left(b_i+\sum_{j=1}^q\sum_{r=1}^q\sum_{v=1,v\neq i}^d\alpha_{ij}\sigma_{jr}\alpha_{vr}y_v\right)\right)}{1+\exp\left(b_i+\sum_{j=1}^q\sum_{r=1}^q\sum_{v=1,v\neq i}^d\alpha_{ij}\sigma_{jr}\alpha_{vr}y_v\right)}, \qquad y_i\in\{0,1\},\end{aligned} \tag{5.2}$$

for arbitrary $y_v\in\{0,1\}$, $v=1,\ldots,i-1,i+1,\ldots,d$ and each $i\in\{1,\ldots,d\}$ (cf. equation (1.33)).

For $n\in\mathbb{N}$ independent observations $\boldsymbol{Y}_1=\boldsymbol{y}_1,\ldots,\boldsymbol{Y}_n=\boldsymbol{y}_n$ from (5.1) we can use (5.2) to formulate a *pseudo likelihood function* (of fully conditional densities) by defining

$$\begin{aligned}\mathcal{L}(\boldsymbol{\vartheta};\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) &\coloneqq \prod_{i=1}^d\prod_{k=1}^n \mathsf{P}(Y_{ik}=y_{ik}\mid Y_{vk}=y_{vk},\, v=1,\ldots,d, v\neq i)\\ &= \prod_{i=1}^d \mathcal{L}_i(\boldsymbol{\vartheta}_i;\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_{i-1},\boldsymbol{\vartheta}_{i+1},\ldots,\boldsymbol{\vartheta}_d,\boldsymbol{\Sigma}),\end{aligned} \tag{5.3}$$

for $\boldsymbol{\vartheta}=(b_1,\ldots,b_d,\sigma_{11},\ldots,\sigma_{1q},\sigma_{22},\ldots,\sigma_{2q},\ldots,\sigma_{qq})^\intercal\in\mathbb{R}^{d+q(q+1)/2}$ and $\boldsymbol{y}_k\in\{0,1\}^d$, $k=1,\ldots,n$ where $\mathcal{L}_i$ is the product of all conditional pmfs in (5.2) for item i with

$$\begin{aligned}&\mathcal{L}_i(\boldsymbol{\vartheta}_i;\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_{i-1},\boldsymbol{\vartheta}_{i+1},\ldots,\boldsymbol{\vartheta}_d,\boldsymbol{\Sigma})\\ &\qquad\coloneqq\prod_{k=1}^n\frac{\exp\left(y_{ik}\left(b_i+\sum_{j=1}^q\sum_{r=1}^q\sum_{v=1,v\neq i}^d\alpha_{ij}\sigma_{jr}\alpha_{vr}y_{vk}\right)\right)}{1+\exp\left(b_i+\sum_{j=1}^q\sum_{r=1}^q\sum_{v=1,v\neq i}^d\alpha_{ij}\sigma_{jr}\alpha_{vr}y_{vk}\right)},\end{aligned} \tag{5.4}$$

and $\boldsymbol{\vartheta}_i=(b_i,\alpha_{i1},\ldots,\alpha_{iq})^\intercal$, $i=1,\ldots,d$. We call each global maximum $\hat{\boldsymbol{\vartheta}}_n\equiv\hat{\boldsymbol{\vartheta}}_n(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n)$ of $\mathcal{L}(\boldsymbol{\vartheta};\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n)$ a *pseudo likelihood estimator*. Anderson et al. (2007) proved that the

PLE is consistent for $n \to \infty$. They moreover discussed that $\mathcal{L}$ can feasibly be maximized using logistic regression software and that the procedure is feasible in cases where the MLE for n independent observations from (5.1) no longer is.[1]

Remark 5.1.1
If $\boldsymbol{\Sigma}$ is positive definite, then model (5.1) is equivalent to the logit IRT model

$$\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{b}, \qquad (\boldsymbol{\eta} \mid \boldsymbol{Y} = \boldsymbol{0}) \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{5.5}$$

(cf. Section 1.2 for a detailed discussion of the equivalence of the two models). The interpretation of model (5.1) in the sense of the IRT model (5.5) is not always available because for model (5.1) it is only required that $\boldsymbol{\Sigma}$ is symmetric but it is not required that $\boldsymbol{\Sigma}$ is positive semi-definite.

Model (5.5) is a Rasch-type model because the matrix $\boldsymbol{A}$ of the factor loadings is a known constant matrix. Model (5.1) can therefore be considered a *generalized Rasch model*. See Subsection 1.2.4 for a discussion of the distributional assumption of $\boldsymbol{\eta}$ in (5.5) and also see Chapter 3 for details regarding its asymptotic justification for $d \to \infty$.

The consistency result of Anderson et al. (2007) cannot be applied if the PLE is used to estimate $\boldsymbol{b}$ for a Rasch-type model with $\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{b}$ and $\boldsymbol{\eta} \sim \mathcal{H}$ for some q-variate distribution $\mathcal{H}$ so that posterior normality of $\boldsymbol{\eta}$ given $\boldsymbol{Y} = \boldsymbol{y}$ for $\boldsymbol{y} \in \{0,1\}^d$ does not hold. If $\mathcal{H}$ satisfies (CS1') or (PCT1)[2], we get asymptotic posterior normality either in the sense of semiproper or proper centering of the posterior for $d \to \infty$ so that we can still justify the application of the PLE for comparably large d. The simulations by Anderson et al. (2007) and our new simulations in Subsection 1.3.3.6 additionally support the application of the PLE to fit Rasch models without assuming posterior normaliy. ◇

Later, Li (2010) extended the PLE to generalized Rasch models (in the sense of Remark 5.1.1) with explanatory variables. Her simulation studies support the use of the PLE for these models and underline that it is a fast procedure.

Anderson and Yu (2007) briefly mentioned the idea of generalizing the PLE to estimate the unknown discrimination parameters of a manifest probabilities model which corresponds to the 2PL model (cf. Example 1.2.5). They sketched an iterative two-step procedure for this estimation where the PLE of Anderson et al. (2007) is repeatedly applied in the first step. Paek (2016) worked this approach out in detail, extended it to polytomous and multidimensional settings and evaluated the approach via a simulation study (cf. also the publication of the results by Paek and Anderson, 2017, PA hereafter).

The generalization of the PLE for models with unknown factor loadings presented by them is derived under the restriction that it is known on which items each latent variable has an effect, i.e. they considered a confirmatory logit model. For formal clarity, we suppose that the zeros in the matrix $\boldsymbol{A}$ of the factor loadings – i.e. the items on which a latent variable does not has any direct effect – are placed in a such way that the only rotation matrix $\boldsymbol{Q} \in \mathbb{R}^{q \times q}$, which ensures that the zero elements of $\boldsymbol{AQ}$ are at exactly the same positions as $\boldsymbol{A}$, is the identity, i.e. $\boldsymbol{Q} = \mathrm{I}_q$. To ensure identifiability of the model parameters further constraints are required. PA considered $\sigma_{jj} = 1$, $j = 1, \ldots, q$, for this purpose. The algorithm which PA discussed to estimate $\boldsymbol{b}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ based on an iid sample $\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n$ from (5.1) is then given as follows.

[1] The MLE for model (5.1) becomes infeasible for increasing d because all *possible* response patterns have to be considered to evaluate the normalization constant a. Evaluating (5.3) remains feasible because it is independent of a.

[2] Cf. p. 112 or 139, respectively.

Algorithm 2 Pseudo Likelihood Estimation for confirmatory logit models by PA (PLE-PA)

Require: response vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \{0,1\}^d$, latent dimension $q \in \mathbb{N}$, number of iterations $T \in \mathbb{N}$, positions $(i_p, j_p) \in \{1, \ldots, d\} \times \{1, \ldots, q\}$, $p = 1, \ldots, z \in \{(q-1)q/2, \ldots, dq - d\}$ where $\alpha_{i_p j_p} = 0$ is fixed, stating value $\boldsymbol{\vartheta}^{(0)} \coloneqq (b_1^{(0)}, \ldots, b_d^{(0)}, \alpha_{i^{(\mathrm{f})}j^{(\mathrm{f})}}^{(0)}, \ldots, \alpha_{i^{(\mathrm{l})}j^{(\mathrm{l})}}^{(0)}, \sigma_{12}^{(0)}, \ldots, \sigma_{(q-1)q}^{(0)})^\intercal \in \mathbb{R}^{d+qd-z+(q-1)q/2}$ where $i^{(s)}j^{(\mathrm{f})}$ and $i^{(l)}j^{(\mathrm{l})}$ denote the first and the last position where α_{ij} is not held fixed

for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, d$ **do**
    set $\boldsymbol{\vartheta}_i^{(t)} = (b_i^{(t)}, \alpha_{i1}^{(t)}, \ldots, \alpha_{iq}^{(t)})^\intercal$ to an arbitrary maximal point of $\mathcal{L}_i$ of (5.4) in

$$\Theta_i \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{q+1} \colon x_k = 0 \text{ if } \alpha_{i(k-1)} = 0 \text{ is known, } x_k \in \mathbb{R} \text{ else } k = 1, \ldots, q+1\}.$$

    where $\boldsymbol{\vartheta}_v = \boldsymbol{\vartheta}_v^{(t)}$, $v = 1, \ldots, i-1$ and $\boldsymbol{\vartheta}_v = \boldsymbol{\vartheta}_v^{(t-1)}$, $v = i+1, \ldots, d$ as well as $\sigma_{jr} = \sigma_{rj} = \sigma_{jr}^{(t-1)}$, $j = 1, \ldots, q-1$, $r = j+1, \ldots, q$ and $\sigma_{jj} = 1$, $j = 1, \ldots, q$ are held fix and $\alpha_{ij}^{(t^*)} \coloneqq 0$ if position ij is held fixed, $t^* = t-1, t$
  end for
  set $(b_1^{(t)}, \ldots, b_d^{(t)}, \sigma_{12}^{(t)}, \ldots, \sigma_{(q-1)q}^{(t)})^\intercal$ to an arbitrary maximum point of $\mathcal{L}$ in (5.3) where $\sigma_{jj} = 1$, $j = 1, \ldots, q$ and

$$\alpha_{ij} = \begin{cases} \alpha_{ij}^{(t)}, & \text{if } \alpha_{ij} \text{ is not held fixed,} \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, d,\ j = 1, \ldots, q$$

  set $\boldsymbol{\vartheta}^{(t)} \coloneqq (b_1^{(t)}, \ldots, b_d^{(t)}, \alpha_{i^{(\mathrm{f})}j^{(\mathrm{f})}}^{(t)}, \ldots, \alpha_{i^{(\mathrm{l})}j^{(\mathrm{l})}}^{(t)}, \sigma_{12}^{(t)}, \ldots, \sigma_{(q-1)q}^{(t)})^\intercal$
end for
return $\boldsymbol{\vartheta}^{(T)}$

Remark 5.1.2

1. Each iteration consists of two primary steps:
   (i) Update $\boldsymbol{b}$ and $\boldsymbol{A}$.
   (ii) Update $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$.

   The algorithm is a two-step algorithm in this sense. Especially, $\boldsymbol{b}$ is updated twice in each iteration.

2. PA discuss that if $q = 1$, the step of updating $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ in each iteration can be dropped.

3. Each maximization can be performed using logistic regression software for which the currently fixed parameters are considered as explanatory variables.

4. In each application of this algorithm, all T iterations are computed. Paek (2016, p. 152–153) discussed to decide about the question if the algorithm converged by monitoring over t if the single item log-likelihoods change, the complete log-likelihood changes or the parameters do not change. If there still are greater changes in any parameter in one of the last iterations, the algorithm can be restarted with an increased T. ◇

Theoretical results on the convergence of this algorithm are not available.[3] During their simulations, PA at least did not observe a sensitivity of the algorithm with respect to the starting values. Neither did they observe any sign of finding local optima which are no global optima. In the cases where the MPMLE (manifest probabilities MLE, i.e. the MLE of model (5.1)) with the same restriction on $\boldsymbol{A}$ as Algorithm 2 was feasible, they observed a remarkable similarity of the estimates achieved with MPMLE to them achieved with PLE ($d = 4$, 6 and $n = 200$, 500 and 1000). In the other cases, they observed high correlations between the pseudo likelihood estimates and the true parameter values (they observed average correlation coefficients of 0.898 to 0.999 between the vector of estimates and the true parameter vector for $d = 20$, 50 and $n = 200$, 500, 1000).[4] Both observations support the use of the PLE to fit manifest probabilities models. Moreover, they mentioned the fast convergence and simple implementation which make the approach an interesting method for the cases of a large d where MPMLE cannot be computed feasibly.

In her conclusion, Paek (2016, p. 149, 152–153) discussed limitations of her implementation of the PLE and suggested improvements of the implementation as a target for future research. Especially, she mentioned the current inclusion of the termination criteria. But more importantly, her program is resource-consuming which can possibly be improved by a new implementation. For the application of the PLE on data sets with 50 items, she needed a computer with 16GB RAM while 8 GB RAM where sufficient for 20 or less items.[5] Moreover, she noted that she needed up to 12 minutes to fit 20 item models with up to 1000 responses while the computation time increased up to 2 hours and 45 minutes for 50 items. Also taking into account the improvement in memory, the current PLE implementation becomes infeasible for many modern computers for data sets with a large number of items. Notice that modern computer development focuses on multiple cores instead of an increased frequency while parallel computing makes the implementation much more complicated. This implies that simply waiting for much more powerful single core CPUs is no solution. A new implementation which avoids this issue is therefore integral in making the PLE applicable for actual high dimensional IRT data.

Another aspect in the case of $q > 1$ using the current approaches is the possibility that $\widehat{\boldsymbol{\Sigma}}$ can be not positive (semi-)definite, meaning that it is not possible to consider the estimates as estimates for the IRT model (5.5).

In the following we discuss our new implementations for the pseudo likelihood estimation which were feasible in all our experiments. As in the previous parts of this thesis, we will consider the exploratory IRT models where it is a-priori completely unknown on which items each latent variable has an effect. In this chapter, this provides us with the benefit that we can discuss an estimation method the results of which always can be considered

[3] Paek and Anderson (2017) indicated it as a open question. An answer has not yet been published to the best of our knowledge.

[4] Paek (2016, p. 62–85) also compared PLE and MPMPLE to the MMLE for the two parameter logit model with $\mathcal{H} = \mathcal{N}(0, 1)$ in the special case of $q = 1$. But she neither mentioned whether she transformed the manifest probabilities maximum likelihood (MPML) and pseudo likelihood (PL) estimates correspondingly nor how she transformed them in this case. In this special case, we always have $\boldsymbol{\Sigma} = 1$ which is positive (definite). Hence, we can interpret the MPML and/or PL estimates according to model (5.5). But the first two moments of $\mathcal{N}(0, 1)$ and the distribution of $\boldsymbol{\eta}$ in (5.5) are always different. Because these moments have a strong influence on the values of the other parameters, it is absolutely necessary to transform the estimates in order to compare them to MMLE with $\mathcal{H} = \mathcal{N}(0, 1)$. Therefore, the results of this comparison are difficult to interpret. In the later publication by PA, this comparison is not included so there is no clarification.

[5] She also mentioned the switch from a 3.5GHz CPU to a 3.6 GHz CPU but this is an almost negligible improvement in most applications.

estimates of the IRT model $\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{b}$, $\boldsymbol{\eta} \sim \mathcal{H}$ where $\mathcal{H}$ is unknown except identification constraints on the first two moments. For the model in this section, additional constraints would be required to ensure that $\widehat{\boldsymbol{\Sigma}}$ is positive definite. These additional constraints would also require a further development of the method.

5.2 Pseudo Likelihood Estimation for Exploratory Logit IRT Models

In this section we develop the pseudo likelihood estimation for model (5.5) in cases where $\boldsymbol{A}$ is entirely unknown. First, we discuss the PLE for the corresponding manifest probabilities model (i.e. (5.1) with unknown $\boldsymbol{A} \in \mathbb{R}^{d \times q}$) and how we can easily ensure that $\boldsymbol{\Sigma}$ (or $\widehat{\boldsymbol{\Sigma}}$, respectively) is positive definite so that the estimates can always be considered as estimates for an IRT model. We will derive two algorithmic approaches, compare them and discuss whether they are improvements with respect to the research perspective which was mentioned by Paek (2016, p. 152–153) and which we summarized in the last section.

Next, we examine how the estimates obtained can be transformed such that we can use them to estimate the parameters $\boldsymbol{A}$ and $\boldsymbol{b}$ of the general exploratory logit IRT model

$$\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{b}, \qquad \boldsymbol{\eta} \sim \mathcal{H}$$

with an unknown $\mathcal{H}$ except the location and scale identification constraints $\mathsf{E}(\boldsymbol{\eta}) = \boldsymbol{0}$ and $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$.

For this model, the rotational indeterminacy of $\boldsymbol{A}$ can be resolved by applying an arbitrary common approach to the obtained $\hat{\boldsymbol{A}}$. In this work, we use the varimax solution of $\hat{\boldsymbol{A}}$, other approaches can be used equivalently.

5.2.1 Fitting the Manifest Probabilities Model

5.2.1.1 Preliminaries

Now we restrict the model parameters to the case that $\boldsymbol{\Sigma}$ is positive definite. Suppose an arbitrary $\boldsymbol{A} \in \mathbb{R}^{d \times q}$ and an arbitrary symmetric positive definite $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ with $q < d$. Then there always is a symmetric positive definite square root matrix $\boldsymbol{\Sigma}^{1/2}$ in such way that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$. In consequence, we get

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\intercal = \boldsymbol{A}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^\intercal = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\intercal,$$

with $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Sigma}^{1/2}$. Moreover, let $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{q \times q}$ be another positive definite matrix, then we get

$$\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\intercal = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\Sigma}}^{-1/2}\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\Sigma}}^{-1/2}\tilde{\boldsymbol{A}}^\intercal = \boldsymbol{A}^*\tilde{\boldsymbol{\Sigma}}\boldsymbol{A}^{*\intercal},$$

with $\boldsymbol{A}^* = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\Sigma}}^{-1/2}$ and, especially, $\boldsymbol{A}^* = \boldsymbol{A}$ if $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$. Therefore, if $\boldsymbol{A}$ is completely unknown in model (5.1), $\boldsymbol{\Sigma}$ is not identifiable in the set of all possible positive definite $q \times q$ matrices. The positive definiteness of $\boldsymbol{\Sigma}$ hence can be ensured by restricting $\boldsymbol{\Sigma}$ to an arbitrary constant positive definite matrix. The most simple case is of course

$$\boldsymbol{\Sigma} = \mathrm{I}_q. \tag{5.6}$$

This especially simplifies the formula of the conditional pmf of Y_i under Y_v, $v \neq i$, (i.e. (5.2)) to the form

$$\begin{aligned} &\mathsf{P}(Y_i = y_i \mid Y_v = y_v,\, v = 1, \ldots, d, v \neq i) \\ &\quad = \frac{\exp\left(y_i \left(b_i + \sum_{j=1}^{q} \sum_{v=1, v\neq i}^{d} \alpha_{ij}\alpha_{vj} y_v\right)\right)}{1 + \exp\left(b_i + \sum_{j=1}^{q} \sum_{v=1, v\neq i}^{d} \alpha_{ij}\alpha_{vj} y_v\right)}, \qquad y_i \in \{0, 1\}, \end{aligned} \tag{5.7}$$

for all $i = 1, \ldots, d$. This also simplifies the implementation and thus can increase the computational efficiency slightly. The single item pseudo likelihood functions $\mathcal{L}_i$ in (5.4) change analogously, i.e. plug in $\boldsymbol{\Sigma}$ of (5.6) in (5.4). The new overall pseudo likelihood function becomes

$$\mathcal{L}(\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \prod_{k=1}^{n} \prod_{i=1}^{d} \frac{\exp\left(y_{ik} \left(b_i + \sum_{j=1}^{q} \sum_{v=1, v\neq i}^{d} \alpha_{ij}\alpha_{vj} y_{vk}\right)\right)}{1 + \exp\left(b_i + \sum_{j=1}^{q} \sum_{v=1, v\neq i}^{d} \alpha_{ij}\alpha_{vj} y_{vk}\right)}, \tag{5.8}$$

with $\boldsymbol{\vartheta} = (b_1, \ldots, b_d, \alpha_{11}, \ldots, \alpha_{dq})^\intercal \in \mathbb{R}^{(q+1)d}$.

Remark 5.2.1
If there is interest mainly in the manifest probabilities model

$$\log(\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y})) = a + \boldsymbol{b}^\intercal \boldsymbol{y} + \frac{1}{2} \boldsymbol{y}^\intercal \boldsymbol{\Gamma} \boldsymbol{y}, \qquad \boldsymbol{y} \in \{0, 1\}^d, \tag{5.9}$$

for a symmetric $(\gamma_{iv})_{i,v=1}^{d,d} = \boldsymbol{\Gamma} \in \mathbb{R}^{d \times d}$ with $\gamma_{ii} = 0$, $i = 1, \ldots, d$, $\boldsymbol{b} \in \mathbb{R}^d$ and the normalization constant $a \in \mathbb{R}$, PLE can also be used to estimate $\boldsymbol{\Gamma}$ and $\boldsymbol{b}$.[6] In this case we can replace the $d - 1$ sums $\sum_{j=1}^{q} \alpha_{ij}\alpha_{vj}$ by the parameters $\gamma_{iv} \in \mathbb{R}$ for $i \neq v$ in the definition of the pseudo likelihood function $\mathcal{L}$ in (5.8). Due to the conditions obtained by Liu (1994), the symmetry of $\boldsymbol{\Gamma}$ is sufficient to ensure that the fully conditional probabilities uniquely imply the marginal distribution of $\boldsymbol{Y}$ in (5.9) (compare Subsection 1.2.3). ◇

The algorithm of PA can be adapted to be applicable here. Additionally, we examine a direct approach to maximize $\mathcal{L}$. We analyze both approaches in the following.

5.2.1.2 A Two-Stage Iterative Algorithm for Exploratory Models

For the adaption of the PLE-PA algorithm to the exploratory case we remove the step of updating $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ in each iteration because $\boldsymbol{\Sigma}$ is now fixed and $b_1, \ldots, b_d$ are updated during the first step of updating $b_i, \alpha_{i1}, \ldots, \alpha_{iq}$ for $i = 1, \ldots, d$. PA also suggested this simplification of the algorithm for the case of $q = 1$ where $\boldsymbol{\Sigma} = \sigma_{11} = 1$. Our multidimensional case has the similarity of a known $\boldsymbol{\Sigma}$ and can use this simplification. The algorithm for multidimensional exploratory models is therefore no longer a two-step algorithm, but it still repeatedly applies an iterative optimization procedure so that it can be regocnized as a *two-stage iterative algorithm*. Additionally, we add termination criteria to improve the

[6] Model (5.9) can be parameterized equivalently if the diagonal of $\boldsymbol{\Gamma}$ is set to zero by reparameterizing $\boldsymbol{b}$ correspondingly. Since the diagonal elements are not identifiable and the corresponding parameters vanish in the formulation of the conditional probabilities, it is easier to formulate model (5.9) with $\gamma_{ii} = 0$, $i = 1, \ldots, d$.

general applicability. We use an absolute and a relative check regarding the d item-wise-steps in the parameter and additionally implement a check regarding the maximal change in $\mathcal{L}_i$ in each iteration. The algorithm terminates if any of these three criteria are satisfied. Other criteria are also possible but in our experiments the algorithm worked fine using these and the termination was always caused by the change in the function value. The overall algorithm is then given as follows (details of our implementation in R are discussed in Section C.3 in the appendix).

