Title
Multidimensional Item Factor Analysis with Semi-Nonparametric Latent Densities

Author
Monroe, Scott Lee

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Multidimensional Item Factor Analysis With Semi-Nonparametric Latent Densities

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Scott Lee Monroe

2014

Abstract of the Dissertation

Multidimensional Item Factor Analysis With Semi-Nonparametric Latent Densities

by

Scott Lee Monroe

Doctor of Philosophy in Education

University of California, Los Angeles, 2014

Professor Li Cai, Chair

Woods and Lin (2009) proposed a unidimensional item response theory (IRT) model where the distribution of the latent variables is estimated using a semi-nonparametric (SNP, Gallant & Nychka, 1987) density. Estimation of the latent variable density can reduce bias in parameter estimates that results from misspecifying the form of the density (Woods & Thissen, 2006; Woods & Lin, 2009). However, application of the Woods and Lin (2009) model is restricted to the unidimensional setting. To address this limitation, the present research generalizes the Woods and Lin (2009) model to multidimensional IRT (MIRT). The resulting model, the SNP-MIRT model, may also be considered a generalization of the "standard" MIRT model, which specifies a normal density for the latent variables.

A secondary focus of this research concerns a new proposal for calculating student growth percentiles (SGP, Betebenner, 2009). In Betebenner (2009), quantile regression (QR, Koenker & Bassett, 1978; Koenker, 2005) is used to estimate the SGPs. However, a shortcoming of the original methodology is that measurement error in the score estimates, which always exists in practice, leads to bias in the SGP estimates (Shang, 2012). One way to address this issue is to estimate the SGPs using a modeling framework that can directly account for the measurement error. MIRT is one such framework, and the one utilized here. To maximize the

generality of the approach, as well as guard against misspecification of the latent variable density, the SNP-MIRT model is used. SNP-MIRT estimates, in turn, are used with the calibrated projection linking methodology (Thissen, Varni, et al., 2011; Thissen, Liu, Magnus, & Quinn, 2014; Cai, in press-a, in press-b) to produce SGP estimates.

Preliminary simulation studies are conducted to investigate the fidelity of the SNP-MIRT model and SGP estimation implementations. The simulation study for the SNP-MIRT model focuses on recovery of the shape of the data-generating latent variable density. The simulation study for the proposed SGP method focuses on comparing the accuracy of the QR, standard MIRT, and SNP-MIRT approaches. Finally, empirical applications are provided to illustrate the new methods.

The dissertation of Scott Lee Monroe is approved.

Ying Nian Wu

Mike Seltzer

Noreen Webb

Li Cai, Committee Chair

University of California, Los Angeles

2014

to Carolyn

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am indebted to my advisor and committee-chair, Professor Li Cai, for his support, guidance, and scholarship. This work would not have been possible without him. I am also grateful for the feedback and advice from the other members of my committee: my advisor Professor Noreen Webb, Professor Mike Seltzer and Professor Ying Nian Wu.

I thank Professors Carol Woods and David Thissen. Their research efforts have inspired this work.

I also thank past and current colleagues from the psychometrics lab at UCLA: Professors Taehun Li and Ji Seung Yang; Drs. Mark Hansen and Carl Falk; and Larry Thomas, Moonsoo Lee, Megan Kuhfeld, and Zhen Li.

Finally, I thank Carolyn, for her support.

Vita

EDUCATION

2000	Bachelor of Arts, University of California, San Diego. Political Science.
2004	Juris Doctor, University of California, Los Angeles.
2007	Masters of Science, Brooklyn College School of Education. Mathematics Education.
2013	Masters of Science, University of California, Los Angeles. Statistics.

WORK

2000-2001	Instructor, The Princeton Review, San Diego, CA.
2005-2007	Mathematics Teacher, Bedford Academy High School, Brooklyn, New York.
2007-2009	Mathematics Teacher, Culver City High School, Culver City, California.
2010-Present	Graduate Student Researcher, University of California, Los Angeles.

Publications

Herman, J., Wang J., Rickles, J., Hsu, V., Monroe, S., Leon, S., & Straubhaar,

R. (2012). *Evaluation of Green Dot's Locke transformation project: from the perspective of teachers and administrators.* CRESST Report No. 815. Los Angeles, CA: UCLA.

Cai, L., & Monroe, S. (2013). IRT model fit evaluation from theory to practice: progress and some unanswered questions. *Measurement: Interdisciplinary Research and Perspectives, 11*, 102-106.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-Curve item response theory model by the Metropolis-Hastings Robbins-Monro Algorithm. *Educational and Psychological Measurement, 74*, 343-369.

Cai, L. & Monroe, S., (2014). *A new test statistic for item response theory models for ordinal data.* CRESST Report No. 839. Los Angeles, CA: UCLA.

Monroe, S., & Cai, L. (2014). Abstract: Automated fitting of MIRT models by a simultaneous perturbation algorithm. *Multivariate Behavioral Research, 49*, 291.

Hansen M., Cai. L., Monroe, S., & Li, Z. (2014). *Limited-information goodness-of-fit testing of diagnostic classification item response theory models.* CRESST Report No. 840. Los Angeles, CA: UCLA.

CHAPTER 1

Introduction

Woods and Lin (2009) proposed a unidimensional item response theory (IRT) model where the distribution of the latent variables is estimated using the semi-nonparametric (SNP) density of Gallant and Nychka (1987). The present research is primarily concerned with the generalization of the Woods and Lin (2009) model to a multidimensional IRT (MIRT) framework. Hereafter, the proposed model will be referred to as the SNP-MIRT model. In particular, this research concerns full-information maximum likelihood (FIML) estimation of the SNP-MIRT model, along with likelihood inference more generally. While the SNP density has been successfully used in other modeling contexts, the exisiting parameterizations are not well-suited for a general confirmatory MIRT model implementation. Thus, this research proposes a new parameterization that facilitates customary MIRT model specification. As such, the SNP-MIRT model may also be considered a generalization of the standard MIRT model which specifies a normal distribution for the latent variable density.

Due to the potentially high-dimensional latent variables, the Metropolis-Hastings Robbins-Monro (MH-RM, Cai, 2010a) algorithm is adopted for estimation. Use of the MH-RM algorithm also facilitates the incorporation of various methods for inference related to the Metropolis-Hastings (M-H, Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) algorithm. To that end, this research also extends the work of Chib and Jeliazkov (2001) to propose a new estimator for the information matrix. Like MH-RM, this estimator eschews quadrature in favor

of Monte Carlo methods. And, since this estimator depends on output from the M-H algorithm, it naturally complements the MH-RM algorithm.

A secondary focus of this research concerns a new proposal for estimating student growth percentiles (SGP, Betebenner, 2009). In the original methodology, quantile regression (QR, Koenker & Bassett, 1978; Koenker, 2005) is used to regress the current year score estimates on previous years' score estimates. In this research, an alternative method is proposed for estimating SGPs, where all data are modeled jointly using MIRT, or its generalization, SNP-MIRT. Specifically, the new proposal capitalizes on the *calibrated projection linking* (Thissen, Varni, et al., 2011; Thissen et al., 2014; Cai, in press-a, in press-b) methodology to produce the SGP estimates.

This research, though technical in nature, is ultimately motivated by expanding the range of questions that may be investigated by social science researchers. Regarding the SNP-MIRT model, researchers may have theory-driven ideas about the form of the latent variable distribution. Theory may hold that this distribution is, for example, normal, skewed, or multimodal. On the other hand, researchers may be unwilling to claim any prior knowledge as to the distributional form. In either case, the shape of the distribution may itself be a point of substantive interest. Regarding the new framework for estimation of SGPs, educational researchers and policymakers likely have numerous questions about the merits and shortcomings of the original SGP methodology (Betebenner, 2009). The framework presented here provides a versatile tool to investgate many of those questions. This issue is further discussed in Chapter 2.

The remainder of this chapter reviews methodological topics particularly relevant to this research. These topics are: modeling latent densities; FIML estimation of MIRT models; and the original QR-based approach to SGPs.

1.1 Modeling Latent Densities

Across numerous modeling frameworks, methodologists have used a variety of densities for latent variables. This section reviews these efforts for generalized linear mixed models (GLMMs), generalized linear latent variable models (GLLVMs), and IRT models. In this review, particular attention is devoted to the use of the SNP density. The technical details of the density, however, are presented in Chapter 4.

An appealing feature of the SNP density is that it is quite flexible. It can approximate a wide range of densities, including skewed densities and those with multiple modes. Figure 1.1 gives an example of an SNP density. Both marginals are clearly nonnormal, as is the joint distribution. However, the density in Figure 1.1 is standardized, with null mean vector and identity covariance matrix. Thus, even when the SNP density is standardized, it can take on a variety of shapes.

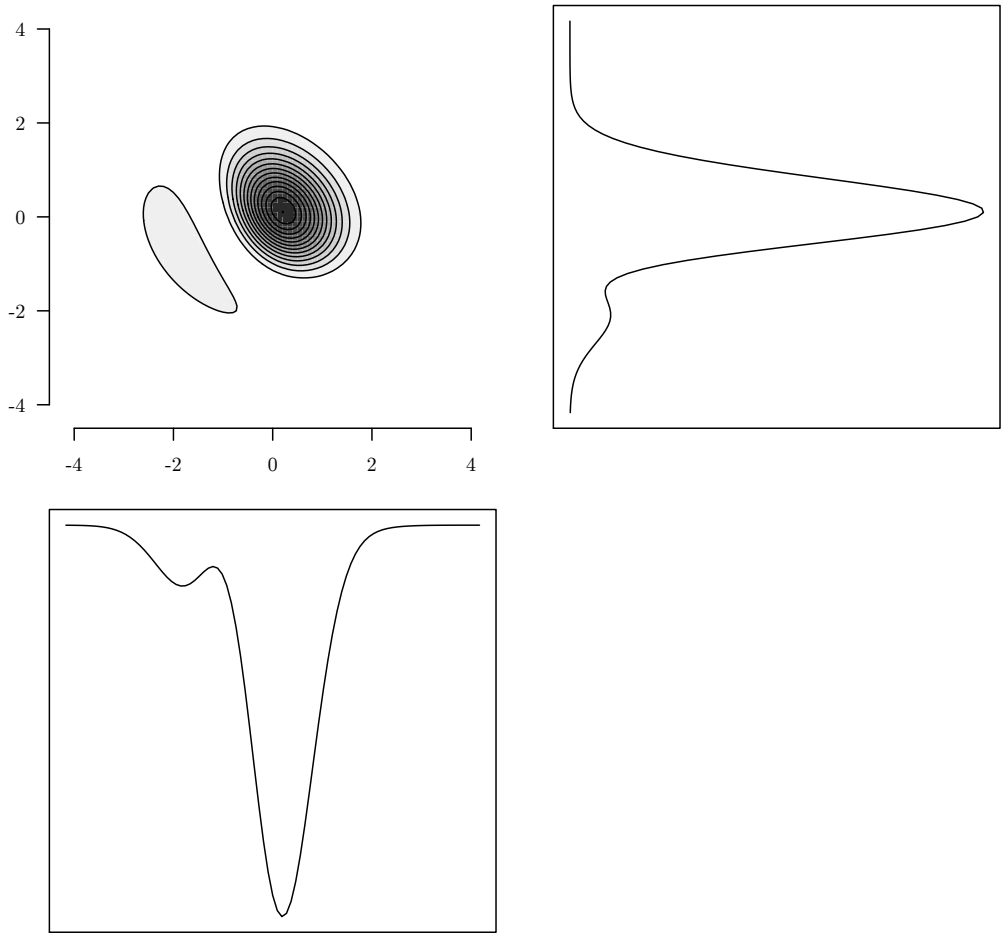
1.1.1 Generalized Linear Mixed Models

Within the GLMM framework, the random effects in the model can be regarded as latent variables. Thus, estimating the distribution of the random effects for GLMMs is analogous to estimating the distribution of the latent variables in IRT. Typically, the random effects density is specified as normal, but researchers have proposed several alternative approaches.

One of these alternatives utilizes the SNP density (e.g., Zhang & Davidian, 2001; Chen, Zhang, & Davidian, 2002; Papageorgiou & Hinde, 2012; Vock, Davidian, & Tsiatis, 2012, 2014). Two notable implementations are given by Zhang and Davidian (2001) and Chen et al. (2002). Zhang and Davidian (2001) is important for a reparameterization that stabilized estimation, while Chen et al. (2002) is noteworthy for its method of estimation (Monte Carlo EM).

Other alternatives include Gaussian mixtures (Verbeke & Lesaffre, 1996; Magder

Figure 1.1: Contour of Bivariate SNP Denisty and Marginal Densities



Note. The SNP density above is standardized, with null mean vector and identity covariance matrix.

& Zeger, 1996), skew normal mixtures (Arellano-Valle, Bolfarine, & Lachos, 2005; Lin & Lee, 2008), and skew ellipticals (Jara, Quintana, & San Martin, 2008). Skew normal mixture densities have also been used in IRT (e.g., Azevedo, Bolfarine, & Andrade, 2011).

With so many competing models, it is natural to wonder how they compare to one another. With this motivation, Ghidey, Lesaffre, and Verbeke (2008) compared several methods through simulation study. The research concluded that no single methodology was best across all conditions, including heavily-skewed and bimodal distributions. Compared to other methods, the SNP approach was found to perform very well in recovering bimodal distributions, and less well in recovering heavily-skewed distributions.

1.1.2 Generalized Linear Latent Variable Models

Density estimation has been incorporated into the estimation of a number of latent variable models besides IRT. Notably, in factor analysis, a Gaussian mixture (Montanari & Viroli, 2010b) and a skew normal approach (Montanari & Viroli, 2010a) have been taken. And recently, Irincheeva, Cantoni, and Genton (2012) proposed a GLLVM using the SNP method for density estimation. Notably, these approaches all adopt the EM algorithm (Dempster, Laird, & Rubin, 1977) or a general Newton-type optimizer for estimation. As will be discussed, these choices are largely impractical for FIML estimation of MIRT.

1.1.3 Item Response Theory Models

Bock and Aitkin (1981) is best known for its implementation of EM to obtain maximum marginal likelihood estimates for item parameters, as the approach is still widely considered the gold-standard for parameter estimation for many IRT models. However, the research also presented the Empirical Histogram (EH)

method (see also Mislevy, 1984) for characterizing a latent variable distribution as a collapsing of all posteriors across the E-step tables. More recently, Knott and Tzamourani (2007) proposed a generalization of the EH method, where the locations of the quadrature points need not be fixed.

Two alternatives to the EH method have been advanced by Carol Woods and colleagues. One of these methods is Ramsay-Curve IRT (RC-IRT, Woods & Thissen, 2006; Woods, 2007, 2008; Monroe & Cai, 2014), which has been implemented for unidimensional models. In its original implementation (Woods & Thissen, 2006), the RC method is incorporated into an EM scheme. Basically, the estimates of the Ramsay-Curve parameters are updated during the Maximization step (M-step) by regressing the proportions given by the EH method onto a basis spline. In a precursor to the current research, Monroe and Cai (2014) implemented an RC-IRT model using the MH-RM algorithm.

The other method is SNP-IRT (Woods & Lin, 2009), originally termed "Davidian Curve IRT," in honor of Marie Davidian, who has been instrumental in developing the SNP methodology.

1.2 FIML Estimation of MIRT Models

In theory, FIML estimation for MIRT can be achieved by the EM algorithm, as in Bock and Aitkin (1981). In other words, there is no deficiency in the theoretical approach of the EM algorithm for estimating parameters of MIRT models. Instead, the difficulty lies in the computational burden of EM for MIRT models, which grows exponentially with the dimensionality of the model. This is because in the Expectation step (E-step), posterior distributions (of response patterns) are evaluated at some specified number of quadrature points per dimension. The use of 49 points for one dimension implies $49^3 > 100,000$ points for three dimensions. In the literature, this phenomenon is often referred to as the "curse of dimension-

ality." Moreover, functions of each response pattern must be evaluated at every point, at every iteration.

1.2.1 Addressing the "Curse of Dimensionality"

The curse of dimensionality has motivated many research efforts designed to avoid the sort of computational burden that attends use of the EM algorithm for high-dimensional MIRT models. Cai (2010a) provides a taxonomy of the different methods and the relationship of the MH-RM algorithm to these various methods. A similarly comprehensive review is not presented here. However, two methods will be briefly described as they are particularly relevant to the SNP-MIRT research. The first of these is Monte Carlo EM (MCEM, Wei & Tanner, 1990). The second is Stochastic EM (SEM, Diebolt & Ip, 1996).

MCEM is of special interest here because it is used to estimate the SNP-GLMM model in Chen et al. (2002). In MCEM, the integration in the E-step is accomplished by Monte Carlo methods as opposed to, for instance, the use of quadrature via E-step tables. This is attractive because, in principle, the Monte Carlo sample size can be relatively small. However, as the likelihood approaches its maximum, the Monte Carlo sample size must also increase so that the Monte Carlo error does not dominate the M-step update (Booth & Hobert, 1999). To the extent that the Monte Carlo sample size must approach the E-step table size to achieve convergence, MCEM remains plagued by the curse of dimensionality.

Unlike MCEM, SEM does not produce ML estimates. However, it is described here because it is related to both MCEM and MH-RM. SEM can be understood as MCEM with a fixed Monte Carlo sample size of 1. As mentioned above, since the sample size does not increase across iterations, the likelihood will not converge to its maximum. However, the sequence of parameter estimates can yield a sample from a stationary distribution where the mean is close to the ML estimate. This

characteristic of the sequence of SEM estimates is itself useful, and MH-RM adopts a similar approach to obtain refined starting values (discussed in Chapter 5).

1.2.2 The Metropolis-Hastings Robbins-Monro Algorithm

Like SEM and MCEM, the MH-RM algorithm (Cai, 2010a, 2010b) iterates between two steps. However, unlike the methods just discussed, there is no maximization step. In this sense, MH-RM is not a variant of the EM algorithm. It is similar, though, to both SEM and MCEM in that Monte Carlo samples are drawn from the posterior predictive distribution of missing data. In other words, latent variable draws "fill-in" the missing data. This step is accomplished via an M-H sampler (Metropolis et al., 1953; Hastings, 1970), and constitutes the "MH" part of the algorithm. In the "RM" part, the Robbins-Monro method (RM, Robbins & Monro, 1951) is used to slowly filter out the noise introduced by the MH samples across the sequence of iterations. Thus, the RM filter obviates the need to increase the Monte Carlo sample size in order to ensure the sequence of estimates converges to the ML solution. To the contrary, in practice, MH-RM often proves stable with sample size 1.

Technical details of the MH-RM algorithm are presented in Chapter 5. It should be noted, though, that the algorithm has been successfully applied to exploratory factor analysis (EFA, Cai, 2010a) and confirmatory factor analysis (CFA, Cai, 2010b), and is available in commercial IRT software (Cai, 2013; Cai, Thissen, & du Toit, 2011). Further, it has been used to estimate a unidimesional RC-IRT model (Monroe & Cai, 2014). Consequently, it is an attractive method to use in estimating the parameters of the SNP-MIRT model.

1.3 The QR-Based Approach to SGPs

The SGP methodology (Betebenner, 2009) is used to locate a student's current score in a conditional distribution based on the student's past score(s). As such, SGPs may be considered conditional status percentiles (Castellano & Ho, 2013). Instead of focusing solely on current achievement, SGPs provide context for that achievement. For example, suppose a student's current achievement is categorized as "below basic." By itself, this evaluation may be considered disappointing. However, if the accompanying SGP is 90, there is reason for encouragement: the interpretation is that this student's current achievement is higher than 90% of students who share the same score history. In this way, SGPs can add to our understanding of how well students are doing, and how they are progressing. Consequently, the SGP methodology has grown in popularity, and is used in numerous states to describe student performance. Moreover, the measure can be aggregated in an effort to describe teacher or school performance. In this latter case, the desired inference is that higher aggregate SGPs indicate higher effectiveness.

For the QR approach to SGPs, estimation of the conditional distribution uses the current year score as an outcome variable and previous years' scores as covariates. The approach uses a B-spline parameterization, as opposed to a linear parameterization, for the conditional quantile functions (Betebenner, 2009). This choice helps to account for non-linearity, skewness, and heteroskedasticity in the data. Just as with linear regression analysis, however, measurement error in the scores introduces bias into the regression parameter estimates, and ultimately the SGP estimates (Shang, 2012). Within the context of QR using a linear parameterization, Shang (2012) applied simulation extrapolation (SIMEX) in an effort to correct the regression parameter bias caused by measurement error in the covariates (i.e., past years' scores).

In this research, the proposed method for SGP estimation accounts for the correlated measurement error in all scores (i.e., past and current years). This method is presented and studied in Chapter 10.

CHAPTER 2

Relevance of Topics to Educational Policy and Practice

This chapter discusses how the methodological topics explored in this research are relevant to education policy and practice. The proposal of the SNP-MIRT model is broadly relevant to any education policy that utilizes IRT modeling in some form. On the other hand, the proposal of the SGP estimation framework is directly relevant to evaluating states' current SGP policies and may help to shape future practice in measuring student growth.

2.1 Relevance of SNP-MIRT Model

Prior research on density estimation in IRT has found that misspecification of the latent variable density can lead to bias in the parameter estimates (Yamamoto & Muraki, 1991; Woods & Thissen, 2006; Woods & Lin, 2009). This phenomenon certainly extends to the multidimensional setting to some extent, and is one of the motivations for the current research. This potential bias is important because in educational testing settings, parameter estimates are used to calculate a multitude of quantities.

For instance, individual latent variable estimates and conditional standard errors of measurement both depend on parameter estimates. Consequently, classification decisions also depend on these estimates. Further, when individual latent variable estimates are aggregated to either the teacher or school level, these ag-

gregated estimates must necessarily depend on parameter estimates. Since the SNP-MIRT model can yield less biased parameter estimates, the method has the potential to improve the validity of inferences that results from any educational test.

As a concrete example to help illustrate the relevance of the SNP-MIRT model to educational practice, consider the assessment being developed by Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC) (see, e.g., Herman & Linn, 2013). The stated goal of the assessment is to measure progress towards college- and career-readiness. While the assessment includes English Language Arts/literacy, this example will focus on mathematics.

For mathematics, the assessment aims to report achievement level descriptors (ALD), aligned with the Common Core State Standards, for each of four Content Claims. These Content Claims are: Concepts and Procedures; Problem-Solving; Communicating Reasoning; and Modeling and Data Analysis. For each of these (related) Content Claims, students will receive a Content ALD, rating the student from 1-4. The relevance of the SNP-MIRT to this assessment can be appreciated by considering the intended structure of the data and the sorts of inferences that Smarter Balanced and PARCC will want to make.

First, the test as designed will likely be multidimensional. In the absence of pilot test data, it is not possible to state whether there will be four factors, as suggested by the number of Content Claims. However, an exploratory item factor model is one tool that could be used to determine the number of empirical factors. More generally, it could be used to provide validity evidence for the claim that the assessment structure conforms to design specifications, including whether subscores will represent designed subdomains. Also, some individual items may involve multiple factors. For instance, temporarily assuming the existence of four factors aligned with the Content Claims, items that involve Communicating Rea-

soning could conceivably involve other Claims. This is because the "reasoning" involved will be due to knowledge of concepts, or maybe problem-solving. In any event, an exploratory item factor model could be used to analyze the dimensionality of individual items, and the pattern of loadings overall.

Second, it may be desirable to conduct differential item functioning (DIF) analyses to compare ethnic groups, or to compare native speakers with English Language Learners (ELL). And for some DIF methods, a multiple group model is necessary. For instance, DIF can be assessed using likelihood ratio tests that compare nested two-group IRT models, as in Thissen, Steinberg, and Wainer (1993). For this method, one group is designated the "reference" group, with fixed mean and variance. In relation to the reference group, the means and variances for the other groups are then be estimated. This sort of DIF analysis could be conducted with the SNP-MIRT model, as it is possible to freely estimate density means and variances.

Third, depending on the constitution of the groups, it may be unreasonable to assume normality of the latent variables. For example, consider a two-group comparison math achievement level between native speakers and ELLs. Further, assume that the native speakers are designated as the reference group. It is at least plausible that the latent variable density for the native speakers is normal. For the ELLs, however, such an assumption would be less plausible. The different stages of language acquisition likely interact with the math Content Claims in unpredictable ways, making an assumption of normality less tenable. Again, the SNP-MIRT model could address this concern.

2.2 Relevance of SGP Estimation Framework

The proposed MIRT-based SGP estimation framework has the potential to significantly influence education policy and practice in two ways. First, it may be

used to better study properties of the existing QR-based methodology. Second, it could potentially be adopted instead of the existing QR-based methodology.

Without an IRT-based SGP estimation framework, it is exceedingly difficult to study the QR-based methodology under realistic conditions. This is because the predominant method of generating realistic item level data in simulation studies in educational research is via an IRT model. Yet, without an IRT-based SGP estimation framework, there is no way to calculate the "true" SGPs of the simulees. And without the "true" SGPs, it is difficult to evaluate the performance of the QR-based SGPs for such simulated data.

However, the proposed MIRT-based framework solves this problem. Data may be simulated from a relatively realistic MIRT model and the "true" model-implied SGPs may be computed. These, in turn, may be used to assess the performance of the QR-based SGPs. The implication is that educational researchers now have a more direct method to pursue research topics concerning QR-based SGPs.

Of course, the proposed MIRT-based SGP framework also makes it possible to produce MIRT-based SGP estimates. These estimates may be compared to both "true" model-implied SGPs and QR-based estimates. It is at least conceivable that the MIRT-based estimates are better, in some sense, than the SGP counterparts, which could lead to changes in future policy or practice.

CHAPTER 3

A Multidimensional Item Response Theory Model with Multiple Groups

This chapter introduces the confirmatory multiple group MIRT model used in this research, as well as various results needed for estimation by the MH-RM algorithm. First, this chapter presents how user-defined constraints and multiple groups may be accomodated. Then, the models for the items, multidimensional generalizations of Samejima's (1969) graded response model and Birnbaum's (1968) three-parameter logistic (3PL) model, are introduced. This is followed by a brief introduction of the models for the densities, which are the multivariate normal density and the SNP density. The details of the SNP density are given in Chapter 4. Finally, this chapter presents the observed and complete data log-likelihood equations, as these are needed for the MH-RM algorithm.

3.1 Some Notation

Let there be $g = 1, \ldots, G$ groups, with $i = 1, \ldots, N_g$ respondents in group g. The total sample size is $N = \sum_{g=1}^{G} N_g$. Also, for group g, let there be $j = 1, \ldots, n_g$ items. For the ith respondent in this group, let y_{ijg} be the observed response to the jth item. The n_g responses for respondent i may then be collected in the $n_g \times 1$ vector $\mathbf{y}_{ig}$. For group g, let $\mathbf{Y}_g$ be the $N_g \times n_g$ matrix with ith row $\mathbf{y}'_{ig}$. Finally, all of the observed responses for the full sample may be collected in $\mathbf{Y} = \{\mathbf{Y}_g\}_{g=1}^{G}$.

For group g, let there be p_g latent variables. Then, for the ith respondent, $\boldsymbol{\theta}_{ig}$

is a $p_g \times 1$ vector of latent variable scores. As with the observed responses, the latent variable scores for group g may be collected in a matrix. Let $\boldsymbol{\Theta}_g$ be this matrix, with ith row $\boldsymbol{\theta}'_{ig}$. Then, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_g\}_{g=1}^{G}$ collects all latent variable scores for the full sample.

Moving on to notation for the parameters of the model, let $\boldsymbol{\zeta}$ be a $d_1 \times 1$ vector of independent item parameters and let $\boldsymbol{\xi}$ be a $d_2 \times 1$ vector of independent group or density parameters. Then, $\boldsymbol{\omega} = (\boldsymbol{\zeta}', \boldsymbol{\xi}')'$ is a $d \times 1$ vector of all freely estimated parameters, where $d = d_1 + d_2$. With this general structure, individual item and group parameters may be expressed as functions of $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$, respectively.

3.2 Item Response Models, Latent Density Models, and Constraints

Generically, let π_{ijk} be the probability that the observed response y_{ij} takes the score k, $k \in \{0, \ldots, K_j - 1\}$. In this research, the K_j responses for item j are ordered, but the modeling framework may accommodate unordered responses (e.g., Thissen, Cai, & Bock, 2011). Here, π_{ijk} is expressed as a function of $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}_i$ using one of two item response models.

The first is a a multidimensional version of Samejima's (1969) graded response model. Let $\boldsymbol{\alpha}(\boldsymbol{\zeta})$ be a $(K_j - 1) \times 1$ vector of item intercepts and let $\boldsymbol{\beta}(\boldsymbol{\zeta})$ be a $p \times 1$ vector of item slopes, where the parentheses emphasize the functional dependence of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ on $\boldsymbol{\zeta}$. The nature of this functional dependence will be discussed below.

Returning to the graded response model, π_{ijk} is defined as

$$\pi_{ijk} = T(y_{ij} = k | \boldsymbol{\alpha}(\boldsymbol{\zeta}), \boldsymbol{\beta}(\boldsymbol{\zeta}), \boldsymbol{\theta}_i) = T_j^+(k) - T_j^+(k+1), \tag{3.1}$$

where $T_j^+(k)$ is a multidimensional version of the familiar two-parameter logistic

model,

$$T_j^+(k) = \frac{1}{1 + \exp(-(\alpha_{jk}(\boldsymbol{\zeta}) + \boldsymbol{\beta}_j(\boldsymbol{\zeta})'\boldsymbol{\theta}_i))}. \tag{3.2}$$

That is, $T_j^+(k)$ is the conditional probability that a response is in category k or higher. Also, define $T^+(0) = 1$ and $T^+(K_j) = 0$ for all j, so that given $\boldsymbol{\theta}_i$, the sum of the conditional probabilities is 1.

The second model is a multidimensional version of the 3PL model. Since the 3PL model is defined for dichotomous responses, $\alpha(\boldsymbol{\zeta})$ is a scalar. Also, let $c(\boldsymbol{\zeta})$ be the guessing parameter. Then, conditional on $\boldsymbol{\theta}_i$, the conditional probability of observing $y_{ij} = 1$ is

$$\begin{aligned} T_j^*(1) &= T(y_{ij} = 1 | c(\boldsymbol{\zeta}), \alpha(\boldsymbol{\zeta}), \boldsymbol{\beta}(\boldsymbol{\zeta}), \boldsymbol{\theta}_i) \\ &= c(\boldsymbol{\zeta}) + \frac{1 - c(\boldsymbol{\zeta})}{1 + \exp(-(\alpha_j(\boldsymbol{\zeta}) + \boldsymbol{\beta}_j(\boldsymbol{\zeta})'\boldsymbol{\theta}_i))}, \end{aligned} \tag{3.3}$$

and the probability of observing $y_{ij} = 0$ is simply

$$T_j^*(0) = T(y_{ij} = 0 | c(\boldsymbol{\zeta}), \alpha(\boldsymbol{\zeta}), \boldsymbol{\beta}(\boldsymbol{\zeta}), \boldsymbol{\theta}_i) = 1 - T_j^*(1). \tag{3.4}$$

Then, for the multidimensional version of the 3PL, the probability of an item response is

$$\pi_{ijk} = T_j^*(1)^{y_{ij}} T_j^*(0)^{1 - y_{ij}}. \tag{3.5}$$

Moving on to the latent density models, in general, let $\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})$ represent a density function for $\boldsymbol{\theta}$ depending on $\boldsymbol{\xi}$. If the density is specified as multivariate normal, it is parameterized by the $p \times 1$ mean vector, $\boldsymbol{\mu}(\boldsymbol{\xi})$, and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\xi})$. In this case, the function will be written as $\phi(\boldsymbol{\theta}; \boldsymbol{\mu}(\boldsymbol{\xi}), \boldsymbol{\sigma}(\boldsymbol{\xi}))$, where $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$ is a $p^* \times 1$ vector of the unique elements of $\boldsymbol{\Sigma}$, $p^* = p(p+1)/2$, and vech$(\cdot)$ returns elements on or below the main diagonal.

Alternatively, the latent denstiy may be modeled by the SNP density function,

in which case it will be written as $h(\boldsymbol{\theta}; \boldsymbol{\psi}(\boldsymbol{\xi}))$, where $\boldsymbol{\psi}$ is a vector of parameters. The dimensionality of $\boldsymbol{\psi}$ depends on both p and a user-defined tuning contant κ. As will be explained in Chapter 4, this research proposes a new parameterization for the SNP density where one set of the parameters defines the "shape" of the density, and a disjoint set defines the location and scale.

Following Cai (2010b), the notation for the dependence of the parameters on $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ is indicated by function notation to elucidate how linear constraints may be imposed to yield a confirmatory model. As an example, suppose $p = 2$ and a bivariate normal density is chosen to model $\boldsymbol{\theta}$. Further, suppose that the only free parameters are the elements of $\boldsymbol{\mu}(\boldsymbol{\xi})$, while $\boldsymbol{\sigma}(\boldsymbol{\xi}) = \text{vech}(\mathbf{I}_p)$, with $\mathbf{I}_p$ a $p \times p$ identity matrix. Assuming the model is otherwise identified, this constraint may be imposed as

$$\begin{pmatrix} \boldsymbol{\mu}(\boldsymbol{\xi}) \\ \boldsymbol{\sigma}(\boldsymbol{\xi}) \end{pmatrix} = \mathbf{u} + \mathbf{L}\boldsymbol{\xi}, \tag{3.6}$$

where, in this case, $\mathbf{u}$ is a $(p + p^*) \times 1$ vector of constants and $\mathbf{L}$ is a $(p + p^*) \times d_2$ matrix of constants that specify the constraint. For this example, the necessary $\mathbf{u}$ and $\mathbf{L}$ are

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\sigma}_{11} \\ \boldsymbol{\sigma}_{21} \\ \boldsymbol{\sigma}_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ 1 \\ 0 \\ 1 \end{pmatrix}. \tag{3.7}$$

In other words, while the bivariate normal may depend on 5 unique quantities, given this particular constraint, it is only a function of 2 free parameters, ξ_1 and ξ_2. In an analogous fashion, arbitrary linear constraints may be imposed on the various item parameters (e.g., Cai, 2010b).

The example also highlights a convenient feature of the normal density, namely, that it is parameterized directly by the mean and variance. As just shown, this feature makes imposing constraints on either of the moments straightforward. This feature also motivates the proposed parameterization of the SNP density, presented in Chapter 4.

3.3 Observed and Complete Data Likelihoods

Given an item model, the conditional distribution of y_{ij} is a multinomial with K_j cells and cell probabilities π_{ijk},

$$f(y_{ij}|\boldsymbol{\zeta}, \boldsymbol{\theta}_{ig}) = \prod_{k=0}^{K_j-1} \pi_{ijk}^{\chi_k(y_{ij})}, \tag{3.8}$$

where $\chi_k(y)$ is an indicator function defined as

$$\chi_k(y) = \begin{cases} 1, & \text{if } y = k, \\ 0, & \text{otherwise} \end{cases}. \tag{3.9}$$

As is common with latent variable models, item responses are assumed independent after conditioning on the latent variable (Lord & Novick, 1968). The implication is that the conditional density of $\mathbf{y}_i$ is simply the product over the n items:

$$f(\mathbf{y}_i|\boldsymbol{\zeta}, \boldsymbol{\theta}_i) = \prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\zeta}, \boldsymbol{\theta}_i). \tag{3.10}$$

Having specified the conditional distribution of $\mathbf{y}_i$, the next step is to specify its marginal density, which is obtained by integrating out the latent variables.

The marginal density is defined as

$$f(\mathbf{y}_i|\boldsymbol{\omega}) = \int \prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\zeta}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})d\boldsymbol{\theta}. \tag{3.11}$$

At this point, the subscript g is re-introduced to define the observed data likelihood for group g,

$$L(\boldsymbol{\omega}|\mathbf{Y}_g) = \prod_{i=1}^{N_g} \left[\int \prod_{j=1}^{n_g} f(y_{ijg}|\boldsymbol{\zeta}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})d\boldsymbol{\theta} \right]. \tag{3.12}$$

For the entire sample, the observed data likelihood may then be expressed as

$$L(\boldsymbol{\omega}|\mathbf{Y}) = \prod_{g=1}^{G} L(\boldsymbol{\omega}|\mathbf{Y}_g). \tag{3.13}$$

Though we seek the maximum of Equation (3.13), it is difficult to directly optimize due to the N integrals, which may be high-dimensional. Instead, it is easier to optimize Equation (3.13) indirectly, by exploiting the relationship between the observed and complete data models. This relationship will be discussed in Chapter 5. For group g, the observed data, $\mathbf{Y}_g$, with the missing data $\boldsymbol{\Theta}_g$ together constitute the complete data, and its likelihood is

$$L(\boldsymbol{\omega}|\mathbf{Y}_g, \boldsymbol{\Theta}_g) = \prod_{i=1}^{N_g} \left[\varphi(\boldsymbol{\theta}_{ig}; \boldsymbol{\xi}) \prod_{j=1}^{n_g} f(y_{ijg}|\boldsymbol{\zeta}, \boldsymbol{\theta}_{ig}) \right], \tag{3.14}$$

which may be factored as

$$L(\boldsymbol{\omega}|\mathbf{Y}_g, \boldsymbol{\Theta}_g) = \left[\prod_{i=1}^{N_g} \varphi(\boldsymbol{\theta}_{ig}; \boldsymbol{\xi}) \right] \left[\prod_{i=1}^{N_g} \prod_{j=1}^{n_g} f(y_{ijg}|\boldsymbol{\zeta}, \boldsymbol{\theta}_{ig}) \right]. \tag{3.15}$$

Then, just as with the observed data likelihood, the complete data likelihood for

the entire sample is obtained by taking the product across groups,

$$L(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) = \prod_{g=1}^{G} L(\boldsymbol{\omega}|\mathbf{Y}_g, \boldsymbol{\Theta}_g). \tag{3.16}$$

Moreover, taking the log of the complete data likelihood reveals that it is the sum of two different parts:

$$\log L(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) = \log L(\boldsymbol{\zeta}|\mathbf{Y}, \boldsymbol{\Theta}) + \log L(\boldsymbol{\xi}|\boldsymbol{\Theta}), \tag{3.17}$$

Furthermore, each of these parts may consist of multiple mutually independent log-likelihoods. With SNP-MIRT, (at least) one of these log-likelihoods is given by the SNP density. The implications of this structure are remarkable. First, note that Equation (3.17) does not contain the N integrals of Equation (3.12). And second, instead of optimizing all of the elements of $\boldsymbol{\omega}$ simultaneously, the complete data formulation allows a piecemeal approach. Consequently, an optimization scheme that updates parameters based on the complete data log-likelihood (e.g., EM) or its derivatives (e.g., MH-RM) can proceed with individual updates for each item and group density.

This chapter has introduced the graded response and 3PL MIRT models, and the likelihood for the item parameters. The associated log-likelihood and derivatives can be found in Appendix B of Cai (2010a). For SNP-MIRT, the log-likelihood and derivatives for the SNP density are also needed. These are presented next.

CHAPTER 4

Semi-Nonparametric Density Estimation

In this chapter, the SNP density is introduced as a model for "observed" latent variables. Examples are provided to demonstrate the flexibility of the SNP model, as well as illustrate minor limitations. Often, the SNP density is applied in situations where the scale and location of the density may not need to be constrained. However, in MIRT, it is often desirable to place constraints on the latent variable density, either to identify the model or to investigate substantive hypotheses, such as differences in latent variable means. Consequently, in this research, a new parameterization is proposed, which is more amenable to a MIRT modeling context.

While the SNP density is more flexible than the normal, it is also more complex. This relative complexity, of course, may introduce some difficulties during estimation of the SNP-MIRT model. How these challenges are addressed will be discussed in Chapter 5.

4.1 Original Parameterization

Let $\mathbf{x}$ be a p-dimensional vector with jth element x_j. The semi-nonparametric (SNP) density is given by

$$g(\mathbf{x};\boldsymbol{\nu}) = P_\kappa^2(\mathbf{x};\boldsymbol{\nu})\phi(\mathbf{x}) \tag{4.1}$$

where the first factor is a p-variate polynomial in $\mathbf{x}$ with $t = \binom{p+\kappa}{\kappa}$ terms, and $\phi(\cdot)$

is a p-variate standard normal density. Let $\mathbf{a}$ be the $t \times 1$ vector of polynomial coefficients. As will be explained, the elements of $\mathbf{a}$ are functions of the $q = t - 1$ elements in $\boldsymbol{\nu}$. The tuning constant, κ, governs the powers of the elements of $\mathbf{x}$, as the sum of these powers may not exceed κ.

For example, let $p = 2$ and $\kappa = 2$. Then the bivariate polynomial may be written as

$$P(\mathbf{x}) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{11}x_1x_2 + a_{02}x_2^2, \tag{4.2}$$

where, for any term, the subscript of the coefficient corresponds to the set of powers of $\mathbf{x}$. Note that for $\kappa = 0$, $P(\mathbf{x}) = a_{00}$, a constant, and the SNP simplifies to a constant multiple of $\phi(\mathbf{x})$.

It is convenient to organize the powers for all terms in a $t \times p$ matrix, where each row consists of the p powers of a term. Let this matrix be $\boldsymbol{\Lambda}$, with (i, j)th element, λ_{ij}. Continuing with the example above,

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix}. \tag{4.3}$$

Then, let $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{x})$ be a $t \times 1$ vector with ith element

$$\eta_i = \prod_{j=1}^{p} x_j^{\lambda_{ij}}. \tag{4.4}$$

With this arrangement, the polynomial may be written more compactly as $P(\mathbf{x}) = \mathbf{a}'\boldsymbol{\eta}$.

For the general presentation above, the SNP function does not integrate to 1 and cannot serve as a proper density. Gallant and Tauchen (1989) impose the necessary constraint by fixing the first coefficient (e.g., $a_{00} \equiv 1$) and dividing $g(\mathbf{x};\boldsymbol{\nu})$ by a normalizing constant. Another approach, presented in Zhang and Davidian (2001), is adopted here.

4.2 The Zhang and Davidian (2001) Parameterization

The parameterization proposed by Zhang and Davidian (2001) does not involve the calculation of a normalizing constant. Instead, it ensures $g(\mathbf{x};\boldsymbol{\nu})$ integrates to 1 by enforcing the contstraint $\mathbf{c}'\mathbf{c} = 1$, where $\mathbf{c} = \mathbf{c}(\boldsymbol{\nu})$ is defined below. Let $\mathbf{U} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ so that $U_1, \ldots, U_p$ are independent standard normals. Then,

$$\begin{aligned} \int g(\mathbf{x};\boldsymbol{\nu})d\mathbf{x} &= \mathrm{E}\left(P_\kappa^2(\mathbf{U})\right) \\ &= \mathbf{a}'\mathrm{E}\left[\boldsymbol{\eta}(\mathbf{U})(\boldsymbol{\eta}(\mathbf{U}))'\right]\mathbf{a} \\ &= \mathbf{a}'\mathbf{A}\mathbf{a}, \end{aligned} \tag{4.5}$$

where $\mathbf{A} = \mathrm{E}\left[\boldsymbol{\eta}(\mathbf{U})(\boldsymbol{\eta}(\mathbf{U}))'\right]$ is a $t \times t$ matrix with (k,l)th element

$$\mathbf{A}_{kl} = \prod_{j=1}^{p} \mathrm{E}\left(U_j^{\lambda_{kj}+\lambda_{lj}}\right). \tag{4.6}$$

In words, the elements of $\mathbf{A}$ are the products of expectations of standard normals raised to specified powers. We can take the product of the expectations as the U_j are independent. Given p and κ, $\mathbf{A}$ is fixed and only needs to be computed once. Also, the expectations involved can be found by recursive formulas (see, e.g., Johnson, Kotz, & Balakrishnan, 1994). Notably, $\mathbf{A}$ is a sparse positive definite matrix, implying that $\mathbf{A} = \mathbf{B}'\mathbf{B}$ for some $\mathbf{B}$. Next, define the $t \times 1$ vector $\mathbf{c} = \mathbf{B}\mathbf{a}$. Finally, the restriction ensuring $g(\mathbf{x};\boldsymbol{\nu})$ is a proper density may be imposed on $\mathbf{c}$

as

$$\mathrm{E}(P_\kappa^2(\mathbf{U})) = \mathbf{a}'\mathbf{A}\mathbf{a} = \mathbf{a}'\mathbf{B}'\mathbf{B}\mathbf{a} = \mathbf{c}'\mathbf{c} = 1. \tag{4.7}$$

Note, however, that $(-\mathbf{c})'(-\mathbf{c}) = 1$ as well, creating an indeterminacy and need for a further constraint. These constraints may be achieved by the reparameterization of Zhang and Davidian (2001):

$$\begin{aligned} c_1 &= \cos(\delta_1) \\ c_2 &= \sin(\delta_1)\cos(\delta_2) \\ &\vdots \\ c_{t-1} &= \sin(\delta_1)\sin(\delta_2)\cdots\sin(\delta_{q-1})\cos(\delta_q) \\ c_t &= \sin(\delta_1)\sin(\delta_2)\cdots\sin(\delta_{q-1})\sin(\delta_q), \end{aligned} \tag{4.8}$$

for the $q \times 1$ vector $\boldsymbol{\delta}$, where $-\pi/2 < \delta_r \leq \pi/2$, $r = 1, \ldots, q$. The constraint on the angles restricts $\mathbf{c}$ to a half-unit sphere of $\mathbb{R}^t$, and prevents the indeterminacy mentioned above.

To allow unconstrained optimization, which is preferable for MH-RM, $\boldsymbol{\delta}$ must be reparameterized. The strategy adopted in this research is to relate $\boldsymbol{\delta}$ to the $q \times 1$ vector $\boldsymbol{\nu}$ through the arctangent function, that is,

$$\delta_r = \arctan(\nu_r), \qquad r = 1, \ldots, q, \tag{4.9}$$

where ν_r is the rth element of $\boldsymbol{\nu}$. Thus, $-\infty < \nu_r < \infty$, and $-\pi/2 < \delta_r(\nu_r) < \pi/2$, $r = 1, \ldots, q$. Note that with this parameterization, $\delta_r(\nu_r)$ may not attain the value $\pi/2$, though it may come arbitrarily close.

Returning to the earlier example (with $p = 2$ and $\kappa = 2$),

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 3 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{2} & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{2} \end{pmatrix},$$

$$\mathbf{B}^{-1} = \begin{pmatrix} 1 & 0 & 0 & -1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\sqrt{2} & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/\sqrt{2} \end{pmatrix},$$

$\mathbf{c}$ is the 6×1 vector following the definition in Equation (4.8), and

$$\mathbf{a} = \mathbf{B}^{-1}\mathbf{c} = \begin{pmatrix} c_1 - c_4/\sqrt{2} - c_6/\sqrt{2} \\ c_2 \\ c_3 \\ c_4/\sqrt{2} \\ c_5 \\ c_6/\sqrt{2} \end{pmatrix}.$$

With small values of κ (e.g., $\kappa = 1$ or 2), the SNP density can generate an impressive variety of shapes. However, these shapes may not match the modeled data in either location or scale. Consequently, it is desirable to introduce location and scale parameters to increase the flexibility of the density. One approach, illustrated by Zhang and Davidian (2001), has gained popularity in the literature on mixed models (e.g., Chen et al., 2002; Vock et al., 2012), and will be now be presented.

Let $\mathbf{m} = \mathrm{E}(\mathbf{x})$ be the mean of $\mathbf{x}$ and let $\mathbf{V} = \mathrm{Var}(\mathbf{x})$ be its variance. Also, let $\mathbf{SS}' = \mathbf{V}$, where $\mathbf{S}$ is a $p \times p$ lower triangular matrix and $\mathbf{s} = \mathrm{vech}(\mathbf{S})$. For the time being, assume these moments can be calculated. Importantly, in general, $\mathbf{m} \neq \mathbf{0}$ and $\mathbf{V} \neq \mathbf{I}_p$. In Zhang and Davidian (2001), the random effects $\mathbf{b}$ are modeled as

$$\mathbf{b} = \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{R}}\mathbf{x}, \tag{4.10}$$

with inverse transformation

$$\mathbf{x} = \tilde{\mathbf{R}}^{-1}(\mathbf{b} - \tilde{\boldsymbol{\mu}}). \tag{4.11}$$

Then $\mathrm{E}(\mathbf{b}) = \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{R}}\mathbf{m}$ and $\mathrm{Var}(\mathbf{b}) = \tilde{\mathbf{R}}\mathbf{V}\tilde{\mathbf{R}}'$. The tilde symbol is used to indicate that $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{R}}$, while location and scale parameters, may not be directly interpreted. Instead, these parameters serve to "adjust" the mean and variance

implied by $\boldsymbol{\nu}$, and increase the flexibility of the location and scale of the SNP density. We may write this more flexible density as

$$\begin{aligned} f(\mathbf{b};\boldsymbol{\nu},\tilde{\boldsymbol{\mu}},\tilde{\mathbf{r}}) &= g(\tilde{\mathbf{R}}_{\mathbf{b}}^{-1}(\mathbf{b}-\tilde{\boldsymbol{\mu}});\boldsymbol{\nu},\tilde{\boldsymbol{\mu}},\tilde{\mathbf{r}})|\tilde{\mathbf{R}}^{-1}| \\ &= g(\mathbf{x};\boldsymbol{\nu})|\tilde{\mathbf{R}}^{-1}|, \end{aligned} \tag{4.12}$$

where $\tilde{\mathbf{r}}$ is defined analogously to $\mathbf{s}$. The parameterization in Equation (4.12) for the random effects $\mathbf{b}$ has been used successfully in the mixed modeling literature (Chen et al., 2002) and implemented in software (Vock et al., 2014). Note, however, that due to the complicated dependence of $\mathrm{E}(\mathbf{b})$ and $\mathrm{Var}(\mathbf{b})$ on $\boldsymbol{\nu}, \tilde{\boldsymbol{\mu}}$, and $\tilde{\mathbf{r}}$, this parameterization does not lend itself to general confirmatory modeling. More specifically, placing constraints on $\mathrm{E}(\mathbf{b})$ and $\mathrm{Var}(\mathbf{b})$ via the various parameters would be cumbersome at best.

Another approach to introducing location and scale parameters, though, is proposed here.

4.3 A Standardized Parameterization

Ultimately, the parameterization proposed in this research depends on a standardized parameterization of $g(\mathbf{x};\boldsymbol{\nu})$. Let

$$\mathbf{z} = \mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) \tag{4.13}$$

with inverse transformation

$$\mathbf{x} = \mathbf{S}\mathbf{z} + \mathbf{m}. \tag{4.14}$$

That is, $\mathbf{z}$ is simply a standardized version of $\mathbf{x}$. It is easy to verify that $\mathrm{E}(\mathbf{z}) = \mathbf{0}$

and $\text{Var}(\mathbf{z}) = \mathbf{I}_p$, as desired. We may write this density as

$$\begin{aligned} h_z(\mathbf{z};\boldsymbol{\nu}) &= g(\mathbf{Sz}+\mathbf{m};\boldsymbol{\nu})|\mathbf{S}| \\ &= g(\mathbf{x};\boldsymbol{\nu})|\mathbf{S}|. \end{aligned} \tag{4.15}$$

The approach is simple, and appears quite similar to the approach of Zhang and Davidian (2001). However, whereas the elements of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{r}}$ are additional parameters to be estimated, the elements of $\mathbf{m}$ and $\mathbf{s}$ are complex functions of $\boldsymbol{\nu}$.

Standardized latent variable densities are used extensively in IRT modeling. However, in some IRT modeling contexts, it is desirable to estimate the location or scale of the latent variable. For instance, a researcher may wish to allow dimensions to freely correlate. As another example, it is common in multiple group modeling to standardize the distribution for a reference group while freely estimating the location and scale (or components thereof) of the comparison group. In the next section, an approach to extend the standardized SNP density is presented.

4.4 SNP-MIRT Parameterization

Consider the variable $\mathbf{y}$, defined as

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Rz}, \tag{4.16}$$

with inverse transformation

$$\mathbf{z} = \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{4.17}$$

Then, $\text{E}(\mathbf{y}) = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{RR}'$. Also, let $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$. Finally, let $\boldsymbol{\psi} = (\boldsymbol{\mu}', \boldsymbol{\sigma}', \boldsymbol{\nu}')'$ collect the various parameters for the density of $\mathbf{y}$. Then, the

density may be written as

$$\begin{aligned} h_y(\mathbf{y};\boldsymbol{\psi}) = h_y(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu}) &= h_z(\mathbf{R}^{-1}(\mathbf{y}-\boldsymbol{\mu});\boldsymbol{\nu})|\mathbf{R}^{-1}| \\ &= h_z(\mathbf{z};\boldsymbol{\nu})|\mathbf{R}^{-1}| \\ &= g(\mathbf{x};\boldsymbol{\nu})|\mathbf{S}\mathbf{R}^{-1}|, \end{aligned} \tag{4.18}$$

where the full dimensionality of $\boldsymbol{\psi}$ is $q+p+p^*$. The density given in Equation (4.18) is the general SNP parameterization proposed in this research and implemented for the SNP-MIRT model.

Several features of this parameterization are worth noting. First, unlike $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{r}}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are themselves mean and variance parameters. Consequently, for purposes of confirmatory modeling in SNP-MIRT, constraints may be imposed just as with a normal density. Second, if $\boldsymbol{\mu}$ is fixed to $\mathbf{0}$ and $\boldsymbol{\Sigma}$ is fixed to $\mathbf{I}_p$, then, $h_y(\mathbf{y};\boldsymbol{\psi})$ simplifies to $h_z(\mathbf{y};\boldsymbol{\nu})$. Third, if $\kappa = 0$, then a multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ is obtained. Consequently, SNP-MIRT can be understood to be a generalization of standard MIRT, where the latent distribution is assumed normal.

Additionally, two practical features of this parameterization should be noted. Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ should be relatively stable, since location and scale parameters tend to have high information. Also, this parameterization facilitates finding starting values for $\boldsymbol{\psi}$ in the following way. Given a sample, say $\boldsymbol{\theta}$, estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ may be found immediately using the sample mean and variance. These estimates may then be fixed and attention can turn to finding starting values for just $\boldsymbol{\nu}$. In contrast, for the random effects parameterization, it is necessary to find starting values for $\boldsymbol{\nu}$, $\tilde{\boldsymbol{\mu}}$, and $\tilde{\mathbf{r}}$ simultaneously. Obviously, the additional dimensionality may complicate the search.

All of this is not to argue that the parameterization proposed here is superior. It is clearly more complex than the established parameterization. However, given

typical IRT modeling, the proposed parameterization is arguably preferable for that context. The more straight-forward parameterization of Zhang and Davidian (2001) may arguably be better-suited for modeling the random effects in mixed models.

4.5 Mean and Variance of Original Parameterization

The standardized implementation depends on finding $\mathbf{m}$ and $\mathbf{V}$. To find $\mathbf{m}$ and $\mathbf{V}$, we proceed as we did to find $\mathbf{A}$. That is, we again utilize $\mathbf{U} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$, since

$$\begin{aligned} \mathrm{E}(\mathbf{x}) &= \int \mathbf{x} P_\kappa^2(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} \\ &= \mathrm{E}(\mathbf{U} P_\kappa^2(\mathbf{U})). \end{aligned} \tag{4.19}$$

For the mth element of $\mathbf{x}$,

$$\begin{aligned} \mathrm{E}(x_m) &= \mathrm{E}(U_m \mathbf{a}' \boldsymbol{\eta}(\mathbf{U}) \boldsymbol{\eta}(\mathbf{U})' \mathbf{a}) \\ &= \mathbf{a}' \mathrm{E}(U_m \boldsymbol{\eta}(\mathbf{U}) \boldsymbol{\eta}(\mathbf{U})') \mathbf{a} \\ &= \mathbf{a}' \mathbf{A}_m^* \mathbf{a}, \end{aligned} \tag{4.20}$$

where the (k, l)th element of $\mathbf{A}_m^*$ is given by

$$\prod_{j=1}^{p} \mathrm{E}\left(U_j^{\lambda_{jk} + \lambda_{jl} + \chi_j(m)} \right), \tag{4.21}$$

where $\chi_j(m)$ is an indicator function defined as

$$\chi_j(m) = \begin{cases} 1, & \text{if } m = j, \\ 0, & \text{otherwise} \end{cases}. \tag{4.22}$$

A similar strategy is used to find $\mathrm{E}(\mathbf{x}\mathbf{x}')$, needed for $\mathrm{Var}(\mathbf{x})$. Proceeding as above, for the mth and nth element of $\mathbf{x}$,

$$\begin{aligned} \mathrm{E}(x_m x_n) &= \mathrm{E}(U_m U_n \mathbf{a}'\boldsymbol{\eta}(\mathbf{U})\boldsymbol{\eta}(\mathbf{U})'\mathbf{a}) \\ &= \mathbf{a}'\mathrm{E}(U_m U_n \boldsymbol{\eta}(\mathbf{U})\boldsymbol{\eta}(\mathbf{U})')\mathbf{a} \\ &= \mathbf{a}'\mathbf{A}^*_{m,n}\mathbf{a}, \end{aligned} \tag{4.23}$$

where the (k,l)th element of $\mathbf{A}^*_{m,n}$ is given by

$$\prod_{j=1}^{p} \mathrm{E}\left(U_j^{\lambda_{kj}+\lambda_{kl}+\chi_j(m)+\chi_j(n)}\right), \tag{4.24}$$

and the additional indicator function is defined analogously to the Equation (4.22). Using Equations (4.20) and (4.23), $\mathbf{m} = \mathrm{E}(\mathbf{x})$ and $\mathbf{V} = \mathrm{E}(\mathbf{x}\mathbf{x}') - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x})'$ can be obtained.

With the expressions for $\mathbf{m}$ and $\mathbf{V}$, the density in Equation (4.15) can be found. The log-likelihood function for the new parameterization is then

$$\begin{aligned} \log L(\boldsymbol{\psi}|\mathbf{Y}) &= \sum_{i=1}^{N} \log h_y(\mathbf{y}_i;\boldsymbol{\psi}) \\ &= \sum_{i=1}^{N} \log P_\kappa^2(\mathbf{x}_i;\boldsymbol{\nu}) + \log\phi(\mathbf{x}_i) + \log|\mathbf{S}| + \log|\mathbf{R}^{-1}| \\ &= N\log|\mathbf{S}\mathbf{R}^{-1}| + \sum_{i=1}^{N} \log P_\kappa^2(\mathbf{x}_i;\boldsymbol{\nu}) + \log\phi(\mathbf{x}_i) \end{aligned} \tag{4.25}$$

As mentioned previously, in practice κ is treated like a tuning parameter, with larger values allowing more flexibility in the shape of the SNP density. If κ is too large relative to the information in the data (i.e., $\boldsymbol{\theta}$), then estimation of elements in $\boldsymbol{\nu}$ will be unstable. In more extreme cases, estimation may fail. However, it is difficult, a priori, to know how much information is provided by $\boldsymbol{\Theta}$ in order

to specify κ appropriately. Therefore, in the interest of stabilizing estimation, a diffuse prior may be placed on the $\boldsymbol{\nu}$ parameters.

4.6 Prior for SNP Parameters

Following Woods and Thissen (2006) and Monroe and Cai (2014), a diffuse q-variate normal prior is placed on the $\boldsymbol{\nu}$ parameters. It can be confirmed from the Zhang and Davidian (2001) parameterization that $g(\mathbf{x}; \mathbf{0}) = \phi(\mathbf{x})$. Consequently, the q-variate normal prior is specified with mean vector $\mathbf{0}$. The covariance matrix for the prior, $\varsigma\mathbf{I}_q$, implies that the marginal univariate priors on the elements of $\boldsymbol{\nu}$ are independent with common dispersion ς. Then, the actual objective function to be maximized is $\log L(\boldsymbol{\psi}|\mathbf{Y}) + \phi(\boldsymbol{\nu}; \mathbf{0}, \varsigma\mathbf{I}_q)$.

The first and second derivatives of this expression are needed for the MH-RM algorithm. The derivatives for the first term are presented in Appendix A. The derivatives for the second term are standard results (see, e.g., Equations (18) to (20), Woods & Thissen, 2006). With the SNP model fully defined, the complete data log-likelihood in Equation (3.17) is entirely specified. We can now turn our attention to estimation of the model, using the MH-RM algorithm.

CHAPTER 5

The Metropolis-Hastings Robbins-Monro Algorithm

5.1 The EM Algorithm as a Point of Comparison

Before presenting the MH-RM algorithm, the EM algorithm will be presented for two purposes. First, the EM algorithm is more well-known, and comparing details of EM to details of MH-RM may make the latter easier to understand. And second, the two algorithms are both motivated by the so-called "Fisher's Identity," (Fisher, 1925).

Let l denote the log-likelihood, so that $l(\boldsymbol{\omega}|\mathbf{Y})$ is the observed data log-likelihood and $l(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta})$ is the complete data log-likelihood. Formally, the goal of MML estimation is to find

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y}). \tag{5.1}$$

Instead of maximizing $l(\boldsymbol{\omega}|\mathbf{Y})$ directly, the EM algorithm finds $\hat{\boldsymbol{\omega}}$ by iteratively maximizing the expectation of $l(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta})$ over $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$, where $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$ is the posterior distribution of missing data, given observed data and current parameter estimates. Taking the appropriate expectation is the E-step, and calculating its maximum is the M-step. These two steps are iterated until some convergence criteria is met.

Thus, the EM algorithm capitalizes on the equality

$$l(\boldsymbol{\omega}|\mathbf{Y}) = \mathrm{E}(l(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta})) \tag{5.2}$$

where the expectation is with respect to $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$. Let $\nabla_{\boldsymbol{\omega}}$ return the $d \times 1$ vector of all partial first derivatives with respect to $\boldsymbol{\omega}$. Then, let

$$\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) = \nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) \tag{5.3}$$

be the complete data score function. As noted by Efron (1977) in the discussion section of Dempster et al. (1977), Equation 5.2 is a reformation of

$$\begin{aligned} \nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y}) &= \int \mathbf{s}(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) \Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega}) d\boldsymbol{\Theta} \\ &= \mathrm{E}(\mathbf{s}(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta})), \end{aligned} \tag{5.4}$$

first presented by Fisher (1925). This connection is important here because Equation (5.4) also drives MH-RM. In this way, EM and MH-RM exploit the same relationship to iteratively find the MLE. The mechanics of how MH-RM capitalizes on Equation (5.4) will now be presented.

5.2 MH-RM in Broad Outline

Notice that the root of the left-hand side of Equation (5.4) is the MLE. The MH-RM algorithm finds this maximum by iteratively updating $\hat{\boldsymbol{\omega}}$, using the right-hand side. The right-hand side, in turn, is an expectation that can be approximated by Monte Carlo, provided samples from $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$ may be obtained. Typically, when computing Monte Carlo approximations, it is preferable to increase the sample size to reduce the Monte Carlo error. With MH-RM, however, minimizing the Monte Carlo error for any iteration is not a priority. Perhaps counter-intuitively,

with MH-RM, a small Monte Carlo sample size is preferable because it is computationally cheap. The RM method serves to filter the noise, and the sequence of estimates converges to $\hat{\boldsymbol{\omega}}$. In the sections below, more details will be provided on how to obtain the samples, compute the updates, and apply the RM filter. These are the key steps in the algorithm.