Algorithm 3 Two-Stage Pseudo Likelihood Estimation for exploratory logit models (2SPLE)

Require: response vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \{0,1\}^d$, latent dimension $q \in \mathbb{N}$, maximal number of iterations $T \in \mathbb{N}$, stating value $\boldsymbol{\vartheta}^{(0)} \coloneqq (b_1^{(0)}, \ldots, b_d^{(0)}, \alpha_{11}^{(0)}, \ldots, \alpha_{dq}^{(0)})^\intercal \in \mathbb{R}^{d+qd}$, termination criterion $\varepsilon > 0$

set $\ell_i^{(0)}(\boldsymbol{\vartheta}_i^{(0)}) \coloneqq -10^{100}$, $i = 1, \ldots, d$
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, d$ **do**
    set $\boldsymbol{\vartheta}_i^{(t)} = (b_i^{(t)}, \alpha_{i1}^{(t)}, \ldots, \alpha_{iq}^{(t)})^\intercal$ to an arbitrary maximal point of $\ell_i^{(t)} \colon \mathbb{R}^{d+1} \to \mathbb{R}$,

$$\boldsymbol{\vartheta}_i \mapsto \sum_{k=1}^{n} \left(y_{ik} \left(b_i + \sum_{j=1}^{q} \alpha_{ij} \sum_{\substack{v=1 \\ v \neq i}}^{d} \alpha_{vj} y_{vk} \right) - \log \left(1 + \exp \left(b_i + \sum_{j=1}^{q} \alpha_{ij} \sum_{\substack{v=1 \\ v \neq i}}^{d} \alpha_{vj} y_{vk} \right) \right) \right), \tag{5.10}$$

    where $\alpha_{vj} \coloneqq \alpha_{vj}^{(t)}$ if $v < i$ and $\alpha_{vj} \coloneqq \alpha_{vj}^{(t-1)}$ if $v > i$
  end for
  set $\boldsymbol{\vartheta}^{(t)} \coloneqq (b_1^{(t)}, \ldots, b_d^{(t)}, \alpha_{11}^{(t)}, \ldots, \alpha_{dq}^{(t)})^\intercal$
  set

$$\Delta_1^{(t)} \coloneqq \max_{i \in \{1,\ldots,d\}} \left\{ \|\boldsymbol{\vartheta}_i^{(t)} - \boldsymbol{\vartheta}_i^{(t-1)}\| - \varepsilon \right\}$$
$$\Delta_2^{(t)} \coloneqq \max_{i \in \{1,\ldots,d\}} \left\{ \|\boldsymbol{\vartheta}_i^{(t)} - \boldsymbol{\vartheta}_i^{(t-1)}\| - \varepsilon \|\boldsymbol{\vartheta}_i^{(t)}\| \right\}$$
$$\Delta_3^{(t)} \coloneqq \max_{i \in \{1,\ldots,d\}} \left\{ |\ell_i^{(t)}(\boldsymbol{\vartheta}_i^{(t)}) - \ell_i^{(t-1)}(\boldsymbol{\vartheta}_i^{(t-1)})| - \varepsilon \right\}$$

  if $\min(\Delta_1^{(t)}, \Delta_2^{(t)}, \Delta_3^{(t)}) < 0$ **then**
    break
  end if
end for
return $\boldsymbol{\vartheta}^{(t)}$

Remark 5.2.2
The maximization of (5.10) can be done using logistic regression software because $\sum_{v=1, v\neq i}^{d} \alpha_{vj} y_{vk}$, $j = 1, \ldots, q$, $k = 1, \ldots, n$ is known in each iteration and for each i. ◇

Next we derive a direct approach to maximize the pseudo likelihood function and then we will discuss both algorithms.

5.2.1.3 A Direct Algorithm for Exploratory Models

The pseudo likelihood function (5.8) implies the following pseudo log-likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\vartheta}) &:= \log\left(\mathcal{L}(\boldsymbol{\vartheta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\right) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{d}\left(y_{ik}\left(b_i + \sum_{j=1}^{q}\sum_{\substack{v=1\\v\neq i}}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right) - \log\left(1+\exp\left(b_i + \sum_{j=1}^{q}\sum_{\substack{v=1\\v\neq i}}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right)\right)\right)\end{aligned} \tag{5.11}$$

for $\boldsymbol{\vartheta} \in \mathbb{R}^{(q+1)d}$. We will compute the gradient of ℓ next. First, the partial derivative with respect to b_z, $z \in \{1, \ldots, d\}$ is given by

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial b_z} &= \sum_{k=1}^{n}\left(y_{zk} - \frac{\exp\left(b_i + \sum_{j=1}^{q}\sum_{v=1,v\neq i}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right)}{1+\exp\left(b_i + \sum_{j=1}^{q}\sum_{v=1,v\neq i}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right)}\right) \\ &= \sum_{k=1}^{n}\left(y_{zk} - \pi_{ik}(\boldsymbol{\vartheta})\right),\end{aligned} \tag{5.12}$$

where

$$\begin{aligned}\pi_{ik}(\boldsymbol{\vartheta}) :=& \frac{\exp\left(b_i + \sum_{j=1}^{q}\sum_{v=1,v\neq i}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right)}{1+\exp\left(b_i + \sum_{j=1}^{q}\sum_{v=1,v\neq i}^{d}\alpha_{ij}\alpha_{vj}y_{vk}\right)} \\ =& \mathsf{P}(Y_{zk} = 1 \mid Y_{ik} = y_{ik},\, i = 1, \ldots z-i, z+1, \ldots, d),\end{aligned} \tag{5.13}$$

for all $i = 1, \ldots, d$, $k = 1, \ldots, n$, $\boldsymbol{\vartheta} \in \mathbb{R}^{(q+1)d}$. For $z \in \{1, \ldots, d\}$ and $h \in \{1, \ldots, q\}$ we get:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \alpha_{zh}} =& \sum_{k=1}^{n}\left(y_{zk}\sum_{\substack{v=1\\v\neq z}}^{d}\alpha_{vh}y_{vk} - \pi_{zk}(\boldsymbol{\vartheta})\sum_{\substack{v=1\\v\neq z}}^{d}\alpha_{vh}y_{vk}\right) \\ &+ \sum_{k=1}^{n}\sum_{\substack{v=1\\v\neq z}}^{d}\left(y_{vk}\alpha_{zh}y_{zk} - \pi_{vk}(\boldsymbol{\vartheta})\alpha_{zh}y_{zk}\right) \\ =& \sum_{k=1}^{n}\sum_{\substack{v=1\\v\neq z}}^{d}\alpha_{vh}\left(2y_{zk}y_{vk} - y_{zk}\pi_{vk}(\boldsymbol{\vartheta}) - \pi_{zk}(\boldsymbol{\vartheta})y_{vk}\right),\end{aligned} \tag{5.14}$$

for all $\boldsymbol{\vartheta} \in \mathbb{R}^{(q+1)d}$. Now, we get $\nabla\ell(\boldsymbol{\vartheta})$ from (5.12) and (5.14) with

$$\nabla\ell(\boldsymbol{\vartheta}) = \left(\frac{\partial\ell(\boldsymbol{\vartheta})}{\partial b_1}, \ldots, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial b_d}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial \alpha_{11}}, \ldots, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial \alpha_{d1}}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial \alpha_{12}}, \ldots \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial \alpha_{dq}}\right)^{\mathsf{T}}, \qquad \boldsymbol{\vartheta} \in \mathbb{R}^{(q+1)d}.$$

This directly provides the pseudo likelihood equations $\nabla\ell(\boldsymbol{\vartheta}) = \mathbf{0}$. Efficient computation of pseudo likelihood estimates can now be obtained by maximizing ℓ using numerical nonlinear

optimization techniques. Since the gradient $\nabla\ell$ can be evaluated appropriately and is continuously differentiable, gradient based methods are recommended for local optimization (cf. e.g. Nocedal and Wright, 2006). We used a Newton-type algorithm for our implementation – further aspects of our implementation in R and the selected algorithm are discussed in Section C.3 in the appendix. A pseudo likelihood estimate is then given by a local maximal point $\hat{\boldsymbol{\vartheta}} = (\hat{b}_1, \ldots, \hat{b}_d, \hat{\alpha}_{11}, \ldots, \hat{\alpha}_{dq})^\intercal \in \mathbb{R}^{(q+1)d}$ of ℓ. In particular, we see that ℓ and $\nabla\ell$ can be evaluated directly for arbitrary d and q. So the method remains feasible in cases where MMLE or manifest probabilities MLE are computationally difficult to apply.

5.2.1.4 Discussion and Comparison

In our experiments, our two newly implemented methods forced the R-session for their application to allocate less than 100 MB additional RAM to estimate the parameters of a model with $d = 120$, $q = 8$ and $n = 10000$. Hence, even the parallel execution of seven R-sessions using PLE to fit models of this dimension on a computer with an eight-code CPU and 8GB RAM is not critical.[7] From this point of view, both of our new implementations are successful improvements to the implementation of PLE by Paek (2016, p. 152–153) in her setting and with regard to the possible future developments mentioned by her in the conclusion.

In our experiments and studies, both PLE implementations were not sensitive with respect to the starting value. For arbitrary and randomly generated starting values, both algorithms always converged to equivalent solutions. This indicates that neither possible local optima nor selecting good starting values are critical here. In the final estimates, we observed minimal differences between the results of both algorithms which were usually slightly better for the direct approach. But these differences can be explained by differences in the termination criteria and were at a level of negligible practical importance.

A more important aspect is the possible difference of the computation time needed to obtain the estimates. We analyze this in a small simulation study. We generate $n \in \{1000, 3000, 5000\}$ responses from the standard logit-normal IRT model

$$\operatorname{logit}\big(P_i(\boldsymbol{\eta})\big) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij}\eta_j, \qquad i = 1, \ldots, d, \qquad \boldsymbol{\eta} \sim \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$$

with $d \in \{20, 40, 60\}$ items and latent vectors of dimension $q \in \{1, 2\}$. The α_{i0}s were drawn independently from $\mathcal{U}([-2, 2])$. The α_{i1}s for $q = 1$ are drawn in such a way that $\log(\alpha_{11}), \ldots, \log(\alpha_{d1}) \overset{\text{iid}}{\sim} \mathcal{N}(-0.5,\, 0.5)$ to ensure that they are all positive – corresponding to the usual assumption –, mostly in a region around one but a few large values are also possible. For $q = 2$ we generated α_{i1}^{pre}, $i = 1, \ldots, 60$, independently from $\mathcal{U}([-0.5, 1])$. We further generated α_{i2}^{pre}, $i = 1, \ldots, 60$, by a random permutation of the α_{i1}^{pre}s. Then the actual loadings are the varimax rotation of $\boldsymbol{A}^{\text{pre}} := (\alpha_{ij}^{\text{pre}})_{i,j=1}^{d,2}$. In multidimensional settings the loadings can be negative. Using this generation, about 1/3 of the loading values are negative. Generally, these parameters ensure that the probability of all responses to single items being 0 (or 1) is negligibly small. However, the actual parameter values are not very important for this analysis. The resulting parameter values used are given in Table 5.2. We generated $N = 100$ data sets for each combination of d, q, n for the analysis. Both

[7] In this example, the eights core is not executing R to ensure the simultaneous fluid use of Windows 10 for other purposes.

i	α_{i0}	α_{i1} $q=1$	α_{i1} $q=2$	α_{i2}	i	α_{i0}	α_{i1} $q=1$	α_{i1} $q=2$	α_{i2}
1	-0.906	0.288	0.559	0.595	31	1.856	0.647	-0.153	0.602
2	-1.337	0.928	-0.163	0.290	32	1.716	0.833	0.176	1.002
3	-0.913	0.581	-0.009	0.259	33	-1.634	0.624	0.364	-0.069
4	0.387	0.249	0.841	1.009	34	-0.224	0.470	-0.131	0.386
5	0.676	0.302	-0.050	0.525	35	1.138	0.724	0.486	0.523
6	-0.084	0.293	0.717	0.054	36	1.641	0.583	0.721	0.271
7	0.138	0.492	0.159	0.326	37	1.169	0.457	0.515	-0.316
8	1.608	0.481	0.969	-0.285	38	0.122	0.848	0.310	-0.077
9	-0.882	0.615	0.245	0.764	39	-0.624	0.789	0.157	0.124
10	1.240	0.182	0.073	0.727	40	0.187	0.921	0.769	-0.052
11	-0.919	1.824	0.480	0.934	41	1.644	0.321	-0.398	-0.172
12	0.216	0.764	-0.372	0.894	42	-0.419	1.101	0.050	0.844
13	1.900	0.338	0.472	0.118	43	1.269	0.521	0.085	0.020
14	0.510	0.869	0.037	0.492	44	-0.848	0.945	0.106	0.061
15	-1.109	0.423	0.226	-0.345	45	-0.473	0.357	0.649	0.973
16	-0.104	0.597	0.142	0.321	46	-1.994	0.336	0.952	-0.399
17	0.458	0.512	-0.212	0.713	47	0.656	0.747	0.588	0.669
18	-1.882	0.925	0.131	0.957	48	-1.358	0.489	0.023	0.348
19	0.801	1.469	-0.398	0.325	49	-0.489	1.100	0.380	0.200
20	1.953	0.529	0.162	-0.384	50	-1.118	1.984	0.001	0.197
21	-1.712	0.512	0.729	0.825	51	0.202	0.793	0.296	0.590
22	1.725	1.133	1.002	-0.069	52	-0.278	0.522	0.842	0.628
23	1.980	0.671	0.610	0.115	53	1.535	1.874	0.551	0.166
24	0.171	0.900	-0.406	0.096	54	-0.140	0.822	-0.491	0.145
25	0.499	0.623	0.049	0.172	55	-0.936	0.273	0.570	-0.250
26	-1.775	0.423	0.228	1.015	56	0.746	0.582	0.921	0.210
27	1.523	0.579	-0.326	-0.398	57	-1.473	0.727	-0.182	0.782
28	1.548	1.474	-0.371	0.056	58	0.058	0.747	0.976	0.137
29	-0.823	0.812	0.560	0.656	59	-0.967	0.650	0.089	-0.405
30	0.822	0.501	0.833	0.087	60	1.095	0.408	-0.415	0.609

Table 5.1: *True parameters for data generation of the pseudo likelihood simulation studies (rounded to three digits).*

algorithms mentioned in the two previous subsections used the same starting value. These starting values were generated by iid values from $\mathcal{U}(0.2, 1.2)$.[8] For the comparison we study the ratio t_1/t_2 where t_1 denotes the time needed for the direct maximization approach and t_2 for the 2SPLE algorithm. The resulting summary statistics of these ratios for all 100 data sets for each combination of d, q, n are given in Table 5.2.

We see that the direct approach is advantageous in the cases of smaller d, n while 2SPLE becomes advantageous for increasing d, n. With regards to estimating the parameters of a multidimensional model for a very large data set, the 2SPLE algorithm is favorable. For further increasing n when $d \leq 20$ is hold fixed, the situation will not change drastically because the response patterns can then be grouped efficiently so that we can expect similar results in computational terms even for $n = 10000$ or $n = 100000$. For one dimensional models, both approaches seem to be comparably efficient for $d = 40$. We therefore implement the overall PLE procedure including both algorithms and we select the direct approach for $d \leq 20$ and for $d \leq 40$ if $q = 1$, otherwise we select 2SPLE.

[8] Recall that the PLE is based on the manifest probabilities model which has another anchor as $\mathcal{H}$. E.g. in an example case for $d = 60$, the PL estimates of $b_1, \ldots, b_d$ stayed between -6.6 and 1 and not between ≈ -2 and ≈ 2.

		$q = 1$			$q = 2$		
d	n	mean	median	sd	mean	median	sd
20	1000	0.569	0.560	0.100	0.505	0.465	0.181
	3000	0.808	0.781	0.150	0.653	0.625	0.149
	5000	0.901	0.867	0.131	0.716	0.682	0.152
40	1000	0.700	0.677	0.149	1.157	1.113	0.251
	3000	1.028	0.966	0.231	1.442	1.373	0.305
	5000	1.226	1.216	0.273	1.602	1.557	0.307
60	1000	0.842	0.822	0.184	1.743	1.682	0.399
	3000	1.363	1.306	0.317	2.118	2.076	0.540
	5000	1.773	1.693	0.413	2.469	2.473	0.634

Table 5.2: *Summary statistics of the ratio t_1/t_2 for 100 data sets in each combination of d, q, n where t_1 denotes the time needed for the direct maximization approach and t_2 for the 2SPLE algorithm.*

Note that the implemented algorithm of Paek (2016, p. 130) took on average 21 minutes for binary data sets with $d = 50$, $q \leq 4$ and $n \leq 1000$ on a 3.6 GHz CPU. For comparison, we needed on average for five estimations for $d = 50$, $q = 4$ and $n = 1000$ on a 3.0GHz CPU 10.3 seconds using the 2SPLE and 12 seconds using the direct approach. In her discussion she came to the conclusion that the PLE is a fast method to estimate the parameters and with the even faster times we observed, we support this notion. Both algorithms and their implementations provided by us are appropriate and well applicable in real settings of large data sets and using multidimensional models. In Section 5.4 we will compare PLE to MMLE for IRT models and also study the computation time required.

5.2.2 Transformation of the Estimates to the Corresponding Logit IRT Model

We aim to estimate $\boldsymbol{A}$ and $\boldsymbol{b}$ of the IRT model

$$\boldsymbol{\lambda}(\boldsymbol{\eta}) = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{b}, \qquad \mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}, \quad \mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q. \tag{5.15}$$

In the last subsection, we received pseudo likelihood estimates $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{A}}$ under the constraints $\mathsf{E}(\boldsymbol{\eta} \mid \boldsymbol{Y}_1 = \mathbf{0}) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y}_1 = \boldsymbol{y}) = \mathrm{I}_q$, $\boldsymbol{y} \in \{0, 1\}^d$. By the laws of total expectation,

$$\mathsf{E}(\boldsymbol{\eta}) = \sum_{\boldsymbol{y} \in \mathbb{Y}} \boldsymbol{A}^\mathsf{T} \boldsymbol{y} \mathsf{P}(\boldsymbol{Y}_1 = \boldsymbol{y}) \tag{5.16}$$

holds under the considered manifest probabilities model which is supposed by the PLE where $\mathbb{Y} = \{0, 1\}^d$. Moreover, the law of total covariance yields for the same manifest probabilities model

$$\mathsf{Cov}(\boldsymbol{\eta}) = \boldsymbol{A}^\mathsf{T} \mathsf{Cov}(\boldsymbol{Y}_1) \boldsymbol{A} + \mathrm{I}_q. \tag{5.17}$$

Clearly, (5.17) implies that $\mathsf{Cov}(\boldsymbol{\eta}) \neq \mathrm{I}_q$ always holds and by equation (5.16) also $\mathsf{E}(\boldsymbol{\eta}) \neq \mathbf{0}$ in general. Hence, we have to transform the obtained estimates in order to use them as estimates for model (5.15). We illustrate this circumstance with the following example.

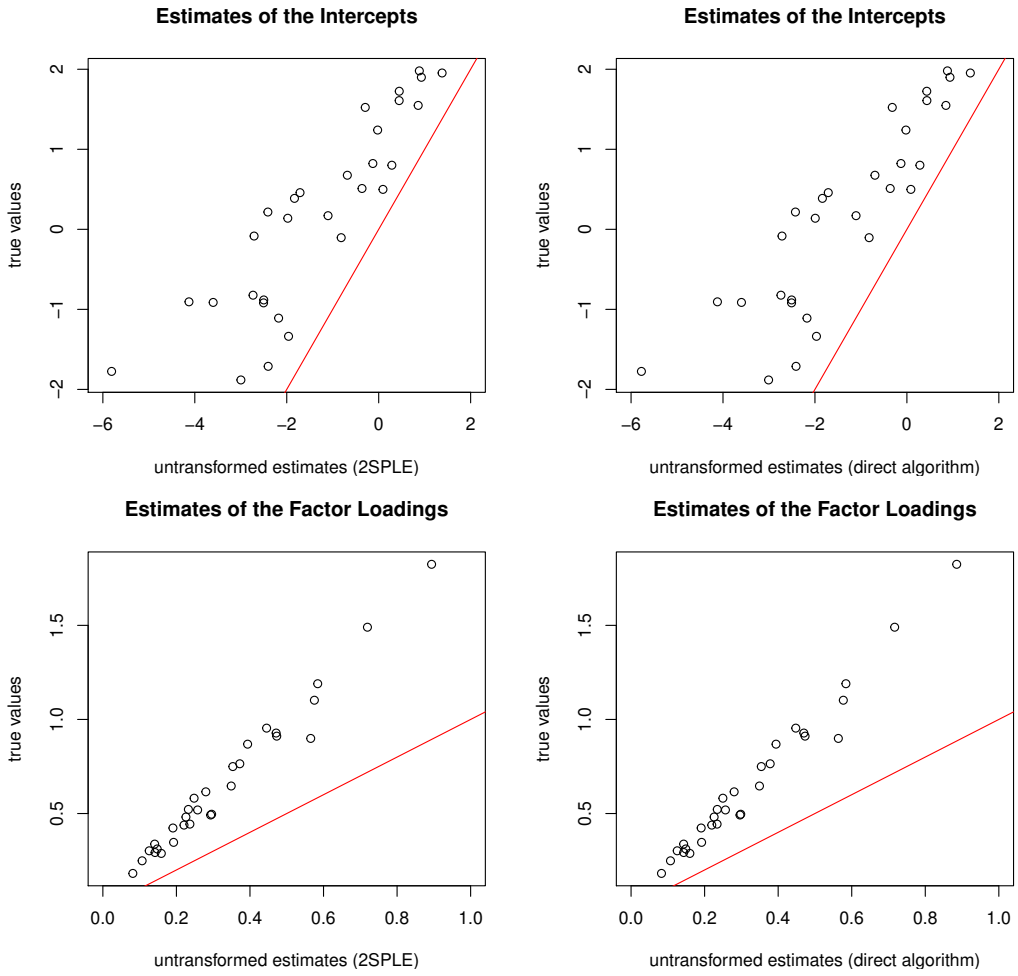

Figure 5.1: *Scatterplot of the estimated model parameters using PLE (left column: 2SPLE, right column: Direct algorithm) for the intercepts (top) and the factor loadings (bottom). The estimates are visualized on the horizontal axis and the true values on the vertical axis. For this example, no transformation of the PLE is applied. Each red lines visualizes the identity and indicates perfect estimates.*

Example 5.2.3

An artificial data set is generated from model (5.15) for $q = 1$, $d = 30$ and $n = 2000$ where $\mathcal{H} = \mathcal{N}(0, 1)$, i.e. the true model is the logit-normal model. The true parameter values are that of Table 5.1 for $i = 1, \ldots, 30$. The parameters are estimated by PLE using the direct algorithm (cf. Subsubsection 5.2.1.3) and using the 2SPLE-algorithm (cf. Algorithm 3). In this case, 2SPLE-algorithm equals the algorithm of Paek and Anderson (2017), i.e. Algorithm 2, with $T = 14$ iterations and without the step of updating $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ as described in Remark 5.1.2 (2.) for $q = 1$. Scatterplots of the resulting parameter estimates are given in Figure 5.1. We see that the obtained estimates by both algorithms are nearly identical. However, the estimates for the factor loadings as well as for the intercepts show a remarkable difference to the true values. This is especially indicated by the difference of all circles to the red lines in the sense that all circles are left to the lines, where a perfect estimate is located directly on the line. ◇

By searching for a transformation, we can replace $\boldsymbol{A}$ with $\hat{\boldsymbol{A}}$. However, neither (5.16) nor (5.17) can be feasibly computed for large d due to the exponential growth of the set of all possible response patterns from which each element has to be evaluated. We therefore suggest the use of approximate transformations to easily obtain the desired estimates of the IRT model. We derive three possible transformations in the following and compare them afterwards.

5.2.2.1 Transformation by Global Standardization (GS)

Assuming $(\boldsymbol{\eta}_k \mid \boldsymbol{Y}_k = \boldsymbol{0}) \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_q)$, $k = 1, \ldots, n$, the posterior expected value of $\boldsymbol{\lambda}(\boldsymbol{\eta}_k)$ is given by

$$\mathsf{E}\big(\boldsymbol{\lambda}(\boldsymbol{\eta}_k) \mid \boldsymbol{Y}_k = \boldsymbol{y}_k\big) = \boldsymbol{A}\mathsf{E}(\boldsymbol{\eta}_k \mid \boldsymbol{Y}_k = \boldsymbol{y}_k) + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{y}_k.$$

Denote by $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{A}}$ the PL estimates obtained in the last section. Considering these estimates as true values, we can estimate $\boldsymbol{\lambda}(\boldsymbol{\eta}_k)$ by

$$\widehat{\boldsymbol{\lambda}(\boldsymbol{\eta}_k)} := \hat{\boldsymbol{b}} + \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_k, \qquad k = 1, \ldots, n.$$

If we apply the aEAP of Section 4.2 to obtain point estimates of the latent variables $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ (i.e. (4.10), (4.11) and (4.12)), we get the estimates $\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n$ of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ with sample mean $\boldsymbol{0}$ and sample covariance matrix I_q. In particular, they have the representation

$$\tilde{\boldsymbol{\eta}}_k = \boldsymbol{S}(\hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_k + \boldsymbol{m}), \qquad k = 1, \ldots, n,$$

for some matrix $\boldsymbol{S}$ and a vector $\boldsymbol{m}$. We can also use this standardization to transform the parameters in a way which does not affect the estimates of $\boldsymbol{\lambda}(\boldsymbol{\eta}_k)$, $k = 1, \ldots, n$, because

$$\hat{\boldsymbol{b}} + \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_k = (\hat{\boldsymbol{b}} - \hat{\boldsymbol{A}}\boldsymbol{m}) + (\hat{\boldsymbol{A}}\boldsymbol{S}^{-1})\Big(\boldsymbol{S}(\hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_k + \boldsymbol{m})\Big).$$

So, overall, a reasonable transformation for $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{A}}$ is given by

$$\hat{\boldsymbol{b}}' := \hat{\boldsymbol{b}} - \hat{\boldsymbol{A}}\boldsymbol{m}, \qquad \hat{\boldsymbol{A}}' := \hat{\boldsymbol{A}}\boldsymbol{S}^{-1}. \tag{5.18}$$

Finally, for matter of interpretation or comparison, $\hat{\boldsymbol{A}}'$ can be rotated in a particular way of interest as usual estimates of the matrix of factor loadings. As mentioned earlier, we use the varimax rotation in our implementation.

5.2.2.2 Transformation by Item-Wise Standardization (IS)

In the pseudo likelihood estimation, rest-score estimates are used instead of the test totals. This can also be considered in the transformation by standardizing the rest-score estimates of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ for item i, $i = 1, \ldots, d$. For any matrix $\boldsymbol{B}$, denote by $\boldsymbol{B}^{(-i)}$ the sub-matrix of $\boldsymbol{B}$ where the ith row is removed and by $\boldsymbol{B}_{(i)}$ its ith row itself, $i = 1, \ldots, d$. The procedure is then given as follows:

For $i = 1, \ldots, d$, do

(i) Determine the vector $\boldsymbol{m}_i$ and the matrix $\boldsymbol{S}_i$ in such a way that

$$\boldsymbol{S}_i\Big(\big(\hat{\boldsymbol{A}}^{(-i)}\big)^\mathsf{T}\boldsymbol{y}_k + \boldsymbol{m}_i\Big), \qquad k = 1, \ldots, n,$$

has sample mean $\boldsymbol{0}$ and sample covariance matrix I_q.

(ii) Set $\hat{\boldsymbol{b}}_i' := \hat{\boldsymbol{b}}_i - \hat{\boldsymbol{A}}_{(i)}\boldsymbol{m}_i$ and $\hat{\boldsymbol{A}}_{(i)}' := \hat{\boldsymbol{A}}_{(i)}\boldsymbol{S}_i^{-1}$.

The matrix $\hat{\boldsymbol{A}}'$ is obtained as the matrix having the rows $\hat{\boldsymbol{A}}_{(i)}'$, $i = 1, \ldots, d$. Again $\hat{\boldsymbol{A}}'$ can be rotated arbitrarily. For $d \to \infty$, the difference between both transformations is negligible.

5.2.2.3 Transformation by Total Standardization (TS)

Let $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$ be symmetric and positive definite and suppose that $\mathsf{Cov}(\boldsymbol{\eta} \mid \boldsymbol{Y}_1 = \boldsymbol{y}) = \boldsymbol{\Sigma}$ for all $\boldsymbol{y} \in \{0,1\}^d$. Then the law of total covariance implies that

$$\mathsf{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Sigma}\mathsf{Cov}(\boldsymbol{X})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}$$

with random vector $\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{Y}_1$. Denote by $\boldsymbol{T}$ the sample covariance of $\hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_1, \ldots, \hat{\boldsymbol{A}}^\mathsf{T}\boldsymbol{y}_n$. Now we can search for a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{q\times q}$ so that

$$\boldsymbol{STS} + \boldsymbol{S} = \mathrm{I}_q,$$

i.e. we estimate the posterior covariance matrix of $\boldsymbol{\eta}$ under $\mathsf{Cov}(\boldsymbol{\eta}) = \mathrm{I}_q$ and the given posterior covariance homogeneity of the manifest probabilities model. The location of $\boldsymbol{\eta}$ can be transformed in the same manner as in the first approach. We then transform analogously to (5.18).

5.2.2.4 Comparison

To get a first impression of the three transformations, we apply them to the obtained PL estimates in Example 5.2.3.

Example 5.2.4 (Example 5.2.3 continued)
An application of the three transformations of Subsection 5.2.2 to the untransformed PL estimates of Example 5.2.3 is performed. The resulting transformed estimates are visualized in Figure 5.2 (intercepts) and in Figure 5.3 (factor loadings). The intercept transformations by GS and TS are identical by definition but still similar to the transformation by IS. All intercept transformations shift the estimates to an close area around the red line, i.e. improve the estimates. The three transformations of the factor loadings are visually more different than the intercept transformations. Additionally, after the transformation of the factor loadings there is still a visual difference to the red line and there is no obviously best transformation among these three transformations, such that a deeper comparison is necessary. The results for 2SPLE (left column of graphics in the figure) and the direct algorithm (right column) are nearly identical. ◇

Two general disadvantages of TS in comparison to GS and IS are:

1. Finding $\boldsymbol{S}$ requires the application of another iterative procedure and
2. the $\boldsymbol{S}$ found can be non-positive definite.

Especially the second disadvantage weights heavily because the justification of this approach is to estimate the posterior covariance of $\boldsymbol{\eta}$ by $\boldsymbol{S}$. To estimate how often this can happen and how well the three approaches perform in comparison, we apply them to 100 simulated data sets for $q = 2$ and $d \in \{20, 40, 60\}$, $n \in \{1000, 3000, 5000\}$. To study the quality of the resulting estimates we compute the AAE (average absolute errors, i.e. $\frac{1}{3d}\sum_{i=1}^{d}\sum_{j=0}^{2}|\hat{\alpha}_{ij} -$

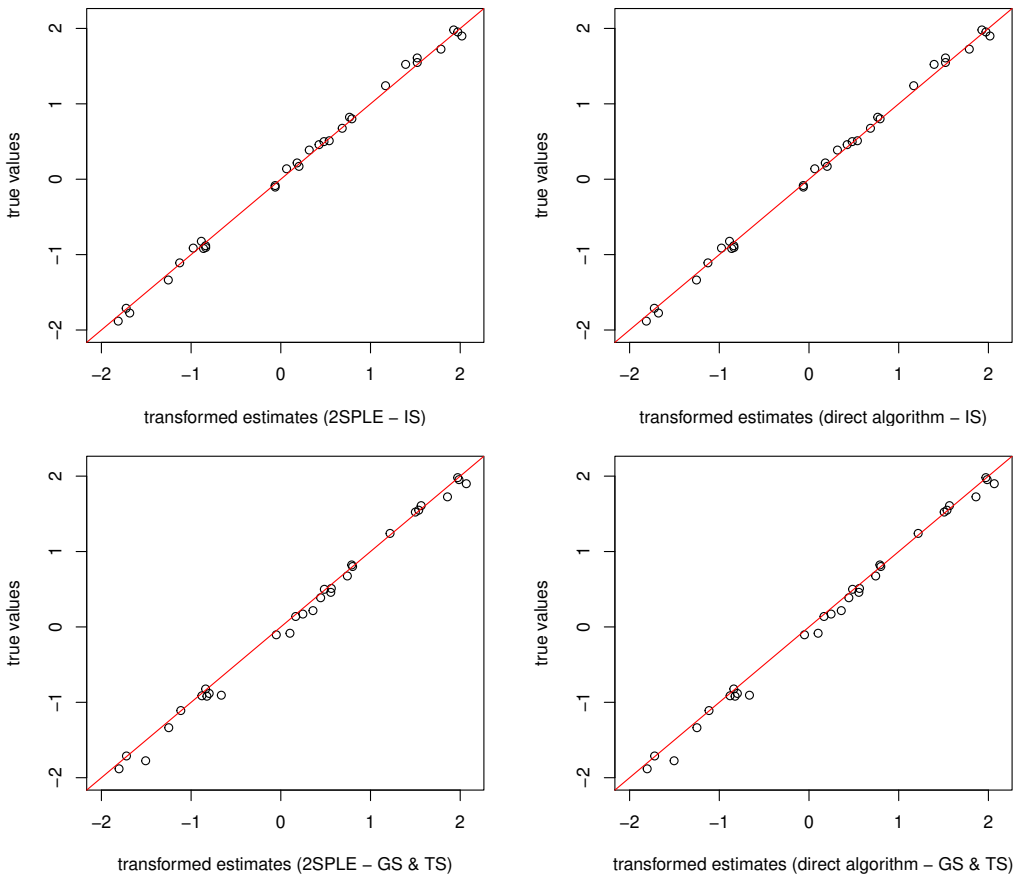

Figure 5.2: *Scatterplot of the estimated intercepts using PLE (left column: 2SPLE, right column: Direct algorithm) for the model parameters. The estimates are visualized on the horizontal axis and the true values on the vertical axis. For this example, the three transformations of Subsection 5.2.2 are applied to the PL estimates of Example 5.2.3 that are visualized in Figure 5.1 without any transformation. Each red lines visualizes the identity and indicates perfect estimates.*

$\alpha_{ij}|$) values averaged over the 100 data sets. Additionally, we count the frequency of the cases in which the AAE value resulting for each transformation is smaller than the AAE value resulting from the other two transformations. Using these measures, we can compare the average precision and if there were a few really good and many less good estimates in comparison to the other two approaches. We also count the frequency of the cases where $\boldsymbol{S}$ found for TS is positive definite to study how critical the possibility of a non-positive definite $\boldsymbol{S}$ is. Shown in Table 5.3 are the resulted values.

First, we can observe that GS is clearly in no case either advantageous in terms of the average AAE or in the proportion of the cases where it was superior to the other approaches. Generally, we see that IS provided the best average precision. Moreover, IS was clearly better than the other two approaches for $d = 40$ and $d = 60$. For $d = 20$, we recognize however that TS was the better transformation in more cases. Since the frequency of the cases where $\boldsymbol{S}$ of TS was positive definite is high for $d = 20$, TS and IS can both be considered viable for these data sets. However, for $d = 40$ and $d = 60$, the found $\boldsymbol{S}$ was

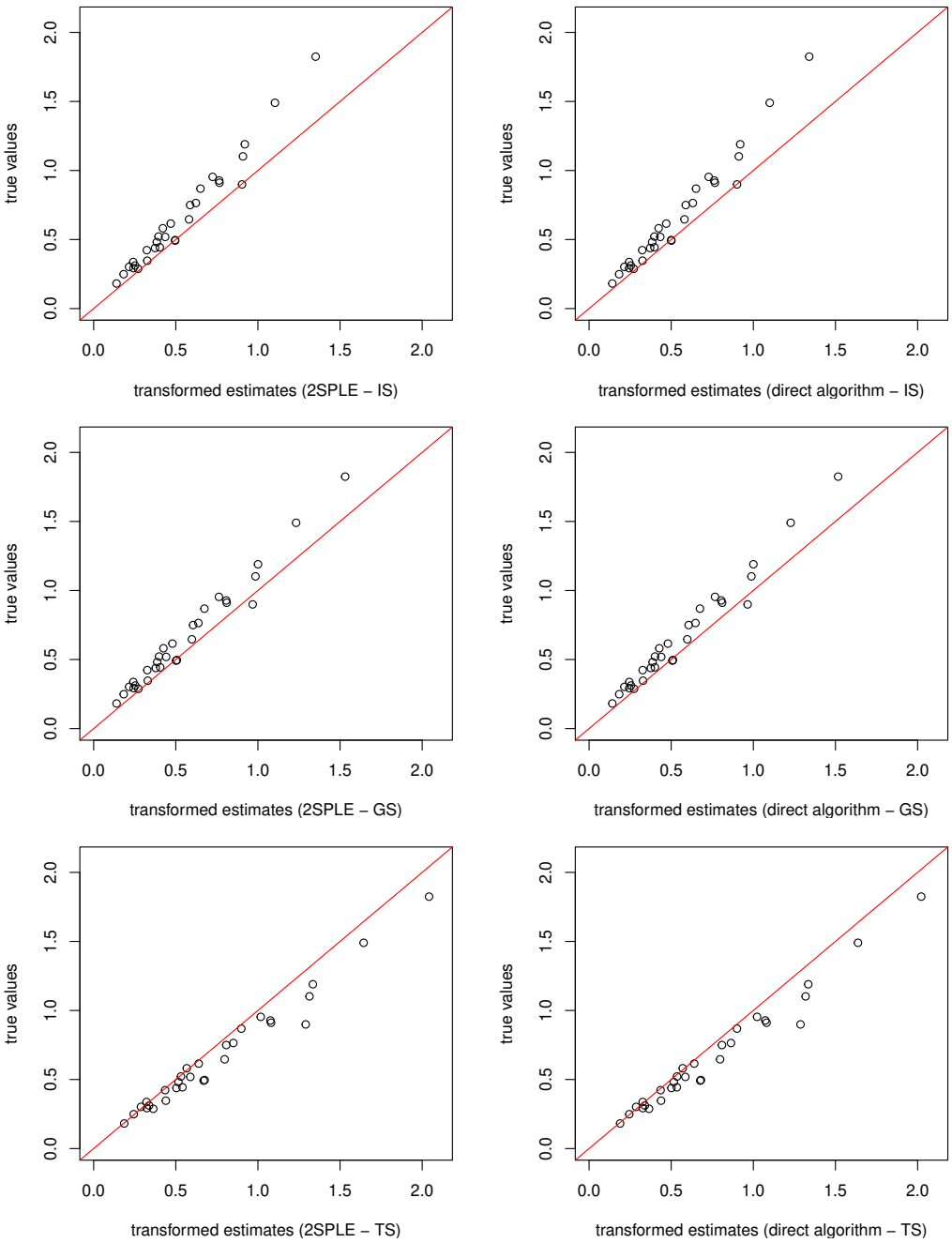

Figure 5.3: *Scatterplot of the estimated factor loadings using PLE (left column: 2SPLE, right column: Direct algorithm) for the model parameters. The estimates are visualized on the horizontal axis and the true values on the vertical axis. For this example, the three transformations of Subsection 5.2.2 are applied to the PL estimates of Example 5.2.3 that are visualized in Figure 5.1 without any transformation. Each red lines visualizes the identity and indicates perfect estimates.*

		GS		IS		TS		
d	n	AAE	FB	AAE	FB	AAE	FB	FPD
20	1000	0.208	1	0.192	56	0.206	43	83
	3000	0.198	0	0.183	28	0.207	72	97
	5000	0.190	0	0.175	20	0.182	80	100
40	1000	0.223	9	0.213	65	0.242	26	15
	3000	0.181	13	0.173	64	0.221	23	1
	5000	0.195	7	0.189	67	0.21	26	1
60	1000	0.195	12	0.189	64	0.237	24	15
	3000	0.188	8	0.18	69	0.209	23	3
	5000	0.193	9	0.187	63	0.221	28	8

Table 5.3: *Results of the simulation study in Subsubsection 5.2.2.4. The average absolute errors (AAE, rounded to three digits) for the parameter estimates for all three transformations, averaged over 100 simulated data sets; the frequency of the cases where a transformation resulted in the smallest AAE of the three transformations (frequency best, FB) and, for TS, the frequency of cases where $\boldsymbol{S}$ was positive definite (frequency positive definite, FPD).*

positive definite only in very few cases. Also the average AAE and FB were comparably bad in these cases. In conclusion, IS has shown to be the overall superior approach of these three approaches. This is especially true given that we aim particularly for an estimation method for high dimensional data sets where MPMLE and MMLE are infeasible. We therefore only consider IS in the following and in our implementation of PLE. However, Figure 5.3 already shows that further research on transformations is important in order to obtain good estimates for the IRT model (5.15). The next section further supports this.

5.3 Analysis of the Bias and the Efficiency of the Estimates

As a parameter estimation method for the manifest probabilities model, we expect that the estimates are consistent and asymptotic normal (however, not asymptotic efficient) for $n \to \infty$ if the manifest probabilities model holds, because it is then a classic composite likelihood approach for a log-multiplicative model (cf. also the discussion for Rasch-type models by Anderson et al., 2007). However, as a estimator for the parameters of a logit IRT model, there are two sources of errors which do not vanish for $n \to \infty$:

1. If $(\boldsymbol{\eta}_1 \mid \boldsymbol{Y}_1 = \boldsymbol{0}) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ does not hold, then the conditional probabilities which are used to formulate the pseudo likelihood function only hold approximately.

2. The standardizations which are used to transform estimates under the manifest probabilities model to the IRT model formulation cannot be expected to be consistent for the first two central moments of $\mathcal{H}$ for fixed d. This is in particular reasonable because the transformation ignores the error of the point estimates of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, which can only be expected to be negligible for $d \to \infty$. We therefore expect a bias in the scaling factor $\boldsymbol{S}^{-1}$ (or $\boldsymbol{S}_i^{-1}$, $i = 1, \ldots, d$, respectively).

Overall, consistency can only be expected to be obtainable for $d, n \to \infty$ simultaneously in appropriate order like JMLE. However, we cannot expect the pseudo likelihood to be

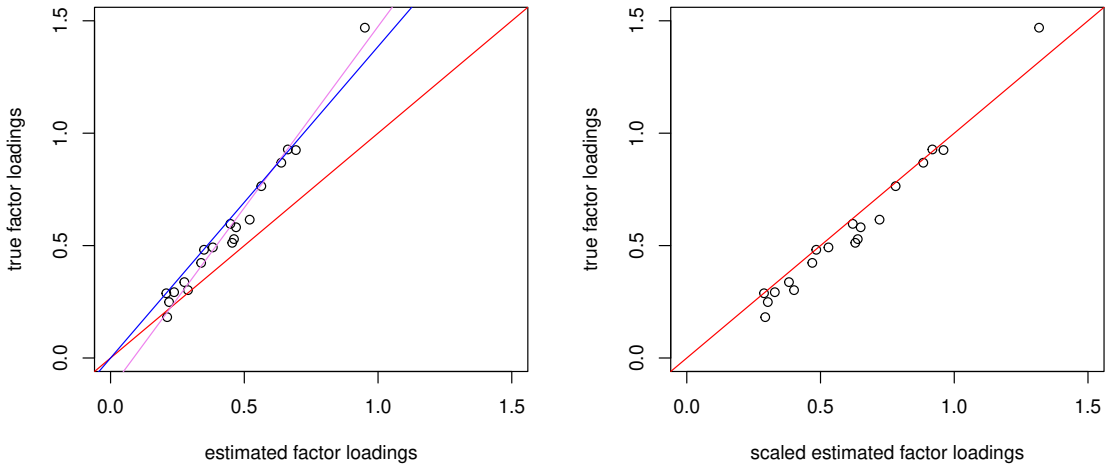

Figure 5.4: *Scatter plot of the estimated factor loadings using PLE for a data set with 20 items and 5000 observed responses by fitting a model with one latent variable. The red line indicates perfect estimates: The closer the circles are to the red line the better estimates are. The blue and the violet line visuale linear regressions with the true parameters as a response and the estimates as explanatory variable (blue without and violet with intercept). The root mean squared error of the estimated factor loadings is* ≈ 0.2229 *and reduces to* ≈ 0.0809 *after scaling using the linear regression without intercept. The root mean squared error of the intercepts reduce from* ≈ 0.06 *to* ≈ 0.0546 *by scaling.*

consistent for $n \to \infty$ when d is held fixed. Due to the asymptotic posterior normality and the increasing precision of the aEAP for $d \to \infty$, we expect the bias to be relatively small for tests with many items. This bias may also depend on other aspects, like the actual distribution of $\boldsymbol{\eta}_1$, the test dimension q, the magnitude of the model parameters or the shape of their "distribution". However, in a practical situation with fixed d and n, pseudo likelihood estimation as a limited information method for IRT models may be a useful approach if classic methods are not feasible.

For a first impression of the quality of pseudo likelihood estimates, we analyze a simulated data set of $d = 20$ items for $q = 1$ latent variable and $n = 5000$ observed response patterns. In the left plot in Figure 5.4, we see a scatter plot of the 20 estimated factor loadings (by item-wise standardization) against the true factor loadings. In a perfect fit, the circles for all $(\alpha_{i1}, \hat{\alpha}_{i1})$ points, $i = 1, \ldots, 20$, are placed on the red diagonal line. For an appropriate estimate, we require that the estimates are spread around the red diagonal with reasonably small variance. In this case, however, that is clearly not the case. Instead, there is a structural error as nearly all dots stay left of the red diagonal. We can draw a plausible line through the circles – we especially observe a high correlation between the acutual parameter values and their estimates analogously to the results of PA. The estimates would improve if we scale them in such a way that a line through the circles is identical to the red diagonal line. We additionally see two linear regressions where the true values are considered as response and the estimates are considered as explanatory variables (using a least squares estimator; blue without intercept, violet with intercept). Since the difference between the

blue and the violet regression line is quite low, we conclude that a major part of the bias in the estimates can be captured by a multiplicative correction of the estimates. In the right plot, we see the scatter plot of the true values and the multiplicatively corrected pseudo likelihood estimates (by the slope ≈ 1.3855 of the linear regression without intercept). This correction cannot be applied in practice due to the unknown true factor loadings and the therefore also unknown value of the slope of this linear regression. A first approach to estimate the correction factor from the data is discussed in Appendix B. In the following we study the factor which we obtain from the linear regression where the true parameters are considered as response and the estimates as explanatory variables. This allows us to discuss the general problem and provides us an estimate of an upper bound of the best possible improvement which a bias correction can have that is based on the estimation of these linear regression coefficients. Using this correction factor, the root mean squared error (RMSE), i.e. $\sqrt{\frac{1}{20}\sum_{i=1}^{20}(\hat{\alpha}_{i1}-\alpha_{i1})^2}$, reduces from ≈ 0.2229 to ≈ 0.0809 which is a remarkable improvement. The situation for the intercepts is similar but the bias is smaller. There, by scaling the estimates with a factor of ≈ 1.0237, we can decrease the RMSE from ≈ 0.06 to ≈ 0.0546. We always observed analogous situations when applying the PLE. On the other side, since the factor is greater than one, it increases the variances of the estimators, so that the overall effect on the mean squared error is of interest. In the following we discuss a simulation study of the bias and its effect on the efficiency of the estimates.

Our aim is to gain knowledge about the evolution of the correction factors for factor loadings and intercepts and the influence of the correction on the quality of the estimates if d grows. By incorporating different numbers of latent variables, distributions of the latent variables and magnitudes of the parameters, we study the influence of these components on the correction factor and the efficiencs of the estimator. Especially, we examine how great the need for a bias correction becomes for long tests and how large the benefit on the overall estimation is. We study estimates of the standard logit IRT model for d items and q latent variables, i.e.

$$\operatorname{logit}\big(P_i(\boldsymbol{\eta})\big) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij}\eta_j, \qquad i = 1, \dots, d.$$

We collect the parameters to an intercept vector $\boldsymbol{\alpha} = (\alpha_{10}, \dots, \alpha_{d0})^\intercal$ and a factor loading matrix $\boldsymbol{A} = (\alpha_{ij})_{i,j=1}^{d,q}$. If $q > 1$, then we use the *varimax* solution of the estimate of $\boldsymbol{A}$. In the simulation, we consider the following three settings:

1. For each d, we take the first d parameter values of Table 5.1 – i.e. the same as for the simulations in Subsection 5.2.1.4 – and set $\mathcal{H} = \mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$.

2. We multiply each parameter value of Table 5.1 by 1.2. The remaining aspects of this setting are identical to the first setting.

3. We take the parameter values as in the first case. For $q = 1$, the latent variables are samples from the distribution of η, where $\eta + 1 \sim \mathrm{Exp}(1)$. If $q = 2$, then the latent vectors are samples from the distribution of $\boldsymbol{\eta} = (\eta_1, \eta_2)^\intercal$, where η_1 and η_2 are independent, $\eta_1 + 1 \sim \mathrm{Exp}(1)$ and η_2 has the following continuous density function:

$$f^{\eta_2}(x) = \frac{1}{2}\phi_1(x; -0.9,\, 0.19) + \frac{1}{2}\phi_1(x; 0.9,\, 0.19), \qquad x \in \mathbb{R},$$

where $\phi_1(\,\bullet\,; \mu, \sigma^2)$ is the density of $\mathcal{N}_1(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

In each case, we generate $N = 100$ data sets for the analysis and for each data set, we generate $n = 5000$ independent responses. As quality measures, we consider the mean absolute bias, the mean variance and the average root mean squared error like in the comparison study in Subsection 1.3.3.6, i.e.

$$\begin{aligned}
\text{maBias} &:= \frac{1}{d(q+1)} \sum_{j=0}^{q} \sum_{i=1}^{d} \left| \frac{1}{N} \sum_{k=1}^{N} \left(\hat{\alpha}_{ij}^{(k)} - \alpha_{ij} \right) \right|, \\
\text{mVar} &:= \frac{1}{d(q+1)} \sum_{j=0}^{q} \sum_{i=1}^{d} \frac{1}{N-1} \sum_{k=1}^{N} \left(\hat{\alpha}_{ij}^{(k)} - \bar{\alpha}_{ij} \right)^2, \\
\text{armse} &:= \frac{1}{d(q+1)} \sum_{j=0}^{q} \sum_{i=1}^{d} \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(\hat{\alpha}_{ij}^{(k)} - \alpha_{ij} \right)^2},
\end{aligned}$$

where N is the number of generated data sets, $\hat{\alpha}_{ij}^{(k)}$ is the (corrected) estimate of α_{ij} and $\bar{\alpha}_{ij}$ the mean estimate of α_{ij} with

$$\bar{\alpha}_{ij} := \frac{1}{N} \sum_{k=1}^{N} \hat{\alpha}_{ij}^{(k)},$$

$i = 1, \ldots, d$, $j = 0, \ldots, q$. The obtained results are given in Table 5.4. In the columns "scale $\hat{\boldsymbol{A}}$" and "scale $\hat{\boldsymbol{\alpha}}$", we give the average correction factor and in brackets the standard deviation. Further, we denote by PLE the pseudo likelihood estimates using item-wise standardization and by cPLE the corrected ones, where the scaling factor is computed for each single data set as the slope of the least squares linear regression of the true value on the PLEs.