5.3 MH-RM in Greater Detail

5.3.1 Constructing an MH Sampler

Again, to approximate Equation (5.4) by Monte Carlo, samples from $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$ are needed. Following Patz and Junker (1999) and Cai (2010a), a Metropolis-within-Gibbs sampling scheme is used to impute these latent variable scores. Consider the posterior predictive density of $\boldsymbol{\theta}_i$ for a generic respondent. By Bayes' rule, the density is

$$\Pi(\boldsymbol{\theta}_i|\mathbf{y}_i, \boldsymbol{\omega}) = \frac{f(\mathbf{y}; \boldsymbol{\theta}_i, \boldsymbol{\zeta})\varphi(\boldsymbol{\theta}_i; \boldsymbol{\xi})}{\int f(\mathbf{y}; \boldsymbol{\theta}_i, \boldsymbol{\zeta})\varphi(\boldsymbol{\theta}_i; \boldsymbol{\xi})d\boldsymbol{\theta}} \propto f(\mathbf{y}; \boldsymbol{\theta}_i, \boldsymbol{\zeta})\varphi(\boldsymbol{\theta}_i; \boldsymbol{\xi}), \tag{5.5}$$

where the constant of proportionality is $f(\mathbf{y}_i|\boldsymbol{\omega})$ from Equation (3.11). By the Metropolis-Hastings algorithm, given the current value $\boldsymbol{\theta}_i$, a candidate value, $\check{\boldsymbol{\theta}}_i$, is drawn from a convenient transition density, $q(\boldsymbol{\theta}_i, \check{\boldsymbol{\theta}}_i)$. The candidate value is accepted with probability

$$\alpha(\boldsymbol{\theta}_i, \check{\boldsymbol{\theta}}_i) = \min\left[\frac{\Pi(\check{\boldsymbol{\theta}}_i|\mathbf{y}_i, \boldsymbol{\omega})q(\check{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i)}{\Pi(\boldsymbol{\theta}_i|\mathbf{y}_i, \boldsymbol{\omega})q(\boldsymbol{\theta}_i, \check{\boldsymbol{\theta}}_i)}, 1\right]. \tag{5.6}$$

Note that since $\Pi(\cdot|\mathbf{y}_i, \boldsymbol{\omega})$ appears in both the numerator and denominator of Equation (5.6), the expression may be simplified by cancelling the (difficult to evaluate) constant of proportionality. Additionally, if a symmetric transition density is chosen, Equation (5.6) may be further simplified. With these simplifications,

the acceptance factor takes the form

$$\alpha(\boldsymbol{\theta}_i, \check{\boldsymbol{\theta}}_i) = \min\left[\frac{f(\mathbf{y}_i|\boldsymbol{\zeta}, \check{\boldsymbol{\theta}}_i)\varphi(\check{\boldsymbol{\theta}}_i; \boldsymbol{\xi})}{f(\mathbf{y}_i|\boldsymbol{\zeta}, \boldsymbol{\theta}_i)\varphi(\boldsymbol{\theta}_i; \boldsymbol{\xi})}, 1\right], \tag{5.7}$$

Conveniently, calculating the acceptance probability for each $\boldsymbol{\theta}_i$ depends on neither the latent variable scores of other persons, nor their item responses. With this sampler, a Monte Carlo approximation to Equation (5.4) may be calculated for any provisional estimate of $\boldsymbol{\omega}$. Next, the process by which these estimates are updated is presented.

5.3.2 Calculating Updates

Let $\boldsymbol{\omega}_k$ be the vector of parameter estimates at the end of the kth iteration. At iteration $k+1$, the update is a function of three components. One of these is an approximation of $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}_k|\mathbf{Y})$. Another is a recursive approximation to the conditional expectation of the complete data information matrix, which will be defined below. The final component is the *gain constant*, γ, which will be explained in the next section. It is also helpful to view the update at each iteration as consisting of three stages: *stochastic imputation*, *stochastic approximation*, and an *RM update*.

For *stochastic imputation*, we draw m_k sets of missing data $\{\boldsymbol{\Theta}_{k+1}^{(j)}; j = 1, \ldots, m_k\}$ from $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega}_k)$. Coupled with $\mathbf{Y}$, these imputations create m_k complete data sets. Again, the RM filter makes large m_k unnecessary. In many IRT applications (e.g., Cai, 2010a; Monroe & Cai, 2014), $m_k \equiv 1$ has proven sufficient.

For *stochastic approximation*, using Equation (5.4), we approximate the observed data gradient, $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}_k|\mathbf{Y})$, by the sample average of complete data gradients,

$$\tilde{\mathbf{s}}^{(k+1)} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{s}(\boldsymbol{\omega}^{(k)}|\mathbf{Y}, \boldsymbol{\Theta}_{k+1}^{(j)}). \tag{5.8}$$

Essentially, $\tilde{\mathbf{s}}_{k+1}$ provides the direction of the update. However, its magnitude may not be optimal. Consequently, convergence may be hastened given an appropriate scaling factor. Following Cai (2010a), we choose the scaling factor to be the conditional expectation of the complete data information, where the expectation is with respect to $\Pi(\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\omega})$. Let the complete data information matrix be

$$\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}) = \frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'}. \tag{5.9}$$

Also, let $(\boldsymbol{\omega}_0, \boldsymbol{\Gamma}_0)$ be initial values, where $\boldsymbol{\Gamma}_0$ is a symmetric positive definite matrix. Then, $\mathrm{E}(\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y}, \boldsymbol{\Theta}))$ may be recursively approximated as

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \left\{ \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{H}(\boldsymbol{\omega}_k|\mathbf{Y}, \boldsymbol{\Theta}_{k+1}^{(j)}) - \boldsymbol{\Gamma}_k \right\}. \tag{5.10}$$

Finally, in the *RM update*, we set the new parameter estimate to

$$\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k + \gamma_k (\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}). \tag{5.11}$$

Due to the independence implied by Equation (3.17), the *stochastic approximation* and *RM update* steps may be carried out separately for the different items and group densities.

5.3.3 Applying the RM Filter

Let $k = 1, 2, \ldots, \infty$ index the iteration for the MH-RM algorithm. The gain constants γ_k, for $k \geq 1$, scale the updates and serve to slowly average out the noise in the updates. For this to occur, the γ_k need to slowly decrease to zero,

which is ensured by the following conditions,

$$\gamma_k \in (0, 1], \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \text{ and } \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty. \tag{5.12}$$

If the γ_k decrease too quickly, then the estimates of $\boldsymbol{\omega}$ may stabilize prematurely, before the MLE is reached. Alternatively, if the γ_k decrease too slowly, the estimates for $\boldsymbol{\omega}$ may never stabilize. Cai (2010a) suggests taking $\gamma = 1/k$, which has proven effective for most MIRT models. The SNP-MIRT model, however, is relatively delicate, and it is desirable to have greater control over the size of the updates. In this research, the gain constant is defined as

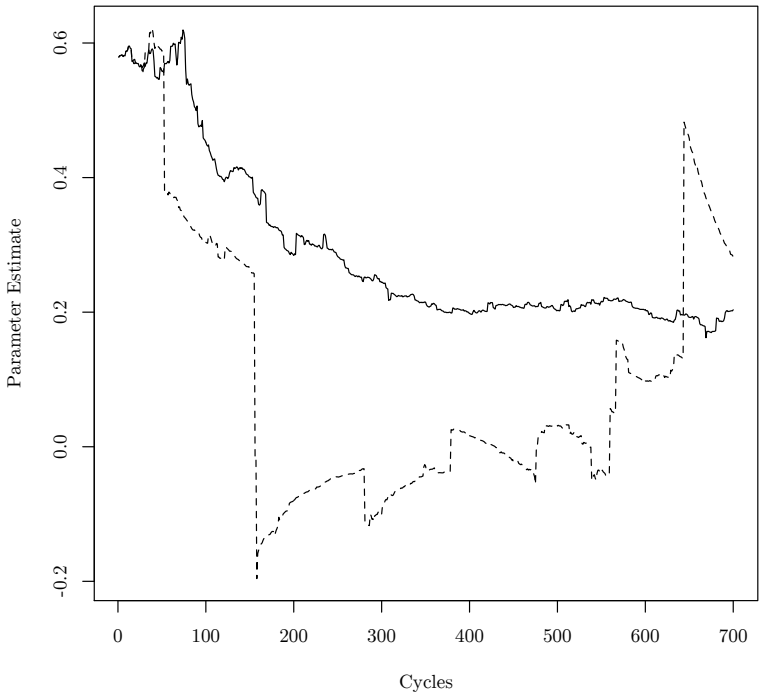
$$\gamma_k = \frac{1}{A + k^{\epsilon}}, \tag{5.13}$$

where $1/2 < \epsilon \leq 1$, and A is some positive integer known as the *stability coefficient* (Spall, 1998). Setting ϵ closer to $1/2$ allows for greater step-sizes at later iterations, thus guarding against premature convergence. Monroe and Cai (2014), following Polyak and Juditsky (1992), found setting $\epsilon = .75$ to be effective, especially for the Ramsay-curve parameters. The purpose of A is to dampen the magnitude of the updates in the early iterations, when the sequence of estimates tends to be least stable. Spall (2003) recommends setting A to approximately 10% of the number of cycles expected to reach convergence.

In the stochastic approximation literature (e.g., Spall, 1997, 1998), it is also common to allow for update *blocking*, meaning that if certain criteria are met, no update is taken for that iteration. As an example, criteria may be based on the absolute magnitude of the proposed update. This research also adopts blocking criteria to stabilize the estimation algorithm, and a graphical example is provided in Figure 5.1.

These tuning features highlight the fact that MH-RM is an incredibly flexible

Figure 5.1: Effect of Blocking on Parameter Estimation



Note. The two trajectories show SNP-MIRT estimation that implements blocking (solid line) and estimation that does not (dashed line). Both histories start at the same value, with the same random number seed. The difference in the trajectories is due to blocking.

framework, and ideas developed in the rich stochastic approximation literature may be incorporated as needed. To the same end, some aspects of the original implementation may be relaxed without compromising the convergence results. In particular, the update in Equation (5.11) is still proper even if the scaling factor $\boldsymbol{\Gamma}_{k+1}^{-1}$ is removed from the expression, provided that the gain constants are chosen with care.

5.3.4 Obtaining Starting Values

The convergence of the sequence of MH-RM estimates to the MLE depends on the use of starting values that are sufficiently close to the MLE. These starting values will be referred to as "Stage III" starting values, for reasons that will soon be apparent. Whether a set of Stage III starting values satisfies this requirement depends on a host of factors, which certainly includes the data and choice of model. Generally, the following two steps, or stages, have been used to obtain Stage III starting values in MH-RM:

1. For any parameter, choose some crude, but not unreasonable, initial value. Run MH-RM cycles, with m_k and γ_k fixed to one for all k, until the estimates stabilize within a "neighborhood."

2. Continuing with these settings, compute an average of the parameter estimates over the next M cycles, where M is sufficiently large so that the average represents the "neighborhood."

The two stages above are often referred to as "Stage I" and "Stage II" (see, e.g., Cai, 2013). The averages from Stage II may then be used as starting values for Stage III, when decreasing gain constants are applied. As noted in Cai (2010a), the MH-RM algorithm, as specified in Stages I and II, is a close relative to Diebolt and Ip's (1996) stochastic EM (SEM) algorithm. Importantly, the SEM-type

iterations move $\hat{\boldsymbol{\omega}}_k$ quickly to the neighborhood of the MLE. While the parameter estimates are not themselves sampled, the first stage is akin to the burn-in of an MCMC scheme. Extending the analogy, the "neighborhood" is somewhat like a target distribution.

In practice, for most item models, as well as the normal density, choosing intial values for Stage I is easy. For example, item slopes may be initialized to 1.0, and item intercepts may be initialized based on sample proportions of observed responses. However, for the SNP shape parameters, $\boldsymbol{\nu}$, there is no obvious choice for the initial values. Setting $\boldsymbol{\nu} = \mathbf{0}$ produces a normal, which would seem like a good choice. However, in many cases, the likelihood surface for $\boldsymbol{\nu}$ is quite irregular. The implication of this irregularity, with regards to MH-RM, is that unless the Stage I initial value for $\boldsymbol{\nu}$ is chosen carefully, the Stage III starting values will not be close to the MLE. For such a condition, the convergence results for MH-RM do not hold. Thus, an alternative strategy is needed for $\boldsymbol{\nu}$.

Other research on the SNP density has recognized the difficulty of finding useful initial values (see, e.g., Irincheeva et al., 2012; Vock et al., 2014). Of particular difficulty is that the parameter space of $\boldsymbol{\nu}$ may be high-dimensional, limiting the practicality of quadrature-based approaches. At the same time, as mentioned above, the likelihood surface tends to be irregular, which limits the effectiveness of Newton-type optimizers. Assume for the moment that an initial estimate of $\hat{\boldsymbol{\Theta}}$ may be obtained, say $\hat{\boldsymbol{\Theta}}_0$. For the present research, the initial value of $\boldsymbol{\nu}$, given $\hat{\boldsymbol{\Theta}}_0$, is found using the differential evolution (DE) algorithm (Storm & Price, 1997). The DE algorithm is a stochastic direct search method for global optimization that is relatively robust to multi-modality. Notably, DE does not guarantee an optimal solution (Storm & Price, 1997). However, an optimal solution is unnecessary for the present purposes, as the initial value for $\boldsymbol{\nu}$ only needs to be "good enough."

In practice, researchers will want to compare the results from the standard MIRT and SNP-MIRT models. With this in mind, the starting values for SNP-MIRT may be found in the following manner. First, obtain $\tilde{\boldsymbol{\omega}}$, the ML estimates for a standard MIRT model. Then, use $\tilde{\boldsymbol{\omega}}$ to obtain latent variable score estimates. Next, treat these score estimates as $\hat{\boldsymbol{\Theta}}_0$, and use the DE algorithm to find the initial value of $\boldsymbol{\nu}$. This initial value, along with $\tilde{\boldsymbol{\omega}}$, may be used for Stage I of MH-RM for the SNP-MIRT model.

CHAPTER 6

On Likelihood Inference for MIRT

Following estimation of an SNP-MIRT model, it is desirable to conduct inference, compare models, examine parameter significance, etc. Two quantities important to likelihood inference are the observed data log-likelihood and the information matrix. The former is routinely used in model comparison, either directly, or through derived statistics such as AIC and BIC. The latter may be used for several purposes, including the assessment of parameter estimate variability, DIF analyses, and construction of overall model fit statistics. Approaches for estimating these quantities in conjunction with Bock and Aitkin's (1981) quadrature-based EM have received a good deal of attention (see, e.g. Cai, 2008b; Tian, Cai, Thissen, & Xin, 2012). However, for high-dimensional models, where MH-RM is the preferred algorithm, methods for estimating these quantities have received less attention.

In this chapter, some current approaches to estimation of the observed data likelihood and information matrix are presented. Then, approaches based on Chib and Jeliazkov (2001) are presented. In particular, extending the marginal likelihood estimator in Chib and Jeliazkov (2001), a new estimator for the observed information matrix is proposed.

As the observed log-likelihood and observed information are summed over respondents, it is sufficient to consider a single respondent, with observed response pattern $\mathbf{y}_i$, and latent variable $\boldsymbol{\theta}_i$. However, to simplify the presentation, the subscript i will be dropped temporarily.

6.1 Current Approaches

6.1.1 Observed Data Likelihood

Restating Equation (3.11) in Chapter 3, the observed likelihood for response pattern $\mathbf{y}$ is

$$f(\mathbf{y}|\boldsymbol{\omega}) = \int f(\mathbf{y}|\boldsymbol{\zeta}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})d\boldsymbol{\theta}, \tag{6.1}$$

where, as a reminder, $\boldsymbol{\theta}$ may be high-dimensional, rendering quadrature-based methods impractical. A straightforward approach, and the current standard, is to estimate Equation (3.11) as a Monte Carlo expectation, using random draws from $\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})$. That is,

$$f(\mathbf{y}|\boldsymbol{\omega}) \approx M^{-1} \sum_{r=1}^{M} f(\mathbf{y}|\boldsymbol{\zeta}, \boldsymbol{\theta}^{(r)}), \tag{6.2}$$

where $\{\boldsymbol{\theta}^{(r)}\}$ is sampled from $\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})$. A benefit of this Monte Carlo approach is that each iteration is fast. To see the potential problem with this approach, suppose that $f(\mathbf{y}|\boldsymbol{\zeta}, \boldsymbol{\theta})$ is nearly 0 for $\boldsymbol{\theta}$ outside some region B. In such a scenario, accurate approximation by Equation (6.2) requires a sufficient number of draws where $\boldsymbol{\theta} \in B$. However, as p and n increase, $P(\boldsymbol{\theta} \in B)$ will decrease. At some point, increasing M to address this issue may not be practical.

This situation is fairly common in Bayesian inference, and a number of variance-reduction methods have been developed to address it. Among these are importance sampling and stratified sampling.

6.1.2 Observed Information

Louis (1982) presented a result relating the observed information to derivatives of the complete data model. The equation is

$$\begin{aligned} -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} &= \mathrm{E}(\mathbf{H}(\boldsymbol{\omega}|\mathbf{Y},\boldsymbol{\Theta})) - \mathrm{E}(\mathbf{s}(\boldsymbol{\omega};\mathbf{Y},\boldsymbol{\Theta})\,[\mathbf{s}(\boldsymbol{\omega};\mathbf{Y},\boldsymbol{\Theta})]') \\ &\quad + \mathrm{E}(\mathbf{s}(\boldsymbol{\omega};\mathbf{Y},\boldsymbol{\Theta}))\mathrm{E}([\mathbf{s}(\boldsymbol{\omega};\mathbf{Y},\boldsymbol{\Theta})]'), \end{aligned} \tag{6.3}$$

where all expectations are with respect to $\Pi(\boldsymbol{\Theta}|\mathbf{Y},\boldsymbol{\omega})$. Note that at the ML estimate, again, due to Fisher (1925), $\mathbf{s}(\hat{\boldsymbol{\omega}};\mathbf{Y},\boldsymbol{\Theta}) = \mathbf{0}$. Thus, evaulated at $\hat{\boldsymbol{\omega}}$, the last term on the right-hand side in Equation (6.3) is a $d \times d$ null matrix.

Cai (2010a) proposed a method to recursively estimate the quantities needed for the Louis formula concurrently with the main MH-RM cycles. The first term on the right-hand side may be estimated using Equation (5.10), while the second and third terms may take analogous forms, using the complete data score in Equation (5.8). A benefit of this approach is that the information matrix estimate is available upon convergence of the MH-RM algortihm. This method is implemented in flexMIRT®(Cai, 2013).

A second approach, also implemented in flexMIRT®and used in Monroe and Cai (2014), is based on a method proposed by Diebolt and Ip (1996). Basically, the strategy uses Monte Carlo integration following convergence of MH-RM. This approach obviously requires more computation than the concurrent approach, but may be slightly more stable, as $\boldsymbol{\omega}$ is fixed at $\hat{\boldsymbol{\omega}}$ for all Monte Carlo samples.

In practice, using either of the above approaches, there tends to be substantial variability in the Monte Carlo estimates, especially when the fraction of missing information is high. In some situations, this can result in negative error variance estimates, which is obviously undesirable.

6.2 Chib and Jeliazkov Estimators

Chib and Jeliazkov (2001) proposed a method to estimate the marginal likelihood of a model, integrating over all free parameters. This quantity is particularly useful in Bayesian model selection, and has been applied to latent variable models (Vitoratou, Ntzoufras, & Moustaki, 2014). In this research, the method is used in a slightly different way, for a slightly different purpose. The goal here is to estimate the marginal likelihood of a response pattern for a fixed $\hat{\boldsymbol{\omega}}$. Thus, instead of integrating over all model parameters, the integration only takes place over $\boldsymbol{\theta}$.

6.2.1 Observed Data Likelihood

Since the observed data marginal is the constant of proportionality for the posterior predictive distribution $\Pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\omega})$, the observed data marginal may be expressed as

$$f(\mathbf{y}; \boldsymbol{\omega}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}; \boldsymbol{\omega})\varphi(\boldsymbol{\theta}; \boldsymbol{\omega})}{\Pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\omega})}. \tag{6.4}$$

This equation holds for any fixed $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^*$. Fixing $\boldsymbol{\theta}$ at this value and taking log of both sides,

$$\log f(\mathbf{y}|\boldsymbol{\omega}) = \log f(\mathbf{y}|\boldsymbol{\theta}^*; \boldsymbol{\omega}) + \log \varphi(\boldsymbol{\theta}^*; \boldsymbol{\omega}) - \log \Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega}). \tag{6.5}$$

As the first two terms on the right-hand side are easy to calculate, the challenge of estimating $f(\mathbf{y}|\boldsymbol{\omega})$ is replaced by the challenge of estimating $\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$. The major contribution of Chib and Jeliazkov (2001) is that it proposes an estimator for this latter quantity when the M-H algorithm is used to sample from $\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$. For MH-RM, this is of course the case.

Recall that $q(\cdot, \cdot)$ and $\alpha(\cdot, \cdot)$ are the transition density and acceptance probability, respectively, defined in Section 5.3. Utilizing properties of the M-H Algorithm,

Chib and Jeliazkov (2001) shows that $\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$ may be written as

$$\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega}) = \frac{\mathrm{E}_\Pi \left[q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \right]}{\mathrm{E}_q \left[\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \right]} \tag{6.6}$$

where the numerator expectation, E_Π, is with respect to $\Pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\omega})$ and the denominator expectation, E_q, is with respect to $q(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. Consequently, a consistent estimate of $\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$ is

$$\hat{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega}) = \frac{C^{-1} \sum_{c=1}^{C} q(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^*) \alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^*)}{J^{-1} \sum_{j=1}^{J} q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)})}, \tag{6.7}$$

where $\{\boldsymbol{\theta}^{(c)}\}$ are (possibly correlated) draws from $\Pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\omega})$, and $\{\boldsymbol{\theta}^{(j)}\}$ are independent draws from $q(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. Let $\bar{u}$ and $\bar{v}$ be the numerator and denominator of the right-hand side of Equation (6.7), respectively. Then, by this method, $\log \hat{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega}) = \log \bar{u} - \log \bar{v}$.

Let the observed data likelihood estimate obtained by this method be $\hat{f}(\mathbf{y}|\boldsymbol{\omega})$. Then, to assess the variability in $\log \hat{f}(\mathbf{y}|\boldsymbol{\omega})$, it is only necessary to assess the variability in $\log \hat{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$, as the other quantities are fixed. This may be accomplished using a bivariate Delta method. Plugging in sample quantities as needed, this estimator is

$$\mathrm{Var}(\log \hat{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})) = \frac{\mathrm{Var}(\bar{u})}{\bar{u}^2} + \frac{\mathrm{Var}(\bar{v})}{\bar{v}^2}. \tag{6.8}$$

The terms on the right are merely variances of sample means. Note, however, that the calculation of $\mathrm{Var}(\bar{u})$ should account for the correlation among the samples $\{\boldsymbol{\theta}^{(c)}\}$. In this research, $\mathrm{Var}(\bar{u})$ is estimated using batch means (see, e.g. Roberts, 1996).

For the full sample, the contributions to the log-likelihod and variability are simply summed across the N respondents, as the observations are assumed inde-

pendent. Let the log-likelihood estimator for the full sample be $l_{CJ}(\mathbf{Y}|\boldsymbol{\omega})$.

6.2.2 Observed Information

The Louis (1982) formula in Equation (6.3) provides one method of approximating the information matrix. Another method, for a sample of size N, is given by

$$\mathcal{F} = \sum_{i=1}^{N} \nabla_{\boldsymbol{\omega}} \log f(\mathbf{y}_i|\boldsymbol{\omega})[\nabla_{\boldsymbol{\omega}} \log f(\mathbf{y}_i|\boldsymbol{\omega})]', \tag{6.9}$$

where the subscript i has been re-introduced. Extending Chib and Jeliazkov (2001), we may take derivatives of Equation (6.5) with respect to $\boldsymbol{\omega}$, yielding

$$\nabla_{\boldsymbol{\omega}} \log f(\mathbf{y}_i|\boldsymbol{\omega}) = \nabla_{\boldsymbol{\omega}} \log f(\mathbf{y}_i|\boldsymbol{\theta}^*;\boldsymbol{\omega}) + \nabla_{\boldsymbol{\omega}} \log \varphi(\boldsymbol{\theta}^*;\boldsymbol{\omega}) - \nabla_{\boldsymbol{\omega}} \log \Pi(\boldsymbol{\theta}^*|\mathbf{y}_i,\boldsymbol{\omega}). \tag{6.10}$$

Consequently, if $\nabla_{\boldsymbol{\omega}} \log f(\mathbf{y}_i|\boldsymbol{\omega})$ can be calculated from Equation (6.10), the quantity may then be used in Equation (6.9) to approximate the observed information. The first two terms on the right-hand side of Equation (6.10) are simply complete data gradients evaluated at $\boldsymbol{\theta}^*$, and the MH-RM code may be re-used. The last term on the right-hand side requires taking derivatives of Equation (6.7) with respect to $\boldsymbol{\omega}$. The necessary derivatives are provided in Appendix B.

Let this estimator of the observed information be $\mathcal{F}_{CJ}$. If priors are imposed on any of the free parameters in $\boldsymbol{\omega}$, then it is necessary to adjust $\mathcal{F}_{CJ}$ to reflect the additional information. This can be done in the following manner. Let ω be an arbitrary element of $\boldsymbol{\omega}$, with prior density $\pi(\omega)$. Then, to reflect the additional information contained in the prior, $-\partial^2\pi(\omega)/\partial\omega^2$ should be added to the corresponding diagonal element of $\mathcal{F}_{CJ}$.

CHAPTER 7

SNP-MIRT Simulation Study Design

This chapter outlines the Monte Carlo simulation study design used to investigate the major features of the proposed methods regarding the SNP-MIRT model. For any study design, the number of factors and corresponding levels must be managed so that the study may be completed in a timely fashion. Such is the case here, and conditions were chosen so that the study results would address the main research questions. However, this necessarily leaves many interesting questions unanswered. To that end, further simulation work should be pursued.

7.1 Goals

There were several goals for the simulation study. First, the simulation was designed to assess the implementation of the SNP-MIRT model. This is primarily a question of parameter recovery, but other features of the implementation, such as the stability of the algorithm, are also of importance. Evidence of accurate parameter recovery may also be taken as evidence that the proposed SNP parameterization is valid. Assuming the model can be reliably estimated, then attention can turn to comparing the SNP-MIRT model with the standard MIRT model, which assumes a normal. These comparisons are facilitated by accurate likelihood and information matrix estimates. Thus, a final goal of the simulation is to evaluate the proposed Chib and Jeliazkov (2001) estimators.

7.2 Simulation Conditions

Overall, there were 8 conditions, which differed in the dimensionality and shape of the latent variable density. For each condition, the dimensionality of the latent variable was $p = 1$, 2, or 3. At each of these levels, one condition corresponded to a normal density. Nonnormal densities were constructed as mixtures of normals. The definitions for all generating densities is given in Table 7.2. At the $p = 2$ level, one condition involved 2 groups. For this condition, the latent variable density of the reference group was defined as normal with null mean vector and identity covariance matrix, whereas the latent variable density of the comparison group was defined as nonnormal with mixture-implied $\boldsymbol{\mu} = (-0.46, -0.12)'$ and $\boldsymbol{\sigma} = (1.57, 0.12, 1.38)'$.

In addition to the mixture definitions, the unidimensional generating densities are displayed in Figure 7.1. Univariate and bivariate marginal plots of the generating multidimensional densities will be provided in the Chapter 8, which presents simulation results.

For all conditions, data were generated for $n = 20$ graded items, with a sample size of $N = 1000$. These figures also apply to each group for the 2-group condition. Further, for the 2-group condition, the same item parameters were used to generate data for both groups. Thus, the two groups only differed in their respective latent variable densities.

The slope parameters were drawn from a normal $(1.8, 0.64)$ distribution, truncated at 0.5 and 4. For each item, the first threshold parameter was drawn from a normal $(-1, 0.5)$. The second threshold was obtained by adding a normal $(1, 0.04)$ to the first threshold. This method was repeated to obtain the remaining thresholds. Finally, intercepts were taken as the negative products of the slopes and thresholds. Table 7.2 presents the item parameter values for all conditions.

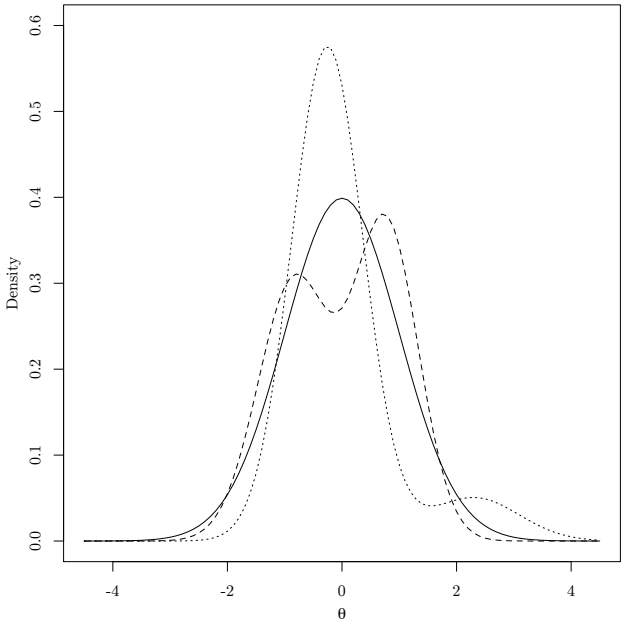
For a given level of p, all conditions implemented the same factor structure.

Table 7.1: Simulation Study: Mixture Definitions for Generating Densities

p	Shape	mp_1	$\boldsymbol{\mu}_1$	$\boldsymbol{\Sigma}_1$	mp_2	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_2$	mp_3	$\boldsymbol{\mu}_3$	$\boldsymbol{\Sigma}_3$
1	Normal	1	0	1						
	Skewed	0.90	-0.26	0.39	0.10	2.30	0.62			
	Bimodal	0.48	-0.83	0.40	0.52	0.77	0.32			
2	Normal	1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & \\ 0.40 & 1\end{pmatrix}$						
	Normal (G1)	1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & \\ 0 & 1\end{pmatrix}$						
	Skewed (G2)	0.65	$\begin{pmatrix}-1.20\\0.20\end{pmatrix}$	$\begin{pmatrix}0.30 & \\ 0 & 1.50\end{pmatrix}$	0.25	$\begin{pmatrix}0.50\\0\end{pmatrix}$	$\begin{pmatrix}0.60 & \\ 0 & 1.20\end{pmatrix}$	0.10	$\begin{pmatrix}2\\0.10\end{pmatrix}$	$\begin{pmatrix}0.30 & \\ 0 & 0.90\end{pmatrix}$
	Bimodal	0.60	$\begin{pmatrix}0.70\\-0.14\end{pmatrix}$	$\begin{pmatrix}0.31 & \\ 0.07 & 1.11\end{pmatrix}$	0.40	$\begin{pmatrix}-1.06\\0.21\end{pmatrix}$	$\begin{pmatrix}0.18 & \\ 0.27 & 0.77\end{pmatrix}$			
3	Normal	1	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1 & & \\ 0.50 & 1 & \\ 0.60 & 0.40 & 1\end{pmatrix}$						
	Bimodal	0.80	$\begin{pmatrix}0.39\\0\\0.19\end{pmatrix}$	$\begin{pmatrix}0.38 & & \\ 0 & 1 & \\ -0.30 & 0 & 0.85\end{pmatrix}$	0	$\begin{pmatrix}-1.57\\0\\-0.76\end{pmatrix}$	$\begin{pmatrix}0.38 & & \\ 0 & 1 & \\ -0.30 & 0 & 0.85\end{pmatrix}$			

Note. p = dimension of density; (G1) = Group 1 for 2-group model; (G2) = Group 2 for 2-group model; mp = mixing proportion; $\boldsymbol{\mu}$ = mean; $\boldsymbol{\Sigma}$ = variance. All subscripts refer to mixture component.

Figure 7.1: Unidimensional Generating Densities



Note. The solid line traces a normal density. The dashed line traces a "bimodal" density. The dotted line traces a "skewed" density.

For the 2-dimensional model, an independent cluster factor pattern was defined, with items $1-10$ loading on θ_1 and items $11-20$ loading on θ_2. For the 3-dimensional model, items $1-6$ load on θ_1, items $7-12$ load on θ_2, and items $13-18$ load on θ_3. Finally, items 19 and 20 load on all 3 latent dimensions.

Table 7.2: Simulation Study: Generating Item Parameter Values

	Intercepts				
Item	1	2	3	4	Slope
1	1.71	−0.32	−2.59	−4.56	1.89
2	2.37	0.88	−0.66	−1.90	1.44
3	3.71	2.50	1.42	0.24	1.45
4	3.20	0.60	−1.24	−2.97	2.18
5	3.28	1.42	−0.25	−2.69	1.81
6	1.07	0.58	0.20	−0.27	0.50
7	0.37	−0.48	−1.56	−2.66	0.89
8	1.00	−0.22	−1.24	−2.37	0.94
9	1.09	−0.32	−1.30	−2.60	1.30
10	2.05	0.20	−1.17	−3.17	1.56
11	2.28	0.51	−1.05	−1.82	1.37
12	5.04	2.11	−2.37	−5.47	3.18
13	3.15	1.05	−1.16	−2.65	2.02
14	3.93	1.97	0.44	−1.23	2.34
15	1.03	−1.06	−2.70	−4.46	1.63
16	1.39	−0.31	−1.33	−2.93	1.41
17	0.62	−0.19	−1.13	−2.22	0.75
18	1.35	−0.37	−2.31	−3.90	1.99
19	1.67	−0.52	−2.60	−4.14	1.85
20	1.06	−0.86	−2.99	−5.14	1.96

Note. For 2-dimensional models, items $1-10$ load on θ_1 and items $11-20$ load on θ_2. For 3-dimensional models, items $1-6$ load on θ_1, items $7-12$ load on θ_2, items $13-18$ load on θ_3, and items 19 and 20 load on θ_1, θ_2, and θ_3.

7.3 Components of a Replication

For each condition, 200 Monte Carlo replications were attempted. Given the multiple goals of the simulation, each replication incorporated several steps. These

were:

1. Fit the standard MIRT model using MH-RM, yielding the ML estimate $\tilde{\boldsymbol{\omega}}$. Obtain CJ estimates of the log-likelihood and information matrix.

2. Obtain EAP score estimates, $\tilde{\boldsymbol{\Theta}}$, using $\tilde{\boldsymbol{\omega}}$.

3. Use $\tilde{\boldsymbol{\Theta}}$ to obtain starting values for the SNP shape parameters, $\boldsymbol{\nu}_0$.

4. Use the starting values, $\tilde{\boldsymbol{\omega}}$ and $\boldsymbol{\nu}_0$, to fit the SNP-MIRT model using MH-RM, yielding the ML estimate $\hat{\boldsymbol{\omega}}$.

5. Obtain CJ estimates of the log-likelihood and information matrix.

The first step is necessary as one of the simulation goals is to compare $\hat{\boldsymbol{\omega}}$ with $\tilde{\boldsymbol{\omega}}$. For the second step, 50 posterior draws (with a thinning interval of 10) were averaged to obtain $\tilde{\boldsymbol{\Theta}}$. Finally, for the CJ estimates, a Monte Carlo sample size of 1000 was used, with no thinning interval.

The SNP tuning parameter, κ was held fixed within a condition. In practice, several values for κ should be fit, with model selection then based on conventional indices, such as AIC, BIC, or HQIC. Such approaches have been found useful in other studies of nonnormal density estimation in IRT (e.g., Woods & Thissen, 2006). However, here, κ is held fixed to limit the scale of the simulation study. The number of free SNP parameters for each condition is given in Table 7.3. Also, to add stability to estimation of the SNP parameters, a multivariate normal prior was implemented, as detailed in Section 4.6. After some trial and error, a value of $\tau = 1$ was selected and used for all conditions.

To obtain starting values for the SNP parameters, the DE algorithm described in Section 5.3.4 was used. In particular, the 'DEoptim' function in the R package of the same name (Mullen, Ardia, Gil, Windover, & Cline, 2011) was used, with default arguments.

Table 7.3: SNP Tuning Constants and Number of Free Parameters

p	Shape	κ	$\dim(\boldsymbol{\nu})$	$\dim(\boldsymbol{\psi})$
1	Normal	5	5	5
	Bimodal	5	5	5
	Skewed	5	5	5
2	Normal	3	9	10
	Bimodal	3	9	9
	2 Group	2	5	10
3	Normal	2	9	12
	Bimodal	2	9	9

Note. p = Number of dimensions; κ = SNP tuning constant.

Also, when the generating model was normal (for all p), several other statistics were collected to study the CJ estimators. All of these statistics were based on the ML estimate from the standard MIRT model, $\tilde{\boldsymbol{\omega}}$. First, the observed information was approximated via the Louis (1982) formula in the fashion of Diebolt and Ip (1996). For this estimate, a Monte Carlo sample size of $5,000$ was used, with a thinning interval of 10. The likelihood was estimated by standard Monte Carlo, with a sample size of $25,000$.

Further, quadrature-based estimates of the information and likelihood were obtained to serve as references for both the CJ and conventional estimators. Specifically, the information matrix was estimated using a quadrature-based gradient cross-product approximation, as in Equation (6.9). The likelihood was estimated using the quadrature-based E-step tables in an EM estimation scheme. All of the quadrature-based estimates used 49 points per dimension, from -6 to 6, and were calculated using flexMIRT®(Cai, 2013). For the 3-dimensional models, these computations were relatively time-consuming, as might be expected.

7.4 Collected Statistics

The bias and root mean square error (RMSE) were collected for the parameter estimates, $\tilde{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\omega}}$. For an arbitrary parameter ω and a corresponding estimate, say $\hat{\omega}$, bias is defined as

$$\text{bias}(\omega) = M^{-1} \sum_{m=1}^{M} (\omega - \hat{\omega}_m), \tag{7.1}$$

where M is the number of Monte Carlo replications. RMSE is defined as

$$\text{RMSE}(\omega) = \sqrt{M^{-1} \sum_{m=1}^{M} (\psi - \hat{\psi}_m)^2}. \tag{7.2}$$

The SNP parameter estimates are not themselves interpretable, but the degree to which the parameters recover the generating density may still be assessed. The most straightforward approach is graphical. Using a quadrature grid, the mean estimate of the SNP density, across replications, may be plotted and compared to the generating density. Similarly, empirical confidence intervals may be constructed. In the results, a 90% confidence interval is used for the SNP estimated densities.

The information matrix estimates were evaluated via the implied standard errors. Let ω be an arbitrary parameter and $\hat{\omega}$ be its estimate. Also, let $se(\hat{\omega})$ be the estimated standard error for $\hat{\omega}$. Then, the Monte Carlo standard deviation of the estimates, $SD(\hat{\omega})$ may be compared to the mean of the standard error estimates $\text{E}(se(\hat{\omega}))$. If the standard error estimates accurately reflect the sampling variability, then the two quantities should be similar. Additionally, the parameter estimate standard errors may be used to construct confidence-intervals. Then, the empirical coverage rate (ECR) of the confidence intervals may be examined. In the results, ECRs for 95% confidence intervals are presented.

CHAPTER 8

Simulation Study Results

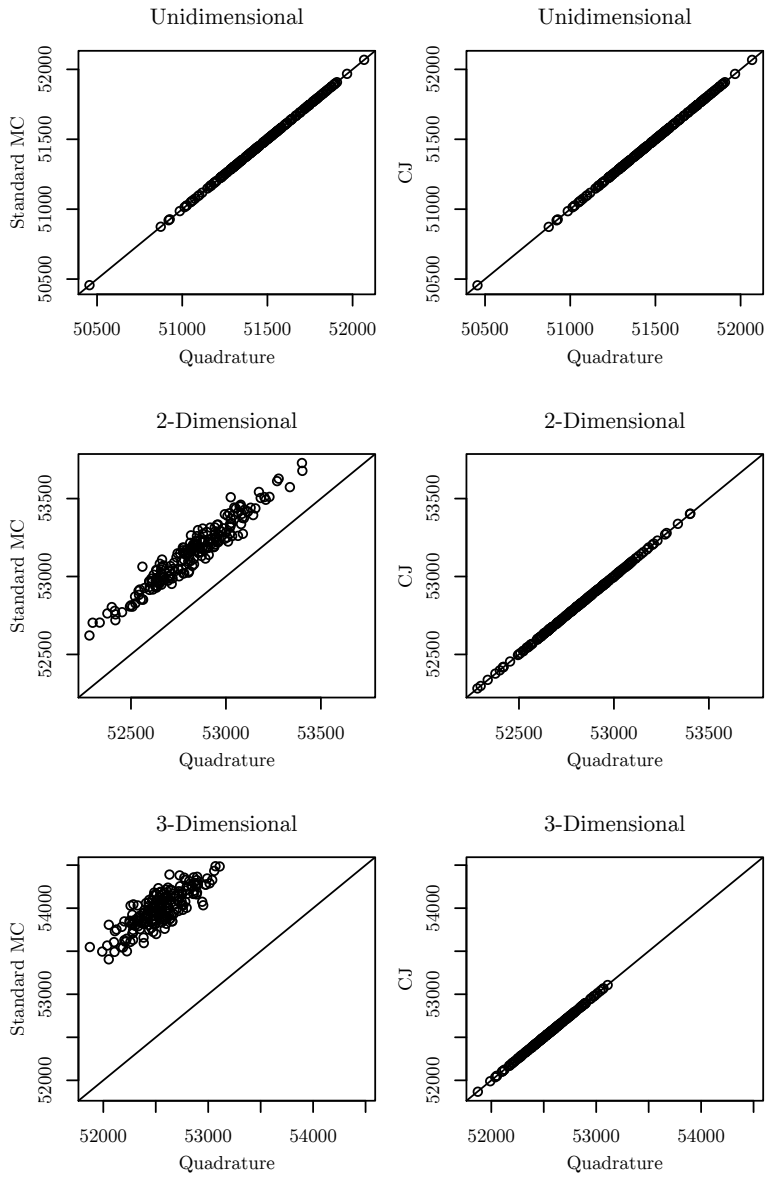
The results of primary interest relate to the SNP-MIRT model. However, results relating to the CJ estimators will be presented first, since the estimators may then be used to evaluate the SNP-MIRT model.

8.1 Chib and Jeliazkov (2001) Likelihood

Like the accuracy of the standard Monte Carlo approach, the accuracy of the quadrature-based approach will deteriorate as the number of items or dimensions grows. However, with $p = 3$, the specification of 49 points per dimension results in over $100,000$ points, a relatively fine grid. (Recall that the Monte Carlo sample size for the standard approach is $25,000$). Consequently, the quadrature-based calculation may be considered reasonably accurate, and both the standard Monte Carlo and CJ approaches may be compared to it. Figure 8.1 makes these comparisons for the three normal generating densities. For the unidimensional model, there is no appreciable difference between any of the three estimators. However, for the 2-dimensional and 3-dimensional models, there are clear differences. The CJ estimates and quadrature-based estimates continue to correspond closely, while the standard Monte Carlo estimator appears positively biased.

Equation (6.8) presented an estimator for the CJ likelihood variance. Its ability to capture the variability can be explored by repeatedly estimating the likelihood for a single model, dataset, and MLE. The results shown in Table 8.1 are based on

Figure 8.1: $-2\times$Log-Likelihood Estimate Comparison



Note. Standard Monte Carlo estimates based on a sample size of $25,000$. CJ estimates based on a sample size of $5,000$, with a thinning interval of 10. Quadrature estimates based on 49 points per dimension.

200 replications. Of note, the bias is relatively small, which supports the intuition gained from Figure 8.1. Also, the MCSD and mean SE values correspond closely. This is even true for the estimates with no "thinning," suggesting that the batch means approach to addressing the correlation in the Monte Carlo draws is effective.

Table 8.1: CJ Estimates of $-2\times$Log-Likelihood for Three Simulated Datasets

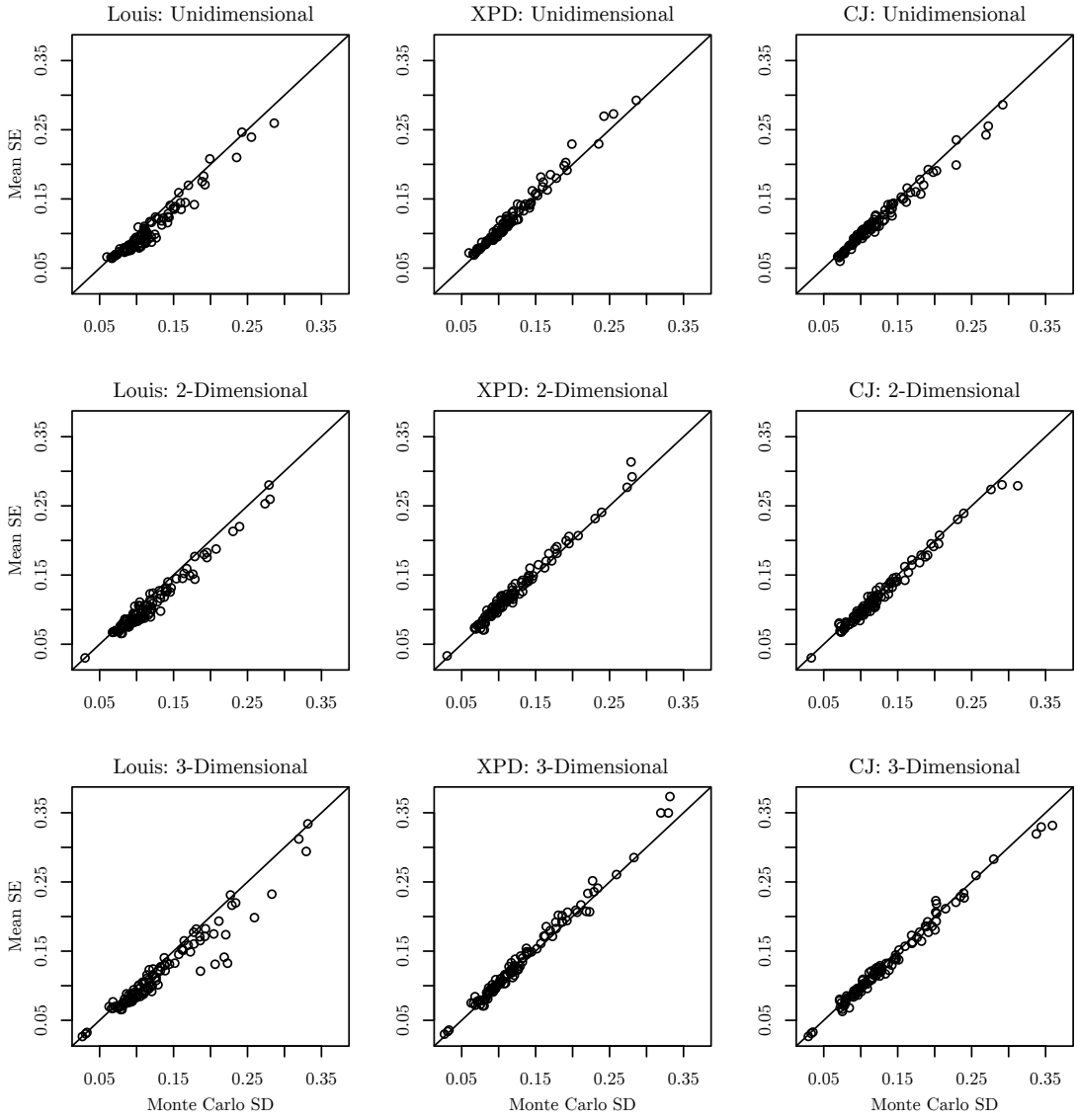
p	$-2 \times \text{LogL}_q$	Thin	Samples	Estimate	Bias	MCSD	Mean SE
1	51481.05	1	5000	51480.87	0.82	1.03	1.06
		10	1000	51482.20	2.16	2.47	2.36
2	52778.68	1	5000	52778.08	1.09	1.23	1.42
		10	1000	52780.62	2.87	2.94	2.94
3	52392.53	1	5000	52390.85	1.99	1.90	2.02
		10	1000	52394.91	3.64	3.67	3.84

Note. p = Number of dimensions; $-2 \times \text{LogL}_q$ is calculated by quadrature; Thin = thinning interval for posterior draws; Samples = Monte Carlo sample size; Estimate = mean of point estimates; Bias = absolute bias; MCSD = standard deviation of point estimates; Mean SE = mean of estimated SEs.

8.2 Proposed Information Estimator

Cai (2008a), appealing to multiple imputation theory, noted that when m_k (i.e., the number of imputations per MH-RM cycle) is small, the parameter estimate standard errors tend to have a downward bias. Simulation results in Cai (2008a) and Monroe and Cai (2014) support this claim, as do results from the present study. Figure 8.2 compares the standard error estimators, plotting $\text{E}(se(\hat{\omega}))$ against $SD(\hat{\omega})$ for the 3 normal generating models. Examining the left column of plots corresponding to the Louis (1982) estimator, the downward bias of the standard errors is slight but clear. This bias becomes more pronounced as the dimensionality of the generating model increases. On the other hand, neither the XPD nor CJ estimator appears to have any notable bias, for any level of p considered here.

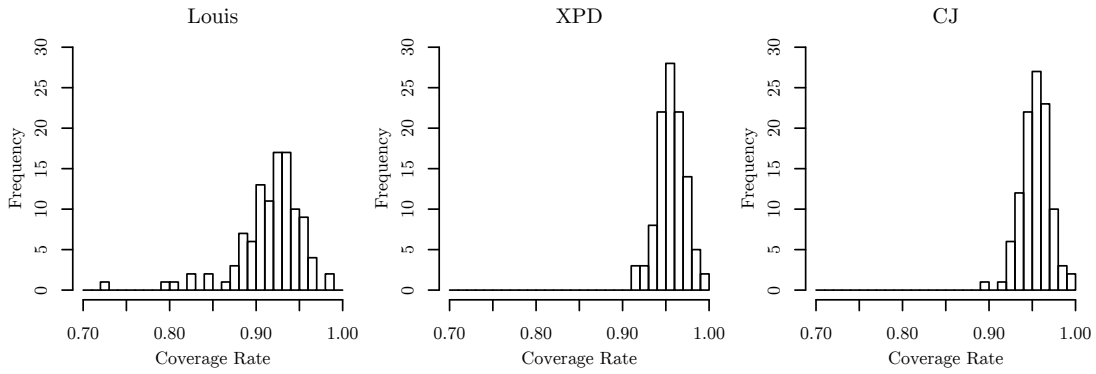
Figure 8.2: Standard Error Comparison for Conditions with a Normal Generating Model



In addition to evaluating E($se(\hat{\omega})$), the variability of $se(\hat{\omega})$ itself may be assessed. Following Tian et al. (2012), let the Monte Carlo standard deviation of the standard error estimates for an arbitrary parameter be $SD(se(\hat{\omega}))$. Then, lower values of $SD(se(\hat{\omega}))$ indicate more stability. Table 8.2 presents mean values of $SD(se(\hat{\omega}))$, under the column heading "E(MCSD)", for the three estimators. ECR statistics are also presented. Surprisingly, the Louis estimator has lower $SD(se(\hat{\omega}))$ values for all levels of p. This may, however, reflect an undesirable sort of stability. In other words, it is possible that the lower $SD(se(\hat{\omega}))$ values are a consequence of the downwards bias in the Louis estimated standard error estimates.

Examining to the ECR statistics, the XPD and CJ estimators have mean ECR rates quite close to the nominal level of 0.95, while the Louis estimator is again biased downwards. This is more apparent for the 3-dimensional model, where the 2.5% empirical percentile is 0.80, compared to 0.93 for both the XPD and CJ estimators. These particular results may be presented graphically using histograms, as in Figure 8.3. The XPD and CJ histograms indicate that the derived standard errors are well-calibrated, particularly in comparison to the Louis standard errors.

Figure 8.3: Empirical Coverage Rates of 95% Confidence Intervals



Note. Generating model is 3-dimensional normal. *Louis* = Louis (1982) SE, XPD = quadrature-based empirical cross-product SE, CJ = Chib and Jeliazkov (2001) SE. Histograms based on 200 replications.

Table 8.2: Information Matrix Simulation Results

			Coverage Rate		
p	Method	E(MCSD)	Mean	2.5%	97.5%
1	XPD	5.28	0.96	0.93	0.98
	Louis	3.91	0.92	0.86	0.96
	CJ	5.28	0.96	0.93	0.98
2	XPD	6.01	0.96	0.93	0.98
	Louis	4.65	0.92	0.88	0.96
	CJ	5.99	0.96	0.93	0.98
3	XPD	7.99	0.96	0.93	0.99
	Louis	6.22	0.92	0.80	0.99
	CJ	7.81	0.96	0.93	0.99

Note. Generating models are normal. p = Number of dimensions; E(MCSD) = mean of Monte Carlo standard deviations of parameter estimate standard errors (multiplied by 1000); Coverage Rate = empirical coverage rate of 95% confidence intervals; Mean = mean parameter ECR; 2.5% and 97.5% are empirical percentiles.

It is perhaps unsurprising that the XPD and CJ results are similar, as both estimators are based on Equation (6.9). The estimators only differ in how the necessary quantities are approximated. This difference, however, is an important distinction for high-dimensional models. The XPD estimator is quadrature-based, and is therefore subject to the "curse of dimensionality." On the other hand, the CJ estimator is "posterior-based" in the sense that the estimator depends on approximations of $\Pi(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$, and its gradient. Importantly, this approximation relies on the M-H sampler, and not quadrature. Consequently, the CJ estimator would seem to be a viable general method for high-dimensional MIRT models in general, and a natural complement to the MH-RM algorithm in particular.

8.3 SNP-MIRT Model

The simulation results for the SNP-MIRT model will be presented in the following order. First, the unidimensional results will be presented. These results should be

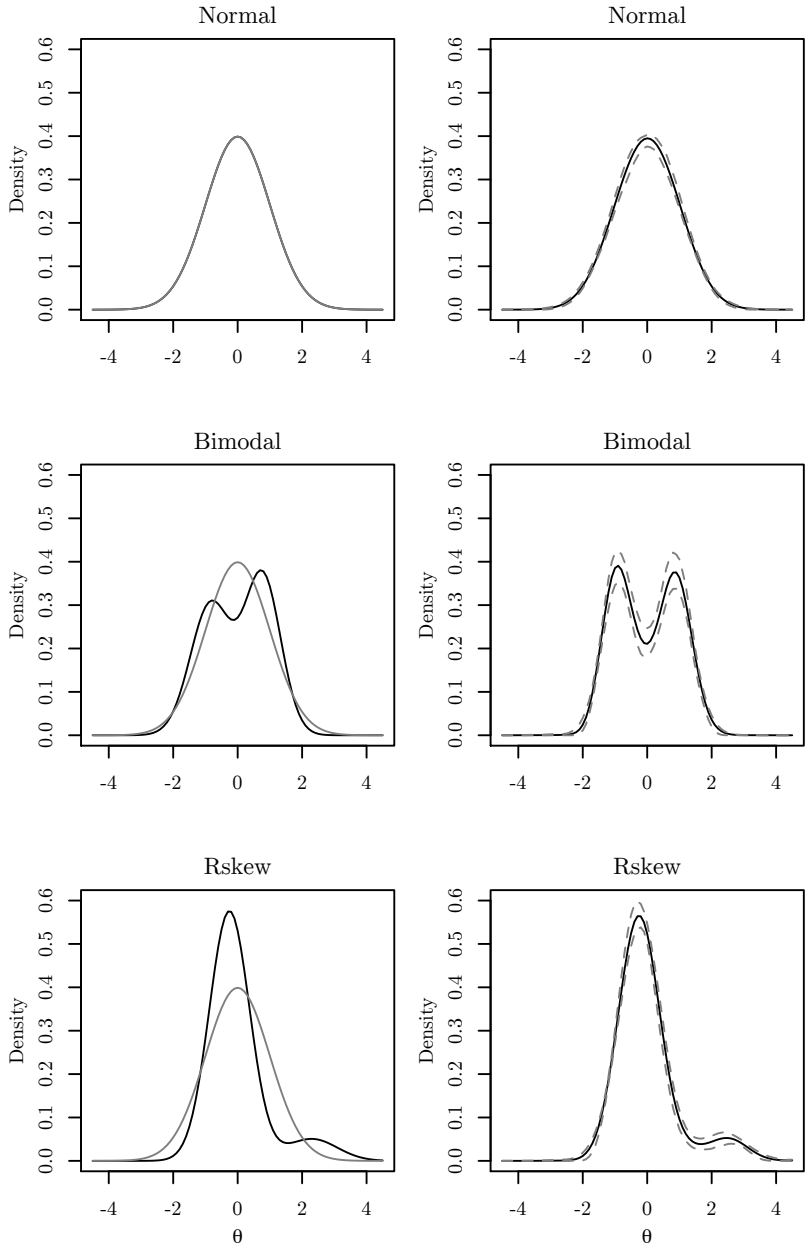
consistent with results from similar research efforts, in particular Woods and Lin (2009). Second, the results for the normal generating density conditions will be presented. Ideally, for these conditions, the SNP-MIRT parameter estimates and log-likelihood values will be comparable to those from the standard MIRT estimation. Also, the parameterization of the mean and variance parameters for the SNP density may be assessed, since the 2-dimensional and 3-dimensional generating densities have correlated dimensions. Finally, the results for the nonnomral multidimensional density conditions will be presented. Most importantly, these conditions provide an opportunity to assess the ability of the SNP-MIRT model to recover the generating densitites.

8.4 Unidimensional Models

Generally, the unidimensional model results correspond to those in Woods and Lin (2009). Figure 8.4 shows the mean estimated SNP curves, as well as 90% empirical confidence intervals. Overall, the estimated densities compare favorably to the generating densities. Notably, the mean estimated bimodal density is more symmetric than the generating density. This lack of accuracy in recovery for the bimodal model is unsurprising, as prior research (e.g., Woods & Lin, 2009; Monroe & Cai, 2014) has produced similar findings. The mean estimate for the skewed, however, closely approximates the generating density.

Table 8.3 presents bias and RMSE statistics for the unidimensional conditions, as well as the log-likelihood values. The bias and RMSE statistics are largely comparable for the SNP-MIRT and standard MIRT models. The most notable difference between the results for the models is seen for the skewed generating density, where the SNP-MIRT parameter estimates have slightly less bias than the standard MIRT counterparts. For both the bimodal and skewed generating densities, the log-likelihood values clearly indicate that, on average, the SNP-

Figure 8.4: Mean SNP Estimated Densities for Unidimensional Models



Note. For plots in the left column, the gray solid line is a normal density, included as a reference. For plots in the right column, the gray dashed lines gives a 90% empirical confidence interval.

MIRT model provides better fit.

Table 8.3: Simulation Results for Unidimensional SNP-IRT

Model			Slopes		Intercepts	
Generating	Estimated	−2×LogL	Bias	RMSE	Bias	RMSE
Normal	Standard	51473.92	0.01	0.10	0.01	0.12
	SNP	51473.09	0.01	0.10	0.01	0.12
Bimodal	Standard	51418.27	0.02	0.09	0.02	0.12
	SNP	51293.78	0.01	0.09	0.01	0.12
Skewed	Standard	51650.37	0.05	0.12	0.04	0.13
	SNP	51407.39	0.01	0.11	0.01	0.13

Note. Bias = absolute bias.

8.5 Multidimensional Normal Generating Models

When the simulated latent variables are sampled from a normal density, then the standard MIRT model is not misspecified. Typically, with a sufficiently large sample and number of items, the model should then fit simulated data well. For it to be useful, the SNP-MIRT model should also fit these data well. Theoretically, this should be the case since the standard MIRT model is nested within the SNP-MIRT model.

Figure 8.5 presents the log-likelihood values for all of the normal generating conditions, for both the SNP-MIRT and standard MIRT models. Clearly, the two models provide nearly identical fit for these conditions. The item parameter bias and RMSE are not presented here, but these statistics are presented for the group parameters. The motivation for presenting these specific results is that they are indictative of the validity of the proposed SNP parameterization. These results are presented in Table 8.4. Surprisingly, the SNP estimates appear to be less biased than those from the standard model, though the absolute difference is likely inconsequential. In any case, the SNP-MIRT model provides accurate

estimates of the correlations among the latent variable dimensions.

Figure 8.5: $-2\times$Log-Likelihood Values for Normal-Generating Models

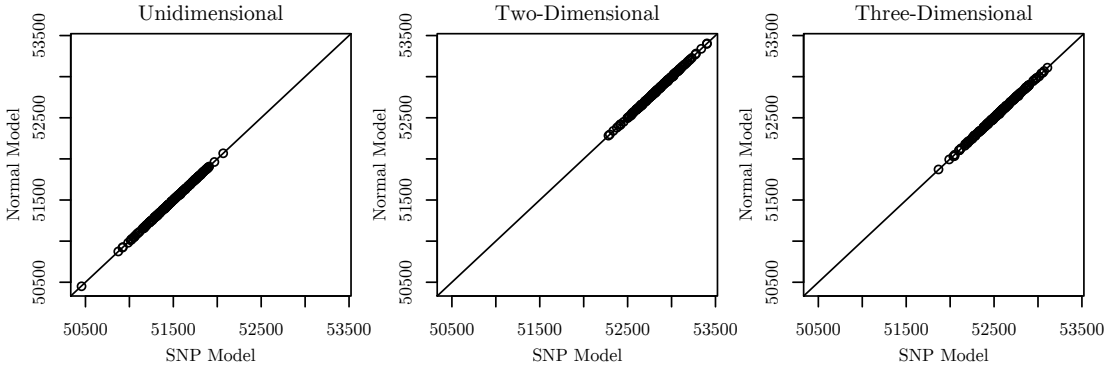


Table 8.4: Correlation Estimates for Normal Generating Models

			Estimated Model			
			Standard		SNP	
p	Parameter	Value	Bias	RMSE	Bias	RMSE
2	σ_{21}	0.40	0.013	0.033	0.001	0.031
3	σ_{21}	0.50	0.012	0.094	0.005	0.110
	σ_{31}	0.60	0.014	0.188	0.005	0.206
	σ_{32}	0.40	0.006	0.033	0.007	0.034

Note. Bias = absolute bias.

8.6 Multidimensional Nonnormal Generating Models

Figure 8.6 presents the generating density (center column) as well as the standard and SNP model estimates for the bimodal 2-dimensional density. Looking at the bivariate contours, the mean SNP estimate is a reasonable approximation of the generating density, clearly capturing the bimodality. Similarly, the univariate plots show the ability of the SNP model to recover the univariate shapes. For

the 3-dimensional skewed density, the generating bivariate contours are presented in Figure 8.7. Again, the SNP model is able to recover the general shape if not the precise features. For the 3-dimensional model, the 3 univariate plots may also be examined, which are given in Figure 8.8. Among the 3 dimensions, only θ_1 is clearly nonnormal.

Table 8.5 presents some results for these two conditions. The SNP models clearly fit better, as evidenced by the log-likelihood values. And, as with other conditions, the SNP parameter estimates exhibit slightly less bias than the standard counterparts. The RMSE values, though, are largely comparable.

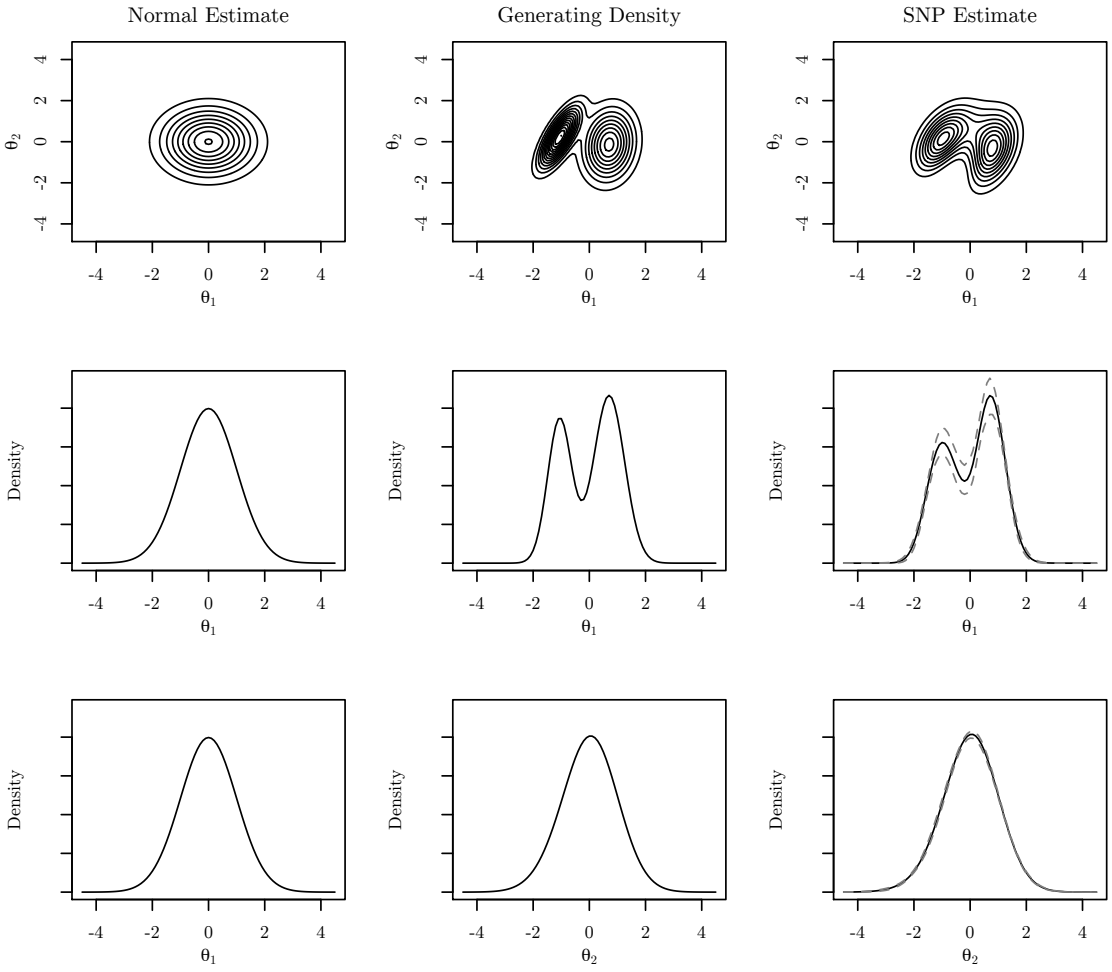
Table 8.5: Simulation Results for Multidimensional Nonnormal Models

	Model			Slopes		Intercepts	
p	Generating	Estimated	$-2\times$LogL	Bias	RMSE	Bias	RMSE
2	Bimodal	Normal	52786.15	0.015	0.107	0.015	0.127
		SNP	52646.65	0.016	0.105	0.010	0.126
3	Nonormal	Normal	52690.32	0.027	0.125	0.026	0.136
		SNP	52572.28	0.022	0.123	0.016	0.132

For the 2-Group condition, Figure 8.9 displays the generating contours and univariate margins, as well as the SNP and standard model estimates. The graphical results are similar to those results presented for the other conditions. However, of note, for this condition, the means and variances of the density were estimated. Though it is difficult to gauge from Figure 8.9, the standard method mean estimate for σ_{11} is biased upwards. The relevant numerical results are presented in Table 8.6. Along with the bias of standard model estimates of σ_{11}, the ECR rates for two standard estimates are relatively low. The ECR for μ_1 is only 0.89, while the ECR for σ_{11} is 0.92. On the other hand, the ECRs for the SNP model are all at the nominal level or better.