We see that the correction of the intercepts is always relatively small, while the correction of the factor loadings is remarkably higher for few items than for 60 items. We also see that the correction leads to a decrease in the armse value, which is much stronger for few than for may items. However, this coincides with the reduced correction for many items. The differences in the results for the different settings and different latent dimensions are small but seem to exist based on the received results. Overall, we see that a bias correction is recommended if the PLE is used for 20-40 items, but we expect that it becomes less important for quite large response data sets.

5.4 Comparison to Marginal Maximum Likelihood Estimation

PLE is a feasible alternative to MMLE for the most established IRT model. The comparison of pseudo likelihood estimation for IRT models is interesting and important because of two primary reasons:

1. It is interesting to compare computational times between PLE and MMLE and quantify the time saved by using PLE to determine whether PLE is an actual alternative to consider in computationally problematic situations.

2. PLE was derived as approximate composite likelihood method and can, therefore, expected to be less efficient than (asymptotically) efficient estimators like MMLE. Additionally, the last section indicated that PL estimates have a bias. It is important

					PLE			cPLE		
setting	q	d	scale $\hat{\boldsymbol{A}}$	scale $\hat{\boldsymbol{\alpha}}$	maBias	mVar	armse	maBias	mVar	armse
1	1	20	1.390 (0.032)	1.030 (0.008)	0.096	0.001	0.109	0.001	0.002	0.062
		25	1.297 (0.024)	1.027 (0.007)	0.084	0.001	0.097	0.001	0.002	0.054
		30	1.249 (0.022)	1.026 (0.006)	0.077	0.001	0.091	0.001	0.002	0.051
		35	1.215 (0.022)	1.023 (0.006)	0.068	0.001	0.083	0.001	0.002	0.049
		40	1.186 (0.019)	1.020 (0.006)	0.061	0.001	0.076	0.001	0.002	0.046
		45	1.168 (0.017)	1.016 (0.005)	0.055	0.001	0.072	0.001	0.002	0.044
		50	1.155 (0.016)	1.014 (0.005)	0.051	0.001	0.069	0.001	0.002	0.046
		55	1.142 (0.017)	1.013 (0.004)	0.047	0.001	0.066	0.001	0.002	0.045
		60	1.132 (0.016)	1.012 (0.005)	0.044	0.001	0.062	0.001	0.001	0.043
	2	20	1.456 (0.047)	1.051 (0.008)	0.161	0.001	0.168	0.001	0.002	0.144
		25	1.231 (0.044)	1.054 (0.007)	0.186	0.001	0.193	0.001	0.002	0.183
		30	1.137 (0.040)	1.044 (0.008)	0.187	0.001	0.196	0.001	0.002	0.190
		35	1.322 (0.026)	1.039 (0.006)	0.120	0.001	0.129	0.001	0.002	0.104
		40	1.310 (0.022)	1.038 (0.006)	0.094	0.001	0.105	0.001	0.002	0.073
		45	1.253 (0.020)	1.033 (0.006)	0.108	0.001	0.118	0.001	0.002	0.100
		50	1.250 (0.019)	1.032 (0.005)	0.075	0.001	0.087	0.001	0.002	0.058
		55	1.207 (0.016)	1.028 (0.005)	0.085	0.001	0.097	0.001	0.002	0.080
		60	1.081 (0.016)	1.026 (0.004)	0.141	0.001	0.151	0.001	0.001	0.147
2	1	20	1.354 (0.025)	1.033 (0.008)	0.108	0.001	0.122	0.001	0.002	0.073
		25	1.263 (0.024)	1.028 (0.007)	0.092	0.002	0.107	0.002	0.002	0.063
		30	1.215 (0.022)	1.026 (0.006)	0.083	0.002	0.098	0.002	0.002	0.059
		35	1.190 (0.017)	1.024 (0.006)	0.076	0.002	0.092	0.002	0.002	0.056
		40	1.162 (0.016)	1.020 (0.006)	0.065	0.002	0.083	0.002	0.002	0.053
		45	1.146 (0.016)	1.016 (0.005)	0.058	0.002	0.077	0.002	0.002	0.050
		50	1.139 (0.015)	1.013 (0.004)	0.055	0.002	0.075	0.002	0.002	0.053
		55	1.128 (0.026)	1.013 (0.009)	0.056	0.005	0.090	0.005	0.006	0.074
		60	1.122 (0.015)	1.011 (0.004)	0.048	0.002	0.069	0.002	0.002	0.049
	2	20	1.075 (0.033)	1.059 (0.008)	0.252	0.008	0.269	0.008	0.009	0.261
		25	1.094 (0.029)	1.062 (0.007)	0.244	0.001	0.251	0.001	0.002	0.242
		30	1.044 (0.027)	1.051 (0.006)	0.238	0.001	0.245	0.001	0.002	0.237
		35	1.300 (0.019)	1.044 (0.005)	0.119	0.002	0.130	0.002	0.002	0.094
		40	1.229 (0.019)	1.042 (0.005)	0.129	0.002	0.140	0.002	0.002	0.115
		45	1.253 (0.018)	1.037 (0.005)	0.086	0.002	0.099	0.002	0.002	0.066
		50	1.160 (0.018)	1.034 (0.005)	0.128	0.002	0.139	0.002	0.002	0.125
		55	1.195 (0.015)	1.031 (0.004)	0.069	0.002	0.085	0.002	0.002	0.055
		60	0.951 (0.013)	1.028 (0.005)	0.216	0.001	0.226	0.001	0.001	0.223
3	1	20	1.476 (0.039)	1.016 (0.008)	0.113	0.001	0.126	0.001	0.002	0.076
		25	1.436 (0.034)	1.023 (0.007)	0.116	0.001	0.130	0.001	0.002	0.081
		30	1.408 (0.035)	1.027 (0.007)	0.116	0.001	0.130	0.001	0.002	0.086
		35	1.381 (0.032)	1.027 (0.006)	0.112	0.001	0.125	0.001	0.002	0.085
		40	1.351 (0.031)	1.024 (0.005)	0.109	0.001	0.121	0.001	0.002	0.080
		45	1.306 (0.026)	1.019 (0.005)	0.096	0.001	0.110	0.001	0.002	0.074
		50	1.281 (0.023)	1.014 (0.005)	0.093	0.001	0.107	0.001	0.002	0.071
		55	1.281 (0.021)	1.017 (0.005)	0.094	0.001	0.108	0.001	0.002	0.075
		60	1.263 (0.022)	1.015 (0.005)	0.088	0.001	0.103	0.001	0.002	0.072
	2	20	1.659 (0.049)	1.053 (0.008)	0.122	0.001	0.132	0.001	0.002	0.080
		25	1.512 (0.041)	1.055 (0.008)	0.139	0.001	0.149	0.001	0.002	0.119
		30	1.238 (0.031)	1.046 (0.006)	0.183	0.001	0.191	0.001	0.002	0.184
		35	1.439 (0.026)	1.042 (0.006)	0.096	0.001	0.108	0.001	0.002	0.066
		40	1.353 (0.025)	1.041 (0.005)	0.114	0.001	0.124	0.001	0.002	0.097
		45	1.354 (0.025)	1.036 (0.005)	0.083	0.001	0.096	0.001	0.002	0.065
		50	0.990 (0.017)	1.034 (0.005)	0.193	0.001	0.202	0.001	0.001	0.199
		55	1.267 (0.016)	1.031 (0.005)	0.077	0.001	0.089	0.001	0.002	0.065
		60	1.242 (0.016)	1.029 (0.005)	0.079	0.001	0.092	0.001	0.002	0.073

Table 5.4: *Analysis of the bias in pseudo likelihood estimates and the bias' reduction (values rounded to three digits).*

for applications of PLE to quantify the loss in precision and efficiency for using PLE instead of MMLE.

Both aspects are difficult to answer and need different, specific treatments. As we discussed in Subsection 1.3.1, there are many different developed algorithms to perform marginal maximum likelihood estimation – from both a conceptual and a computational point of view. The actual parameter estimation depends on several specific aspects, such as starting values or termination criteria, including the values in the decision rule, while the performance of the selected algorithm can depend on the scale of the problem and the distance of the starting value to the optimal value. Generally, various implementations of the same algorithm can be of various computational efficiency. Specifically for the MMLE, there is always the trade-off between the integral approximation quality and computation time. For the PLE on the other side, the optimal point for the initial estimates depends on the identification criteria, which we selected according to computational reasons. In consequence, an arbitrary starting point is of varying degrees of convenience for PLE than for MMLE, e.g. the true parameter values will have a smaller distance to the optimal point of the objective function for MMLE than that for PLE. These aspects should be kept in mind when reading the following results and their discussion. In order to obtain at least some sort of objectivity, we use for MMLE the implemented algorithm of the R-package **ltm** with its default settings as the reference implementation.

For computation time comparisons, both methods got the same randomly selected starting vector. The actual time needed is not of interest but we want to forecast the relative computation time for huge data sets. Thus, we computed the ratio of the time needed for PLE to the time needed for MMLE for various settings. The results are given in Table 5.6. PLE took less time than MMLE in all cases. The ratio decreases for an increasing n for all d and q. The ratios are notably smaller for $q = 2$ than for $q = 1$. For $q = 1$, the ratios are comparably constant for an increasing d. For $q = 2$, the ratios are slightly increasing for an increasing d.

It is plausible that there are cases for $d \to \infty$ where the ratio of the times needed by PLE and MMLE are close to one or even greater than one when n, q are held fixed. Extrapolating from the values of the ratio in the table, such a d can be expected to be very large. However, we are primarily interested in cases where d, n and q are large. Extrapolating from the given values, we expect that PLE will take much less time than MMLE in large, high dimensional settings. We will study a single example for the case of $d = 120$, $q = 8$ and $n = 10000$ later in this section. This example will support this expectation.

For a simulation study comparing the efficiency of PLE and MMLE, we take the same settings for data generation as in the bias analysis simulation study in Section 5.3. To consider the dependence of the method on n, we study $n \in \{1000, 3000, 5000\}$. As we are now less interested in the specific evolution of some quantity for an increasing d, we now focus on the cases $d \in \{40, 60\}$. For each setting, we generated 100 data sets and computed the same measures of efficiency as in Section 5.3. In several cases, the finite-differences-approximation of the Hessian of the marginal log-likelihood function was numerically not at full rank. We counted the proportion of these cases, but still considered all cases valid because the point estimates seemed appropriate in control samples and several refits by ltm under a higher number of quadrature points yielded the same result. That the numerical approximation of the Hessian was numerically non-positive-definite in several cases should be kept in mind because it might have affected the point estimates.

The results of this simulation study are given in Table 5.6. We see that for $q = 1$,

an appropriate scaling of the PLE improves the estimates remarkably. In particular, the estimates become of comparable quality to the MMLE. For $q = 2$, the situation changes partially. The effect of the scaling only becomes noteworthy in the cases with greater d or n – in particular for $d = 60$ and $n = 5000$ in the settings 2 and 3. In these two cases, the scaled estimates come closer to the MMLE, but do not become equal. In conclusion, PLE is much faster than MMLE of ltm at the costs of less efficient estimates.

Finally, we compare the PLE to the MMLE using the aMCEM method from Subsection 4.3.2 on the large data example from Subsection 4.3.3. In two applications of PLE and aMCEM using the same starting values, aMCEM took 324734.09 seconds (i.e. $\approx$ 3 d 18 h 12.2 min) and 292151.1 seconds (i.e. $\approx$ 3 d 9 h 9.2 min). The direct approach for PLE took 3819.78 seconds (i.e. $\approx$ 1h 3.7min) and 2896.49 seconds (i.e. $\approx$ 48.3 min). This means that aMCEM took between 76.5 and 112.1 times as long as the direct approach PLE. The 2SPLE algorithm only took 648.39 and 629.39 seconds (i.e. $\approx$ 11 and 10 min). This means that aMCEM took between 464.2 and 516 times as long as 2SPLE. We therefore recognize that using PLE saves an immence amount of time, similarly to the smaller data examples in Table 5.6. Analogously to Subsection 4.3.3, we use the root mean squared error (RMSE) as a quality measure for a single data set. We observe an RMSE of ≈ 0.1296 for the factor loadings (aMCEM ≈ 0.061) and ≈ 0.0798 for the intercepts (aMCEM ≈ 0.041)[9] – meaning the errors are roughly twice as large. Compare also Figure 5.5 for scatter plots of the estimates of PLE to the true values. For comparison, revisit Figure 4.4 with the values of aMCEM. Hence, similar to the cases before, the saving of time comes at the cost of a remarkable decrease in estimation precision. Generally, we see in Figure 5.5 that the estimates of PLE are strongly correlated with the true values but again have a slightly wrong scale. In this case, we obtain the scaling factor ≈ 1.165 for the loadings and ≈ 1.064 for the intercepts. Using these, the RMSE decreases to ≈ 0.109 for the loadings and ≈ 0.0388 for the intercepts. The errors of the loadings stay in their order of scale and the intercept reduces to the order of aMCEM. As an interesting aspects, we see that the scaling factors are similar to the scaling factors of 40–60 items for $q = 1, 2$.

5.5 Summary and Outlook

We extended the pseudo likelihood method to the fitting of manifest probabilities models under the constraint that the model always can be interpreted as exploratory logit IRT models. We derived two algorithmic approaches and our implementations of these algorithms solve the issues which were pointed out by Paek (2016). Our simulations showed that the further development of the two-stage algorithm is faster than the direct approach for larger data sets, and both are improvements to the implementation by Paek (2016). We also derived three approaches to transform PL estimates for manifest probabilities models to estimates for logit IRT models under the most common identification constraints. A comparison of their application for simulated data sets revealed that the item-wise standardization of the pseudo likelihood estimates is the best approach to transform the estimates so that the usual identification constraints of exploratory IRT models hold.

[9] These values are obtained using the direct approach, our implementation of the 2SPLE algorithm returned estimates whose RMSEs are ≈ 0.0802 for the intercepts and ≈ 0.1338 for the factor loadings – this is slightly larger than those from the direct approach but of a similar size. Die average absolute difference of the estimated factor loadings obtained by both PLE algorithms is ≈ 0.0085 and the average absolute difference of the estimated intercepts is ≈ 0.0028. Both PLE algorithms result in practically equivalent estimates.

				PLE			cPLE			MMLE			prop
setting	q	d	n	maBias	mVar	armse	maBias	mVar	armse	maBias	mVar	armse	instable
1	1	40	1000	0.060	0.007	0.110	0.019	0.008	0.091	0.008	0.009	0.093	0.000
			3000	0.059	0.002	0.082	0.019	0.003	0.055	0.004	0.003	0.053	0.000
			5000	0.061	0.001	0.076	0.020	0.002	0.046	0.004	0.002	0.041	0.000
		60	1000	0.043	0.007	0.100	0.017	0.008	0.089	0.007	0.008	0.090	0.000
			3000	0.044	0.002	0.071	0.017	0.003	0.055	0.004	0.003	0.052	0.000
			5000	0.043	0.001	0.063	0.017	0.002	0.044	0.003	0.002	0.040	0.000
	2	40	1000	0.111	0.034	0.210	0.095	0.037	0.206	0.010	0.012	0.107	0.460
			3000	0.162	0.002	0.176	0.153	0.003	0.170	0.005	0.004	0.060	0.500
			5000	0.180	0.001	0.189	0.172	0.001	0.184	0.004	0.002	0.046	0.500
		60	1000	0.079	0.007	0.123	0.059	0.009	0.116	0.009	0.010	0.098	0.430
			3000	0.155	0.002	0.171	0.151	0.002	0.169	0.006	0.003	0.056	0.480
			5000	0.113	0.001	0.124	0.100	0.001	0.115	0.004	0.002	0.043	0.480
2	1	40	1000	0.063	0.008	0.119	0.023	0.009	0.099	0.009	0.011	0.101	0.000
			3000	0.065	0.003	0.090	0.026	0.003	0.063	0.006	0.004	0.058	0.000
			5000	0.067	0.002	0.084	0.026	0.002	0.053	0.004	0.002	0.045	0.000
		60	1000	0.046	0.009	0.110	0.021	0.009	0.098	0.008	0.011	0.099	0.000
			3000	0.049	0.003	0.079	0.021	0.003	0.060	0.005	0.003	0.056	0.000
			5000	0.047	0.002	0.068	0.022	0.002	0.050	0.004	0.002	0.043	0.000
	2	40	1000	0.129	0.037	0.225	0.111	0.041	0.219	0.010	0.013	0.111	0.570
			3000	0.087	0.003	0.108	0.026	0.004	0.068	0.007	0.004	0.063	0.500
			5000	0.204	0.002	0.213	0.193	0.002	0.205	0.004	0.003	0.049	0.560
		60	1000	0.102	0.009	0.147	0.087	0.010	0.141	0.013	0.011	0.102	0.280
			3000	0.110	0.003	0.129	0.092	0.004	0.118	0.005	0.004	0.058	0.270
			5000	0.064	0.001	0.080	0.024	0.002	0.051	0.004	0.002	0.046	0.510
3	1	40	1000	0.106	0.007	0.147	0.062	0.009	0.118	0.064	0.009	0.120	0.000
			3000	0.104	0.002	0.123	0.059	0.003	0.086	0.061	0.003	0.088	0.000
			5000	0.106	0.001	0.119	0.060	0.002	0.079	0.063	0.002	0.081	0.000
		60	1000	0.085	0.007	0.132	0.053	0.008	0.112	0.051	0.008	0.111	0.000
			3000	0.090	0.002	0.110	0.054	0.003	0.082	0.055	0.003	0.082	0.000
			5000	0.088	0.001	0.103	0.053	0.002	0.073	0.053	0.002	0.072	0.000
	2	40	1000	0.147	0.009	0.184	0.130	0.012	0.178	0.035	0.012	0.117	0.550
			3000	0.114	0.002	0.131	0.085	0.004	0.111	0.030	0.004	0.074	0.440
			5000	0.173	0.023	0.226	0.170	0.023	0.222	0.029	0.002	0.062	0.580
		60	1000	0.102	0.007	0.141	0.087	0.009	0.136	0.027	0.010	0.104	0.470
			3000	0.188	0.002	0.202	0.186	0.002	0.200	0.024	0.003	0.065	0.520
			5000	0.074	0.001	0.088	0.049	0.002	0.068	0.024	0.002	0.054	0.480

Table 5.5: *Analysis of the efficiency of the PLE in comparison to MMLE (values rounded to three digits).*

		$q = 1$			$q = 2$		
d	n	mean	median	sd	mean	median	sd
20	1000	0.637	0.614	0.126	0.069	0.064	0.024
	3000	0.570	0.555	0.113	0.038	0.036	0.009
	5000	0.497	0.479	0.077	0.031	0.030	0.006
40	1000	0.549	0.525	0.133	0.098	0.098	0.021
	3000	0.488	0.466	0.099	0.067	0.067	0.013
	5000	0.472	0.463	0.093	0.049	0.048	0.010
60	1000	0.567	0.554	0.110	0.103	0.101	0.021
	3000	0.534	0.523	0.128	0.059	0.059	0.014
	5000	0.519	0.504	0.113	0.057	0.055	0.013

Table 5.6: *Statistics of the ratio of the computational effort of the PLE compared to the MMLE (values rounded to three digits).*

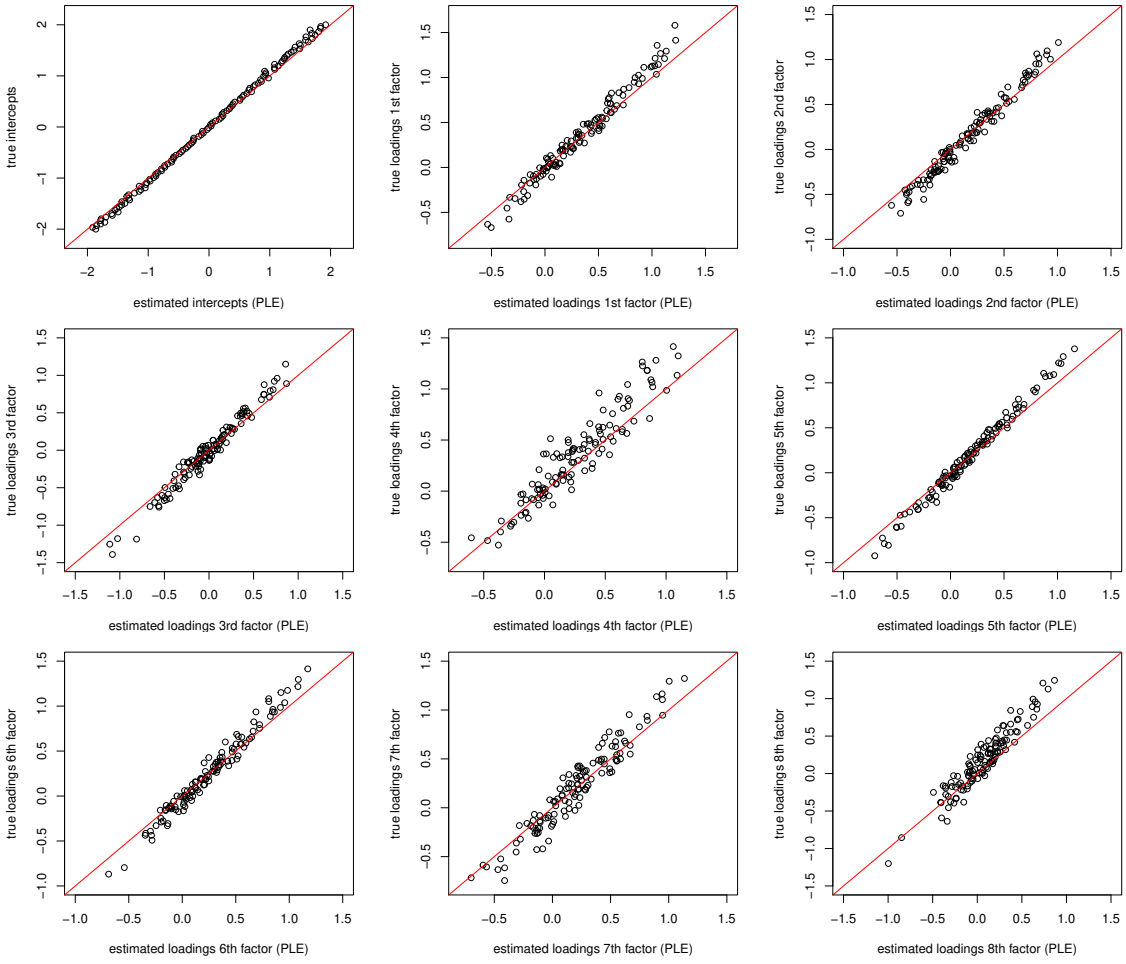

Figure 5.5: *Application of PLE to the same simulated data set to which aMCEM was applied in Subsection 4.3.3 with $d = 120$ items, $q = 8$ factors and $n = 10000$ observed response patterns. The red lines indicate perfect estimates.*

Our further analysis revealed that the resulting PLE for logit IRT models has a bias which depends mostly linearly on the absolute value of the parameters. A simulation study regarding the bias and the best possible effect of a correction of this effect on the efficiency showed that the bias decreases for $d \to \infty$. A comparison to MMLE showed that the default PLE is notably less efficient than MMLE. The comparison also showed that corrected estimates for $q = 1$ can reach the efficiency of MMLE and that the effect for multidimensional models is smaller. On the other hand, our studies showed that both new implementations of PLE are remarkably faster than the method by Paek (2016), the algorithm of the R-package **ltm** to compute MML estimates, and aMCEM from Section 4.3.2. This is especially true for the two-stage algorithm. In an example of the case for $d = 120$, $n = 10000$ and $q = 8$, this algorithm is up to ≈ 516 times faster than the aMCEM algorithm for computing the MMLE.