Figure 8.10 plots the σ_{11} estimates for the SNP and standard estimates, which clearly show the bias associated with the standard model. Finally, in Figure 8.10,

Figure 8.6: Density Estimates for Bimodal Two-Dimensional Model



Note. Standard and SNP contours and univariate densities based on means of replication estimates.

Figure 8.7: Density Estimates for Skewed Three-Dimensional Model

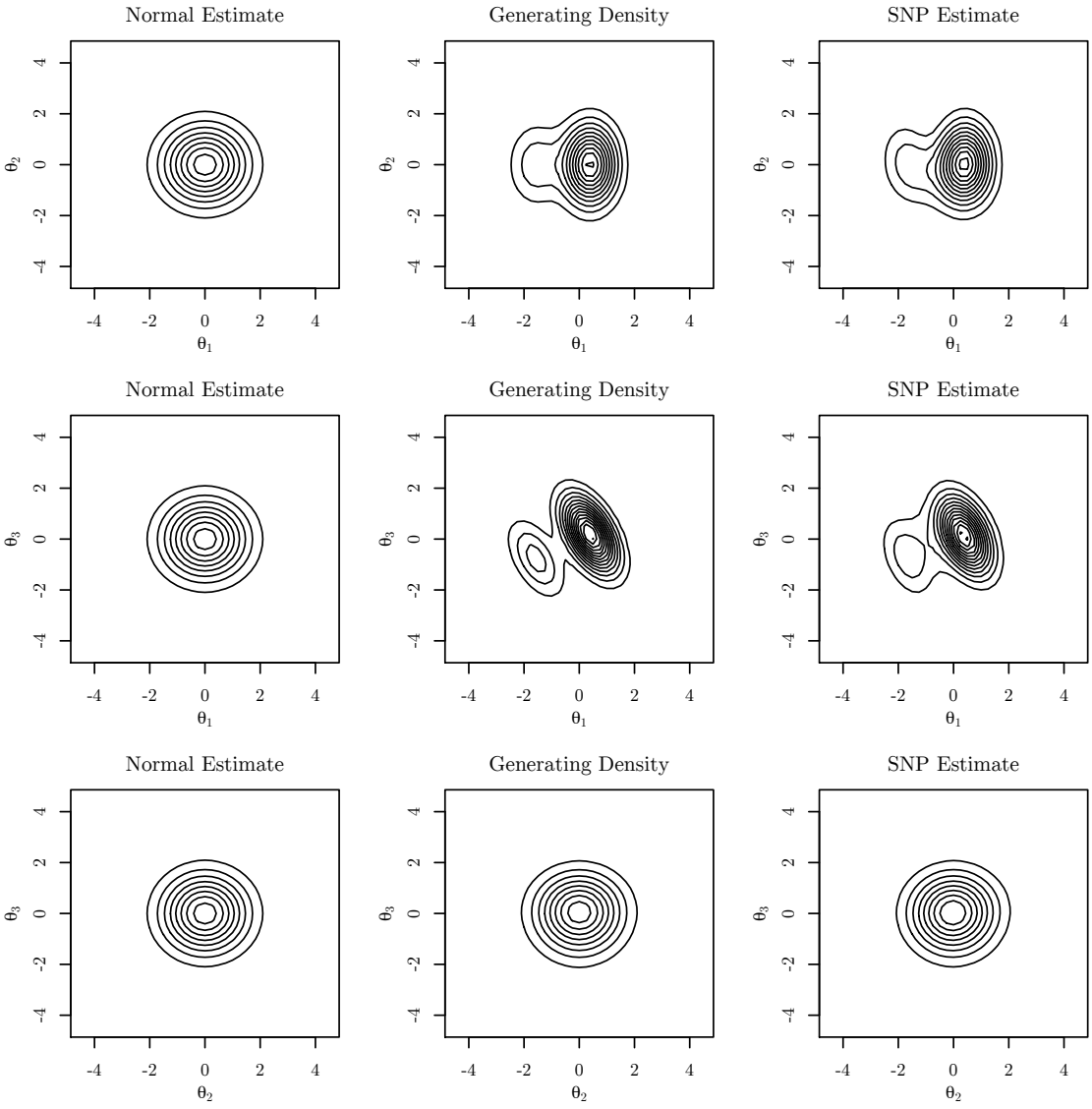


Figure 8.8: Univariate Margins for Skewed Three-Dimensional Density

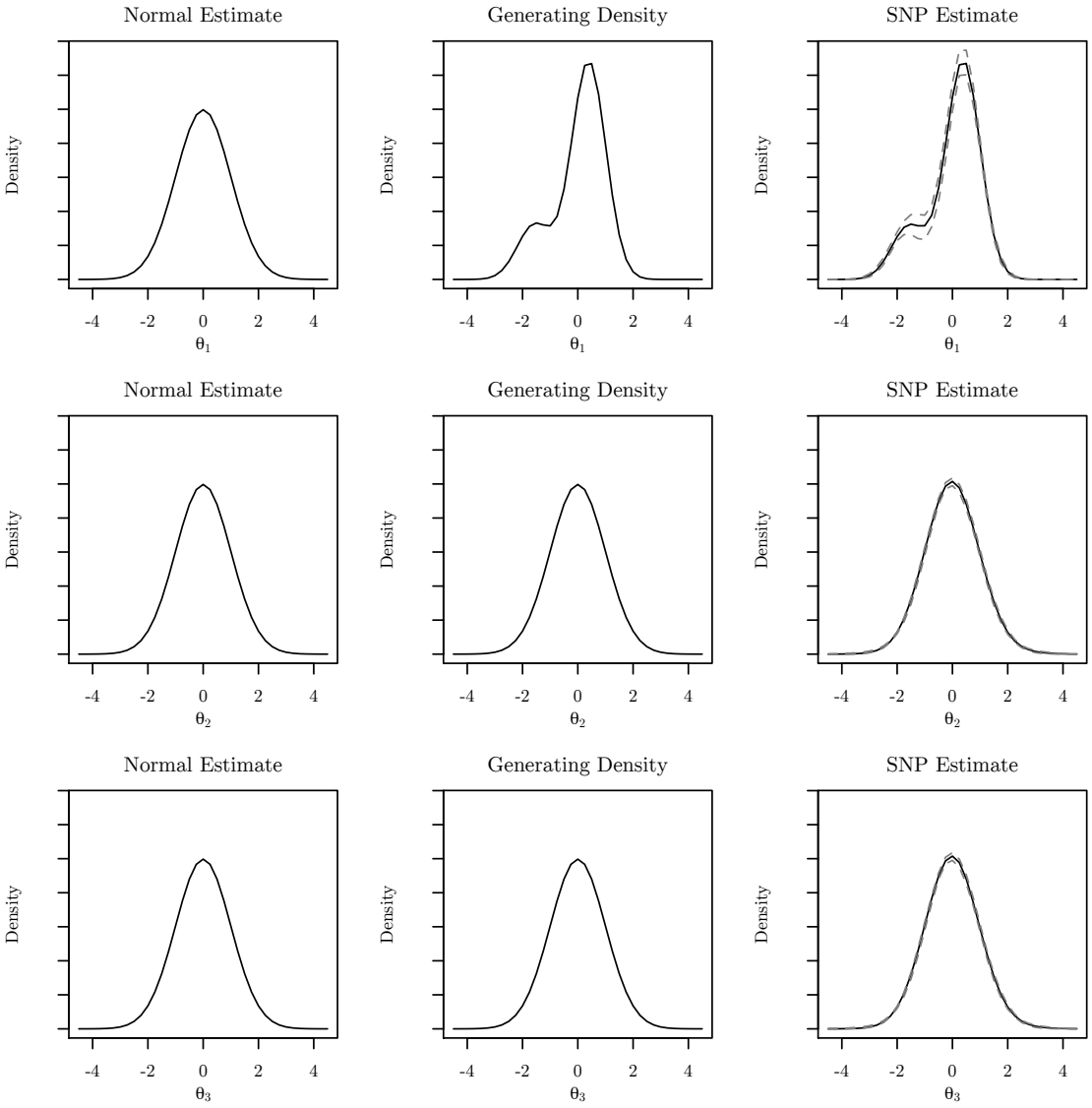


Figure 8.9: Contour Plots for Group 2 Skewed 2-Dimensional Density

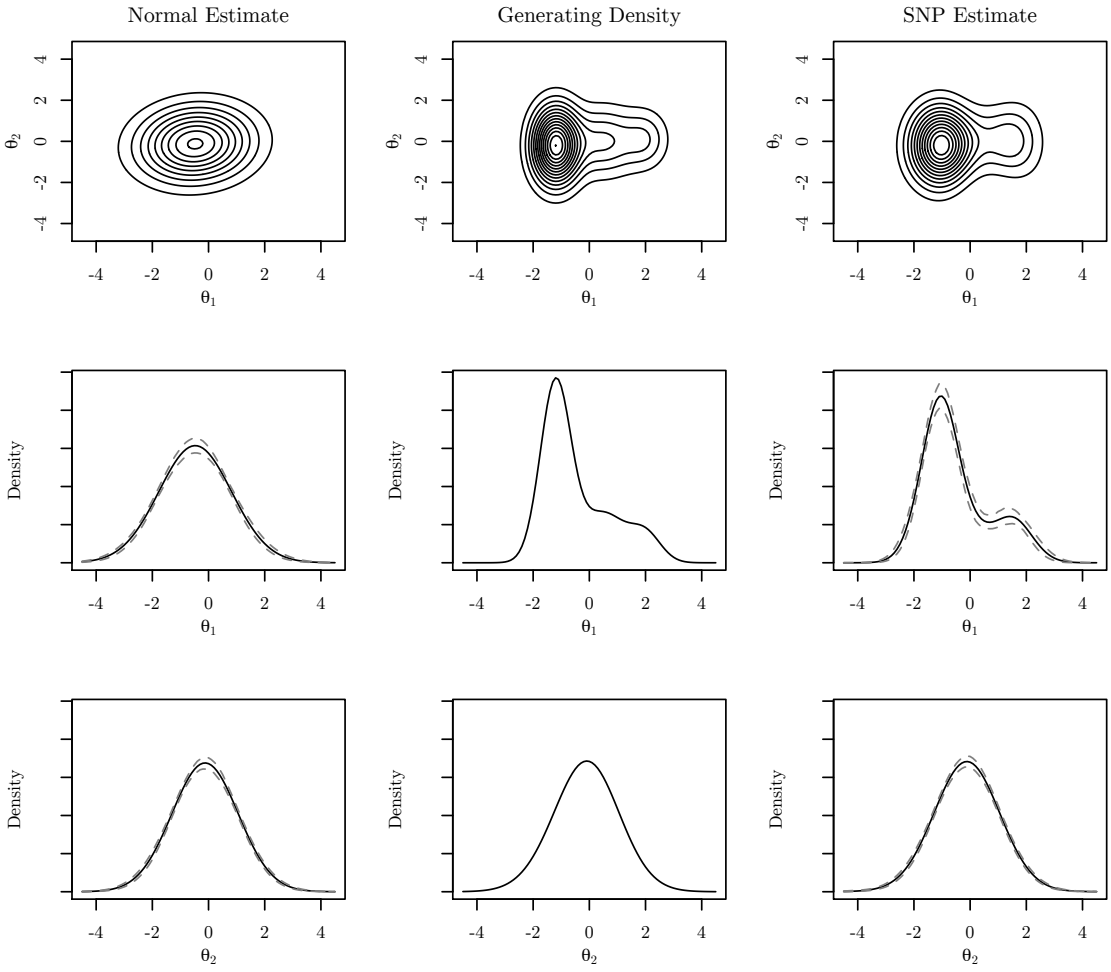


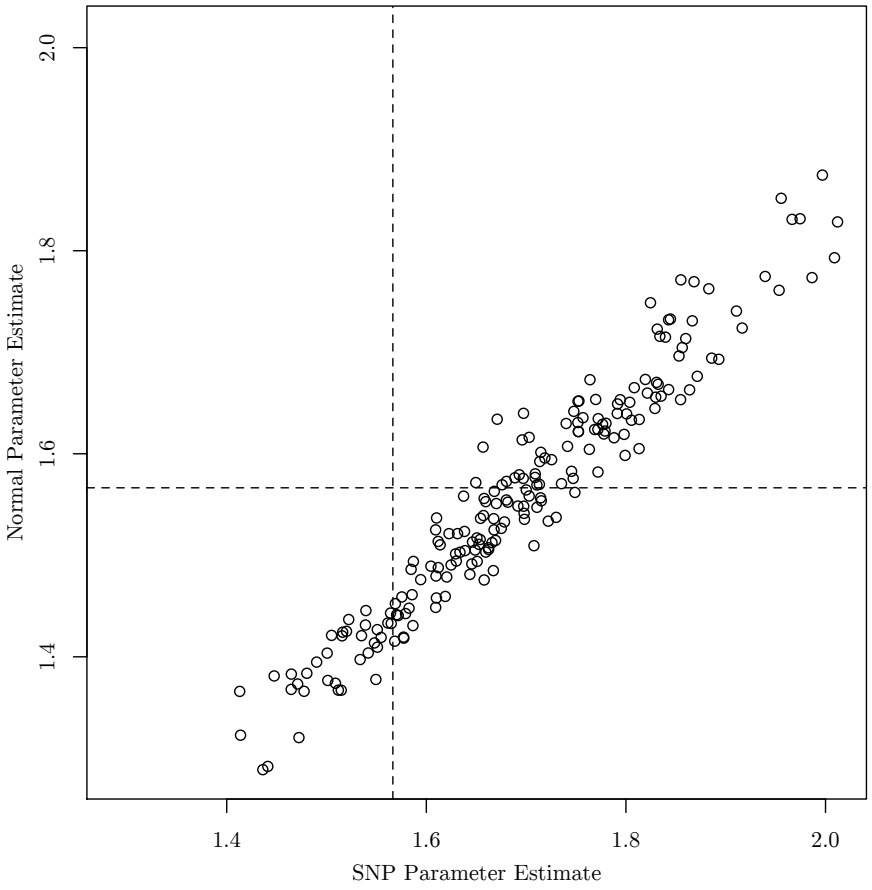
Table 8.6: Parameters Estimates: 2-Group 2-Dimensional Model

		Estimated Model					
		Standard			SNP		
Parameter	Value	Bias	RMSE	ECR	Bias	RMSE	ECR
Slopes		0.011	0.079	0.939	0.007	0.077	0.954
Intercepts		0.011	0.098	0.950	0.005	0.097	0.955
μ_1	-0.455	0.019	0.047	0.885	0.009	0.041	1.000
μ_2	-0.120	0.004	0.033	0.995	0.003	0.033	1.000
σ_{11}	1.566	0.125	0.180	0.915	0.010	0.118	0.950
σ_{21}	0.121	0.002	0.048	0.995	0.010	0.048	0.985
σ_{22}	1.378	0.019	0.076	0.990	0.005	0.071	1.000

Note. Bias = absolute bias. ECR = empirical coverage rate of 95% confidence intervals.

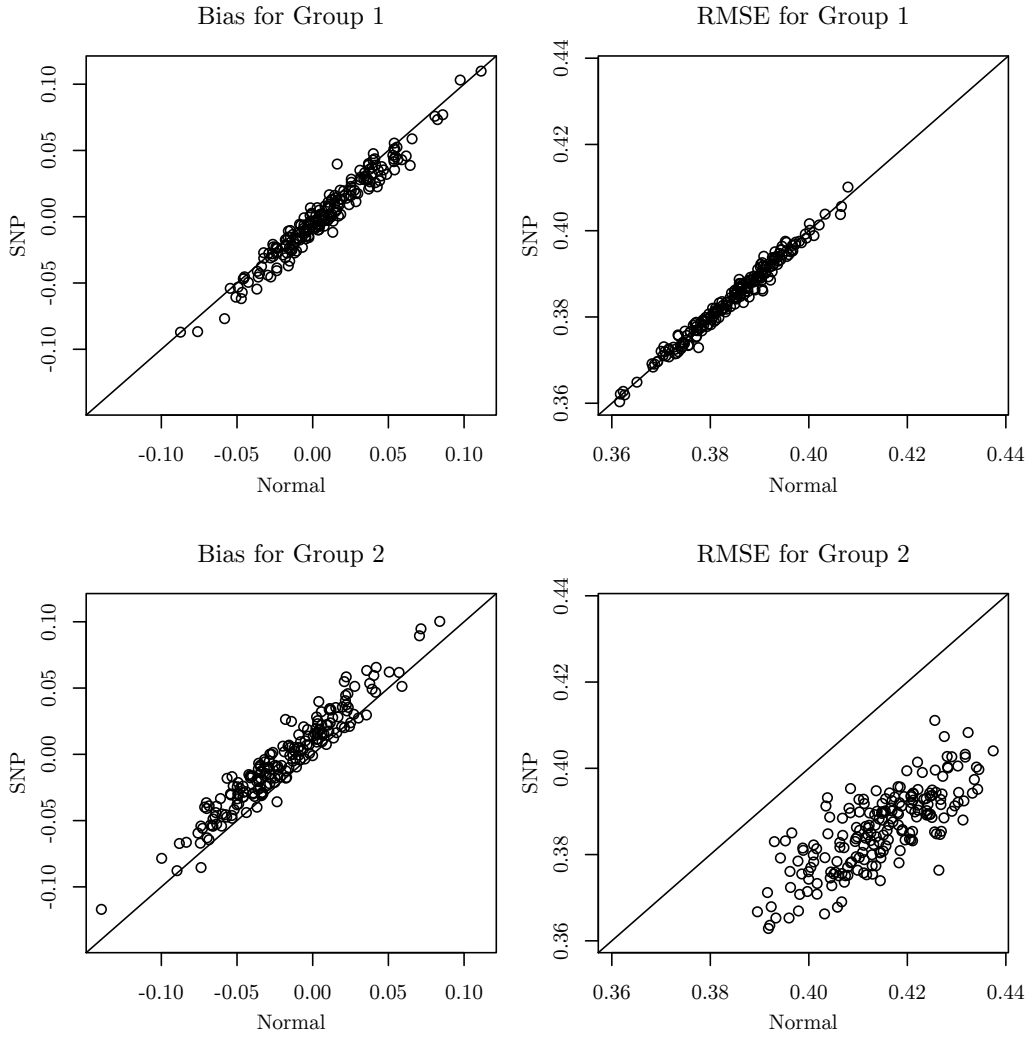
EAP estimates for the 2 groups are compared, by replication. EAP estimates were computed using the sample average of 100 draws from the posterior, with a thinning interval of 10. For Group 1, the scores are of similar quality. However, for Group 2, the RMSE values for the EAP estimates from the standard model are clearly greater than the RMSE values from the SNP model. Clearly, the standard model did not recover the Group 2 density particularly well, which led to poorer estimates for several quantities of interest.

Figure 8.10: Estimates of σ_{11} for Group 2 in 2-Group Model



Note. Dashed lines indicate true parameter value.

Figure 8.11: EAP Scoring for θ_1: 2-Group Model



CHAPTER 9

Empirical Application for the SNP-MIRT Model

Differences between genders in terms of mathematics achievement and attitude towards the subject have received a great deal of attention in educational and psychological research (Else-Quest, Hyde, & Linn, 2010). In the United States, standardized tests indicate that, on average, girls perform as well as boys (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). However, recent meta-analyses have shown consistent gender differences in attitudes towards the subject (Else-Quest et al., 2010). Typically, boys report greater intrinsic motivation and less math anxiety. One data source used in the meta-analysis of Else-Quest et al. (2010) is the PISA 2003 student questionnaire (Adams, 2005). In that analysis, students scores are averaged within a construct. These mean scores are then used in the meta-analysis.

Here, using the same PISA data, the gender difference in mathematics attitudes and affect is explored using a 2-group SNP-MIRT model. The data are responses to 17 items intended to measure three constructs: Mathematics Self-Efficacy (MSE); Mathematics Anxiety (ANX); and Mathematics Self-Concept (MSC). The item stems and wordings are presented in Table 9.1. Of the 2,705 female and 2,750 male respondents, 1,000 of each group were randomly sampled for the analysis. Also, to keep the example relatively simple, the analysis ignores both the hierarchical nature of the data, and the possibility that any of the items are characterized by differential item functioning (DIF). Consequently, all item parameters were constrained to be equal across groups.

Table 9.1: Items for PISA 2003 Self-Related Cognition in Mathematics Scale

Construct		*Stem* and Item Wordings
		How confident do you feel about having to do the following mathematical tasks?
MSE	1.	Using a train timetable to work out how long it would take to get from one place to another.
	2.	Calculating how much cheaper a TV would be after a 30% discount.
	3.	Calculating how many square feet of tile you need to cover a floor.
	4.	Understanding graphs presented in newspapers.
	5.	Solving an equation like $3x + 5 = 17$.
	6.	Finding the actual distance betwee two places on a map with a $1 : 100$ scale.
	7.	Solving an equation like $2(x + 3) = (x + 3)(x - 3)$.
	8.	Calculating the gas mileage of a car.
		To what extent do you agree with the following statements?
ANX	9.	I often worry that it will be difficult for me in mathematics classes.
	10.	I get very tense when I have to do mathematics homework.
	11.	I get very nervous doing mathematics problems.
	12.	I feel helpless when doing a mathematics problem.
	13.	I worry that I will get poor grades in mathematics.
MSC	14.	I am just not good at mathematics.
	15.	I get good grades in mathematics.
	16.	I have always believed that mathematics is one of my best subjects.
	17.	In my mathematics class, I understand even the most difficult work.

Note. MSE = Mathematics Self-Efficacy; ANX = Mathematics Anxiety; MSC = Mathematics Self-Concept. Response options for the first stem are: *Very confident, Confident, Not very confident,* or *Not at all confident.* Response options for the the second stem are: *Strongly agree, Agree, Disagree,* and *Strongly disagree.* All items were inverted for scaling except for item 14.

On the other hand, according to the literature, the latent variable distributions for males and females are expected to be different, at least in terms of location. To explore this issue, the males are designated as the reference group, with null mean vector and latent variable variances fixed to 1.0. Note, though, that the correlations are not fixed to 0, as previous research has found sizable correlations between constructs similar to those studied in this analysis. Then, in order to compare the groups, the elements of the mean vector and covariance matrix for the females may be freely estimated. This standard MIRT model was estimated, assuming both latent distributions were normal, using MH-RM. For convenience, this model will be referred to as STD.

Additionally, the shapes for both distributions could be analyzed, using SNP-MIRT. Specifying the tuning parameter as $\kappa = 2$ for both distributions resulted in 18 additional parameters. Let this model be the SNP2 model. A compromise between the specifications of the STD and SNP2 models is to constrain the SNP shape parameters to be equal between the groups. This particular configuration was not explored in the simulation study, but is arguably appropriate in the present context. Let this be the SNP1 model, which adds an additional 9 parameters to the STD model (but has 9 fewer parameters than the SNP2 model). Conveniently, the three models are nested, facilitating model comparison. All 3 models were estimated using the same settings as in the simulation study, with the following exception. The proposal standard deviation was reduced to 0.25 to maintain acceptable levels of the M-H sampler acceptance ratio.

All three models converged properly, and the results are presented in Table 9.2. Using a likelihood ratio test, SNP1 is preferred to SNP2 ($\chi^2_9 = 13.36, p = 0.15$), while both SNP models are clearly preferable to the STD model. These conclusions are also supported by the HQIC values. Additionally, Table 9.2 gives the parameter estimates for the female mean vector. The estimates are all quite similar. However, substantively, the results are consistent with those in the literature

indicating that females have higher levels of mathematics anxiety and lower levels of mathematics self-concept.

Table 9.2: $-2\times$Log-Likelihood and HQIC for PISA 2003 Data

				Group 2 Estimates (SE)		
Model	d	$-2\times$LogL (CI)	HQIC	μ_1	μ_2	μ_3
STD	80	66808.84 (2.84)	67133.36	$-.11$ (.05)	.21 (.05)	$-.24$ (.05)
SNP1	89	65302.88 (2.93)	65663.91	$-.11$ (.04)	.21 (.05)	$-.24$ (.05)
SNP2	98	65289.52 (2.92)	65687.06	$-.09$ (.05)	.22 (.05)	$-.21$ (.05)

Note. d = number of free parameters; STD = standard MIRT model; SNP1 = SNP-MIRT model with constraints among SNP parameters; SNP2 = SNP-MIRT model without constraints among SNP parameters; μ = mean of latent variable.

Turning to the density estimates, the bivariate contours of the males for the STD and SNP1 models are displayed in Figure 9.1. While the SNP contour plots for the males and females look quite similar, careful examination reveals that they are not identical. Though the elements of $\boldsymbol{\nu}$ (the "shape" parameters) are constrained to be equal across groups, the implied standardized SNP density is transformed via the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ (the mean and variance), which differs by group. Finally, figure 9.3 shows the univariate marginal distributions for the females. For θ_2 and θ_3, it is difficult to perceive any difference between the distributions. For θ_1, the SNP estimated distribution appears slightly more peaked than the STD distribution, and is also slightly left-skewed.

Figure 9.1: PISA 2003: Estimated Contour Plots for Males

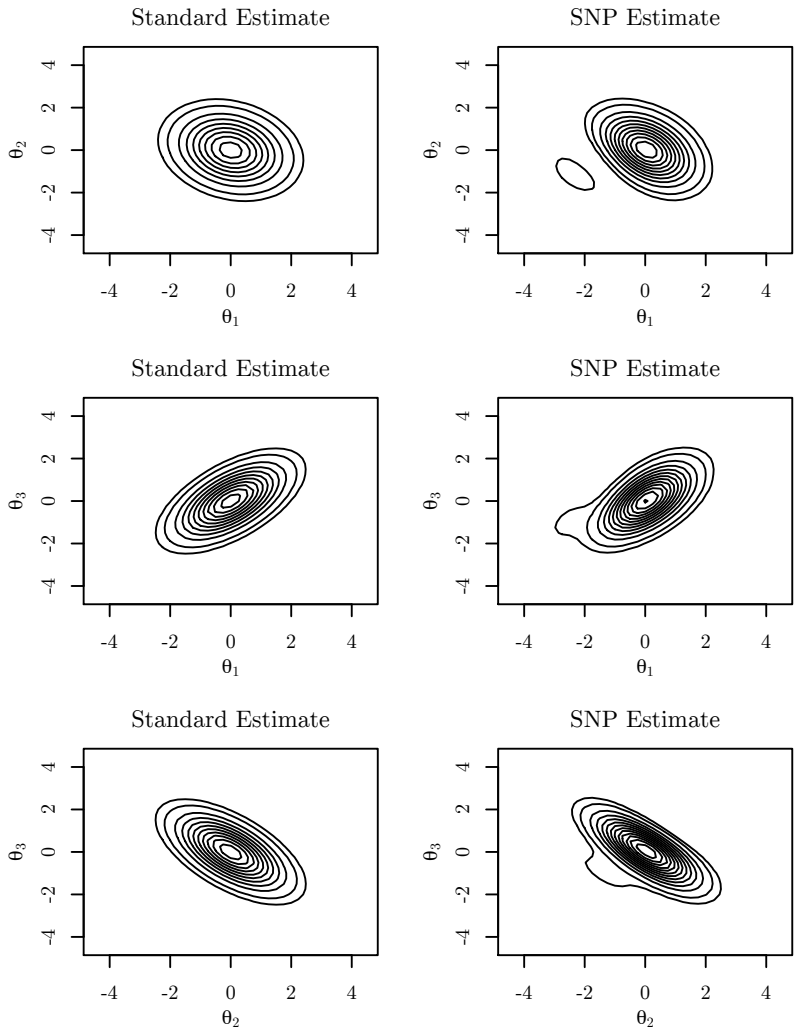


Figure 9.2: PISA 2003: Estimated Contour Plots for Females

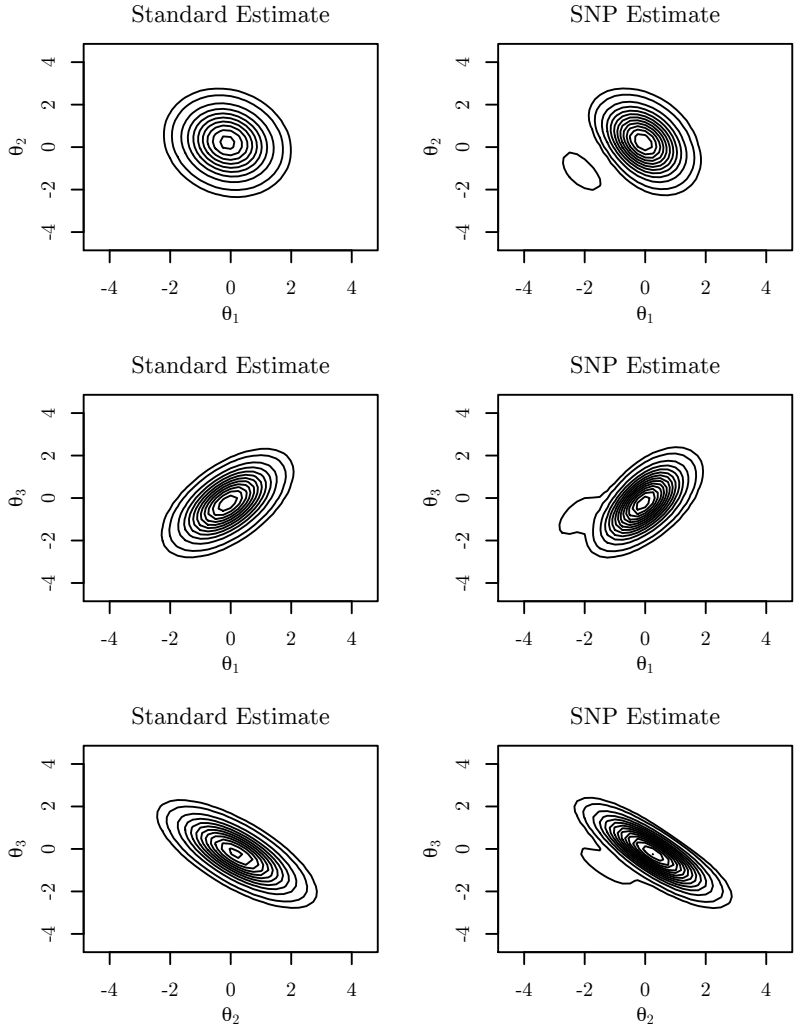
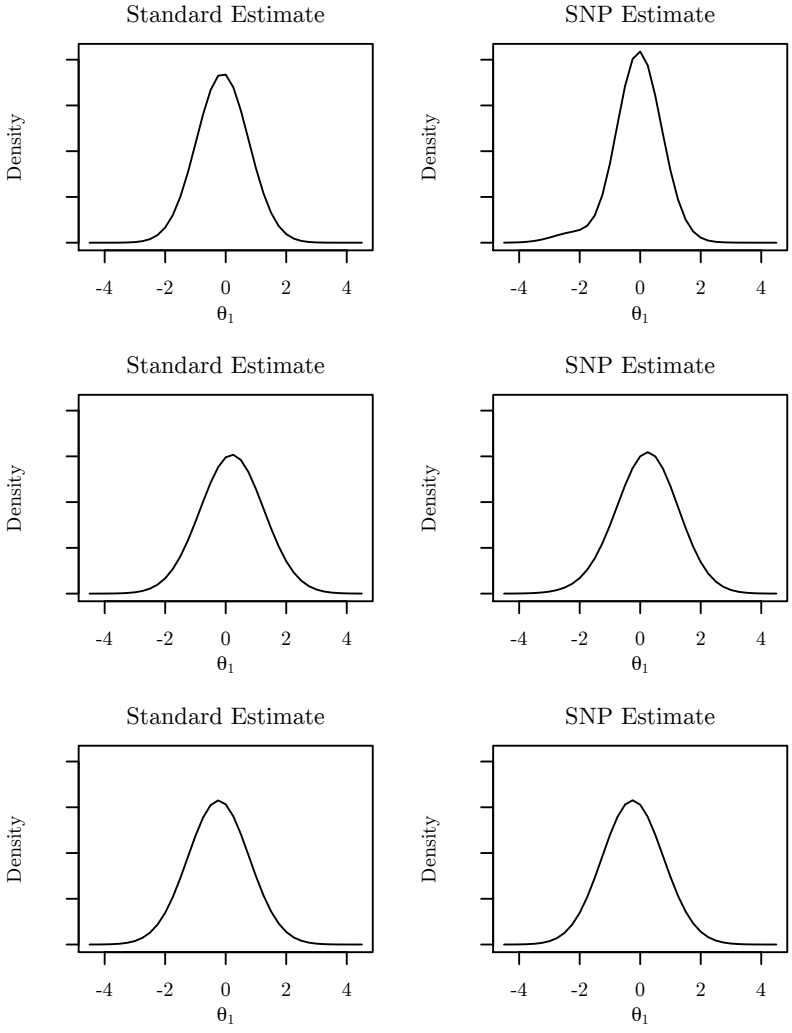


Figure 9.3: PISA 2003: Estimated Univariate Marginal Densities for Females



CHAPTER 10

Student Growth Percentiles

In this chapter, the SGP approach discussed in Chapter 1 is presented in detail. First, the approach that utilizes the standard MIRT model is presented. Then, a generalization is presented, where the SNP-MIRT model is used. This chapter also contains a small simulation study with two conditions to evaluate the proposed method and compare it with the original QR-based approach. Finally, an empirical application is presented, where SGPs are calculated for longitudinal achievement data.