Since both approaches for PLE can easily be applied for arbitrary d, q, n and the estimates are highly correlated with the true values, PLE is an interesting option for high dimensional settings where MMLE as well as MPMLE are infeasible. Especially, both PLE approaches avoid various hyperparameters like the number of quadrature points so that PLE can be attractive for practitioners without much experience in numerical analysis. A drawback of PLE is the bias of the estimates and the higher variance compared to MMLE while the latter was expected because PLE is based on composite likelihood approaches. Future research could focus on approaches which lead to consistent PL estimators for MIRT models – or at least a substantial reduction of the bias to negligible scale – for making reliable inference. Possible approaches for the bias reduction of PL estimates are given in (i) other transformations of PL estimates for manifest probabilities models to IRT models and (ii) bias correction techniques. We sketched out a possible solution of the problem of the bias of PLE in Appendix B.

Combining several methods with their different strong and weak points is a potential new way to obtain good estimates while still quickening the overall procedure. For example, one could use a roughly bias reduced PLE to obtain good starting values for aMCEM so that its computation time is reduced. A combination of several approaches such as a later switch of aMCEM to a MMLE algorithm using a fast convex optimization method for high-dimensional settings could be computationally attractive.

Conclusions and Future Research

After an review of modeling and estimation approaches in item response theory, we first provided an extension of results regarding estimation for the Rasch model, namely a proof of the asymptotic efficiency of the marginal maximum likelihood estimator, and a simulation study which illustrated the similarity of various estimators for the Rasch model.

We also introduced an extension of the generalized linear latent variable model class with respect to binary response variables and continuous latent vectors by including parametric link functions functions and structural models (EGLLVM). For this model class, and consequently also for the included models with fixed link functions or structural model, we proved the existence, consistency, asymptotic normality and efficiency of the marginal maximum likelihood estimator. This result is also new for the nested classic models. We further proved the asymptotic chi-square distribution of the Wald, score and likelihood-ratio tests for this model class. By comparing two models with parametric and fixed link functions (single, several or all items) or structural models, tests for these model components are obtained. A simulation study showed that IRT models with skewness parameters for single items require thousands of responses for reliable estimates. Another simulation study revealed that models with 1–2 additional parameters for a common skewness of all items and/or the structural model provide a good trade-off between model flexibility and parsimony to obtain good ICC and latent trait estimates. These models only have slighly worse results than a probit model if the probit model holds such that the loss in estimation quality which is caused by the increased number of parameters is not critical. A third simulation study showed that the AIC is suitable for the selection of the link function and structural model. This study moreover shows that LR tests are reliable for nested model comparison of EGLLVMs and, especially, for testing the link function against a parametric alternative.

The assumption of (asymptotic) posterior normality of latent vectors is important for several modeling approaches and derived fitting algorithms but had up to this only been assumed in the multidimensional IRT literature. In this thesis, we contributed two theories which provide variously strong results regarding the asymptotic posterior normality of latent vectors and their point estimators under variously strong conditions. We provide a comprehensive discussion of this topic in IRT and extended the current results regarding the asymptotics of the posterior of a single latent variable in the IRT literature with respect to the multivariate case but also the type of the convergence, the efficiency of the estimators and further extensions of the theory for a single latent variable.

One primary usage of the asymptotic posterior normality is the derivation of algorithms for marginal maximum likelihood estimation for many items and a high latent dimension. By applying the new theory, we derive a simple approximation of the posterior expected value of a set of latent vectors. Further applying the APN theory and this approximation of the posterior expectation, we derive a new EM-algorithm for MMLE of high dimensional logit models which is easy to implement for any latent dimension by simplifying steps of efficient adaptive algorithms for high dimensional settings. By using an adaptive Monte Carlo integration, a smaller number of quadrature points can be used than for the state-

of-the-art adaptive Gauss–Hermite quadrature EM-algorithms are required. Especially, the new algorithm is interesting for IRT practitioners with little programming experience when fitting high dimensional logit models without access to implemented procedures for their setting and software system.

Finally, we further developed the pseudo likelihood estimation method for logit IRT models so that it can be ensured that the fitted model can be interpreted as an IRT model which is not the case in other approaches. We discussed two algorithms to compute the estimates that both solve computational issues of a current algorithm for non-type-Rasch models. Especially, one of them requires much less computation time than our reference MMLE implementations. This procedure therefore is interesting for fitting high dimensional models to large data sets. We derived three transformations of the estimates to obtain estimates for logit IRT models in the common parameterization. One of them (which is based on an item-wise standardization of the resulting rest-score estimates of latent vectors) proved to be clearly superior to the other two approaches. However, the resulting estimator for the IRT model had a notable bias, which decreased for an increasing number of items. For a fixed number of items, the estimator seems to be inconsistent because the bias did not approach zero for an increasing sample size.

Future research regarding EGLLVM could analyze the performance of the tests for a smaller sample size to increase the knowledge regarding the general applicability of the tests for link functions and structural models regardless of the sample size and in generally different settings. We designed the simulation studies regarding EGLLVM for a single latent variable to focus on the influence of the link function and the structural model. Future research could also analyze the corresponding MIRT settings in simulation studies.

One important application of actual posterior normality is the derivation of integral free representations of logit models using the Dutch Identity. Another important application of APN theories therefore is the asymptotic justification of this approach. An interesting future research objective is a rigorous clarification of the related Dutch Identity Conjecture – which contains this asymptotic justification – using the APN theories.

Our theoretical discussions and the simulation examples regarding the new EM-algorithm for logit models illustrated its performance and, especially, its easy applicability. An interesting future research objective would however be a comprehensive analysis regarding its comparison to other approaches in high dimensional settings.

Finally, the pseudo likelihood approach has proved to be remarkably fast, but its bias leads to limited applicability unless the number of items is very large. We discussed a heuristic simulation based algorithm to reduce this bias in Appendix B, but this approach increased the required computation time while there is no guaranteed convergence. An important question for future research therefore is if there is a possible transformation of the pseudo likelihood estimates to the IRT model which leads to none or at least a practically negligible bias without drastically increasing the computation time.

Appendix A

Proof of the Asymptotics of the Marginal Maximum Likelihood Estimator

In the following we discuss the proofs of Theorems 1.3.5, 2.3.2, Theorem 2.4.1 and Remark 1.3.6, which we omitted in Chapters 1 and 2.

A.1 The Constrained Rasch Model

First, we prove Theorem 1.3.5. Before we start, notice that we observe n iid random experiments with finitely many possible outcomes. Therefore we can treat each response vector $\boldsymbol{Y}_k$ as real valued random variable. Also note that the distribution is absolutely continuous with respect to the counting measure. We denote the counting measure by μ and the Radon–Nikodym density by $P(\,\cdot\,;\boldsymbol{\beta})$ with

$$P(\boldsymbol{y};\boldsymbol{\beta}) \coloneqq \mathsf{P}(\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\beta}) = \int_{\mathbb{R}} \prod_{i=1}^{d} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\,\mathcal{H}(\mathrm{d}\eta), \qquad \boldsymbol{y}\in\mathbb{Y}=\{0,1\}^d,\ \boldsymbol{\beta}\in\mathbb{R}^d,$$

where $\boldsymbol{y}=(y_1,\dots,y_d)^\intercal$. It is therefore sufficient to show the usual regularity conditions (cf. e.g. Lehmann and Casella, 1998, Chapter 6, Theorem 5.1).

Obviously we have a parametric distribution family with a common support and the distributions for $\boldsymbol{\beta}_1\neq\boldsymbol{\beta}_2$ always are different. Since the parameter space is $\mathbb{R}^d$, there is an open subset $\omega\subset\mathbb{R}^d$ in which the true parameter vector $\boldsymbol{\beta}$ is an interior point.

First, for $k=1,\dots,d$, we want to prove the equation

$$\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k} = \int_{\mathbb{R}} \frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \prod_{i\neq k} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\,\mathcal{H}(\mathrm{d}\eta). \tag{A.1}$$

As a first step, we take a look on the conditional probabilities for a single response given $\eta\in\mathbb{R}$

$$\mathsf{P}(Y_k=y_k\mid\eta;\boldsymbol{\beta}) = \mathsf{P}(Y_k=y_k\mid\eta;\beta_k) = \frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}, \qquad y_k\in\{0,1\},\beta_k\in\mathbb{R}.$$

With some calculus, it follows

$$\begin{aligned}\frac{\partial \mathsf{P}(Y_k = y_k \mid \eta; \beta_k)}{\partial \beta_k} =& \frac{-y_k \exp(y_k(\eta-\beta_k))(1+\exp(\eta-\beta_k))}{(1+\exp(\eta-\beta_k))^2} \\ & - \frac{\exp(y_k(\eta-\beta_k))(-\exp(\eta-\beta_k))}{(1+\exp(\eta-\beta_k))^2} \\ =& \frac{-y_k \exp(y_k(\eta-\beta_k)) - y_k \exp(y_k(\eta-\beta_k)+\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \\ & + \frac{\exp(y_k(\eta-\beta_k)+\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2}. \end{aligned} \tag{A.2}$$

For $y_k = 0$, we get

$$\frac{\partial \mathsf{P}(Y_k = 0 \mid \eta; \beta_k)}{\partial \beta_k} = \frac{\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \in (0,1)$$

and for $y_k = 1$, the derivative equals

$$\begin{aligned}\frac{\partial \mathsf{P}(Y_k = 1 \mid \eta; \beta_k)}{\partial \beta_k} =& \frac{-\exp(\eta-\beta_k) - \exp(2(\eta-\beta_k)) + \exp(2(\eta-\beta_k))}{(1+\exp(\eta-\beta_k))^2} \\ =& -\frac{\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \in (-1,0).\end{aligned}$$

Therefore, we can summarize (A.2) to

$$\frac{\partial P(Y_k = y_k|\eta; \beta_k)}{\partial \beta_k} = \frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \in (-1,1)\backslash\{0\},$$

since $y_k \in \{0,1\}$. Analogously, we get for the second and third order derivative

$$\begin{aligned}\frac{\partial^2 \mathsf{P}(Y_k = y_k \mid \eta; \beta_k)}{\partial \beta_k^2} &= \frac{(-1)^{y_k}(\exp(\eta-\beta_k)-1)\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^3} \in (-1,1)\backslash\{0\}, \\ \frac{\partial^3 \mathsf{P}(Y_k = y_k \mid \eta; \beta_k)}{\partial \beta_k^3} &= \frac{(-1)^{y_k}(\exp(3(\eta-\beta_k)) - 4\exp(2(\eta-\beta_k)) + \exp(\eta-\beta_k))}{(1+\exp(\eta-\beta_k))^4} \\ &\in (-1,1)\backslash\{0\}.\end{aligned}$$

Since

$$\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \eta; \boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)} = \left(\prod_{i\neq k} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right) \frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}$$

holds for any $\eta \in \mathbb{R}$, we get immediately

$$\frac{\partial \mathsf{P}(\boldsymbol{Y} = \boldsymbol{y}|\eta; \boldsymbol{\beta})}{\partial \beta_k} = \left(\prod_{i\neq k} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right) \frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2}. \tag{A.3}$$

So, we have to show that

$$\begin{aligned}\frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \int_{\mathbb{R}} \prod_{i=1}^{d} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)} \, \mathcal{H}(\mathrm{d}\eta) \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial \beta_k} \prod_{i=1}^{d} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)} \, \mathcal{H}(\mathrm{d}\eta).\end{aligned} \tag{A.4}$$

From the definition of the partial derivatives, we get directly

$$\begin{aligned}
&\frac{\partial}{\partial\beta_k}\int_{\mathbb{R}}\prod_{i=1}^{d}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\,\mathrm{d}\mathcal{H}(\eta)\\
&\quad=\lim_{\Delta\to 0}\int_{\mathbb{R}}\left(\prod_{i\neq k}\frac{\exp(x_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\\
&\qquad\qquad\times\frac{1}{\Delta}\left(\frac{\exp(y_k(\eta-(\beta_k+\Delta)))}{1+\exp(\eta-(\beta_k+\Delta))}-\frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}\right)\mathcal{H}(\mathrm{d}\eta).
\end{aligned}$$

From the mean value theorem, we get for one β_k^* between β_k and $\beta_k+\Delta$

$$\begin{aligned}
\frac{\exp(y_k(\eta-(\beta_k+\Delta)))}{1+\exp(\eta-(\beta_k+\Delta))}-\frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}&=\Delta\cdot\frac{\partial}{\partial\beta_k}\frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}\bigg|_{\beta_k=\beta_k^*}\\
&=\Delta\cdot\frac{(-1)^{2-y_k}\exp(\eta-\beta_k^*)}{(1+\exp(\eta-\beta_k^*))^2}.
\end{aligned}$$

Now, let K be an open interval around β_k, then we get from the last equation

$$\begin{aligned}
&\left|\frac{1}{\Delta}\left(\prod_{i\neq k}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\left(\frac{\exp(y_k(\eta-(\beta_k+\Delta)))}{1+\exp(\eta-(\beta_k+\Delta))}-\frac{\exp(y_k(\eta-\beta_k))}{1+\exp(\eta-\beta_k)}\right)\right|\\
&\qquad\leq\sup_{\beta_k\in K}\left|\left(\prod_{i\neq k}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2}\right|.
\end{aligned}$$

Each factor in the first product on the right-hand side takes values in $(0,1)$ and the further factor takes values in $(-1,1)\backslash\{0\}$ for each $\eta\in\mathbb{R}$, $\boldsymbol{\beta}\in\mathbb{R}^d$ and $\boldsymbol{y}\in\mathbb{Y}$. Hence, the right-hand side is lower than one. Since $\int_{\mathbb{R}}1\,\mathrm{d}\mathcal{H}=1$, we conclude (A.4) by an application of Lebesgue's theorem of dominated convergence. Finally, (A.3) and (A.4) imply (A.1). Additionally, we conclude that $\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k}$ exists for each $\boldsymbol{\beta}\in\mathbb{R}^d$ and $\boldsymbol{y}\in\mathbb{Y}$ and takes values between -1 and 1.

Similarly, we obtain for all $k,p,q=1,\ldots,d$ with $k\neq p$, $p\neq q$, $k\neq q$ and $\boldsymbol{y}\in\mathbb{Y}$:

$$\begin{aligned}
\frac{\partial^2 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k\partial\beta_p}&=\int_{\mathbb{R}}\left(\prod_{i\neq k,p}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\frac{\partial\mathsf{P}(Y_k=y_k\mid\eta;\beta_k)}{\partial\beta_k}\\
&\qquad\times\frac{\partial\mathsf{P}(Y_p=y_p\mid\eta;\beta_p)}{\partial\beta_p}\,\mathcal{H}(\mathrm{d}\eta)\in(-1,1),\\
\frac{\partial^2 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k^2}&=\int_{\mathbb{R}}\left(\prod_{i\neq k}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\frac{\partial^2\mathsf{P}(Y_k=y_k\mid\eta;\beta_k)}{\partial\beta_k^2}\,\mathcal{H}(\mathrm{d}\eta)\in(-1,1),\\
\frac{\partial^3 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k\partial\beta_p\partial\beta_q}&=\int_{\mathbb{R}}\left(\prod_{i\neq k,p,q}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\frac{\partial\mathsf{P}(Y_k=y_k\mid\eta;\beta_k)}{\partial\beta_k}\\
&\qquad\times\frac{\partial\mathsf{P}(Y_p=y_p\mid\eta;\beta_p)}{\partial\beta_p}\frac{\partial\mathsf{P}(Y_q=y_q\mid\eta;\beta_q)}{\partial\beta_q}\,\mathcal{H}(\mathrm{d}\eta)\in(-1,1),\\
\frac{\partial^3 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial\beta_k^2\partial\beta_p}&=\int_{\mathbb{R}}\left(\prod_{i\neq k,p}\frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\right)\frac{\partial^2\mathsf{P}(Y_k=y_k\mid\eta;\beta_k)}{\partial\beta_k^2}
\end{aligned}$$

$$\times \frac{\partial \mathsf{P}(Y_p = y_p \mid \eta; \beta_p)}{\partial \beta_p} \, \mathcal{H}(\mathrm{d}\eta) \in (-1, 1),$$

$$\frac{\partial^3 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k^3} = \int_{\mathbb{R}} \left(\prod_{i \neq k} \frac{\exp(y_i(\eta - \beta_i))}{1 + \exp(\eta - \beta_i)} \right) \frac{\partial^3 \mathsf{P}(Y_k = y_k | \eta; \beta_k)}{\partial \beta_k^3} \, \mathcal{H}(\mathrm{d}\eta) \in (-1, 1).$$

Therefore all first, second and third derivatives exist and are bounded for all $\boldsymbol{y} \in \mathbb{Y}$ and all $\boldsymbol{\beta} \in \mathbb{R}^d$.

Denoting by $\mathsf{E}_{\boldsymbol{\beta}}(\,\cdot\,)$ the expectation with respect to the distribution of $\boldsymbol{Y}$ with parameter value $\boldsymbol{\beta}$. Then the equations

$$\mathsf{E}_{\boldsymbol{\beta}} \left(\frac{\partial \log P(\boldsymbol{Y}; \boldsymbol{\beta})}{\partial \beta_k} \right) = 0, \quad k = 1, \ldots, d,$$

hold, because

$$\mathsf{E}_{\boldsymbol{\beta}} \left(\frac{\partial \log P(\boldsymbol{Y}; \boldsymbol{\beta})}{\partial \beta_k} \right) = \sum_{\boldsymbol{y} \in \mathbb{Y}} \frac{\frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k}}{P(\boldsymbol{y}; \boldsymbol{\beta})} P(\boldsymbol{y}; \boldsymbol{\beta}) = \sum_{\boldsymbol{y} \in \mathbb{Y}} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k}$$
$$= \frac{\partial}{\partial \beta_k} \sum_{\boldsymbol{y} \in \mathbb{Y}} P(\boldsymbol{y}; \boldsymbol{\beta}) = \frac{\partial 1}{\partial \beta_k} = 0, \quad k = 1, \ldots, d.$$

Therefore, we have shown that the conditions are satisfied, which ensure the existence of a consistent estimation sequence with probability tending to one.

Next, we prove the asymptotic normality. We have to show, that for each j, k, q there exist a function M_{jkq} such that

$$\left| \frac{\partial^3 \log P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_q} \right| \leq M_{jkq}(\boldsymbol{y}), \quad \text{for all } \boldsymbol{\beta} \in \omega, \, \boldsymbol{y} \in \mathbb{Y},$$

where $\mathsf{E}_{\boldsymbol{\beta}}(M_{jkq}(\boldsymbol{Y})) < \infty$. On the left-hand-side we take the absolute value of

$$\frac{\partial^3 \log P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_q} = \frac{\frac{\partial^3 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_q}}{P(\boldsymbol{y}; \boldsymbol{\beta})} - \frac{\frac{\partial^2 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k \partial \beta_q} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j}}{P(\boldsymbol{y}; \boldsymbol{\beta})^2} - \frac{\frac{\partial^2 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_q} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k}}{P(\boldsymbol{y}; \boldsymbol{\beta})^2} - \frac{\frac{\partial^2 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_q}}{P(\boldsymbol{y}; \boldsymbol{\beta})^2}$$
$$+ 2 \frac{\frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k} \frac{\partial P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_q}}{P(\boldsymbol{y}; \boldsymbol{\beta})^3}.$$

Each numerator takes values between -1 and 1 and each denominator takes values greater than zero. Therefore a upper bound for the absolute value is given by

$$\left| \frac{\partial^3 \log P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_q} \right| \leq \sup_{\boldsymbol{\beta} \in \omega} \frac{6}{\min\limits_{\boldsymbol{y} \in \mathbb{Y}} P(\boldsymbol{y}; \boldsymbol{\beta})} =: m =: M_{jkq}(\boldsymbol{y}).$$

Since M_{jkq} is constant, it holds $\mathsf{E}_{\boldsymbol{\beta}}(M_{jkq}(\boldsymbol{Y})) = m < \infty$ for each j, k, q.

The additional the regularity condition,

$$\int \frac{\partial^2 P(\boldsymbol{y}; \boldsymbol{\beta})}{\partial \beta_k^2} \, \mu(\mathrm{d}\boldsymbol{y}) = \frac{\partial^2}{\partial \beta_k^2} \int P(\boldsymbol{y}; \boldsymbol{\beta}) \, \mu(\mathrm{d}\boldsymbol{y}) = 0, \quad k = 1, \ldots, d,$$

holds, since we get analogously to the first derivative

$$\begin{aligned}\int \frac{\partial^2 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_k^2}\,\mu(\mathrm{d}\boldsymbol{y}) &= \sum_{\boldsymbol{y}\in\mathbb{Y}} \frac{\partial^2 P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_k^2} \\ &= \frac{\partial^2}{\partial \beta_k^2} \sum_{\boldsymbol{y}\in\mathbb{Y}} P(\boldsymbol{y};\boldsymbol{\beta}) = \frac{\partial^2}{\partial \beta_k^2} \int P(\boldsymbol{y};\boldsymbol{\beta})\,\mu(\mathrm{d}\boldsymbol{y}) = 0,\end{aligned}$$

for $k = 1, \ldots, d$. Therefore, we get for the component (k, ℓ) of the Fisher-Information $\mathcal{I} = \mathcal{I}(\boldsymbol{\beta})$ at point $\boldsymbol{\beta} \in \mathbb{R}^d$

$$\begin{aligned}\mathcal{I}_{k\ell} =& \mathsf{E}_{\boldsymbol{\beta}}\left(-\frac{\partial^2 \log P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial \beta_k \partial \beta_\ell}\right) = \mathsf{E}_{\boldsymbol{\beta}}\left(\frac{\partial \log P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial \beta_k}\frac{\partial \log P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial \beta_\ell}\right) \\ =& \mathsf{E}_{\boldsymbol{\beta}}\left(\frac{\frac{\partial P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial \beta_k}\frac{\partial P(\boldsymbol{Y};\boldsymbol{\beta})}{\partial \beta_\ell}}{P(\boldsymbol{Y};\boldsymbol{\beta})^2}\right) = \sum_{\boldsymbol{y}\in\mathbb{Y}} \frac{\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_k}\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_\ell}}{P(\boldsymbol{y};\boldsymbol{\beta})^2} P(\boldsymbol{y};\boldsymbol{\beta}) \\ =& \sum_{\boldsymbol{y}\in\mathbb{Y}} \frac{\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_k}\frac{\partial P(\boldsymbol{y};\boldsymbol{\beta})}{\partial \beta_\ell}}{P(\boldsymbol{y};\boldsymbol{\beta})}.\end{aligned}$$

Recall, for each $\boldsymbol{y} \in \mathbb{Y}$ and $\boldsymbol{\beta} \in \omega$, it holds $0 < P(\boldsymbol{y};\boldsymbol{\beta})$. Furthermore, both factors in the numerator are finite and so we get that the Fisher-Information matrix is finite. Since we assume in Theorem 1.3.5 that the Fisher information is regular, we get the asymptotic normality and, thus, the asymptotic efficiency of the MMLE for the Rasch model with a fixed but arbitrary $\mathcal{H}$.

Last, we derive the likelihood equations in the formulation of equation (1.45). The derivative of the log-likelihood with respect to β_k can be reformulated by

$$\begin{aligned}\frac{\partial}{\partial \beta_k} \log\left(\mathcal{L}_{\text{marginal}}(\boldsymbol{\beta})\right) &= \sum_{k=1}^n \frac{\partial}{\partial \beta_k} \log(P(\boldsymbol{y}_k;\boldsymbol{\beta})) = \sum_{k=1}^n \frac{1}{P(\boldsymbol{y}_k;\boldsymbol{\beta})} \frac{\partial P(\boldsymbol{y}_k;\boldsymbol{\beta})}{\partial \beta_k} \\ &= \sum_{k=1}^n \frac{1}{P(\boldsymbol{y}_k;\boldsymbol{\beta})} \int_{\mathbb{R}} \frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \prod_{i\neq k} \frac{\exp(y_i(\eta-\beta_i))}{1+\exp(\eta-\beta_i)}\,\mathcal{H}(\mathrm{d}\eta) \\ &= \sum_{k=1}^n \int_{\mathbb{R}} \frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \frac{1}{\mathsf{P}(Y_k = y_k \mid \eta;\boldsymbol{\beta})} \\ &\qquad \times \frac{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \eta;\boldsymbol{\beta})}{P(\boldsymbol{y}_k;\boldsymbol{\beta})}\,\mathcal{H}(\mathrm{d}\eta).\end{aligned}$$

By Bayes' theorem, it holds

$$\frac{\mathsf{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \eta;\boldsymbol{\beta})\,\mathcal{H}(\mathrm{d}\eta)}{P(\boldsymbol{y}_k;\boldsymbol{\beta})} = \mathcal{H}(\mathrm{d}\eta \mid \boldsymbol{Y} = \boldsymbol{y}_k).$$

Additionally, we get

$$\begin{aligned}\frac{(-1)^{2-y_k}\exp(\eta-\beta_k)}{(1+\exp(\eta-\beta_k))^2} \frac{1}{\mathsf{P}(Y_k = y_k \mid \eta;\boldsymbol{\beta})} &= \frac{\partial}{\partial \beta_k} \log\left(\mathsf{P}(Y_k = y_k \mid \eta;\boldsymbol{\beta})\right) \\ &= y_k - \mathsf{P}(Y_k = y_k \mid \eta;\boldsymbol{\beta}),\end{aligned}$$

where the last follows identically to the derivation of the likelihood equations of logit GLM. It remains to set $\nabla \log \mathcal{L}_{\text{marginal}}(\boldsymbol{\beta}) = \mathbf{0}$ and go back to the random variables Y_k for y_k. The version of the MMLE of the Rasch model with parametric $\mathsf{Var}(\eta) = \sigma^2$ in Remark 1.3.6 follows directly from the next section.

A.2 Extended Generalized Linear Latent Variable and Structural Models

The primary arguments to conclude that the regularity conditions are satisfied are analogous to the previous section. So, it remains to discuss the derivatives of the log-likelihood and their regularity.

First, we discuss

$$\frac{\partial P(\boldsymbol{y};\boldsymbol{\vartheta})}{\partial \vartheta_v},$$

for all $1 \leq v \leq (q+1)d + b + z^d$, $\boldsymbol{\vartheta} \in \Theta$ and $\boldsymbol{y} \in \{0,1\}^d$. We first suppose $1 \leq v \leq d$, i.e. $\vartheta_v = \alpha_{v0}$ for an $v \in \{1,\ldots,d\}$. We consider the version of (A.4), i.e. we show:

$$\begin{aligned}\frac{\partial P(\boldsymbol{y};\boldsymbol{\vartheta})}{\partial \alpha_{v0}} =& \frac{\partial}{\partial \alpha_{v0}} \int_\Theta \left(\prod_{i=1}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\ =& \int_\Theta \frac{\partial}{\partial \alpha_{v0}} \left(\prod_{i=1}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta})\end{aligned}$$

Let $o \in \{1,\ldots,d\}$ and without restriction $y_{ok} = 1$ (the case $y_{ok} = 0$ is analogous as in the previous, Rasch model, case). We define for $\Delta \in \mathbb{R}$:

$$\begin{aligned}f_\Delta^{(1)} :=& \frac{1}{\Delta} \left(F\left(\alpha_{i0} + \Delta + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i\right) - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i) \right) \\ & \times \prod_{\substack{i=1 \\ i \neq o}}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i}\end{aligned}$$

Then it holds with $\omega = (\alpha_{i0} - \varepsilon, \alpha_{i0} + \varepsilon)$:

$$\left| f_\Delta^{(1)} \right| \leq \sup_{\delta \in \omega} f(\delta + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i),$$

since it exists a $\delta* \in (\alpha_{o0}, \alpha_{o0} + \Delta)$ due to the mean value theorem so that

$$f_\Delta^{(1)} = f(\delta^* + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i) \prod_{\substack{i=1 \\ i \neq o}}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i}, \tag{A.5}$$

while simultaneously

$$\left| \prod_{\substack{i=1 \\ i \neq o}}^d F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right| \leq 1. \tag{A.6}$$

So, (A.6) can neglected in (A.5). Additionally, by computing the supremum and considering ε sufficiently large, the dependence of δ^* of Δ vanishes. So, analogously to the Rasch model case, we can interchange limit for $\Delta \to 0$ and the integral due to Lebesgue's theorem of dominated convergence with the majorant c_1 of equation (2.9) of condition (PLM3).

Next, we want to show

$$\begin{aligned}\frac{\partial P(\boldsymbol{y};\boldsymbol{\vartheta})}{\partial \alpha_{vj}} =& \frac{\partial}{\partial \alpha_{vj}} \int_\Theta \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\ =& \int_\Theta \frac{\partial}{\partial \alpha_{vj}} \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}),\end{aligned}$$

for all $v \in \{1, \ldots, d\}$ and $j \in \{1, \ldots, q\}$, $\boldsymbol{\vartheta} \in \Theta$ and $\boldsymbol{y} \in \{0,1\}^d$. This time, we select

$$\begin{aligned}f_\Delta^{(2)} :=& \frac{1}{\Delta} \left(F(\alpha_{i0} + \Delta \cdot \eta_j + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i) - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i) \right) \\ & \times \prod_{\substack{i=1 \\ i \neq o}}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i}\end{aligned}$$

We can proceed analogously with

$$|f_\Delta^{(2)}| \leq |\eta_j c_1|,$$

which is a integrable majorant due to (PLM1).

For the derivative with respect to the link function parameters, i.e.

$$\begin{aligned}\frac{\partial P(\boldsymbol{y};\boldsymbol{\vartheta})}{\partial \gamma_{oj}} =& \frac{\partial}{\partial \gamma_{oj}} \int_\Theta \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\ =& \int_\Theta \frac{\partial}{\partial \gamma_{oj}} \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}),\end{aligned}$$

for all $o \in \{1, \ldots, d\}$ and $j \in \{1, \ldots, z\}$, $\boldsymbol{\vartheta} \in \Theta$ and $\boldsymbol{y} \in \{0,1\}^d$, we can proceed as in the first case: Due to equation (2.10) in condition (PLM3), the sequence

$$\begin{aligned}f_\Delta^{(3)} :=& \frac{1}{\Delta} \left(F(\alpha_{o0} + \boldsymbol{\alpha}_o^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_o + \boldsymbol{e}_j \cdot \Delta) - F(\alpha_{o0} + \boldsymbol{\alpha}_o^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_o) \right) \\ & \times \prod_{\substack{i=1 \\ i \neq o}}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i},\end{aligned}$$

for $\Delta \in \mathbb{N}$ and the js standard euclidean unit vector $\boldsymbol{e}_j$, satisfies

$$|f_\Delta^{(3)}| \leq c_2,$$

for c_2 of (2.10). It remains the case of the parameters of the structural model, i.e.

$$\begin{aligned}\frac{\partial P(\boldsymbol{y};\boldsymbol{\vartheta})}{\partial \delta_j} =& \frac{\partial}{\partial \delta_j} \int_\Theta \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}) \\ =& \int_\Theta \left(\prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \right) \frac{\partial}{\partial \delta_j} \mathfrak{h}(\boldsymbol{\eta}, \boldsymbol{\delta})\, \mathscr{G}(\mathrm{d}\boldsymbol{\eta}),\end{aligned}$$

for all $j \in \{1, \dots, b\}$, $\boldsymbol{\vartheta} \in \Theta$ and $\boldsymbol{y} \in \{0,1\}^d$. Since the measurement model satisfies

$$0 \leq \prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \leq 1,$$

we can estimate this product by 1 in the majorant. So, to estimate

$$\begin{aligned} f_\Delta^{(4)} :=& \frac{1}{\Delta} \left(\mathfrak{h}_k(\boldsymbol{\eta}, \boldsymbol{\delta} + \boldsymbol{e}_j \Delta) - \mathfrak{h}_k(\boldsymbol{\eta}, \boldsymbol{\delta})\right) \prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i} \\ =& \partial_{\delta_j} \mathfrak{h}_k(\boldsymbol{\eta}, \boldsymbol{\delta} + \boldsymbol{e}_j \Delta^*) \prod_{i=1}^{d} F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)^{y_i} \left(1 - F(\alpha_{i0} + \boldsymbol{\alpha}_i^\intercal \boldsymbol{\eta};\, \boldsymbol{\gamma}_i)\right)^{1-y_i}, \end{aligned}$$

we define

$$\sup_{\Delta \in (-\varepsilon, \varepsilon)} \left|\partial_{\delta_j} \mathfrak{h}_k(\boldsymbol{\eta}, \boldsymbol{\delta} + \boldsymbol{e}_j \Delta)\right| =: G(\boldsymbol{\eta}), \qquad \boldsymbol{\eta} \in \Theta.$$

This is a majorant for all $\Delta < \varepsilon$ and integrable by (2.7) of condition (PLM2).

The strengthened form (PLM4)–(PLM6) allow an analogous argumentation for the derivatives up to third order. For example ensure the conditions (PLM4) and (PLM6) that there is a constant $c' > 0$ so that $c'|\eta_j^3|$ is an integrable majorant for the sequence constructing the derivarive of third order with respect to α_{vj}, $v \in \{1, \dots, d\}$, $j \in \{1, \dots, q\}$. In particular, we can see that the derivatives of up to third order are locally finite. So, they are finite in a neighborhood of $\boldsymbol{\vartheta}_0$. The remaining regularity conditions follow by the same arguments as for the constrained Rasch model and are omitted.

For the proof of Theorem 2.4.1, we have to check that the distribution family is *quadratic mean differentiable*. Due to Theorem 12.2.2 of Lehmann and Romano (2005, p. 488), this is satisfied if

(i) $P(\boldsymbol{y}, \bullet) \in C^1(\Theta, \mathbb{R})$ for all $\boldsymbol{y} \in \mathbb{Y}$ and

(ii) the Fisher information exists and is continuous in $\boldsymbol{\vartheta}$.

Since we yet proved that the gradient is an integral of differentiable mappings, it is continuous. The Fisher information is a finite sum of finite products of continuous mappings. So, the family is quadratic mean differentiable.

Appendix B

Heuristic Iterative Simulation Based Bias Reduction of Pseudo Likelihood Estimates

The simulation study in Section 5.3 illustrates that a multiplicative scaling can provide a remarkable bias reduction and overall improvement of the estimates. Since the used factor is unknown, we have to estimate it in practice. Combining ideas of the bootstrap (cf. e.g. Efron and Tibshirani, 1994) and numerical solving of nonlinear equations, we derived a method to estimate the scaling factor.[1] Bear in mind that this method has only a heuristic justification at the current state.

Suppose that we observed an sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $n \in \mathbb{N}$, of responses to d items and denote by $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\alpha}}$ the pseudo likelihood estimates for an corresponding logit IRT model of dimension q. Now observe that if we would have

1. ideal scaling factors c_1 and c_1 (for $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\alpha}}$) and
2. a model for $\mathcal{H}$,

then we could generate a sample $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$ from the IRT model using $c_1\hat{\boldsymbol{A}}$ and $c_2\hat{\boldsymbol{\alpha}}$ as parameters, from which new obtained pseudo likelihood estimates $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{\alpha}}$ are close to $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\alpha}}$. Consequently, we expect that if the IRT model is correctly specified, then $c_1\hat{\boldsymbol{A}}$ and $c_2\hat{\boldsymbol{\alpha}}$ are close to the true $\boldsymbol{A}$ and $\boldsymbol{\alpha}$, because they lead to nearly the same pseudo likelihood estimates. Additionally, if c_1' (or c_2') is to high/to low, then a linear regression line with response $c_1'\hat{\boldsymbol{A}}$ (or $c_2'\hat{\boldsymbol{\alpha}}$) and explanatory variable $\tilde{\boldsymbol{A}}$ (or $\tilde{\boldsymbol{\alpha}}$) has a slope $\tilde{c}_1$ (or $\tilde{c}_2$) which is lower/greater than one. Hence, our target is to find that values of c_1 and c_2, such that

$$1 - \tilde{c}_1 \approx 0 \quad \text{and} \quad 1 - \tilde{c}_2 \approx 0$$

hold. This means that we search the root of a function $\mathbb{R}^2 \to \mathbb{R}^2$, which can be done by an iterative algorithm e.g. by Broyden's method (Broyden, 1965).

Another aspect is the generation of values of the latent variables, from which the responses can be generated. If one supposes a model $\mathcal{H}$ for the distribution of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, then this information can be used for generation. If there is instead no a-priori given $\mathcal{H}$, an alternative is given by estimating $\mathcal{H}$ and use this estimate for generation. Corresponding to the approach of the pseudo likelihood, we can use the approximate EAP with the pseudo likelihood estimated parameters for point estimates of the true values of $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ underlying the data. Then, we can either sample from these point estimates or make a further non-parametric density estimation. While the last may be a further source of errors,

[1] Notice, most literature focuses on the reduction of the finite sample bias of consistent methods, such as MacKinnon and Smith (1998). Here, we discuss a bias reduction of an inconsistent method with the target of receiving errors for finite samples, which are acceptable and comparable to that of classical methods.

it incorporates the continuity of the latent variable and "gives their distribution variance back", which was lost by the point estimation.

Putting everything together, we receive the following iterative algorithm for bias reduction of the pseudo likelihood estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\alpha}}$ based on n response patterns on a test with d items and supposed or estimated distribution $\mathcal{H}$ of the latent variables:

Algorithm 4 Heuristic Iterative Simulation Based Bias Reduction for PLE

Require: $M \in \mathbb{N}$, $\boldsymbol{c}^{(1)} = \left(c_1^{(1)}, c_2^{(1)}\right)^\intercal \in \mathbb{R}^2$, $\boldsymbol{c}^{(0)} = \left(c_1^{(0)}, c_2^{(0)}\right)^\intercal \in \mathbb{R}^2$, $T \in \mathbb{N}$, $\delta > 0$

for $m = 1, \ldots, M$ **do**

Generate a sample $\boldsymbol{\eta}_1^{(m)}, \ldots, \boldsymbol{\eta}_n^{(m)}$ from $\mathcal{H}$

Generate a sample $\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_n^{(m)}$ from the IRT model which is based on $\boldsymbol{\eta}_1^{(m)}, \ldots, \boldsymbol{\eta}_n^{(m)}$ with the parameters $c_1^{(0)}\hat{\boldsymbol{A}}$ for the factor loadings and $c_2^{(0)}\hat{\boldsymbol{\alpha}}$ for the intercepts.

Obtain pseudo likelihood estimates $\tilde{\boldsymbol{A}}^{(m)}, \tilde{\boldsymbol{\alpha}}^{(m)}$ from $\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_n^{(m)}$.

end for

Replicate $\hat{\boldsymbol{A}}$ M times and rewrite it as a single vector of length dqM. Denote this vector by $\boldsymbol{v}_1$. Replicate $\hat{\boldsymbol{\alpha}}$ also M times to a single vector of length dM and denote it by $\boldsymbol{v}_2$. Collect all $\tilde{\boldsymbol{A}}^{(M)}, \ldots, \tilde{\boldsymbol{A}}^{(M)}$ to a single vector $\boldsymbol{v}_1'$ of length dqM, such that on each position in $\boldsymbol{v}$, a $\hat{\alpha}_{ij}$ corresponds to a $\tilde{\alpha}_{ij}$ in $\boldsymbol{v}'$ for the same position, $i = 1, \ldots, d$ and $j = 1, \ldots, q$. Analogously, collect all $\tilde{\boldsymbol{\alpha}}^{(M)}, \ldots, \tilde{\boldsymbol{\alpha}}^{(M)}$ to a single vector $\boldsymbol{v}_2'$.

Compute the least squares estimator $\tilde{c}_h^{(0)}$ of the parameter c_h in the linear model $\boldsymbol{v}_h = c_h\boldsymbol{v}_h' + \boldsymbol{\varepsilon}_h$, for some error term $\boldsymbol{\varepsilon}_h$ and $h = 1, 2$.

for $t = 1, 2 \ldots, T$ **do**

for $m = 1, \ldots, M$ **do**

Generate a sample $\boldsymbol{\eta}_1^{(m)}, \ldots, \boldsymbol{\eta}_n^{(m)}$ from $\mathcal{H}$

Generate a sample $\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_n^{(m)}$ from the IRT model which is based on $\boldsymbol{\eta}_1^{(m)}, \ldots, \boldsymbol{\eta}_n^{(m)}$ with the parameters $c_1^{(t)}\hat{\boldsymbol{A}}$ for the factor loadings and $c_2^{(t)}\hat{\boldsymbol{\alpha}}$ for the intercepts.

Obtain pseudo likelihood estimates $\tilde{\boldsymbol{A}}^{(m)}, \tilde{\boldsymbol{\alpha}}^{(m)}$ from $\boldsymbol{y}_1^{(m)}, \ldots, \boldsymbol{y}_n^{(m)}$.

end for

Construct vectors $\boldsymbol{v}_h$ and $\boldsymbol{v}_h'$, $h = 1, 2$ from the original and the bootstrap PLEs.

Compute the least squares estimator $\tilde{c}_h^{(t)}$ of the parameter c_h in the linear model $\boldsymbol{v}_h = c_h\boldsymbol{v}_h' + \boldsymbol{\varepsilon}_h$, for some error term $\boldsymbol{\varepsilon}_h$ and $h = 1, 2$. Denote the estimated standard error of $\tilde{c}_h^{(t)}$ by s_h, $h = 1, 2$.

if $\max(|\tilde{c}_1^{(t)} - 1|, |\tilde{c}_2^{(t)} - 1|) \leq \delta$ **then**

break

end if

if $t = 1$ **then**

set

$$\boldsymbol{J}_0^{-1} = \begin{pmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{pmatrix},$$

with $j_{12} = j_{21} = 0$ and

$$j_{hh} = \frac{c_h^{(1)} - c_h^{(0)}}{\tilde{c}_h^{(0)} - \tilde{c}_h^{(1)}}, \qquad h = 1, 2.$$

end if

set

$$\begin{aligned} \boldsymbol{f}_t :=& (1 - \tilde{c}_1^{(t)}, 1 - \tilde{c}_2^{(t)})^\intercal, \quad \Delta_{\boldsymbol{f},t} := \boldsymbol{f}_t - \boldsymbol{f}_{t-1}, \quad \Delta_{c,t} := -\boldsymbol{J}_{t-1}^{-1}\boldsymbol{f}_{t-1} \\ \boldsymbol{J}_t^{-1} :=& \boldsymbol{J}_{t-1}^{-1} + \frac{\Delta_{c,t} - \boldsymbol{J}_{t-1}^{-1}\Delta_{\boldsymbol{f},t}}{\Delta_{c,t}^\intercal \boldsymbol{J}_{t-1}\Delta_{\boldsymbol{f},t}} \Delta_{c,t}^\intercal \boldsymbol{J}_{t-1}^{-1}, \quad \boldsymbol{c}^{(t+1)} = \boldsymbol{c}^{(t)} + \Delta_{c,t}, \end{aligned}$$

where $\boldsymbol{c}^{(t+1)} = (c_1^{(t+1)}, c_2^{(t+1)})^\intercal$

end for

return $\boldsymbol{c}^{(t+1)}$

Initial estimates $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(0)}$ could be selected by orientating on the results of the bias analysis simulation study in Table 5.4. Generally, it is appropriate to select $\boldsymbol{c}^{(0)} = (1, 1)^\intercal$, i.e. no scaling of the PLE. Our experiments and simulation studies using PLE with item-wise standardization to transform the estimates suggest that the required scaling is commonly larger than one and typically a smaller rescaling is required for the intercepts than for the factor loadings. Selecting both components of $\boldsymbol{c}^{(1)} = (c_1, c_2)$ to be greater than one is therefore appropriate. Especially for large tests, selecting c_2 only slightly larger than 1 is reasonable due to the small correction that can be expected to be required. It is likely that the required correction factor for the factor loadings is larger than the required correction factor for the intercepts (cf. Table 5.4). This relation of the required correction factors can be used for selecting c_1 such that $c_1 > c_2$. In our tests, we received good results in commonly up to 6 iterations. However, we often observed that the algorithm starts to diverge after several iterations. An explanation of this convergence issue is the smoothness requirement on the function surface of gradient based iterative update rules for solving nonlinear systems or nonlinear optimization. In our case, the random generation of artificial responses in each iteration can contain some 'bad luck' in the sense that the new responses push the new step in a direction that differs slighly from the previous one only due to the pseudo random numbers while this small change in the direction leads the step away from the direction of finding the solution. In unlucky cases, this push was critical and cannot be captured by the deterministic structure of the update rule. This is similar to the circumstance that many Newton-type methods divergence if the starting values are sub-optimal and the update rules direct the algorithm away from the solution. By monitoring the values of $\boldsymbol{f}_t$ for each iteration t we can see if a good correction was reached before the algorithm starts to diverge. This means that the value of $\boldsymbol{c}^{(t)}$ can be taken as resulting correction factor if the corresponding value of $\boldsymbol{f}_t$ is sufficiently small (i.e. only slightly larger than the selected δ). If no such iteration was reached before the algorithm started to diverge, the

algorithm can be restarted. Due to the random generation of artificial responses via each iteration, trying a new start of the algorithm is reasonable before changing the starting values or the parameters of the algorithm. If this is not sufficient, increasing or decreasing the starting value $\boldsymbol{c}^{(1)} = (c_1, c_2)$ can lead to a better result. Additionally, increasing M can increase the stability of the iterations because the random fluctuations between all iterations decrease. After the following remark we will discuss practical aspects of selecting M and δ. Additionally we give an illustrative example that shows that the non-guaranteed convergence does not prevent the applicability of the algorithm and that a few iterations of this algorithms can improve the PL estimates substantially.

Remark B.0.1

1. If one prefers another linear bias correction, e.g. with intercept, then this can be included here by enforcing that the intercept of the corresponding regression in the simulation step is approximately zero. Higher order terms should cancel out, too, because we want to find the original pseudo likelihood estimates.

2. In step 2. a) ii., the original PL estimates provide good starting values of the bootstrap PLEs.

3. After final scaling factors $\boldsymbol{c}$ are obtained, we can generate a further sample of M' bootstrap pseudo likelihood estimates as in step (a). Then, the variance of $c_1\tilde{\alpha}_{ij}^{(1)}, \ldots, \tilde{\alpha}_{ij}^{(M')}$ and of $c_2\tilde{\alpha}_{i1}^{(1)}, \ldots, \tilde{\alpha}_{ij}^{(M')}$ can be used as estimates of the variances of the corrected pseudo likelihood estimates. This approach will lead to a slight underestimation of the variance since the variance in $\boldsymbol{c}$ is not taken into account. Instead of computationally intensive repeating the full approach several times, it might be sufficient to iterate several steps after the first acceptable $\boldsymbol{c}$ was obtained and taking them. This could be investigated in future research.

4. In many cases, the final estimation can be several times as computational intensive as the not bias reduced PLE.

5. There are some different ways of starting the method, instead of selecting $\boldsymbol{J}_0^{-1}$ completely arbitrary, we prefer to select it in step 5. e) such that

$$\boldsymbol{c}^{(1)} - \boldsymbol{c}^{(0)} = \boldsymbol{J}_0^{-1}\Big(\boldsymbol{f}_1 - \boldsymbol{f}_0\Big),$$

   where $\boldsymbol{f}_0 = (1 - \tilde{c}_1^{(0)}, 1 - \tilde{c}_2^{(0)})^\intercal$ and $\boldsymbol{f}_1$ corresponds to a second starting vector.

6. A foundation for the specification of the termination criterion can be the formulation of approximate confidence intervals of $1 - \tilde{c}_1^{(t)}$ and $1 - \tilde{c}_2^{(t)}$. In the linear regressions for the computation of $\boldsymbol{c}^{(t+1)}$, the pseudo likelihood estimated parameters of each bootstrap sample are dependent. Hence, even if M is quite large, we cannot apply the classical asymptotic theory for linear models. However, in simulations we received that the distribution of $\frac{\tilde{c}_h - \bar{c}_h}{s_h}$, $h = 1, 2$, still can be well approximated by a t_{b_h-1} distribution[2], where s_h denotes the estimate of the standard deviation of $\tilde{c}_h$ from the linear regression, t_{df} denotes Student's t-distribution for df degrees of freedom, b_h is the number of parameter estimates from which we computed $\tilde{c}_h$ and $\bar{c}_h$ denotes

[2]Hence, it also can be approximated by a $\mathcal{N}(0, 1)$ distribution since the degree of freedom $b_h - 1$ is commonly quite large (e.g. $b_h - 1 = 999$ for $h = 1, 2$ if $d = 20$, $q = 1$, $M = 50$).

the sample mean of a sample of $\tilde{c}_h$s: In a simulation of 100 $\tilde{c}_h$s for $d = 20$, $q = 1$, $n = 1000$ and $M = 10$ we got the p-values of 0.7912 ($h = 1$) and 0.3564 ($h = 2$) for Kolmogorov-Smirnov-tests of the distribution of the sample of the $\frac{\tilde{c}_h - \bar{c}_h}{s_h}$s against the $t_{b_h - 1}$ distribution. For $M = 50$, we got the p-values 0.8628 ($h = 1$) and 0.4837 ($h = 2$). This is no comprehensive analysis but it supports the claim that the $\tilde{c}_h$s are asymptotically normal distributed for $M \to \infty$ and that their distribution can be well approximated by a t distribution for moderate M. ◇

It is absolutely reasonable to allow comparably large deviations of the $\tilde{c}_h$s from 1 in comparison to the common termination criteria of purely deterministic methods. First of all, the (pseudo) random generation of $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$ in each iteration to compute $\tilde{c}_1$ and $\tilde{c}_2$ will cause a random fluctuation which cannot be avoided and only reduced by selection M really large. But selection M large enough to prevent any notable random fluctuation will cause a large computational demand in each iteration due to the computation of pseudo likelihood estimates for each artificial data set. Moreover, the algorithm will commonly need many iterations to terminate for small δ. To ensure that the cost-value ratio of estimate improvement to the loss of computation time is acceptable, we recommend the selection of comparably small M (e.g. $1 \leq M \leq 100$) and comparably large δ (e.g. $\delta = 0.1$).