10.1 The Proposed Method

This section presents the proposed method for calculating MIRT-based SGPs. The method may be applied to latent scores based on response patterns. However, the use of full response patterns is inconvenient for introducing the method, since the number of patterns is exponential in the number of test items. Instead, latent scores based on summed scores are used, as this facilitates the presentation. Further, this choice is not inappropriate, as numerous states utilize latent scores based on summed scores in reporting. Additionally, we limit the number of prior years to 1, again to facilitate the presentation. The proposed method, though, may accommodate multiple prior years.

At this point, it is convenient to introduce some limited notation. Let n_1 and n_2 be the numbers of items in year 1 (i.e., last year) and year 2 (i.e., current year),

respectively. For any student, let $\mathbf{y}_1$ and $\mathbf{y}_2$ be the analogously defined vectors of observed response patterns. Similarly, let $s_1 \in \{0, \ldots, S_1\}$ and $s_2 \in \{0, \ldots, S_2\}$ be the summed scores, where S_1 and S_2 are the respective maximum summed scores. Finally, let θ_1 and θ_2 be the latent achievement scores. For convenience, we assume all students take both tests and that there are no missing data, though both assumptions may be relaxed.

Before presenting the proposed method, we review the QR approach as implemented in this research, which serves as a point of comparison. First, two separate unidimensional IRT models are calibrated, one for each year. The item responses are modeled using the 3PL model in Equation (3.5), but note that here, θ is scalar-valued. Next, the respective IRT scaled scores based on summed scores are calculated. This simulation uses EAP score estimates. For the QR, the estimated scores from year 2 are the dependent variable, and the estimated scores from year 1 are the covariate. Finally, the 'SGP' R package (Betebenner, VanIwaarden, Domingue, & Shang, 2014) is used to compute the QR-based SGPs. For all package functions, the default arguments were used.

Next, we turn to the newly proposed method. Together, the prior and current years imply a two-dimensional MIRT model, where each dimension is measured by one year's items. Consequently, for all year 1 items, $\beta_2 = 0$; for all year 2 items, $\beta_1 = 0$. Again, Equation (3.5) is used, but in this case $\boldsymbol{\theta}$ is two-dimensional.

To present the method, we introduce various conditional distributions with the generic notation $\Pi(\cdot|\cdot)$. An example is $\varphi(\theta_1, \theta_2)$, the unconditional distribution (i.e., prior population distribution) for the latent variable scores. And, as a reminder, $\varphi(\cdot)$ is generic notation for the latent variable prior density. Let the correlation of $\varphi(\theta_1, \theta_2)$ be ρ. Also, to illustrate the calculation of various quantities used in the proposed method, we define a model with $n_1, n_2 = 40$, so that $s_1, s_2 \in \{0, \ldots, 40\}$. For the illustrations, $\rho = 0.85$.

Calculation of MIRT-based SGPs begins with estimation of all MIRT model

parameters, including ρ. Then, for any combination of s_1 and s_2, perform the following:

Step 1. Latent Variable Score Estimation. Estimate the current latent score, based on s_1 and s_2. This estimate (e.g., an EAP score) is denoted $\hat{\theta}_2|s_1, s_2$.

Step 2. Calibrated Projection. Using calibrated projection, find the *reference conditional distribution*, $\Pi(\theta_2|s_1, s_2)$. This distribution is based on ρ and only s_1.

Step 3. SGP Estimation. Calculate the SGP using $\hat{\theta}_2|s_1, s_2$ and the cumulative distribution function of $\Pi(\theta_2|s_1, s_2)$. The location of the current score estimate within this conditional distribution gives the MIRT-based SGP.

These steps are explained in greater detail in the 3 subsections to follow.

10.1.1 Latent Variable Score Estimation

For Step 1, $\hat{\theta}_2|s_1, s_2$ is based on $\int p(\theta_1, \theta_2|s_1, s_2)d\theta_1$. Recognizing that the MIRT model used here is a special case of the two-tier item factor model (Cai, 2010c), $\Pi(\theta_1, \theta_2|s_1, s_2)$ may be found using a modified version of the Lord-Wingersky algorithm (Cai, in press-a). This algorithm makes it possible to calculate $\Pi(\theta_1, \theta_2|s_1, s_2)$ without first calculating $\Pi(\theta_1, \theta_2|\mathbf{y}_1, \mathbf{y}_2)$ for all $\mathbf{y}_1$ and $\mathbf{y}_2$. Again, the numbers of possible response patterns for $\mathbf{y}_1$ and $\mathbf{y}_2$ grow exponentially with n_1 and n_2. Thus, any method requiring calculations for all possible $\mathbf{y}_1$ and $\mathbf{y}_2$ will have practical limitations. On the other hand, the Lord-Wingersky algorithm is constructed so that the number of calculations is approximately linear in n_1 and n_2. This makes calculation of $\Pi(\theta_1, \theta_2|s_1, s_2)$ feasible for very large numbers of items. As an aside, in the case of response pattern scoring, Step 1 is simpler, as it is then only necessary to calculate $\Pi(\theta_1, \theta_2|\mathbf{y}_1, \mathbf{y}_2)$ for all *observed* $\mathbf{y}_1$ and $\mathbf{y}_2$.

Before proceeding, we note that the posterior distribution $\Pi(\theta_1, \theta_2|s_1, s_2)$ reflects the correlated measurement errors for the latent dimensions. While estima-

tion of separate unidimensional IRT models, as in the QR approach to SGPs, may be used to estimate $\Pi(\theta_2|s_1)$ and $\Pi(\theta_2|s_2)$, such an approach cannot account for the association between the measurement errors. Additionally, $\hat{\theta}_2|s_1, s_2$, obtained via the MIRT framework, will be a more efficient estimate than $\hat{\theta}_2|s_2$, obtained via unidimensional IRT (Cai, 2010c).

10.1.2 Calibrated Projection

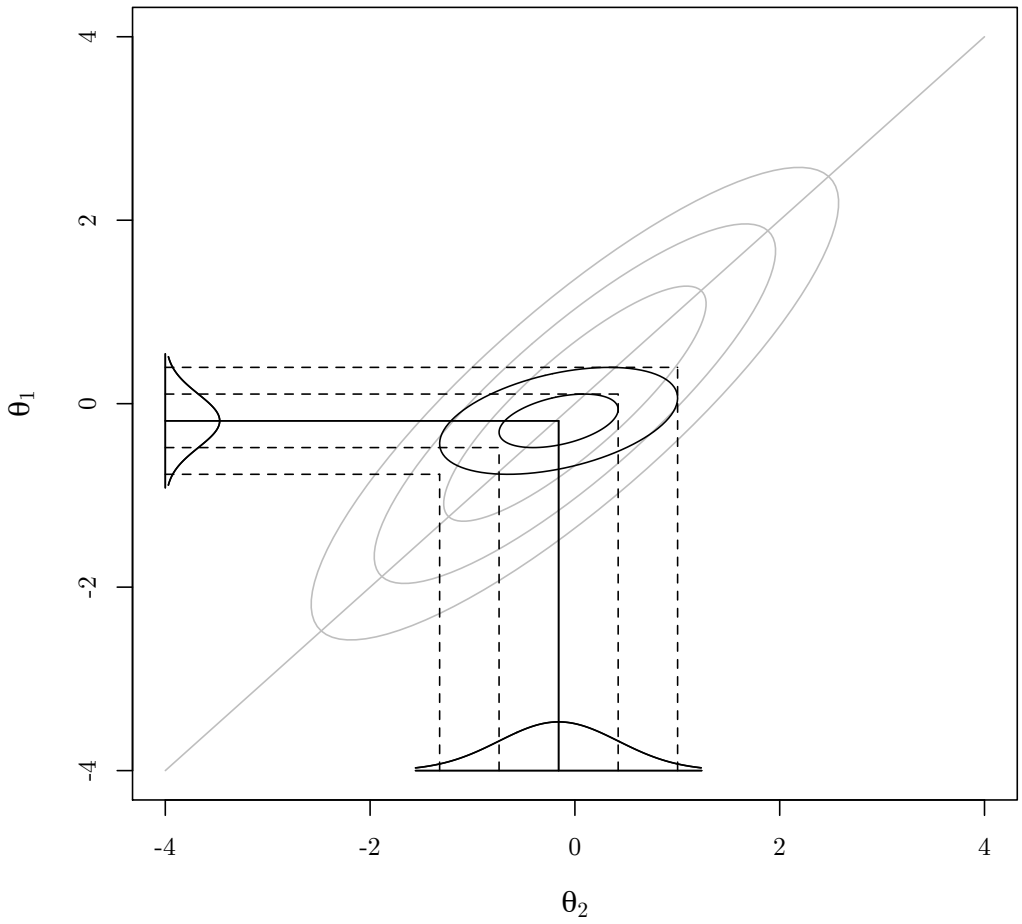
The development of calibrated projection (Thissen, Varni, et al., 2011) was motivated by the need to link two highly similar, though not identical, constructs for the purposes of producing a scoring cross-walk. Utilizing a MIRT framework, calibrated projection provides a means to use item responses from one instrument to produce scores on the scale of a second instrument. These scores are summaries (e.g., EAPs) of the posterior distribution $\Pi(\theta_2|s_1)$. In the original application of calibrated projection, $\Pi(\theta_2|s_1)$ is needed because estimates of θ_2 are desired for people who have not taken test 2.

In contrast, in the current application of calibrated projection, we assume s_2 exists for all students. Still, $\Pi(\theta_2|s_1)$ is a key quantity for MIRT-based SGPs: it represents the conditional distribution of the current latent score for all students with identical score histories. In other words, it is the reference conditional distribution from which an SGP may be estimated. We now provide an example of how calibrated projection may be used to find this reference distribution.

Consider Figure 10.1, which is akin to figures in Thissen, Varni, et al. (2011) and Cai (in press-a). Given a specified s_1 (e.g., 20), the MIRT model implies a distribution on θ_1, shown on the y-axis. This distribution is then projected through the relationship between θ_1 and θ_2 to imply a distribution on θ_2, shown on the x-axis. This latter distribution, $\Pi(\theta_2|s_1)$, is the distribution for all students with identical score histories (here, $s_1 = 20$). It is used in Step 3 to estimate an

SGP.

Figure 10.1: Calibrated Projection Linking



Note. Given a specified s_1 (here, $s_1 = 20$), the MIRT model implies a distribution on θ_1, shown on the y-axis. This distribution, $\Pi(\theta_1|s_1)$, is then projected through the relationship between θ_1 and θ_2 to imply a distribution on θ_2, shown on the x-axis. The dark gray central ellipses approximate $\Pi(\theta_1, \theta_2|s_1)$. The light gray central ellipses represent the prior distribution of latent scores, $\varphi(\theta_1, \theta_2)$.

Other features of Figure 10.1 are worth mentioning. First, the light gray central ellipses represent $\varphi(\theta_1, \theta_2)$, the prior distribution of the latent variables. In Figure 10.1, $\varphi(\theta_1, \theta_2)$ is bivariate normal with $\rho = 0.85$. Second, the dark gray central ellipses represent $\Pi(\theta_1, \theta_2|s_1)$, which does not condition on s_2. As a result, $\Pi(\theta_1, \theta_2|s_1)$ is more variable for θ_2 than for θ_1. Finally, this relative uncertainty

is projected onto the x-axis in $\Pi(\theta_2|s_1)$. Given the estimated parameters, the location and scale of $\Pi(\theta_2|s_1)$ are completely determined by s_1 and $\varphi(\theta_1, \theta_2)$.

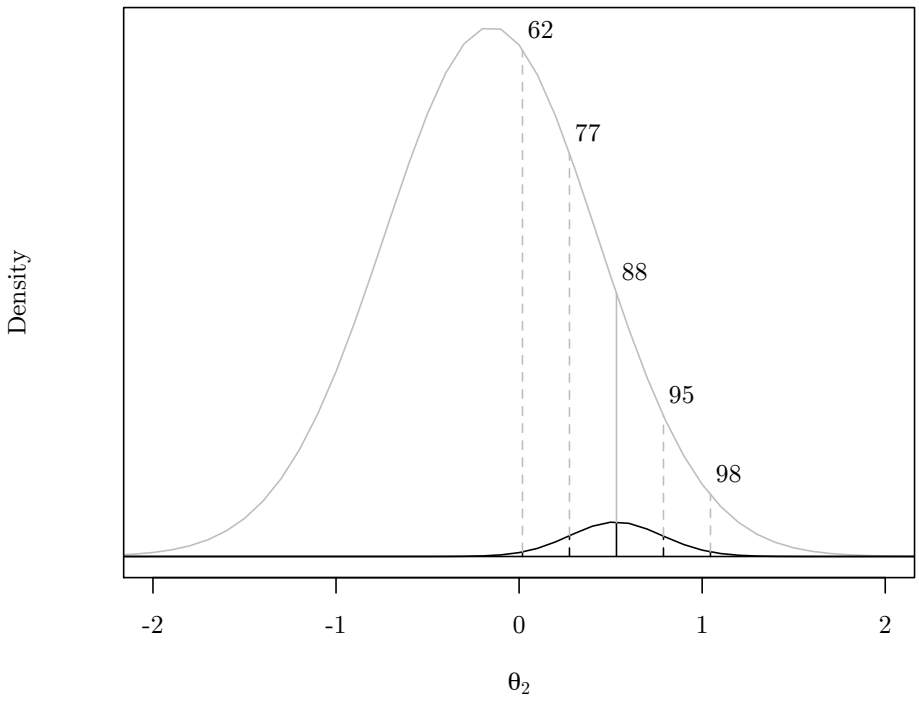
10.1.3 SGP Estimation

Step 3 is conceptually the most straightforward. Let $q_{\theta_2|s_1}(\theta_2)$ be the cumulative distribution function of $\Pi(\theta_2|s_1)$, the reference conditional distribution from Step 2. Then, the MIRT-based SGP estimate is $q_{\theta_2|s_1}(\hat{\theta}_2|s_1, s_2)$, where $\hat{\theta}_2|s_1, s_2$ is the score estimate from Step 1. As an example, consider Figure 10.2, which shows $\Pi(\theta_2|s_1 = 20)$ as the large light gray distribution, and $\Pi(\theta_2|s_1, s_2 = 30)$ as the small black distribution. The EAP of $\Pi(\theta_2|s_1 = 20, s_2 = 30)$, $\hat{\theta}_2|s_1, s_2$, is marked by the solid black vertical line segment. Its position within $\Pi(\theta_2|s_1 = 20)$ is marked by the solid light gray vertical line segment, and corresponds to an SGP estimate of 88 for this score combination. That is, $q_{\theta_2|s_1}(\hat{\theta}_2|s_1 = 20, s_2 = 30) = 88$.

Figure 10.2 also shows how uncertainty in the SGP estimate is directly related to uncertainty in $\hat{\theta}_2|s_1, s_2$. The dashed vertical lines in Figure 10.2 correspond to ± 1 and ± 2 standard errors of measurement for $\hat{\theta}_2|s_1, s_2$. Like $\hat{\theta}_2|s_1, s_2$, these values of θ_2 correspond to percentiles of $\Pi(\theta_2|s_1)$, which are displayed in Figure 10.2. For any given s_1, the uncertainty in the SGP estimate will vary as a function of s_2.

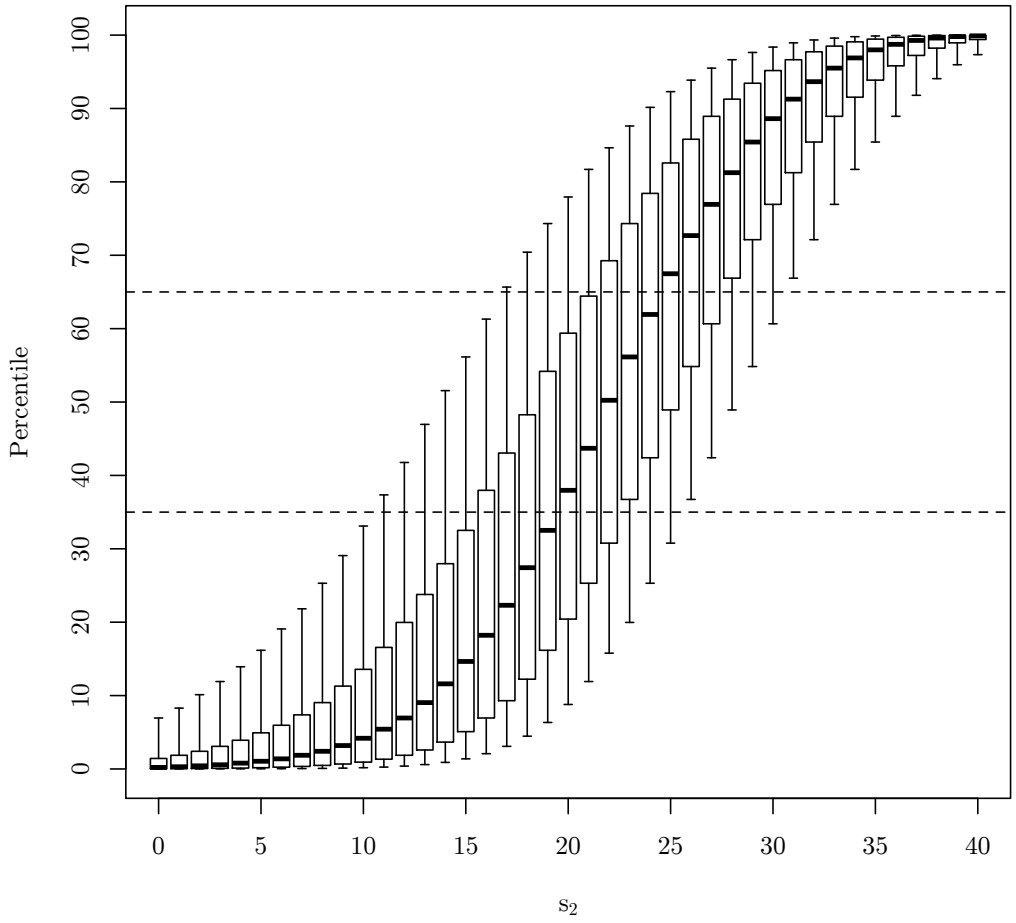
This phenomenon is presented graphically in Figure 10.3, for $s_1 = 20$. For each s_2, the boxplot demarcates the SGP estimates corresponding to $\hat{\theta}_2|s_1, s_2$, and ± 1 and ± 2 standard errors of measurement for $\hat{\theta}_2|s_1, s_2$. Note that the boxplots for s_2 values near $s_1 = 20$ are relatively large. This is because for these values, $\Pi(\theta_2|s_1 = 20, s_2)$ is centrally located in relation to $\Pi(\theta_2|s_1 = 20)$, where small changes in θ_2 lead to large changes in the SGP estimate. At least for this example, there is considerable uncertainty for most of the SGP estimates.

Figure 10.2: Illustration of MIRT-Based SGP Calculation



Note. The dominating light gray curve is $\Pi(\theta_2|s_1)$ (here, $s_1 = 20$). The smaller dark gray curve is $\Pi(\theta_2|s_1, s_2)$ (here, $s_2 = 30$). The 5 vertical line segments demark the expectation of $\Pi(\theta_2|s_1, s_2)$, as well as ± 1 and ± 2 standard errors of measurement. The extended line segments (light gray) correspond to percentile values for $\Pi(\theta_2|s_1)$. Here, $s_1 = 20$ and $s_2 = 30$ yields an SGP point estimate of 88.

Figure 10.3: Uncertainty of MIRT-Based SGPs



Note. The boxplots are of MIRT-based SGPs corresponding to EAP scores and ± 1 and ± 2 standard errors of measurement for $s_1 = 20$ and all possible s_2. The horizontal dotted lines correspond to hypothetical SGP cut-values of 35 and 65. Many boxplots span all 3 "classifications." The boxplot above $s_2 = 30$ corresponds to Figure 10.2.

10.2 Generalizing the Proposed Method with SNP-MIRT

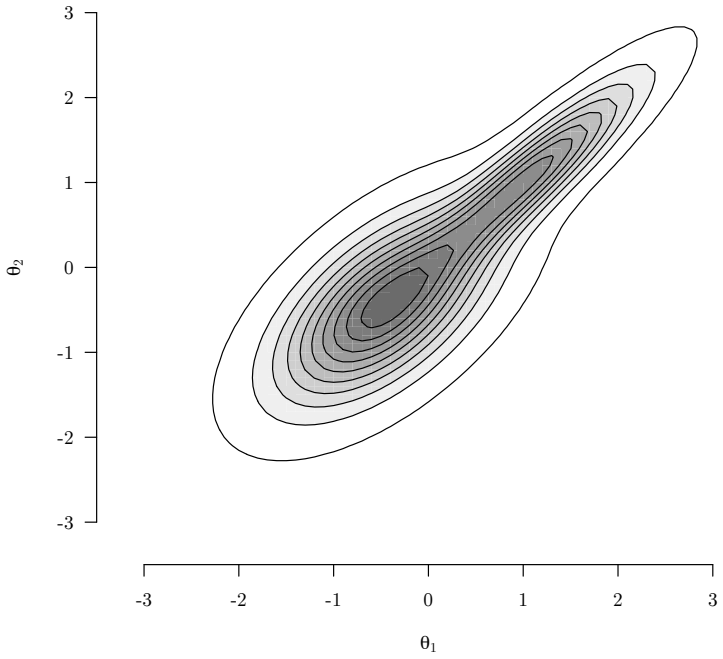
Recall the calibrated projection example given in Figure 10.1. Given the specified MIRT model and estimated parameters, the reference conditional distribution, $\Pi(\theta_2|s_1)$ (shown on the x-axis), depends on two quantities: s_1 and $\varphi(\theta_1, \theta_2)$. Consequently, for a given s_1 and s_2, different specifications of $\varphi(\theta_1, \theta_2)$ may lead to different SGP estimates. In other words, the MIRT-based SGP estimates may be sensitive to the specification of the prior density for the latent scores.

As an example of this potential sensitivity we consider two different specifications for $\varphi(\theta_1, \theta_2)$, holding all other aspects of the model constant. The first is the bivariate normal from the examples in Section 10.1, with null mean vector, unit variances, and $\rho = 0.85$. The second is a nonnormal specification, created using a mixture of normals, shown in Figure 10.4. This distribution likewise has null mean vector, unit variances, and $\rho = 0.85$. For a given s_1, each of these distributions may lead to a different reference conditional distribution.

Figure 10.5 shows the reference conditional distributions formed by using the normal (dashed gray curves) and nonnormal (solid black curves) priors for $s_1 = 20$ (left plot) and $s_2 = 35$ (right plot). For $s_1 = 20$, the two reference conditional distributions are quite similar, and the corresponding SGP estimates would likely be comparable. On the other hand, for $s_2 = 35$, the two reference conditional distributions are clearly different, suggesting that the specification of $\varphi(\theta_1, \theta_2)$ can impact the MIRT-based SGP estimates.

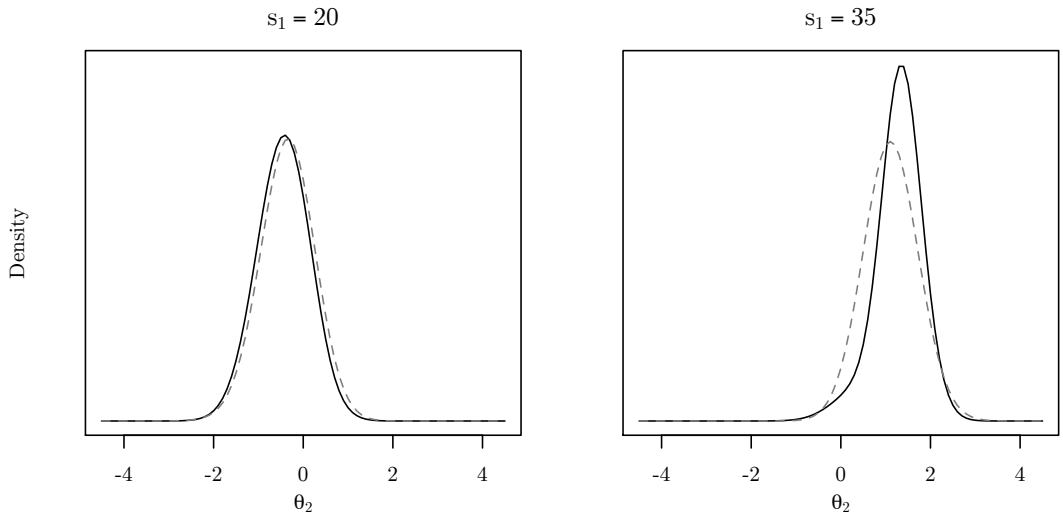
The potential sensitivity of the proposed SGP calculation to the specificaiton of $\varphi(\theta_1, \theta_2)$ suggests the use of SNP-MIRT in lieu of standard MIRT. The only change to the proposed framework is that all of the calculations (i.e., Steps 1-3) are perfomed using the ML estimates from the SNP-MIRT model instead of the standard MIRT model.

Figure 10.4: Example of Bivariate Nonnormal Density



Note. The variance for each dimension is 1, and the correlation between dimensions is $\rho = 0.85$. Each marginal distribution is standardized, but skewed right.

Figure 10.5: Examples of Posterior Distributions of θ_2 Given s_1



Note. The plots are of $\Pi(\theta_2|s_1)$ for $s_1 = 20$ (left plot) and $s_1 = 35$ (right plot) using a normal prior distribution (light gray, dashed curve) and the nonnormal distribution from Figure 4 (black solid curve). For both prior distributions, $\rho = 0.85$.

10.3 SGP Simulation Study

A limited simulation study was conducted to evaluate the proposed methods and to compare SGP approaches. To focus the investigation, only 2 conditions were studied. In Condition 1, latent variable scores were generated from a bivariate normal. Then, SGPs were estimated using the QR and MIRT approaches. In Condition 2, latent variable scores were generated from the normal mixture shown in Figure 10.4. In this latter condition, SGPs were estimated using the QR, MIRT, and SNP-MIRT approaches. Generally, data-generating parameter values were chosen to be representative of large-scale state assessments.

10.3.1 Data Generation and SGP Estimation

For both conditions, the generating density had a null mean vector, variances equal to 1, and a correlation of $\rho = 0.85$. For Condition 2, the latent variable density

was a mixture of 2 bivariate normals with parameters: $\mu_1 = (-0.47, -0.47)'$, $\mu_2 = (1.09, 1.09)'$, $\sigma_1 = (0.48, 0.29, 0.48)'$, $\sigma_2 = (0.48, 0.44, 0.48)'$, $mp_1 = 0.7$, and $mp_2 = 0.3$, where "mp" stands for mixing proportion. This is also the definition of the density in Figure 10.4.

Each of the two dimensions was measured by 40 3PL items. Thus, $s_1, s_2 \in \{0, \ldots, 40\}$. Slopes were drawn from a truncated normal, with mean = 1.5 and standard deviation = 0.5, truncated at 0.5 and 3. Intercepts were drawn from a normal with mean = 0 and standard deviation = 1. Finally, guessing parameters were drawn from a normal with mean = 0.25 and standard deviation = 0.05, truncated at 0.1 and 0.35. The logits of the guessing parameters, used in estimation, are denoted $logit(c)$. All of the data-generating item parameters are presented in Table 10.1.

Let π_{s_1,s_2} be the true model-implied probability for the summed-score combination given s_1 and s_2. For all combinations of s_1 and s_2, π_{s_1,s_2} may be calculated using a modified version of the Lord-Wingersky algorithm (Cai, in press-a). Figure 10.6 presents bubble plots of these probabilities for Condition 1 (left plot) and Condition 2 (right plot), with larger bubbles corresponding to greater probabilities. Although the overall patterns are similar, differences can be detected, in particular for score combinations where both s_1 and s_2 are large.

For each replication, SGPs based on QR, MIRT, and SNP-MIRT were estimated as described in Section 10.1. All SGP estimates were compared to "true" SGPs, calculated using the data-generating parameter values and MIRT model.

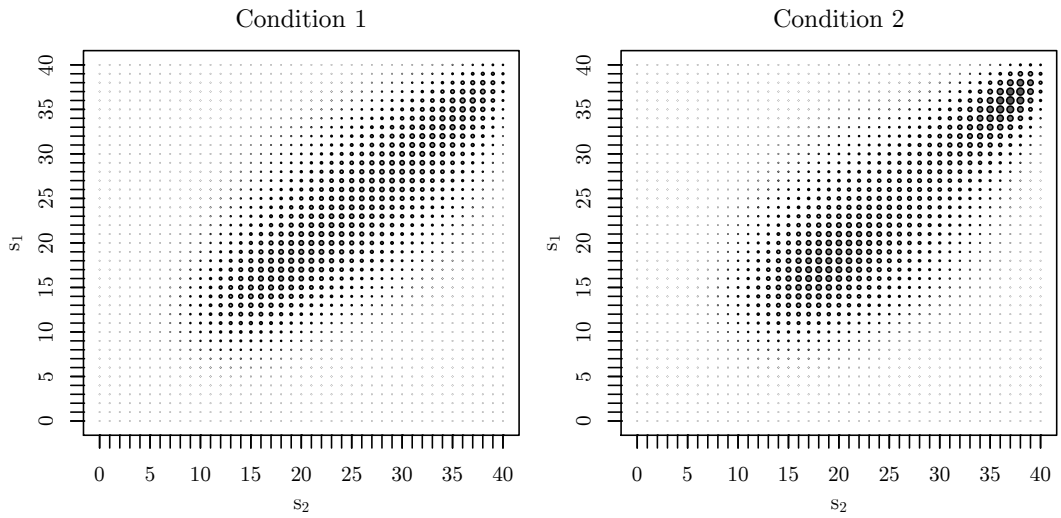
10.3.2 Collected Statistics

The SGP estimates are evaluated using several measures of accuracy. Momentarily suppressing reference to s_1 and s_2, let ψ be a true SGP and $\hat{\psi}$ its corresponding estimate. Two measures used, bias and RMSE, are defined in Equations (7.1)

Table 10.1: Generating Parameters for SGP Simulations

item	logit(c)	α	β	item	logit(c)	α	β
1	−1.41	−0.71	1.22	41	−0.94	0.70	1.15
2	−0.79	0.26	1.38	42	−0.82	−0.26	1.40
3	−1.19	−0.25	2.28	43	−0.92	−1.57	0.87
4	−1.34	−0.35	1.54	44	−1.20	−1.51	2.58
5	−1.16	−0.95	1.56	45	−1.08	−1.60	2.10
6	−1.15	−0.05	2.36	46	−1.30	−0.53	0.94
7	−0.82	−0.78	1.73	47	−1.30	−1.46	1.30
8	−1.08	−1.67	0.87	48	−0.87	0.69	1.27
9	−0.91	−0.38	1.16	49	−1.39	2.10	1.89
10	−1.24	0.92	1.28	50	−0.63	−1.29	1.46
11	−1.19	0.61	1.68	51	−1.12	0.79	1.63
12	−1.04	−0.58	2.11	52	−1.04	0.77	1.49
13	−1.07	−1.62	1.70	53	−1.31	0.33	1.48
14	−1.35	−0.06	1.56	54	−1.26	−1.01	2.18
15	−1.49	0.52	1.22	55	−1.49	−0.12	1.39
16	−0.62	0.30	2.39	56	−1.15	−0.28	2.26
17	−0.94	0.11	1.75	57	−0.99	0.56	0.73
18	−1.47	−0.64	0.52	58	−1.01	−0.37	1.79
19	−1.27	−0.85	1.85	59	−1.32	0.98	1.56
20	−1.45	−1.02	1.26	60	−1.32	−0.37	1.61
21	−0.77	0.12	0.97	61	−1.24	1.05	1.69
22	−1.17	−0.95	1.39	62	−0.73	−1.05	1.25
23	−0.96	−0.49	0.99	63	−1.43	−1.26	1.33
24	−1.21	−0.26	1.14	64	−1.15	3.24	0.99
25	−1.23	−1.84	1.19	65	−0.64	−0.42	0.96
26	−1.32	−0.65	0.66	66	−1.13	0.30	1.65
27	−1.26	0.24	1.92	67	−1.50	0.64	1.72
28	−0.70	0.08	1.58	68	−1.28	−0.48	1.53
29	−1.11	−0.96	0.93	69	−0.97	0.52	1.96
30	−1.07	−0.07	2.13	70	−1.20	0.37	2.53
31	−1.03	1.44	1.71	71	−1.25	−0.22	1.25
32	−0.79	0.45	1.35	72	−1.19	0.07	2.00
33	−1.24	0.04	1.95	73	−1.07	−0.03	1.15
34	−1.38	−0.42	1.94	74	−0.71	2.13	1.16
35	−0.69	−2.05	1.91	75	−1.12	−0.74	2.01
36	−1.22	1.13	1.84	76	−0.83	−1.10	1.36
37	−1.30	−1.46	1.78	77	−0.94	0.04	0.89
38	−1.46	−0.74	1.47	78	−1.13	0.31	1.59
39	−1.48	−1.91	1.35	79	−1.56	0.44	1.43
40	−1.26	−1.44	1.31	80	−1.24	−0.46	1.50

Figure 10.6: Bubble Plots of True Model-Implied Probabilities



Note. Larger bubbles correspond to greater model-implied probabilities.

and (7.2), respectively. Another measure used is the integrated absolute bias. The absolute bias is defined as $\delta = M^{-1} \sum_{m=1}^{M} |\psi - \hat{\psi}_m|$, where M is the number of Monte Carlo replications.

Then, for a given s_1, the integrated absolute bias is

$$\bar{\delta}_{s_1} = \sum_{s_2=0}^{S_2} \delta_{s_1,s_2} W_{s_1,s_2}, \tag{10.1}$$

where $W_{s_1,s_2} = \pi_{s_1,s_2} / \sum_{s_2=0}^{S_2} \pi_{s_1,s_2}$. In words, this measure gives the expected absolute bias for a given s_1, averaged across all possible s_2 values.

The SGP estimates were also evaluated using correct classification rates (CCR). Given a set of cut-percentiles, such as $(0, 35, 65, 100)$, the rate is defined as

$$\text{CCR} = M^{-1} N^{-1} \sum_{m=1}^{M} \sum_{i=1}^{N} \mathbf{1}_{\tau(\psi_{im})}(\tau(\hat{\psi}_{im})) \tag{10.2}$$

where $\tau(\cdot)$ maps an SGP to a classification and $\mathbf{1}_{\tau(\psi_{im})}(\tau(\hat{\psi}_{im}))$ is an indicator

function that returns a 1 if and only if $\tau(\psi_{im})$ is equal to $\tau(\hat{\psi}_{im})$, and 0 otherwise. The CCR is simply the proportion of estimated classifications that agree with true classifications.

Bias and RMSE statistics (see Equations (7.1) and (7.2)) were also collected for the item parameter estimates. For Condition 1, since the models are correctly specified, the estimates should be approximately unbiased. However, for Condition 2, all fitted models are misspecified, since the true latent variable density is a normal mixture. Thus, the bias and RMSE statistics can help to measure the sensitivity of the models to this misspecification. Finally, the log-likelihoods and HQIC values of the MIRT and SNP-MIRT models were collected to make comparisons of overall fit. These results, focusing on parameter estimation and model fit, may therefore supplement the simulation results from Chapter 8.

10.3.3 Condition 1 Results: Normal Latent Variable Density

Table 2 presents some parameter recovery results for Condition 1. As expected, the estimates for all parameter types are approximately unbiased. The RMSE values from the MIRT model are slightly smaller than those from the two separate unidimensional IRT models used for the QR approach. This is to be expected since the correlation between dimensions in the MIRT model leads to an increase in efficiency of the parameter estimates. Table 10.2 also serves as a point of reference for parameter recovery results for Condition 2, when the fitted models are misspecified.

Figures 10.7 and 10.8 present the bias in SGP estimates for all summed-score cross-classifications for the QR and MIRT-based approaches, respectively. The MIRT-based estimates are nearly unbiased, and much less biased than the QR-based estimates. Further, the magnitude and direction of bias for the QR-based estimates clearly depend on s_1 and s_2. One notable trend is that there is relatively

Table 10.2: Condition 1: Parameter Recovery

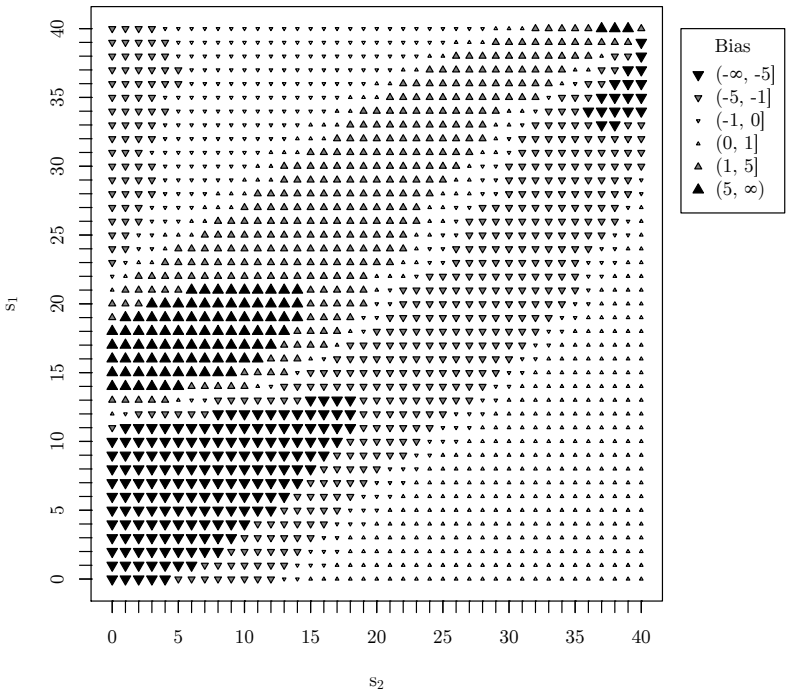
	Bias			RMSE		
	Parameter Type			Parameter Type		
Method	logit(c)	α	β	logit(c)	α	β
QR	0.01	−0.01	0.01	0.15	0.12	0.09
MIRT	0.01	−0.01	0.01	0.14	0.11	0.09

Note. "QR" refers to calibration via two separate unidimensional IRT models, one for year/test 1, one for year/test 2. logit(c) = logit of the guessing parameter; α = intercept; β = slope.

little bias in SGP estimates for $s_1 \approx s_2$, at least when each summed score is around 15 or greater. However, for score combinations near this diagonal where $s_1 > s_2$, there is a clear pattern of positive bias. In contrast, near this diagonal, when $s_1 < s_2$, there is a clear pattern of negative bias. Another trend is that for score combinations where both s_1 and s_2 are relatively small, the QR-based estimates are consistently negatively biased.

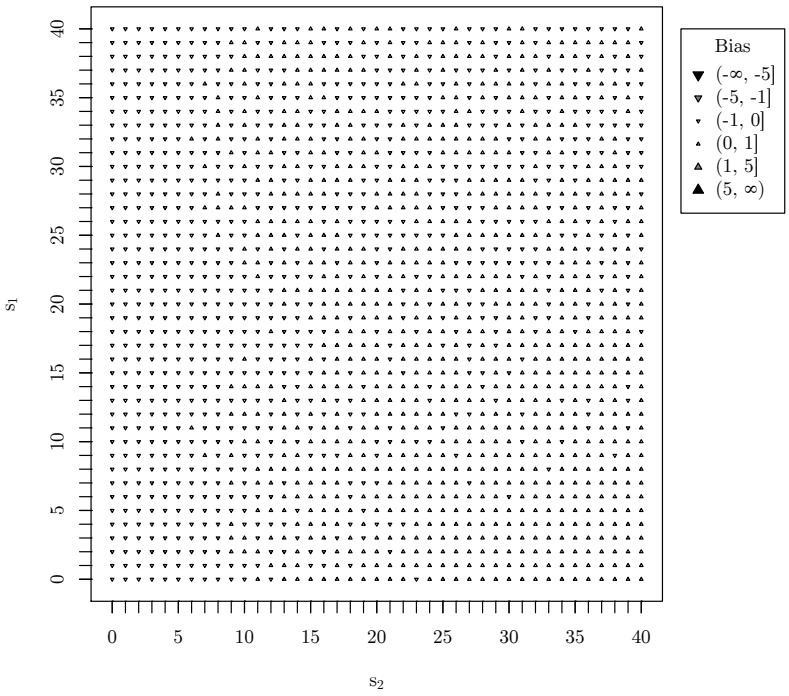
A shortcoming of the bivariate plots in Figures 10.7 and 10.8 is that they do not incorporate the model-implied probabilities for each summed score combination (see Figure 10.6). For example, the bias corresponding to the combination $s_1 = 0$ and $s_2 = 0$ may not be particularly important, since π_{s_1,s_2} is practically 0 in that case. One way to focus our attention is to identify the most probable summed score combinations. Here, the 99% Highest Density Region (HDR, Rosa, Swygert, Nelson, & Thissen, 2001) of combinations is identified, which comprises the minimum number of most probable combinations sufficient to account for 99% of the probability mass. In other words, the least probable combinations are ignored. Figure 10.9 plots the mean SGP estimates against the true SGP values for the 99% HDR score combinations, for both the QR (left plot) and MIRT (right plot) approaches. Again, the MIRT-based estimates are nearly unbiased, while the QR-based estimates are systematically biased. Generally, the QR-based SGP estimates are negatively biased for smaller values and positively biased for greater

Figure 10.7: Condition 1: Bias for QR Approach



Note. Bias is defined as the average estimate across all 100 replications minus the true value.

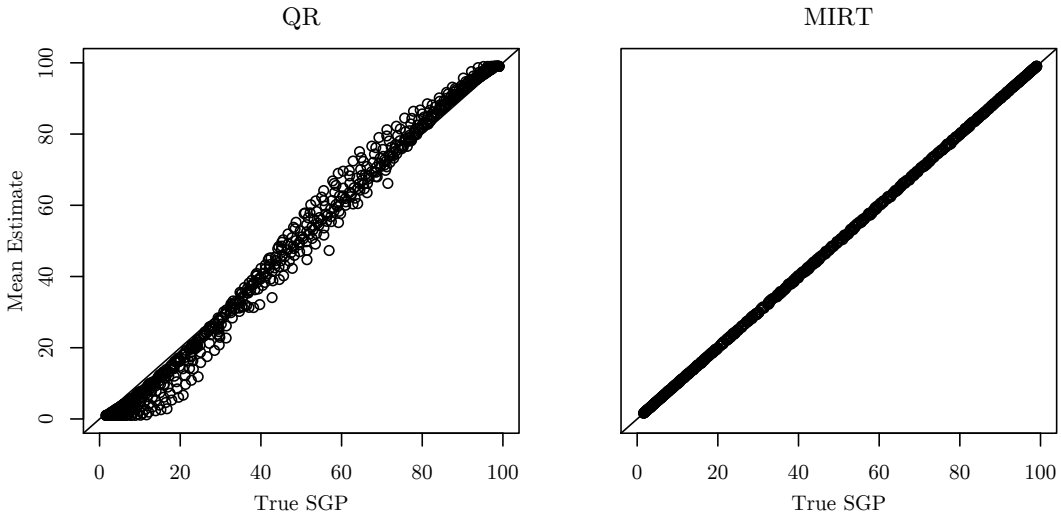
Figure 10.8: Condition 1: Bias for MIRT Approach



Note. Bias is defined as the average estimate across all 100 replications minus the true value.

values. The implication is that the QR approach tends to "exaggerate" SGPs.

Figure 10.9: Condition 1: Estimated and True SGPs

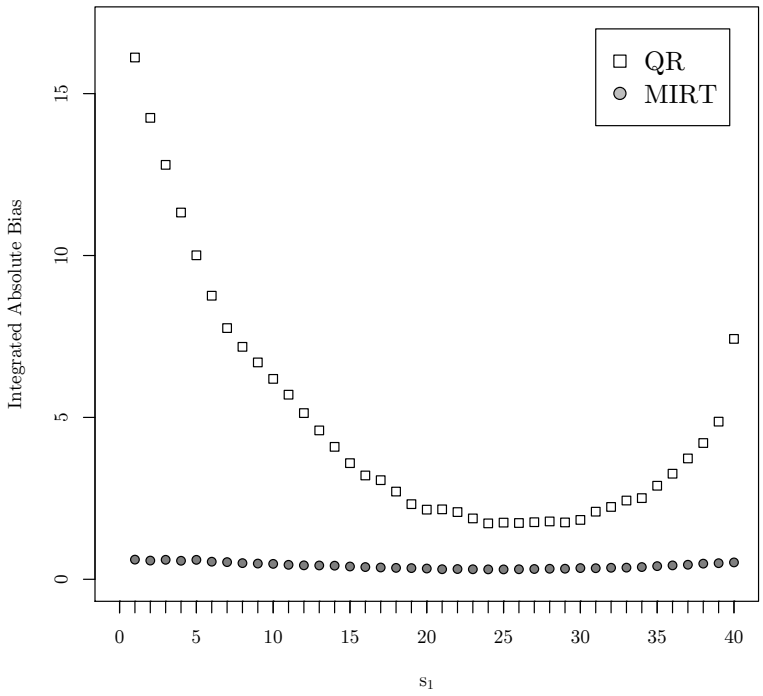


Note. Estimates are averages over all 100 replications. Only the 99% Highest Density Region cross-classifications are displayed.

Another way to focus our attention is to look at the integrated absolute bias, given s_1. Figure 10.10 shows the integrated absolute bias for both QR and MIRT approaches, with the MIRT approach again outperforming the QR approach. Two other features of Figure 10.9 are worth mentioning. First, for this condition, the QR approach performs better near the middle of the s_1 range and worse near the extremes. Second, while the integrated absolute bias for low values of s_1 is relatively great, those values of s_1 have low model-implied probabilities (see Figure 10.6).

Finally, Table 10.3 presents CCRs for the two approaches for several sets of cut-percentiles. For the set with 3 classes, both approaches are quite accurate, with rates of 0.95 and 0.99 for the QR and MIRT approaches, respectively. As expected, as the number of classes increases, the accuracies for both approaches decrease. However, the accuracy for the MIRT approach decreases relatively slowly with an accuracy rate of 0.97 for 10 classes.

Figure 10.10: Condition 1: Integrated Absolute Bias



Note. For each replication, for each s_1, the absolute bias is integrated over the true model-implied probabilities for all s_2. These values are then averaged over all 100 replications.

Table 10.3: Condition 1: Correct Classification Rates

		Correct Classification Rate	
Classes	Cut-Percentiles	QR	MIRT
3	$(0, 35, 65, 100)$	0.949	0.989
4	$(0, 25, 50, \ldots, 100)$	0.916	0.988
5	$(0, 20, 40, \ldots, 100)$	0.893	0.985
10	$(0, 10, 20, \ldots, 100)$	0.749	0.966

Note. Figures based on simulation study with $N = 10,000$ and 100 replications

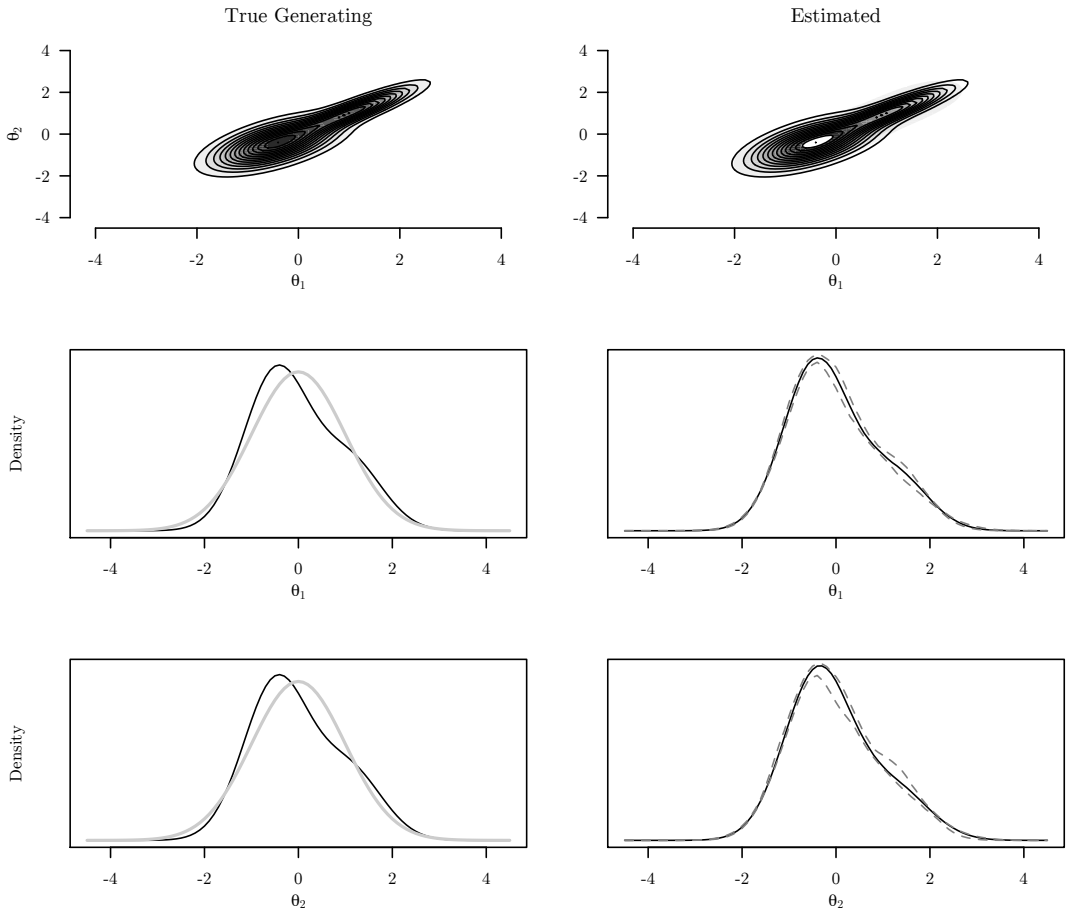
Based on the measures considered, the MIRT-based approach performed extremely well in Condition 1. These results, however, represent the unrealistic situation where the model, including the model for the latent variable density, is exactly correctly specified. This correct specification, in conjunction with the ML estimator, leads to asymptotically unbiased estimates. Further, the large sample sizes in the replications ($N = 10,000$) resulted in highly efficient parameter estimates. And, since the MIRT-based SGPs are a function of the latent variable density and parameter estimates, it should not be surprising that the MIRT approach performed so strongly. Condition 2, however, presents a more challenging case where the latent variable density is misspecified.

10.3.4 Condition 2 Results: Nonnormal Latent Variable Density

The plots in Figure 10.11 show the true generating latent variable density (left column) and the SNP-MIRT estimated density for Condition 2. The top row displays the true bivariate contour and the median of the estimated SNP densities. The resemblance between the two plots indicates the SNP-MIRT model was, to some degree, effective in estimating the shape of the latent variable density. The middle and bottom rows show the univariate marginal densities for θ_1 and θ_2, respectively. In the left column for these plots, the true generating density is represented by the black curve, while the gray curve is a standard normal, provided as a reference. In the right column for these plots, the median of the estimated SNP densities is shown in black, while the dashed gray curves provide an empirical 90% confidence interval. Again, the right column (SNP estimate) resembles the left column (true generating).

Table 10.4 presents parameter recovery results disaggregated by parameter type for the different estimation approaches, including SNP-MIRT. For all parameter types, the SNP-MIRT model yields estimates with the least bias and the lowest RMSE values. And in contrast to the results from Condition 1 (see Table

Figure 10.11: Condition 2: Estimated Densities



Note. Condition 2 results: plots of true generating (left column) and SNP-estimated (right column) prior latent variable densities. The top row shows the bivariate distributions, $\varphi(\theta_1, \theta_2)$. The middle row shows the univariate marginal for θ_1. On the left, the light gray distribution is a standard normal, shown for reference. On the right, the dashed light gray curves give a 90% empirical confidence interval. The bottom row provides the same information as the middle row, but for θ_2.

10.2), Table 10.4 suggests that the QR and standard MIRT approaches lead to biased parameter estimates. Also, for every replication, the SNP-MIRT model was preferred over the standard MIRT model based on both $-2\times$log-likelihood and HQIC values.

Table 10.4: Condition 2: Parameter Recovery

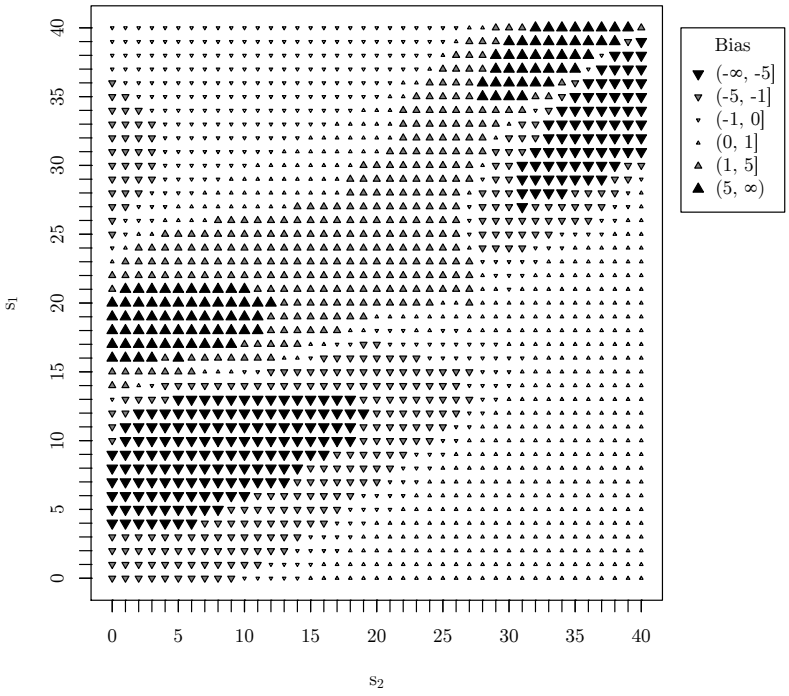
	Bias			RMSE		
	Parameter Type			Parameter Type		
Method	logit(c)	α	β	logit(c)	α	β
QR	0.22	−0.25	0.23	0.26	0.28	0.26
MIRT	0.20	−0.19	0.18	0.24	0.23	0.21
SNP-MIRT	−0.04	0.05	−0.00	0.17	0.13	0.10

Note. "QR" refers to calibration via two separate unidimensional IRT models, one for year/test 1, one for year/test 2. logit(c) = logit of the guessing parameter; α = intercept; β = slope.

To summarize the results for Condition 2 thus far, the SNP-MIRT model was fairly successful in estimating the shape of the true nonnormal latent variable density, and performed the best among the methods in terms of parameter recovery. Since the SGP estimation approach presented in this research depends on the latent variable density and item parameter estimates, we should expect the SGP estimates based on the SNP-MIRT model to be more accurate than those based on the standard MIRT model. We now turn to those results.

Figures 10.12-10.14 present the bias in SGP estimates for all summed score cross-classifications for the QR, MIRT, and SNP-MIRT approaches, respectively. The pattern of bias for the QR-based estimates in Figure 10.12 is similar to the corresponding pattern from Condition 1 in Figure 10.7. Specifically, the pattern of positive and negative bias is similar. One apparent difference, though, is the magnitude of the bias, in particular for the highest summed score combinations. The bias in the QR-based estimates for these combinations is greater in Condition 2 than the corresponding bias in Condition 1.

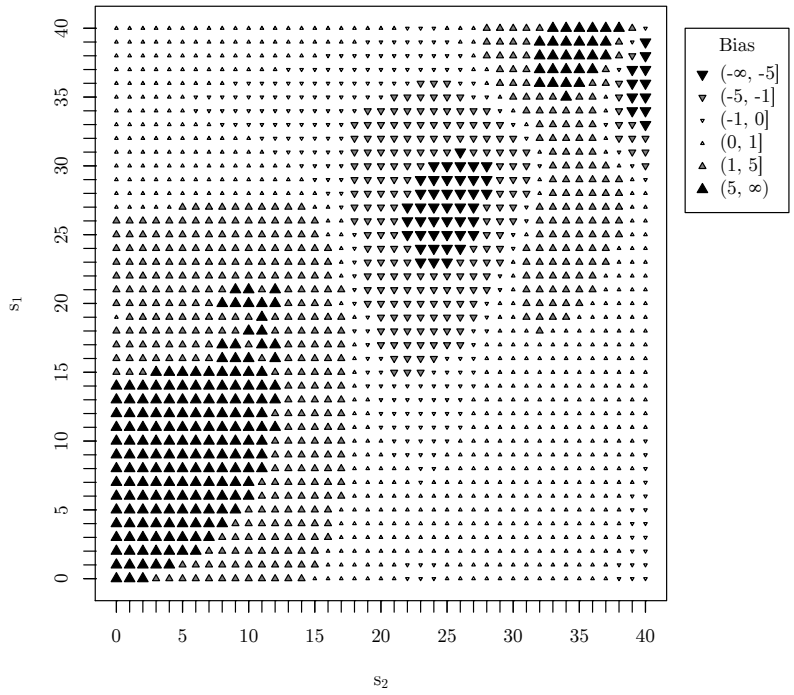
Figure 10.12: Condition 2: Bias for QR Approach



Note. Bias is defined as the average estimate across all 100 replications minus the true value.

Recall Figure 10.8 from Condition 1, which showed that the MIRT-based SGP estimates were approximately unbiased when the latent variable density was correctly specified as normal. Figure 10.13 presents a sharp contrast, as there is considerable bias in the MIRT-based SGP estimates for many cross-classifications. Also, the pattern of bias appears to be mostly a function of s_2. For lower and higher values of s_2, the bias tends to be positive, whereas for more central values of s_2, the bias tends to be negative. This pattern of bias likely depends on the shape of $\varphi(\theta_1, \theta_2)$. In any event, Figures 10.8 and 10.13 makes clear that the proposed method of estimating SGPs based on standard MIRT is clearly sensitive to the misspecification of the latent variable density.

Figure 10.13: Condition 2: Bias for MIRT Approach

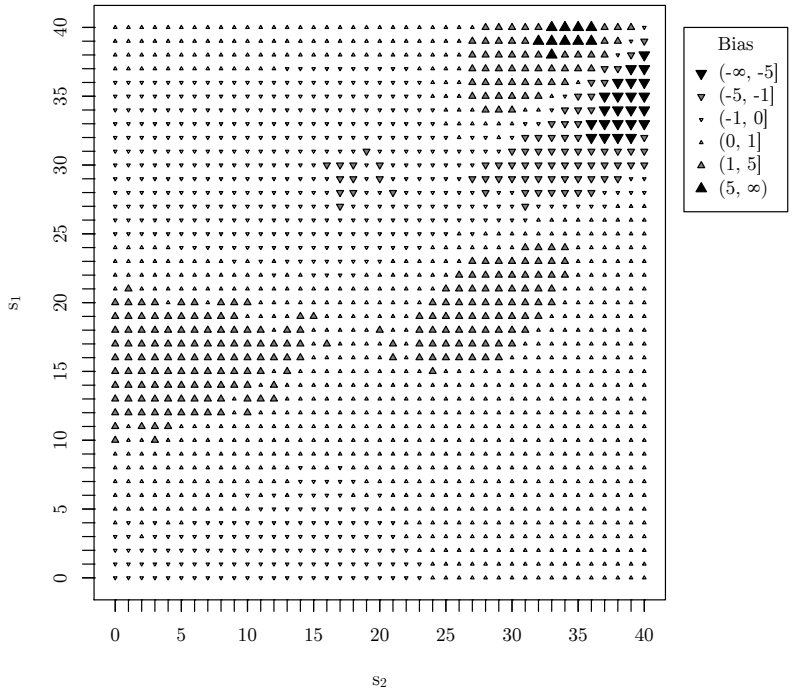


Note. Bias is defined as the average estimate across all 100 replications minus the true value.

Finally, Figure 10.14 presents the bias by cross-classification for the SGP estimates based on the SNP-MIRT model. Overall, the SNP-MIRT approach results in much less bias than the QR and MIRT approaches (see Figures 10.12 and 10.13).

At the same time, the approach results in more bias than the MIRT approach in Simulation 1 (see Figure 10.8). This can be explained by the small amount of bias in the item parameter and density estimation for the SNP-MIRT approach.

Figure 10.14: Condition 2: Bias for SNP-MIRT Approach

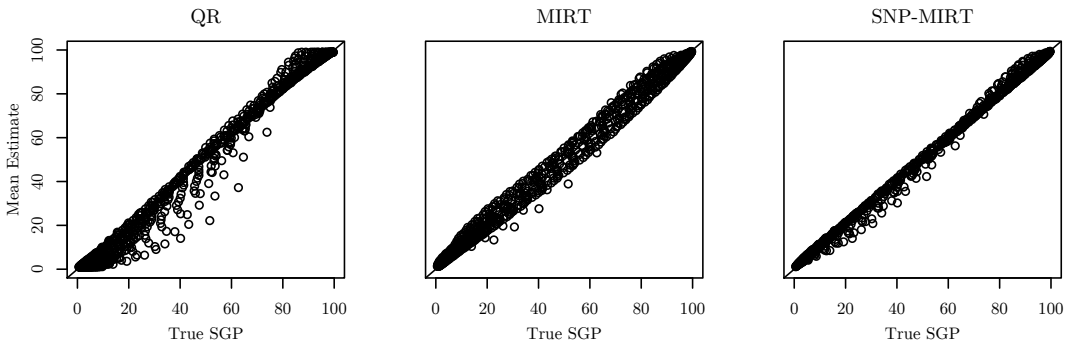


Note. Bias is defined as the average estimate across all 100 replications minus the true value.

As in Condition 1, we also present plots of mean SGP estimates against true SGP values for the 99% HDR score combinations. Figure 10.15 shows these plots for the QR (left plot), MIRT (center plot), and SNP-MIRT (right plot) approaches. As in Condition 1, the QR approach seems to exaggerate the SGP estimates at the high and low ends. The MIRT and SNP-MIRT estimates, in comparison, are better aligned with the 45° line, and it is more difficult to discern any pattern in the plot. Comparing just the MIRT and SNP-MIRT estimates, the latter shows a tighter correspondence with the true SGP values.

The final plot for the results of Condition 2 is Figure 10.16, which displays

Figure 10.15: Condition 2: Estimated and True SGPs



Note. Estimates are averages over all 100 replications. Only the 99% Highest Density Region cross-classifications displayed.