We discuss the example application on a single simulated data set to get an impression of the quality of the estimated correction factors with $\delta = 0.1$ and $M = 100$. In this example, we have a data set of $d = 30$ items, $n = 2000$ responses and a test dimension of $q = 1$. The true latent variable distribution is $\mathcal{H} = \mathcal{N}_1(0, 1)$, which we also used in the scaling factor estimation. With the starting values $\boldsymbol{c}^{(0)} = (1,\, 1)^\intercal$ and $\boldsymbol{c}^{(1)} = (1.5,\, 1.2)^\intercal$, $M = 100$ and $\delta = 0.1$, we received in 2 iterations the scaling factors $\boldsymbol{c}^{(3)} \approx (1.302,\, 1.053)^\intercal$. As described earlier, $\boldsymbol{c}^{(0)}$ corresponds to no scaling of the PL estimates and $\boldsymbol{c}^{(1)}$ was arbitrarily selected by taking larger values that are not unrealistically large. However, the result neither depends on these values specifically nor on the fact that $\boldsymbol{c}^{(1)}$ is component-wise larger than $\boldsymbol{c}^{(3)}$. In fact, by using the alternative starting value $\boldsymbol{c}^{(1)} = (1.1, 1.01)^\intercal$ the algorithm terminated in several trials after two iterations and diverged in none of our trials.[3] Using the obtained correction factor $\boldsymbol{c}^{(3)}$, we reduce the RMSE of the PL estimate from ≈ 0.1242 to ≈ 0.0769 which is a remarkable improvement of the estimate. The RMSE would be ≈ 0.0656 if the correction factor would be computed using the actual (unknown) parameter values and the correction factor would be $\approx (1.245, 1.015)^\intercal$. An MML estimate for the same data set has a RMSE of ≈ 0.0673, hence, the corrected PLE using the factors in $\boldsymbol{c}^{(3)}$ is only slightly worse than PLE with the best possible correction using the correction factor obtained from the true parameters and the MMLE. If $\delta < 0.05875$ would be selected, then a third iteration would be computed. In this case, a third iteration would have again a larger deviation of the $\tilde{c}_h$s from 1. Hence, we see that the algorithm is not perfectly stable. However, any further iteration requires a large extra amount of computation time while we cannot expect to obtain correction factors which reduce the MAE by more than single percents. It is therefore reasonable to stop the improvement of the PLE at this point. We saw in Section 5.4 that the MMLE can require multiple hundred times more computation time than PLE in high dimensional settings such that a few iterations of the iterative improvement algorithm for PLE with a small or moderate M can be spend to improve the PL estimates while PLE still would be seriously faster. For a visual impression in this example, we see the scatter plot of the true against the pseudo likelihood estimated values in Figure B.1 where the

[3] In these cases we obtain correction factors of ≈ 1.23 to ≈ 1.24 for the factor loadings and ≈ 1.02 for the intercepts. These correction factors result in RMSE values of about 0.0658 to 0.0661

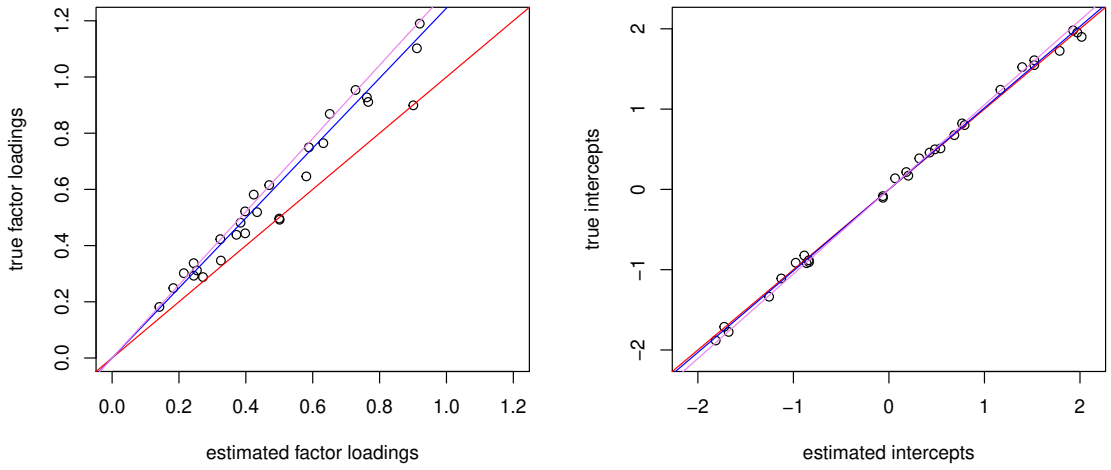

Figure B.1: *Scatter plot of true model parameters against pseudo likelihood estimates with linear regression using the true values (blue), the estimated correction line (violet) and the identity (red).*

red line is the identity, the blue line is the linear regression with least squares and without intercept and the violet line is its estimate using the proposed algorithm. Additionally, we see in Figure B.2 the true ICC of six example items (black) together with their MML estimates (red), their PL estimates (green) and their improved PL estimates (blue). While their is a notable deviation of the PL estimated ICCs from the actual ones, MMLE and the improved PLE are comparably good estimates of the actual ICCs.

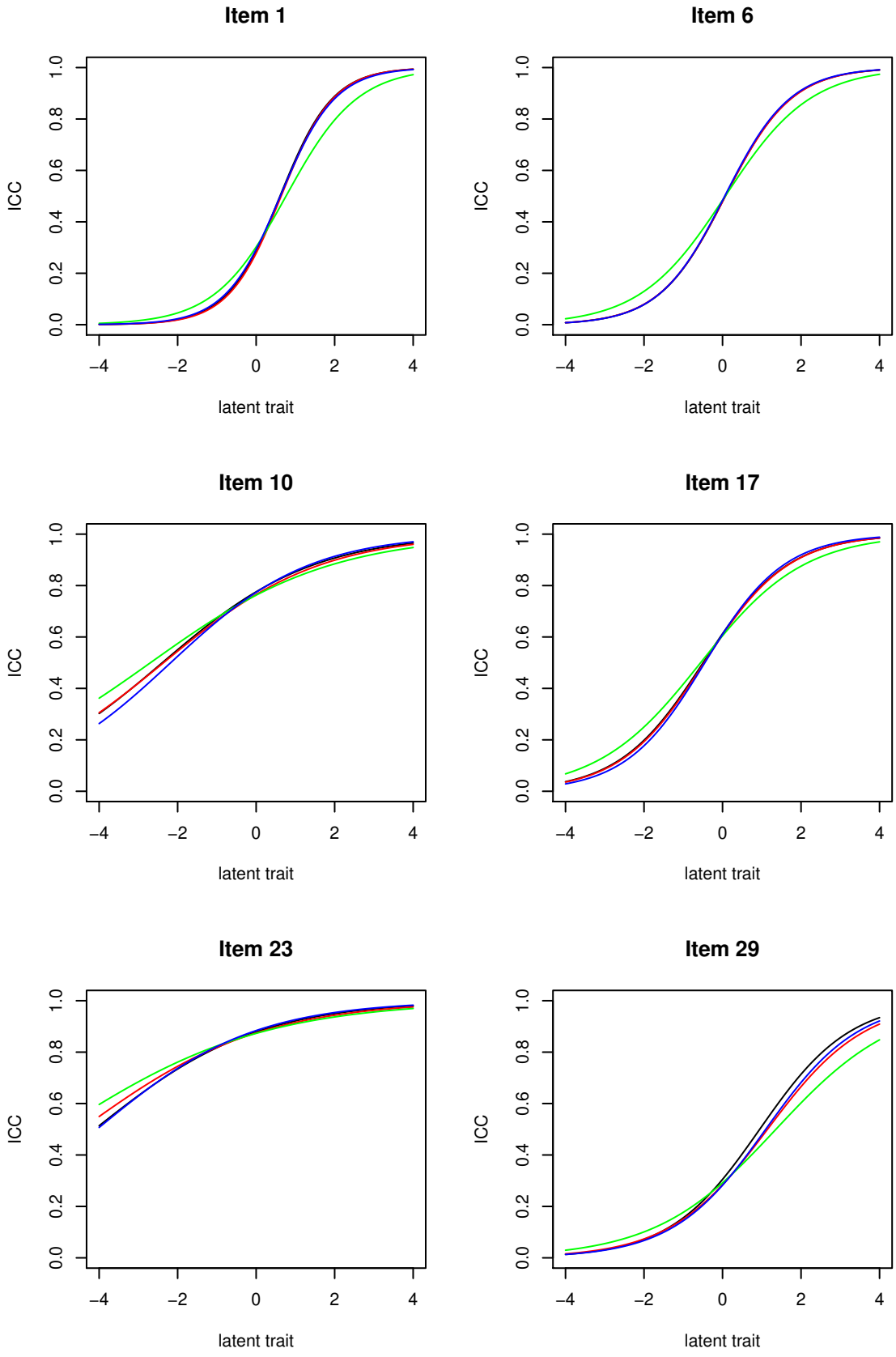

Figure B.2: *The true ICC of six example items (black), the MMLE estimated ICCs (red), the PLE estimated ICCs (green) and the iteratively corrected PLE estimated ICCs (blue).*

Appendix C

Aspects of the Implementation in R

Here, we will shortly sketch some important aspects for the implementation in R (cf. R Core Team, 2017). Relevant source code is available on request.

C.1 Chapter 2: EAP and MMLE for EGLLVM and Generation of Data

For the approximation of the integrals in the log-likelihood and the EAP, we take Gauss–Hermite quadrature and computed the weights and quadrature points via the function `gauss.hermite` from the R-package **rmutil** (cf. Swihart and Lindsey, 2019). Using the vectorization, this can be done by avoiding most of the corresponding loops in the R code. For the optimization of the skew probit models, we take the BFGS-algorithm of the R-base function `optim`. For the guessing models, we take the L-BFGS-B method of `optim` as it allows the inclusion of the corresponding parameter bounds. For the computation of the gradients of the log-marginal likelihood function for this models, we take a finite-differences approximation. This is of acceptable precision and allows an easy adaption to various models. Also, the necessary computation time is acceptable for models of relatively small data sets and, therefore, for the conducted data analyses and simulation studies.

To evaluate the cdf of the standard skew normal distribution, we take the representation of Definition 1.1.7 and the function `pmvnorm` of the R-package **mvtnorm** (cf. Genz and Bretz, 2009; Genz et al., 2020). To accelerate the computation, we tabulated the function values for $x, \gamma \in (-10, 10)$ and interpolated linearly between steps of length ≈ 0.006. While this is of lower importance in MMLE and EAP, it accelerates the generation of the observations during the simulation study a lot. To generate skew normal distributed random numbers, we take the rejection method with a Laplace distribution as proposal. For our parameter choices, we received a acceptance rate of $\approx \frac{2}{3}$, which can be accepted for the considered purposes. The generation of the other needed distributions is easily available in R. The pdf can directly be evaluated using base functions and (1.11). The presented new link functions for guessing in Subsection 2.2.1 can be evaluated using R-base functions.

We specified the most important R-functions as follows.

R-**Function 1** Log-Likelihood function for EGLLVM for binary data

---

Name: `comp.log.lik`

Description: Computes the *negative* value of the log-likelihood function for marginal maximum likelihood estimation of an one dimensional EGLLVM for binary data. There is one version for the guessing model with link function (2.3) with $c = 1$ and for with the skew-probit link function. Both suppose that the

latent variable has a standard skew-normal distribution with one skewness parameter.

Input: `param` A vector of length $3d + 1$ of all parameters. The first d elements are the intercept parameters, the next d the factor loadings, then the parameters for the d link functions and, finally, the parameter of the structural model.

`control` A list with the elements `data` (a matrix with all observed response pattern), `counts` (the number of times which each response pattern was observed), `gh.points.weights` (a two column matrix with the Gauss–Hermite quadrature evaluation points in the first and their weights in the second column, by default 30 points are initiated)

Output: A numeric vector of length one with the value of the negative log-likelihood function

Remarks: The input argument `control` is only contained in the guessing model version as an example for a general purpose function. The function of the skew-probit model uses corresponding variables on the highest level as a specialized single purpose function for the simulation studies. The version for the simulation studies also uses a separation of the parameters into the following input variables:

`a0` A vector the intercept parameters

`a1` A vector of all factor loadings

`gamma` A vector of all link function parameters

`struct_skewness` The parameter of the structural model

R-**Function 2** EAP-Estimator for the latent variables behind all response vectors

Name: `compute.EAP`

Description: Computes the EAP for each observed response vector and given parameters of an one dimensional EGLLVM for binary data.

Input: `a0` A vector the intercept parameters

`a1` A vector of all factor loadings

`gamma` A vector of all link function parameters

`struct_skewness` The parameter of the structural model

Output: A numeric vector of all EAP estimates

Remarks: The function is only used in the simulation studies and uses the data and GH-information in global variables analogously to the log-likelihood function for skew-probit models.

C.2 Chapter 4: aEAP and MMLE for Logit Models by aMCEM

The approximate EAP (aEAP) of (4.12) can be well computed by R-base functionality. For the computation of the information matrices at theses points in (4.17), the use of loops

cannot be avoided. Since this was no time critical place in the algorithm, we implemented this step in R. The following step of transforming $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$ to the samples of the indiviual posterior distributions in (4.20) can be accelerated by implementing it in RCpp, we use additionally the library **RcppArmadillo** for a efficient implementation in RCpp (cf. Eddelbuettel and Sanderson, 2014). The standardization of this sample can again be well done in base R. For the computation of the item "likelihoods" and their gradients by (4.21) and (4.22), we take base R, but here could an acceleration by **RcppArmadillo** be useful. For the M-step, we used again BFGS of `optim` and the parameters of the last step as starting values. We specified the R-function for the aMCEM-algorithm as follows.

R-**Function 3** aMCEM-Algorithm for MMLE of logit IRT models

---

Name: `aMCEM`

Description: Computes the MML estimates for a logit model and a given data set using Algorithm 1.

Input:

`data` A $d \times n$ matrix of the full observed data set, d and n are extracted from this input variable

`q` The latent dimension $q \in \mathbb{N}$

`theta_0` The start vector of length $d(q+1) \in \mathbb{N}$ or `NA`. By default `NA` and selected randomly as described in Section 4.3.3 if $q > 1$. If $q = 1$, then the factor loadings are initiated by a vector of ones.

`MC_N_half` The half of the finally used number of MC quadrature points, by default 20.

`max_iter` Maximal number of iterations, by default 1000.

`eps` $\varepsilon > 0$ for the termination of the algorithm while `max_iter` is not reached, by default `eps=sqrt(.Machine$double.eps)`

`use_antithetic_vars` `TRUE` or `FALSE`, by default `TRUE`, to decide wether the antithetic variable approach is used to specify the set of all quadrature points for MC-integration. If not, `MC_N=2*MC_N_half` iid vectors from $\mathcal{N}_q(\mathbf{0}, \mathrm{I}_q)$ are generated

Output: A list with the following elements:

`A_0` The start value of the matrix of factor loadings

`beta_0` The start value of the vector of intercepts

`A` The final matrix of factor loadings

`beta` The final vector of intercepts

`latent_k` The $q \times n$ matrix of all final aEAP estimated latent vectors (i.e. the estimates for $k = 1, \ldots, n$)

`information_k` The corresponding array of the estimated expected information matrices for all latent vector estimates

`converged` An indicator if the convergence criterion was satisfied at termination or if `max_iter` was reached

`iter` The number of the last iteration

C.3 Chapter 5 and Appendix B: Pseudo Likelihood Estimation

If there are only a few items, then the number of maximal response patterns is limited. Grouping the responses by their response pattern can accelerate the computation if there is mention worthy accumulation. If there are only comparably few response pattern which are observed several times, then we can instead reduce the number of floating point operations if we do not group because we do not have to multiply with the number of counts for each pattern. In trials, we received noteworthy reduction for $d \leq 20$ and the n we considered in this work.

To hold the implementation of the direct approach for PLE flexible, we use grouping for $d \leq 20$ and else, we simply take all response patterns as they are. A critical point in the computation of a function value of the pseudo log-likelihood function in the direct approach is the evaluation of

$$(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{B})\boldsymbol{y}_k, \qquad k = 1, \ldots, n,$$

where $\boldsymbol{B}$ is a diagonal matrix with the diagonal values of $\boldsymbol{A}\boldsymbol{A}^\top$ (for simplicity, suppose n is the number of *unique* patterns). The principally easy way would be the direct computation by:

1. Compute $\boldsymbol{A}\boldsymbol{A}^\top$.
2. Set the diagonal to zero.
3. Multiply by all $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.

But, in this case we multiply a $d \times d$ with a $d \times n$ matrix, which is computationally inefficient. It is actually computational more efficient to proceed as follows:

1. Compute $\boldsymbol{V}_k := \boldsymbol{A}^\top \boldsymbol{y}_k$, $k = 1, \ldots, n$.
2. Compute $\boldsymbol{A}\boldsymbol{V}_k$, $k = 1, \ldots, n$, and subtract the diagonal elements of $\boldsymbol{A}\boldsymbol{A}^\top$ on the corresponding places.

On this way, we reduce the computation essentially to the products of a $d \times q$ and a $q \times n$ matrix (after the inital product of a $q \times d$ and a $d \times n$ matrix), while it holds in all interesting cases $q \ll d$ (i.e. 1 to 8 against 20 to 120 during this work). These products and the final computation of the pseudo log-likelihood and its gradient (i.e. (5.11), (5.12) and (5.14)) can be implemented well in RCpp using **RcppArmadillo**. For the optimization, we use a preconditioned inexact truncated Newton algorithm of the function `nloptr` of the equally named R-package **nloptr**, since it was more efficient than **optim** in our trials (cf. Dembo and Steihaug, 1982; Johnson, 2019).

For the two-stage algorithm, we compute for each $i = 1, \ldots, d$ the values of all $\sum_{v=1, v \neq i}^{d} \alpha_{vj} y_{vk}$, $j = 1, \ldots, q$, $k = 1, \ldots, n$ via RCpp. Using these values, the function value and the gradient of the ℓ_is in (5.10) can be fast computed in another RCpp function. The maximization of the ℓ_is is again be done using the same preconditioned inexact truncated Newton algorithm of the function `nloptr` as in the direct approach.

We specified the most important R-functions as follows.

R-**Function 4** Direct approach PLE

Name: `ple.nloptr`

Description: Computes the pseudo likelihood estimates via the direct approach

Input:
- `data` The $d \times n$ matrix of the complete binary response data set with all observed responses
- `theta_0` A numeric vector of length $d(q+1)$ with the start value of the optimization, by default `NA`
- `q` The latent dimension, by default `NA`
- `eps` The value of $\varepsilon > 0$ to decide if the algorithm terminates, by default `sqrt(.Machine$double.eps)`
- `d_try_shorten` The maximal $d \in \mathbb{N}$ for which the response data will be grouped, by default 20

Output: A list with the following elements:
- `optim_ret` The return value of `nloptr` which includes the estimate for the manifest probabilities model in the element `optim_ret$solution`
- `param` The item-wise transformed parameter estimates

Remarks: Either `theta_0` or `q` has to be specified, if not specified the other argument will be determined automatically (random start value or automatic computation of the value of q based on the length of `theta_0`)

R-**Function 5** 2SPLE – two-stage PLE algorithm

Name: `ple.two_stage`

Description: Computes the pseudo likelihood estimates via Algorithm 3

Input:
- `data` The $d \times n$ matrix of the complete binary response data set with all observed responses
- `theta_0` A numeric vector of length $d(q+1)$ with the start value of the optimization, by default `NA`
- `q` The latent dimension, by default `NA`
- `eps` The value of $\varepsilon > 0$ to decide if the algorithm terminates, by default `sqrt(.Machine$double.eps)`
- `max_iter` The maximal number of iterations $T \in \mathbb{N}$, by default 50

Output: A list with the following elements:
- `objective` The numeric vector of the values of $\ell_i^{(t)}(\boldsymbol{\vartheta}_i^{(t)})$, $i = 1, \ldots, d$ in the last iteration t
- `param` The parameter estimates for the manifest probabilities model
- `param_transformed` The item-wise transformed parameter estimates

`converged` An indicator if and by which criterion the algorithm terminated until the maximal number of iterations was reached

`iterations` The number of iterations until the termination of the algorithm

Remarks: Either `theta_0` or `q` has to be specified, if not specified the other argument will be determined automatically (random start value or automatic computation of the value of q based on the length of `theta_0`)

R-**Function 6** General purpose PLE

Name: `ple`

Description: Computes the pseudo likelihood estimates. If $d \leq 20$ or if $d \leq 40$ and $q = 1$, then the direct approach is used else the two-stage approach is used.

Input:

`data` The $d \times n$ matrix of the complete binary response data set with all observed responses

`theta_0` A numeric vector of length $d(q+1)$ with the start value of the optimization, by default `NA`

`q` The latent dimension, by default `NA`

`eps` The value of $\varepsilon > 0$ to decide if the algorithm terminates, by default `sqrt(.Machine$double.eps)`

Output: The return value of R-Function 5 or 6.

Remarks: Either `theta_0` or `q` has to be specified, if not specified the other argument will be determined automatically (random start value or automatic computation of the value of q based on the length of `theta_0`)

R-**Function 7** PLE bias reduction

Name: `estim.bias.red`

Description: Computes the pseudo likelihood bias reduction via Algorithm 4.

Input:

`ple_erg` Result of a non-bias reduced pseudo likelihood estimation

`d` Number of items

`n` Number of observed responses

`scal_fact_0` First start value $\boldsymbol{c}^{(0)}$

`scal_fact_1` Second start value $\boldsymbol{c}^{(2)}$

`M` Bootstrap sample size, by default 50

`diff` Value of the termination criterion, by default 0.1

`max_iter` The maximal number of iterations $T \in \mathbb{N}$, by default 15

Output: A list with the following elements:

`scal_fact` The final correction factor

`iterations` The number of iterations when the algorithm terminated

Remarks: The latent dimension is automatically determined from the input pseudo likelihood estimate. The algorithm uses the direct approach of the PLE.

List of Figures

1.1 Example 1.1.6: Departure in ICCs 12
1.2 Standard skew normal distributions 14
1.3 Aproximation of the TIEV distribution by a skew normal distribution . . . 15
1.4 Example of reflective and formative models 22

2.1 Left heavy tailed distributions for guessing models 1 64
2.2 Left heavy tailed distributions for guessing models 1 65
2.3 Simulation Example in Subsection 2.2.2: True and estimated ICCs 67
2.4 Real Data Example in Subsection 2.2.2: Estimated ICCs 69
2.5 Simulation study in Subsection 2.5.1: Boxplots of Results Setup 1 82
2.6 Simulation study in Subsection 2.5.1: Boxplots of Results Setup 2, γ_1 . . . 84
2.7 Simulation study in Subsection 2.5.1: Boxplots of Results Setup 2, γ_2 . . . 85
2.8 Simulation study in Subsection 2.5.1: Boxplots of Results Setup 2, γ_{struct} . 85
2.9 Simulation study in Subsection 2.5.2: Example estimates of latent traits . . 93

3.1 Example of Subsection 3.4.4 for CS' conditions: Expected log likelihod ratio 118
3.2 Example of Subsection 3.4.4 for CS' conditions: Minimal eigenvalues 119
3.3 Example of Subsection 3.5.4 for PCT conditions: Minimal eigenvalues . . . 144

4.1 Simulations in Subsection 4.2.2: AAE and RMSE values 180
4.2 Simulations in Subsection 4.2.2: Scatter plot of example estimates 181
4.3 First simulation example Sectiom 4.3: Scatter plot of parameter estimates . 189
4.4 Second simulation example Sectiom 4.3: Scatter plot of parameter estimates 190

5.1 Scatterplot of Example 5.2.3 on the need to transform the PLE 209
5.2 Scatterplot of Example 5.2.4 on the three transformations for the PLE (intercepts) 212
5.3 Scatterplot of Example 5.2.4 on the three transformations for the PLE (factor loadings) 213
5.4 Section 5.3: Example of bias of PLE 215
5.5 Scatter plot of PLE estimates for the simulated large data example of Sectiom 4.3 222

B.1 Simulation example Appendix B: Scatter plot for estimated bias reduction . 240
B.2 Simulation example Appendix B: True and estimated ICCs 241

List of Tables

1.1 Example 1.3.7: Ability separated data 46
1.2 Simulation study of Subsection 1.3.3.6: True parameter values 49
1.3 Simulation study of Subsection 1.3.3.6: ARMSE values 51
1.4 Simulation study of Subsection 1.3.3.6: maBias values 52
1.5 Simulation study of Subsection 1.3.3.6: mVar values 53

2.1 Simulation Example in Subsection 2.2.2: True and estimated parameter values 66
2.2 Simulation Example in Subsection 2.2.2: Goodness of fit 66
2.3 Simulation Example in Subsection 2.2.2: Data set and estimated expected frequencies 67
2.4 Real Data Example in Subsection 2.2.2: Goodness of fit 68
2.5 Real Data Example in Subsection 2.2.2: LSAT data set 68
2.6 Simulation study in Section 2.5: True parameter values 81
2.7 Simulation study in Subsection 2.5.1: Summary Statistics of Results Setup 1 82
2.8 Simulation study in Subsection 2.5.1: Summary Statistics of Results Setup 2 83
2.9 Simulation study in Subsection 2.5.2: Results of ICC estimation 95
2.10 Simulation study in Subsection 2.5.2: Results of latent trait estimation - Part I 96
2.11 Simulation study in Subsection 2.5.2: Results of latent trait estimation - Part II 97
2.12 Simulation study in Subsection 2.5.3: Results of Model Selection - Part I . . 100
2.13 Simulation study in Subsection 2.5.3: Results of Model Selection - Part II . 101

3.1 Example of Subsection 3.4.4 for SCT conditions: Example parameter values 117

5.1 Simulation studies of Chapter 5: True parameter values 207
5.2 Summary statistics for the time comparison of the PLE algorithms. 208
5.3 Results of the comparison of the PLE transformations 214
5.4 Simulation study of Section 5.3: Estimated correction factor and quality measures 218
5.5 Simulation study of Section 5.4: Quality measrures of PLE and MMLE . . 221
5.6 Simulation study of Section 5.4: Computational efficiency PLE vs MMLE . 222

Notation and Acronyms

The following notations are used throughout this thesis. In particular cases, we provide the page number in order to refer to the corresponding definition, explanation or first usage.

Notation	Explanation	Page
Symbols		
$\mathbb{R}$	set of real numbers	
$\mathbb{N}$	natural numbers (without 0)	
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$	
$\mathbb{1}_A$	indicator on set A	37
$\int \dots \mathrm{d}\boldsymbol{\eta}$	measure-integral with respect to the q-dimensional Lebesgue measure	
$\int \dots \mathscr{G}(\mathrm{d}\boldsymbol{\eta})$	measure-integral with respect to the measure $\mathscr{G}$	
o, $\mathcal{O}$	small and big Landau symbols	
o_{P}, $\mathcal{O}_{\mathsf{P}}$	small and big probabilistic Landau symbols with respect to the probability measure P	
$\lvert a \rvert$	absolute value of $a \in \mathbb{R}$	
$\lVert \bullet \rVert$	standard Euclidean norm for vectors in $\mathbb{R}^b$ and the corresponding operator norm for matrices in $\mathbb{R}^{a \times b}$, $a, b \in \mathbb{N}$, if not explicitly specified otherwise	
$B_\delta(\boldsymbol{a})$	$\{\boldsymbol{x} \colon \lVert \boldsymbol{x} - \boldsymbol{a} \rVert < \delta\}$	121
$\overline{B}_\delta(\boldsymbol{a})$	$\{\boldsymbol{x} \colon \lVert \boldsymbol{x} - \boldsymbol{a} \rVert \leq \delta\}$	
$\partial B_\delta(\boldsymbol{a})$	$\overline{B}_\delta(\boldsymbol{a}) \setminus B_\delta(\boldsymbol{a})$	
$\overline{M}$	(topological) closure of the set M	
∂M	(topological) boundary of the set M	
∇, $\nabla_{\boldsymbol{\vartheta}}$	nabla-operator, nabla operator with respect to the elements of $\boldsymbol{\vartheta}$	
$\nu_{\min}$, $\nu_{\max}$	smallest/largest eigenvalue	
$(a, b]$	interval from a to b where b is included and a not	
$(a, b]^2$	$(a, b] \times (a, b]$	
E	expectation functional	
Var	variance functional	
Cov	covariance operator	

(continued)

(continued)

Notation	Explanation	Page
$\mathsf{E}(\bullet \mid \boldsymbol{\eta})$, $\mathsf{E}_{\boldsymbol{\eta}}$, $\mathsf{E}(\bullet \mid \boldsymbol{\eta} = \boldsymbol{\eta}_0)$, $\mathsf{E}_{\boldsymbol{\eta}_0}$	conditional expectation given the σ-algebra which is induced by the random variable $\boldsymbol{\eta}$ or given the event $\boldsymbol{\eta} = \boldsymbol{\eta}_0$	
$\mathsf{P}_{\boldsymbol{\eta}}$, $\mathsf{P}(\bullet \mid \boldsymbol{\eta})$	conditional probability measure given $\boldsymbol{\eta}$	
$\mathsf{Var}_{\boldsymbol{\eta}}$, $\mathsf{Cov}_{\boldsymbol{\eta}}$	conditional (co-)variance given $\boldsymbol{\eta}$	
$\mathcal{B}$, $\mathcal{B}^q$	(q-dimensional) Borel-σ-algebra	
$C^k(M_1, M_2)$	k-times continuously differentiable mappings $M_1 \to M_2$	
d	usually: number of items	6
n	usually: number of responding subjects	6
q	usually: latent dimension	6
$(\Omega, \mathcal{A}, \mathsf{P})$	Probability space with sample space Ω, event space/σ-algebra $\mathcal{A}$ and probability measure P	6
$\intercal$	transpose	6
$\boldsymbol{Y}$	response vector with values in $\{0,1\}^d$	6
$\boldsymbol{y}$	a realization of $\boldsymbol{Y}$	6
$\lvert\boldsymbol{y}\rvert$	$\sum_{i=1}^d y_i$, the total score	37
$\boldsymbol{Y}^{(d)}$	$(Y_1, \ldots, Y_d)^\intercal$ for a sequence $\{Y_i\}_{i\in\mathbb{N}}$	105
$\boldsymbol{\eta}$	a latent vector, usually continuous	6
$\mathcal{H}$	structural model, distribution of $\boldsymbol{\eta}$	7
Θ	usually: $\operatorname{supp}(\mathcal{H}) \subseteq \mathbb{R}^q$, the support of $\mathcal{H}$	7
$\mathfrak{h}$	if $\mathcal{H} \ll \mathscr{L}^q$ for the q-dimensional Lebesgue-measure $\mathscr{L}^q$, then the $\mathscr{L}^q$-density of $\mathcal{H}$	7
$\mathcal{H}(\bullet \mid \boldsymbol{Y})$, $\mathcal{H}(\bullet \mid \boldsymbol{Y} = \boldsymbol{y})$	posterior distribution of $\boldsymbol{\eta}$ given $\boldsymbol{Y}$ (random measure as measure valued random element) and given the observed responses $\boldsymbol{Y} = \boldsymbol{y}$	8
$h(\bullet \mid \boldsymbol{y}) := \mathfrak{h}(\bullet \mid \boldsymbol{Y} = \boldsymbol{y})$, $h(\bullet \mid \boldsymbol{y}^{(d)}) := \mathfrak{h}(\bullet \mid \boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)})$	posterior pdf of $\boldsymbol{\eta}$ given $\boldsymbol{Y} = \boldsymbol{y}$ or $\boldsymbol{Y}^{(d)} = \boldsymbol{y}^{(d)}$	105
$h^*(\bullet \mid \boldsymbol{y})$	posterior pdf of $\mathcal{I}^{(d)}(\boldsymbol{\eta}_0)^{1/2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)$ given $\boldsymbol{Y} = \boldsymbol{y}$	142
$h(\bullet \mid \boldsymbol{Y})$, $h^*(\bullet \mid \boldsymbol{Y})$	$h(\bullet \mid \boldsymbol{y})\big\rvert_{\boldsymbol{y}=\boldsymbol{Y}}$, $h * (\bullet \mid \boldsymbol{y})\big\rvert_{\boldsymbol{y}=\boldsymbol{Y}}$	
iid, $\overset{\text{iid}}{\sim}$	independent and identically distributed	7
$\mathscr{B}(n,p)$	binomial distribution with $n \in \mathbb{N}$ responses and success probability $p \in (0,1)$	7
$\operatorname{Exp}(k)$	exponential distribution with expected value $k > 0$	
$\mathcal{U}(A)$	uniform distribution on the set A	

(continued)

(continued)

Notation	Explanation	Page
$\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	q-variate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^q$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$	13
Φ	cumulative distribution function of $\mathcal{N}_1(0,1) \equiv \mathcal{N}(0,1)$	9
$\tilde{\Phi}_q$	Gaussian measure on $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$	13
ϕ_q	probability density function of Φ_q	13
Φ_γ	standard skew normal cumulative distribution function with skewness parameter $\gamma \in \mathbb{R}$	13
$\boldsymbol{Y} \sim \mathcal{P}(\boldsymbol{\eta})$, $\{Y_i\}_{i \in \mathbb{N}} \sim \mathcal{P}(\boldsymbol{\eta})$	measurement model, $\{Y_i\}_{i \in \mathbb{I}}$, $\mathbb{I} = \{1, \ldots, d\}$ or $\mathbb{I} = \mathbb{N}$, conditionally independent given $\boldsymbol{\eta}$ with $P_i(\boldsymbol{\eta}) := \mathsf{P}(Y_i = 1 \mid \boldsymbol{\eta})$, $\boldsymbol{\eta} \in \Theta$, $i \in \mathbb{I}$	7, 105
X^2, G^2	Pearson and Likehood-ratio statistics	
$\mathfrak{g}$	generic link function $(0,1) \to \mathbb{R}$ (continuous and strictly increasing)	9
logit	$(0,1) \to \mathbb{R}$, $x \mapsto \log(x/(1-x))$	9
Φ^{-1}	probit link, quantile function of $\mathcal{N}(0,1)$	9
λ_i	$\mathrm{logit}(P_i(\bullet))$, item logit function of item i	18
ssp	standard skew-probit link function	13
$\mathcal{M}(\Theta_1)$, $\mathcal{M}(\Theta', \boldsymbol{\varphi})$, $\mathcal{M}(\Theta', \boldsymbol{\varphi}, \mathbb{N})$	EGLLVMs	59
$\boldsymbol{X}$	usually: Component vector $\boldsymbol{A}^\intercal \boldsymbol{Y}$ for a matrix of factor loadings $\boldsymbol{A} \in \mathbb{R}^{d \times q}$	173
Acronyms		
IRT	item response theory	1
MIRT	multidimensional IRT	1
LR	likelihood-ratio	2
AIC	Akaike Information Criterion	3
APN	asymptotic posterior normality	2
MLE	maximum likelihood estimator/estimation	3
PL	pseudo likelihood	4
GLLVM	generalized linear latent variable model	5
ICC	item characteristic curve, i.e. $\eta \mapsto P_i(\eta)$ for $q = 1$	7
CAT	computerized adaptive tests	8
pdf	probability density function	7
cdf	cumulative distribution function	9
GLM	generalized linear model	11

(continued)

(continued)

Notation	Explanation	Page
MGF	moment generating function	18
(A1)–(A2), (B1)–(B2)	conditions of manifest probabilities models	19,21,23
JMLE	joint MLE	31
CMLE	conditional MLE	32
MMLE	marginal MLE	33
EM	Expectation-Maximization	34
MC	Monte Carlo	35
MCEM	EM using MC-integration	35
MPMLE	manifest probabilities MLE	35
PLE	pseudo likelihood estimator	48
EAP	expected a-posteriori estimator	55
WLE	weighted likelihood estimator	54
MAP	maximum a-posteriori estimator	54
MeAP	median a-posteriori estimator	55
EGLLVM	extended GLLVM and structural model	57
GoF	goodness of fit	66
(PLM1)–(PLM7)	conditions for asymptotics of EGLLVM	70,74
LT	latent trait (latent variable)	103
CS	Chang and Stout (1993)	104
GGS	Ghosal et al. (1995)	104
SCT	semiproper centering theory	104
PCT	proper centering theory	104
(CS1)–(CS5)	conditions of CS-theory	107
(GGS1)–(GGS3), (GGS1')	conditions of GGS-theory	110, 111
(CS1')–(CS5')	conditions of semiproper centering theory, the generalization of CS-theory to $q > 1$	112
(PCT1)–(PCT3)	conditions of proper centering theory	138
(M), (M')	conditions for aEAP and aMCEM	173, 175
aEAP	linear approximated EAP for logit IRT models	177
aMCEM	posterior normal approximated MCEM-algorithm for MMLE	187
NIS	normal importance sampling	193

(continued)

(continued)

Notation	Explanation	Page
PA	Paek and Anderson (2017)	199
2SPLE	two-stage PLE algorithm	204

Bibliography

Adams, R. J., Wilson, M., and Wang, W.-c. (1997a). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, 21(1):1–23.

Adams, R. J., Wilson, M., and Wu, M. (1997b). Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, 22(1):47–76.

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scand J Statist*, 20:63–71.

Agresti, A. (2013). *Categorial Data Analysis*. Probabiliy and Statistics. John Wiley & Sons, Hoboken, NJ, third edition.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Probability and Statistics. John Wiley & Sons, New York, NY.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38(1):123–140.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42(1):69–81.

Anderson, C. J. (2013). Multidimensional item response theory models with collateral information as poisson regression models. *Journal of Classification*, 30(2):276–303.

Anderson, C. J., Li, Z., and Vermunt, J. K. (2007). Estimation of models in a Rasch family for polytomous items and multiple latent variables. *Journal of Statistical Software*, 20.

Anderson, C. J., Verkuilen, J., and Peyton, B. L. (2010). Modeling polytomous item responses using simultaneously estimated multinomial logistic regression models. *Journal of Educational and Behavioral Statistics*, 35(4):422–452.

Anderson, C. J. and Vermunt, J. K. (2000). Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, 30:81–121.

Anderson, C. J. and Yu, H.-T. (2007). Log-multiplicative association models as item response models. *Psychometrika*, 72(1):5–23.

Anderson, C. J. and Yu, H.-T. (2014). Log-multiplicative association models as multidimensional item response models for polytomous items. Working Draft.

Anderson, C. J. and Yu, H.-T. (2017). Properties of Second-Order Exponential Models as Multidimensional Response Models. In: van der Ark L.A., Wiberg M., Culpepper S., Douglas J.A., Wang WC. (eds). *Quantitative Psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics*, vol 196. Springer, Cham.

Antal, T. (2007). On the latent regression model of item response theory. *ETS Research Report Series*, 2007(1):i–19.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*. Springer, New York, NY.

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

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, Inc., New York, NY, second edition.

Bartholomew, D. J. (1998). Scaling unobservable constructs in social science. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):1–13.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, volume 904 of *Probabiliy and Statistics*. John Wiley & Sons, Chichester, third edition.

Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8.

Bazán, J. L., Bolfarine, H., and Branco, M. D. (2004). A new family of asymmetric models for item response theory: A skew-normal irt family. Technical Report RT-MAE-2004-17, Universidade de São Paulo, Instituto de Matemática e Estatística.

Bazán, J. L., Bolfarine, H., and Branco, M. D. (2010). A framework for skew-probit links in binary regression. *Communications in Statistics – Theory and Methods*, 39(4):678–697.

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

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

Bianconcini, S., Cagnone, S., Rizopoulos, D., et al. (2017). Approximate likelihood inference in generalized linear latent variable models based on the dimension-wise quadrature. *Electronic Journal of Statistics*, 11(2):4404–4423.

Birnbaum, A. (1968). Some latent trait models. In Lord, F. M. and Novick, M. R., editors, *Statistical Theories Of Mental Test Scores*, pages 397–424. Reading, MA: Addison-Wesley.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2):179–197.

Bock, R. D. and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4):431–444.

Bolfarine, H. and Bázán, J. L. (2010). Bayesian estimation of the logistic positive exponent irt model. *Journal of educational and behavioral statistics*, 35(6):693–713.

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93(441):262–272.

Bradley, R. A. and Gart, J. J. (1962). The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593.

Butler, S. M. and Louis, T. A. (1997). Consistency of maximum likelihood estimators in general random effect models for binary data. *The Annals of Statistics*, 25(1):351–377.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., and Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology*, 59(1):173–194.

Cao, J. and Stokes, S. L. (2008). Bayesian irt guessing models for partial guessing behaviors. *Psychometrika*, 73(2):209.

Chang, H.-H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika*, 61(3):445–463.

Chang, H.-H. and Stout, W. (1991). The asymptotic posterior normality of the latent trait in an IRT model. Technical Report ONR Research Report 91-4, Department of Statistics, University of Illinois at URbana-Champaign.

Chang, H.-H. and Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58(1):37–52.

Chen, Y., Li, X., and Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146.

Chen, Y.-H. (2006). Computationally efficient monte carlo em algorithms for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, 76(9):817–828.

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics*, pages 317–328.

Cressie, N. and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, 48(1):129–141.

Dembo, R. S. and Steihaug, T. (1982). Truncated newton algorithms for large-scale optimization. *Math. Programming*, 26:190–212.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063. http://dx.doi.org/10.1016/j.csda.2013.02.005.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, FL.

Epskamp, S., Maris, G. K. J., Waldorp, L. J., and Borsboom, D. (2018). Network psychometrics. *arXiv:1609.02818v2*.

Forster, O. (2012). *Analysis 3. Maß- und Integrationstheorie, Integralsätze im IRn und Anwendungen*. Vieweg+Teubner, Wiesbaden.

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

Fu, Z.-H., Tao, J., and Shi, N.-Z. (2009). Bayesian estimation in the multidimensional three-parameter logistic model. *Journal of Statistical Computation and Simulation*, 79(6):819–835.

Genter, F. C. and Farewell, V. T. (1985). Goodness-of-link testing in ordinal regression models. *Canadian Journal of Statistics*, 13(1):37–44.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2020). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-12.

Ghosal, S. (1997). Normal approximation to the posterior distribution for generalized linear models with many covariates. *Mathematical Methods of Statistics*, 6(3):332–348.

Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics*, 23(6):2145–2152.

Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Topics V*, pages 183–199. Springer, New York, NY.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5(5):815–841.

Hessen, D. J. (2012). Fitting and testing conditional multinormal partial credit models. *Psychometrika*, 77(4):693–709.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, 55(1):5–18.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society*, 66(4):893–908.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York, NY.

Joe, H. and Liu, Y. (1996). models for a multivariate binary response with covariates based on compatible coconditional specified logistic regressions. *Statistics and Probability Letters*, 31:113–120.

Johnson, S. G. (2019). The NLopt nonlinear-optimization package. http://github.com/stevengj/nlopt. Accessed: 2020-02-06.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, 56(2):255–278.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Statistics for Industry and Technology. Birkäuser/Springer, New York, NY.

Kateri, M. and Agresti, A. (2010). A generalized regression model for a binary response. *Statistics & Probability Letters*, 80(2):89–95.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Kim, J. K. and Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, 58(4):587–599.

Klauer, K. C. (1990). Asymptotic properties of the ML estimator of the ability parameter when item parameters are known. *Methodika*, 4:23–26.

Knott, M. and Albanese, M. (1993). Conditional distributions of a latent variable and scoring for binary data. *Brazilian Journal of Probability and Statistics*, pages 171–188.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 61(3):273–287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 62(1):74–82.

Lee, S. and Bolt, D. M. (2018). An alternative to the 3pl: using asymmetric item characteristic curves to address guessing effects. *Journal of Educational Measurement*, 55(1):90–111.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, NY, second edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer, New York, NY, third edition.

Li, Z. (2010). *Loglinear models as item response models*. PhD thesis, University of Illinois at Urbana-Champaign.

Liu, Y. (1994). A model for multivariate binary data with covariates based on compatible coconditional specified logistic regressions. Master's thesis, University of British Columbia.

Lord, F. M. (1952). *A Theory of Test Scores*. Number 7 in Psychometric Monographs. Psychometric Society, Richmond, VA. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Hillsdale, N.J.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48(2):233–245.

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

Luzardo, M. and Rodríguez, P. (2015). A Nonparametric Estimator of a Monotone Item Characteristic Curve. In: van der Ark L., Bolt D., Wang WC., Douglas J., Chow SM. (eds). *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics*, vol 140. Springer, Cham.

Lyness, J. (1983). When not to use an automatic quadrature routine. *SIAM Review*, 25(1):63–87.

MacKinnon, J. G. and Smith, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, 85:205–230.

Mair, P., Hatzinger, R., and Maier, M. J. (2019). *eRm: Extended Rasch Modeling*. R package version 1.0-0.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., Maas, H. v. d., and Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate behavioral research*, 53(1):15–35.

Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91(435):1254–1267.

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

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

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80(392):993–997.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65(3):391–411.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer, New York, NY.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing, Paris.

Ogasawara, H. (2012). Asymptotic expansions for the ability estimator in item response theory. *Computational Statistics*, 27:661–683.

Ogasawara, H. (2013). Asymptotic properties of the Bayes and pseudo Bayes estimators of ability in item response theory. *Journal of Multivariate Analysis*, 114:359–377.

Paek, Y. (2016). *Pseudo-Likelihood Estimation of Multidimensional Polytomous Item Response Theory Models*. PhD thesis, University of Illinois at Urbana-Champaign.

Paek, Y. and Anderson, C. J. (2017). Pseudo-likelihood Estimation of Multidimensional Response Models: Polytomous and Dichotomous Items. In: van der Ark L.A., Wiberg M., Culpepper S., Douglas J.A., Wang WC. (eds). *Quantitative Psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics*, vol 196. Springer, Cham.

Pashley, P. J. (1991). An alternative three-parameter logistic item response model. *ETS Research Report Series*, 1991(1):i–35.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen. Fore- and Afterword of D. B. Wright, 1980.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17.

Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3):491–501.

Rizopoulos, D. and Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61(2):415–438.

Robitzsch, A. (2020). *sirt: Supplementary Item Response Theory Models*. R package version 3.9-4.

Rose, N., von Davier, M., and Nagengast, B. (2017). Modeling omitted and not-reached items in irt models. *psychometrika*, 82(3):795–819.

Rüschendorf, L. (2014). *Mathematische Statistik*. Springer, Berlin, Heidelberg.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Number 17 in Psychometric Monographs. Psychometric Society, Richmond, VA. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 42(2):163–191.

Samejima, F. (1980). Is bayesian estimation proper for estimating the individual's ability? Technical Report 80-3, Tennessee University Knoxville. Department of Psychology.

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

Samejima, F. (1999). Usefulness of the logistic positive exponent family of models in educational measurement. Paper presented at the 1999 Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

San Martín, E., Del Pino, G., and De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, 30(3):183–203.

Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York, NY.

Schilling, S. and Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70(3):533–555.

Sinharay, S. (2015). The asymptotic distribution of ability estimats: beyond dichotomous items and unidimensional IRT models. *Journal of Educational and Behavioral Statistics*, 40(5):511–528.

Swihart, B. and Lindsey, J. (2019). *rmutil: Utilities for Nonlinear Regression and Repeated Measurements Models*. R package version 1.1.3.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative poisson model. *Scand J Satist*, 9:23–30.

Tsutakawa, R. K. and Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics*, 13(2):117–130.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Venturelli, J. (2015). *A Bayesian Skew Mixture Item Response Model*. PhD thesis, Universidade Federal de Minas Gerais.

von Davier, M. (2009). Is There Need for the 3PL Model? Guess What? *Measurement: Interdisciplinary Research and Perspectives*, 7(2):110–114.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(1):80–88.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, 80(2):428–449.

Wang, C. and Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika*, 76(3):363–384.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3):427–450.

Witting, H. and Müller-Funk, U. (1995). *Mathematische Statistik II*. Teubner, Stuttgart.

Zhang, J. and Stout, W. (1997). On Holland's Dutch identity conjecture. *Psychometrika*, 62(3):375–392.