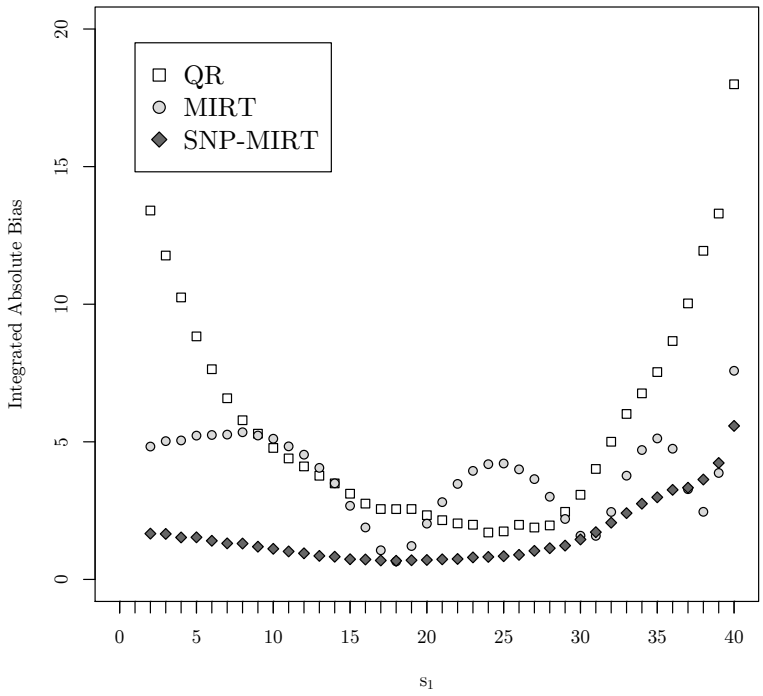
the integrated absolute bias, given s_1, for all three methods. Comparing the QR and SNP-MIRT methods is straightforward: both have higher values at the extremes of s_1, but the bias for the SNP-MIRT approach is always smaller. The results corresponding to the MIRT method, however, are not easily summarized. Overall, the bias values for the MIRT method tend to fall between the SNP-MIRT approach and QR approach, but there are exceptions. For instance, for $s_1 = 38$, the MIRT method produces the smallest bias, while for $s_1 = 25$, it produces the largest bias. This variability across s_1 is likely due to the shape of the latent variable density.

As with Condition 1, we can also examine CCRs for the 3 methods, presented in Table 10.5. The SNP-MIRT approach is the most accurate, regardless of which set of cut-percentiles is used. The other two methods, QR and MIRT, are comparable to one another, but clearly less accurate than the SNP-MIRT approach.

10.4 Empirical Application

To illustrate the proposed SGP estimation methodology, we use longitudinally-matched student achievement data from the 2011-2012 and 2012-2013 academic

Figure 10.16: Condition 2: Integrated Absolute Bias



Note. For each replication, for each s_1, the absolute bias is integrated over the true model-implied probabilities for all s_2. These values are then averaged over all 100 replications.

Table 10.5: Condition 2: Correct Classification Rates

		Correct Classification Rate		
Classes	Cut-Percentiles	QR	MIRT	SNP-MIRT
3	$(0, 35, 65, 100)$	0.922	0.917	0.967
4	$(0, 25, 50, \ldots, 100)$	0.882	0.893	0.947
5	$(0, 20, 40, \ldots, 100)$	0.835	0.848	0.932
10	$(0, 10, 20, \ldots, 100)$	0.677	0.705	0.854

Note. Figures based on simulation study with $N = 10,000$ and 100 replications

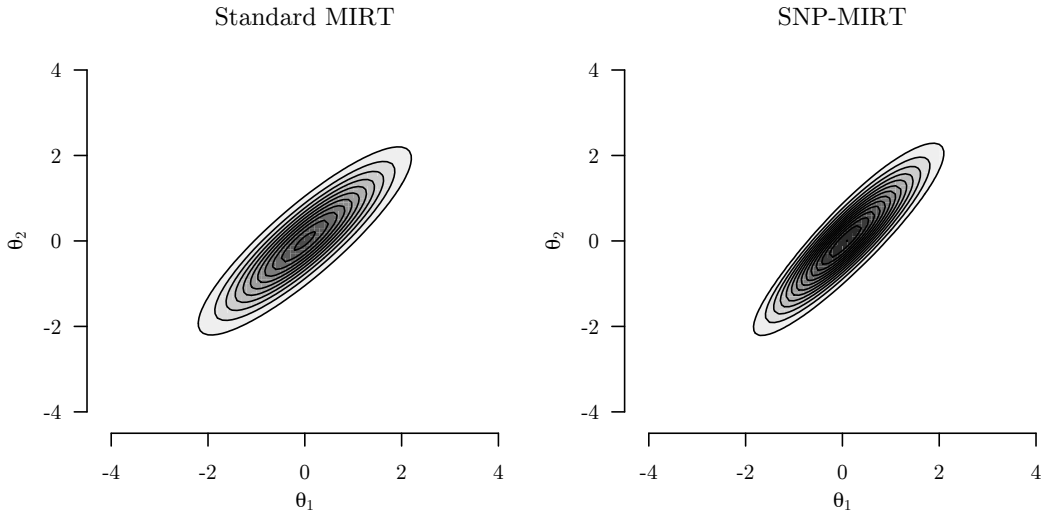
years. The data are from a state's mathematics assessments, but due to confidentiality agreements, the state is not identified. For each year, 44 dichotomous items were analyzed. These items do not constitute a vertical scale. A random sample of 10,000 complete cases was drawn.

SGPs based on the QR, MIRT, and SNP-MIRT approach were calculated as described and illustrated in Section 10.1. As before, a tuning constant of 2 is used for the SNP density, leading to 5 shape parameters in the SNP-MIRT model. Also, as before, the 3PL model, or its multidimensional version, was used for all items.

Figure 10.17 shows the contour plots of the latent variable density for the standard MIRT (left plot) and SNP-MIRT (right plot) models. The estimated SNP density appears approximately normal, although the estimated correlation of 0.86 is slightly smaller than the estimate of 0.88 for the standard MIRT model. Figure 10.18 shows the estimated univariate marginals for the SNP-density (solid black curves) along with normal densities (dashed gray curves) provided for reference. For θ_1, the estimated SNP density is slightly peaked and left-skewed in relation to the normal. A practical interpretation of this is that a greater proportion of students in this sample had lower latent achievement levels in 2011-2012 than would be expected using a normal distribution. On the other hand, for θ_2, the estimated SNP density cannot be easily distinguished from a normal.

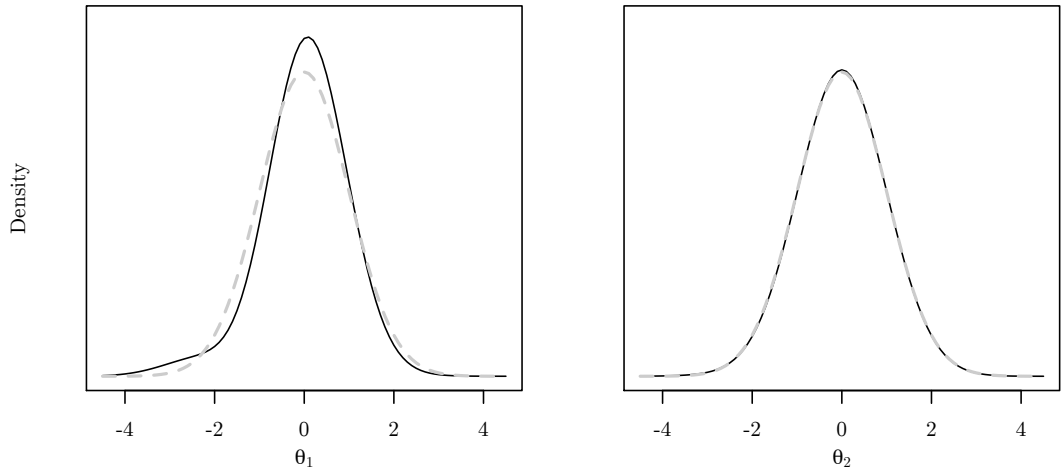
Turning to model comparison, Table 10.6 provides $-2\times$log-likelihood and HQIC values for the different models. The QR approach, which relies on two separate unidimensional IRT models, is formally equivalent to a two-dimensional IRT model where ρ is constrained to 0. Among the competing models, the SNP-MIRT model is preferred by both criteria. Also, since the MIRT and SNP-MIRT models are nested, a likelihood ratio test can be used to judge whether the additional constraints placed on the shape parameters by the standard MIRT model lead to a significant decrement in model fit. Since the test statistic is highly sig-

Figure 10.17: Empirical Application: Bivariate Latent Variable Density



Note. The estimated correlations are $\rho = 0.88$ for the normal and $\rho = 0.86$ for the SNP density.

Figure 10.18: Empirical Application: Univariate Latent Variable Densities



Note. The light gray dashed curves are normal from the standard MIRT estimation; the black solid curves are from the SNP-MIRT estimation.

nificant ($\chi^2_5 = 452.5, p < 0.001$), the conclusion is that the standard MIRT model does not fit the data as well as the SNP-MIRT model.

Table 10.6: State Achievement Data: Model Comparisons

Model	N	Parameters	$\hat{\rho}$	$-2\times$LogL	HQIC
QR	10,000	264	0	987218.00	988390.30
MIRT	10,000	265	0.88	978034.28	978545.35
SNP-MIRT	10,000	270	0.86	977581.80	978102.51

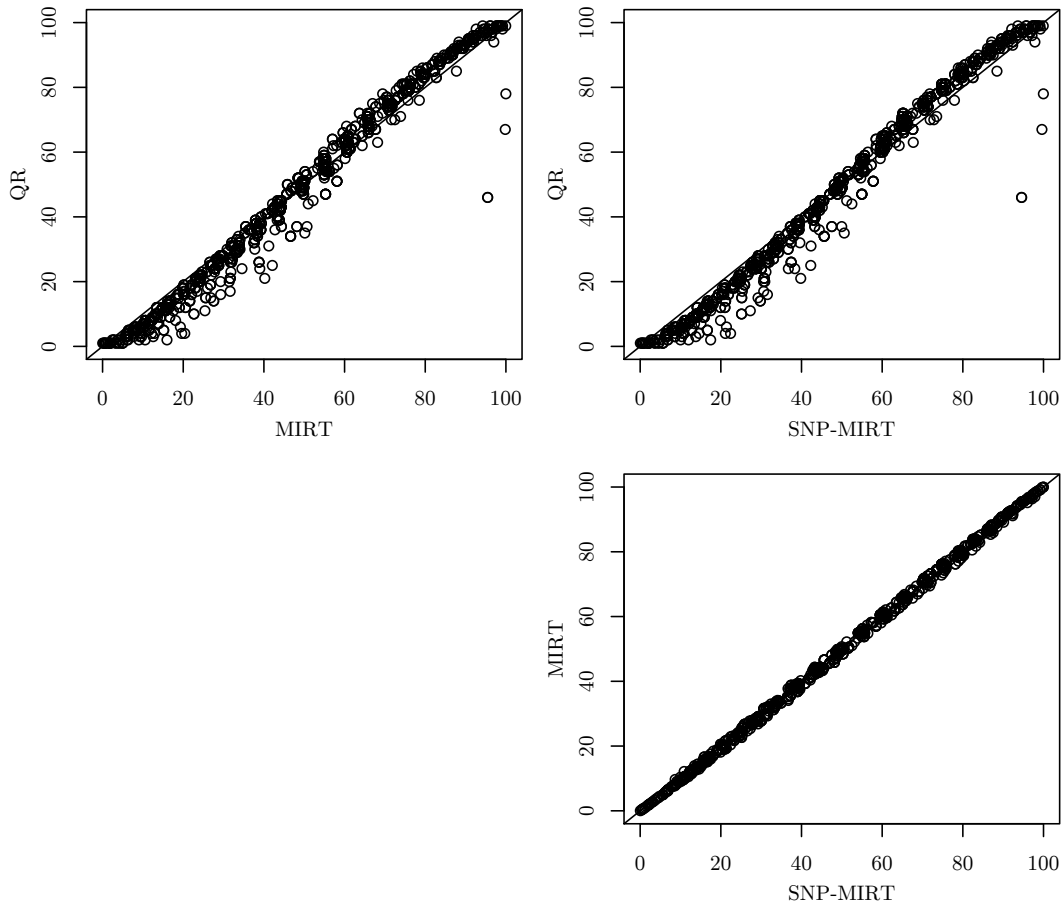
Note. "QR" refers to calibration via two separate unidimensional IRT models, one for year/test 1, one for year/test 2. This is formally equivalent to a two-dimensional IRT model with ρ constrained to 0.

Next, we can compare the SGP estimates from the 3 methods. Figure 10.19 displays bivariate plots of SGP estimates for a random subsample of 1,000 students. The estimates based on the MIRT and SNP-MIRT approaches are highly similar, as evidenced by the correspondence of estimates in the lower-right plot. As for the QR-based estimates, they tend to be more extreme than the estimates based on the other two methods. Interestingly, this is the same pattern exhibited by the QR-based estimates in Condition 1 (see Figure 10.7) and Condition 2 (see Figure 10.12). However, here, the QR-based estimates are being plotted against the MIRT and SNP-MIRT-based estimates, as opposed to the true values.

Table 10.7 also measures the similarity in the 3 sets of estimates by looking at the pairwise classification agreement rates. For the set of cut-percentiles with 3 classes, all methods produce similar results, with agreement rates of 0.94 and higher. However, for the set of cut-percentiles with 10 classes, the results depend substantially on the method. In particular, the QR-based method has low classification agreement rates (< 0.68) with the MIRT and SNP-MIRT methods, while the latter two methods have a relatively high agreement rate (0.92).

The empirical example results provoke several questions and ideas. The SNP-MIRT method yielded a density estimate that was nonnormal, but only slightly

Figure 10.19: Empirical Application: SGPs for All Approaches



Note. Plots use the same random sub-sample of 1,000 students from the full sample of 10,000.

Table 10.7: State Achievement Data: SGP Classification Agreement Rates

		Classification Agreement Rate		
Classes	Cut-Percentiles	QR/MIRT	QR/SNP-MIRT	MIRT/SNP-MIRT
3	$(0, 35, 65, 100)$	0.953	0.946	0.982
4	$(0, 25, 50, \ldots, 100)$	0.907	0.890	0.959
5	$(0, 20, 40, \ldots, 100)$	0.846	0.837	0.953
10	$(0, 10, 20, \ldots, 100)$	0.676	0.660	0.915

Note. Figures based on sample of $N = 10,000$. Classification Agreement Rates are pairwise.

so. Consequently, SGP estimates based on the SNP-MIRT and standard MIRT approaches were highly similar. This suggests that the SNP-MIRT approach may serve as a type of sensitivity analysis for the standard MIRT approach. It is unclear, however, how different the SGP estimates need to be to justify the use of the more complex SNP-MIRT model. Further, the empirical latent variable density suggests that the specified density in Simulation 2 may have been too extreme in its nonnormality. Additional empirical datasets should be analyzed to develop a better understanding of typical density shapes for large-scale longitudinal achievement data.

CHAPTER 11

Summary and Future Directions

11.1 Summary

The present research implemented a new multidimensional item response theory model, the SNP-MIRT model, within a general confirmatory modeling framework. Given that the SNP-MIRT model is parameterized with both mean and variance parameters, it may be substituted freely for the normal density in any confirmatory setting. Simulation studies were carried out to verify the implementation, and the SNP-MIRT model performs admirably in terms of both item parameter and density shape recovery. Notably, even when the true generating distribution is normal, the SNP-MIRT model seems to perform as well as the standard MIRT model. And when the true generating distribution is nonnormal, the SNP-MIRT model typically performs better than the standard model (though the advantage was generally small).

All estimation was performed using the MH-RM algorithm, with a few new wrinkles added to prior implementations. Since the algorithm is firmly rooted in both the stochastic approximation and MCMC methodologies, it can capitalize on innovations from both areas. For instance, in this research, a more nuanced form of gain constant was adopted from the SA literature, as it proved convenient given the complexity of the SNP-MIRT model. As another example, using methodological tools developed for Bayesian inference and MCMC, this research proposed and implemented the two new CJ estimators. The estimators provide approximations

for the log-likelihood and observed information matrix of a model, which are not straightforward to compute in high-dimensional settings.

This research also proposed and studied a new method to calculate SGPs within a MIRT framework, capitalizing on recent research on calibrated projection. The calibrated projection technique can be used to find the reference conditional distribution, which is necessary for SGP estimation. The proposed methods performed well in the simulation study and compared favorably to the original QR-based approach to SGPs. For Condition 1, when the true latent variable density was specified as normal, the MIRT-based approach produced nearly unbiased SGP estimates, while the QR-based approach led to exaggerated SGP estimates. For Condition 2, when the true latent variable density was nonnormal, the results were more mixed. However, the SNP-MIRT approach was effective in estimating the latent variable density shape, and led to the most accurate SGP estimates. Finally, an empirical data analysis of longitudinal achievement data demonstrated that the choice of method affects estimated SGPS, as well as derived classifications.

11.2 Future Directions

Having demonstrated that the SNP-MIRT model is a viable alternative to standard MIRT, future efforts should focus on when the difference in model choice matters. While the simulation studies showed that the SNP-MIRT model is correctly implemented, the relatively large amount of information contributed by the items (20 graded items with 5 categories) limited the influence a misspecified population distribution might have on item and person parameter estimates. Consequently, some simulation work should be devoted to conditions where the items provide less information.

Regarding the algorithmic implementation, a number of research avenues may

be pursued. The CJ estimators show great promise, but its unclear how they compare when timing issues are considered. Compared to the current standard methods, the estimators require more computation per cycle. With that being said, there may be a way to estimate some of the necessary quantities concurrently with the MH-RM algorithm, like the calculation of the Louis (1982) based information approximation in Cai (2010a). More generally, it will be interesting to see what other methods from the SA and MCMC fields eventually prove useful for MIRT modeling.

For the new SGP framework, there are numerous topics for future research. Some of these topics involve the generality of the proposed methods. Theoretically, the framework accommodates scaled scores based on response patterns (as opposed to summed scores) as well as multiple prior years of achievement data. In some sense, using response patterns makes calculating SGPs easier. At the same time, however, the exponentially increasing number of possible response patterns makes studying the methods more challenging. Regarding the incorporation of multiple years of prior data, the MH-RM algorithm (Cai, 2010a, 2010b) provides a way to obtain the required parameter estimates. Also, Thissen et al. (2014) has already demonstrated that calibrated projection can be generalized to more than 2 dimensions. Finally, the framework can theoretically accommodate residual dependencies among items across years (Cai, in press-a) that may result from, for instance, the use of highly similar items in consecutive years. For the reasons just stated, the proposed SGP framework is quite general. However, future research should demonstrate some this generality.

Other potential research topics involve the quality of the SGP estimates themselves. The focus of this preliminary research was on the accuracy of the SGP point estimates. Another obvious evaluation criteria, hinted at by Figure 10.3, is the uncertainty in the SGP estimates. This is a very practical issue, since greater levels of uncertainty would seem to threaten the utility and validity of SGPs, at

least at the student level. On a related note, the new framework also enables more systematic examination of the effects of aggregating SGPs, either to the teacher or school level. Research on these last two topics, in particular, would be of great interest to policymakers.

APPENDIX A

Derivatives for the SNP Density

The log-likelihood of the new parameterization of the SNP density is

$$\begin{aligned} l = \log h_y(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}) &= 2 \log P_\kappa(\mathbf{x}; \boldsymbol{\nu}) + \log \phi(\mathbf{x}) + \log |\mathbf{S}| + \log |\mathbf{R}^{-1}| \\ &= l_1 + l_2 + l_3 + l_4. \end{aligned} \tag{A.1}$$

This Appendix presents the first and second derivatives of Equation (A.1), and is organized as follows. First notational conventions will be introduced. Then, a number of partial first and second derivatives will be given. Finally, the derivatives for Equation (A.1) will be presented.

A.1 Notational Conventions

Let $\mathbf{E}$ be the $p^* \times p^2$ elimination matrix; let $\mathbf{K}$ be the $p^2 \times p^2$ commutation matrix; and let $\mathbf{I}$ be the $p \times p$ identity matrix.

Also, let $\mathbf{s}^{-1} = \text{vech}(\mathbf{S}^{-1})$ and, analogously, let $\mathbf{r}^{-1} = \text{vech}(\mathbf{R}^{-1})$

Generally, let the '∇' symbol indicate derivatives with respect to $\boldsymbol{\nu}$. For example, $\nabla\psi_i = \partial\psi_i/\partial\nu_i$, noting that the argument index also serves to index $\boldsymbol{\nu}$. Similarly, let ∇^2 be defined analogously such that $\nabla^2\psi_i = \partial^2\psi_i/\partial\nu_i^2$

Generally, let the 'dot' and 'double-dot' symbols indicate derivatives with respect to the $t \times 1$ vector $\mathbf{c}$. As an example, consider the $p \times 1$ vector $\mathbf{x}$. Then, $\dot{x}_k$ is a $t \times 1$ vector with ith element $\partial x_k/\partial c_i$, while $\dot{\mathbf{x}}$ is a $p \times t$ matrix with (i, j)th

element $\partial x_i/\partial c_j$. Also, $\ddot{x}_k$ is a $t \times t$ matrix with (i,j)th element $\partial x_k/(\partial c_i \partial c_j)$.

A.2 Some Partial Derivatives

A.2.1 Derivatives of c With Respect to $\boldsymbol{\nu}$

Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be $q \times 1$ vectors such that $\alpha_i = \cos(\delta_i)$ and $\beta_i = \sin(\delta_i)$. Then, the $t \times 1$ vector $\mathbf{c}$ is defined as

$$\begin{aligned} c_1 &= \alpha_1 \\ c_2 &= \beta_1 \alpha_2 \\ c_3 &= \beta_1 \beta_2 \alpha_3 \\ \vdots &= \vdots \\ c_{t-1} &= \beta_1 \beta_2 \cdots \beta_{q-1} \alpha_q \\ c_t &= \beta_1 \beta_2 \cdots \beta_{q-1} \beta_q. \end{aligned} \tag{A.2}$$

Define the following derivatives: $\nabla \delta_i = 1/(1+\nu_i^2)$; $\nabla \alpha_i = -\beta_i \nabla \delta_i$; and $\nabla \beta_i = \alpha_i \nabla \delta_i$.

Then, let $\partial \mathbf{c}/\partial \boldsymbol{\nu}$ be the $t \times q$ matrix of first derivatives, where the (i,j)th element is

$$\frac{\partial c_i}{\partial \nu_j} = \begin{cases} 0 & \text{for } j > i \\ \nabla \alpha_i \prod_{l=1}^{i-1} \beta_l & \text{for } j = i < r \\ \alpha_i \nabla \beta_j \prod_{l \neq j}^{i-1} \beta_l & \text{for } j < i < r \\ \nabla \beta_j \prod_{l \neq j}^{q} \beta_l & \text{for } i = r. \end{cases} \tag{A.3}$$

Moving on to the second derivative, $\nabla^2 \delta_i = -2\nu_i/(1+\nu_i^2)^2$; $\nabla^2 \alpha_i = -\alpha_i \nabla \delta_i^2 - \beta_i \nabla^2 \delta_i$; and $\nabla^2 \beta_i = -\beta_i \nabla \delta_i^2 + \alpha_i \nabla^2 \delta_i$.

Finally, let $\partial^2 c_k/(\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}')$ be the $q \times q$ matrix of second derivatives for c_k. These

matrices are necessarily symmetric, and only the lower triangular elements will be defined. For $k < t$, the (i, j)th element is

$$\frac{\partial^2 c_k}{\partial \nu_i \partial \nu_j} = \begin{cases} \nabla^2 \beta_i \alpha_k \prod_{l \neq i}^{k-1} \beta_l & \text{for } j = i < k \\ \nabla \beta_i \nabla \beta_j \alpha_k \prod_{l \neq i,j}^{k-1} \beta_l & \text{for } j < i < k \\ \nabla \beta_j \nabla \alpha_k \prod_{l \neq j}^{k-1} \beta_l & \text{for } j < i = k \\ \nabla^2 \alpha_k \prod_{l=1}^{k-1} \beta_l & \text{for } j = i = k. \end{cases} \tag{A.4}$$

For $k = t$, the (i, j)th element is

$$\frac{\partial^2 c_k}{\partial \nu_i \partial \nu_j} = \begin{cases} \nabla^2 \beta_i \prod_{l \neq i}^{k-1} \beta_l & \text{for } j = i \\ \nabla \beta_i \nabla \beta_j \prod_{l \neq i,j}^{k-1} \beta_l & \text{for } j < i. \end{cases} \tag{A.5}$$

A.2.2 Derivatives of m With Respect to c

The first derivatives of elements of $\mathbf{m}$ are defined as

$$\dot{m}_i = \frac{\partial m_i}{\partial \mathbf{c}} = 2(\mathbf{B}^{-1})' \mathbf{A}_i^* \mathbf{a}, \tag{A.6}$$

and let $\dot{\mathbf{m}}$ be the $p \times t$ matrix with ith row given by $\dot{m}_i'$. Let the matrix of second derivatives be

$$\ddot{m}_i := \frac{\partial^2 m_i}{\partial \mathbf{c} \partial \mathbf{c}'} = 2(\mathbf{B}^{-1})' \mathbf{A}_i^* \mathbf{B}^{-1}. \tag{A.7}$$

A.2.3 Derivatives of v With Respect to c

Let v_k be the kth element of the $p^* \times 1$ vector $\mathbf{v}$. Further, let v_k correspond to the (i, j)th element of $\mathbf{V}$. Then,

$$\dot{v}_k = \frac{\partial v_k}{\partial \mathbf{c}} = 2(\mathbf{B}^{-1})' \mathbf{A}_{ij}^* \mathbf{a} - (m_j \dot{m}_i + m_i \dot{m}_j). \tag{A.8}$$

Let $\dot{\mathbf{v}}$ be the $p^* \times t$ matrix

$$\dot{\mathbf{v}} = \left[\dot{v}_1 \cdots \dot{v}_{p^*}.\right]' \tag{A.9}$$

The matrix of second derivatives is

$$\ddot{v}_k = \frac{\partial^2 v_k}{\partial \mathbf{c} \partial \mathbf{c}'} = 2(\mathbf{B}^{-1})' \mathbf{A}^*_{ij} \mathbf{B}^{-1} - (m_j \ddot{m}_i + m_i \ddot{m}_j + \dot{m}_i \dot{m}'_j + \dot{m}_j \dot{m}'_i). \tag{A.10}$$

A.2.4 Derivatives of s With Respect to c

For $\mathbf{s}$, the $p^* \times t$ matrix $\dot{\mathbf{s}}$ is defined as

$$\dot{\mathbf{s}} = \frac{\partial \mathbf{s}}{\partial \mathbf{c}'} = \frac{\partial \mathbf{s}}{\partial \mathbf{v}} \dot{\mathbf{v}}, \tag{A.11}$$

and the matrix of second derivatives is

$$\ddot{s}_k = \frac{\partial^2 s_k}{\partial \mathbf{c} \partial \mathbf{c}'} = \dot{\mathbf{v}}' \frac{\partial^2 s_k}{\partial \mathbf{v} \partial \mathbf{v}'} \mathbf{v} + \sum_{i=1}^{p^*} \frac{\partial s_k}{\partial v_i} \ddot{v}_i. \tag{A.12}$$

A.2.5 Derivatives of x With Respect to c

Let $\mathbf{U}$ be the $p \times p^*$ matrix $\mathbf{U} = (\mathbf{z}' \otimes \mathbf{I})\mathbf{E}'$. Then, let $\dot{\mathbf{x}}$ be the $p \times t$ matrix with ith row $\dot{x}'_i$. It is defined as

$$\dot{\mathbf{x}} = \left[\dot{x}_1 \cdots \dot{x}_p\right]' = \mathbf{U}\dot{\mathbf{s}} + \dot{\mathbf{m}}. \tag{A.13}$$

Let the $p^* \times 1$ vector $\mathbf{u}^{(k)}$ be formed from the kth row of $\mathbf{U}$. Then,

$$\ddot{x}_k = \frac{\partial^2 x_k}{\partial \mathbf{c} \partial \mathbf{c}'} = \ddot{m}_k + \sum_{i=1}^{p^*} \mathbf{u}_i^{(k)} \ddot{s}_i. \tag{A.14}$$

A.2.6 Derivatives of $\boldsymbol{\eta}$ With Respect to c

For $\boldsymbol{\eta}$,

$$\frac{\partial \eta_i}{\partial x_j} = \frac{\lambda_{ij}\eta_i}{x_j}, \tag{A.15}$$

and let $\partial\eta_i/\partial\mathbf{x}$ be the $p \times 1$ vector with jth element $\partial\eta_i/\partial x_j$. Then,

$$\dot{\eta}_k = \frac{\partial \eta_k}{\partial \mathbf{c}} = \dot{\mathbf{x}}' \frac{\partial \eta_k}{\partial \mathbf{x}}, \tag{A.16}$$

and the $t \times t$ matrix $\dot{\boldsymbol{\eta}}$ has in its ith row $\dot{\eta}_i'$, that is

$$\dot{\boldsymbol{\eta}} = \left[\dot{\eta}_1 \cdots \dot{\eta}_t\right]'. \tag{A.17}$$

The matrix of second derivatives, with respect to elements of x, has typical element

$$\frac{\partial^2 \eta_k}{\partial x_i \partial x_j} = \begin{cases} \dfrac{\lambda_{ki}\lambda_{kj}\eta_k}{x_i x_j} & \text{if } i \neq j \\ \dfrac{(\lambda_{ki}-1)\lambda_{ki}\eta_k}{x_i^2} & \text{if } i = j. \end{cases} \tag{A.18}$$

Then,

$$\ddot{\eta}_k = \frac{\partial^2 \eta_k}{\partial \mathbf{c} \partial \mathbf{c}'} = \dot{\mathbf{x}}' \frac{\partial^2 \eta_k}{\partial \mathbf{x} \partial \mathbf{x}'} \dot{\mathbf{x}} + \sum_{i=1}^{p} \left(\frac{\partial \eta_k}{\partial \mathbf{x}}\right)_i \ddot{x}_i. \tag{A.19}$$

A.2.7 Derivatives of x With Respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

For first derivatives,

$$\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}} = -\mathbf{S}\mathbf{R}^{-1}, \tag{A.20}$$

and

$$\frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}} = \frac{\partial \mathbf{x}}{\partial \mathbf{r}^{-1}} \frac{\partial \mathbf{r}^{-1}}{\partial \mathbf{r}} \frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}}, \tag{A.21}$$

$$\frac{\partial \mathbf{x}}{\partial \mathbf{r}^{-1}} = ((\mathbf{y} - \boldsymbol{\mu})' \otimes \mathbf{S})\mathbf{E}', \tag{A.22}$$

$$\frac{\partial \mathbf{r}^{-1}}{\partial \mathbf{r}} = \mathbf{E}(-(\mathbf{R}^{-1})' \otimes \mathbf{R}^{-1})\mathbf{E}', \tag{A.23}$$

and $\frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}}$ is defined in Smith (1995).

For second derivatives,

$$\frac{\partial^2 x_k}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = \sum_{i=1}^{p^*} \left(\frac{\partial x_k}{\partial \mathbf{r}^{-1}} \right)_i \frac{\partial^2 r_i^{-1}}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}, \tag{A.24}$$

where

$$\frac{\partial^2 r_i^{-1}}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}} \right)' \frac{\partial^2 r_i^{-1}}{\partial \mathbf{r} \partial \mathbf{r}'} \left(\frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}} \right) + \sum_{j=1}^{p^*} \left(\frac{\partial r_i^{-1}}{\partial \mathbf{r}} \right)_j \frac{\partial^2 r_j}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}, \tag{A.25}$$

and $\frac{\partial^2 r_j}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}$ is defined in Smith (1995). In Equation (A.25), $\frac{\partial^2 r_i^{-1}}{\partial \mathbf{r} \partial \mathbf{r}'}$ may be obtained from

$$\mathbf{G} = (\mathbf{I} \otimes \mathbf{K} \otimes \mathbf{I})(((\mathbf{R}^{-1})' \otimes \mathbf{R}^{-1}) \otimes \mathbf{K}\text{vec}(\mathbf{R}^{-1}) + \text{vec}(\mathbf{R}^{-1}) \otimes (\mathbf{K}(\mathbf{R}^{-1})' \otimes \mathbf{R}^{-1}))). \tag{A.26}$$

The $p^4 \times p^2$ matrix $\mathbf{G}$ is structured as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} \\ \vdots \\ \mathbf{G}_{p1} \\ \vdots \\ \mathbf{G}_{1p} \\ \vdots \\ \mathbf{G}_{pp} \end{bmatrix}, \tag{A.27}$$

where

$$\mathbf{G}_{mn} = \frac{\partial^2 \mathbf{R}_{mn}^{-1}}{\partial \text{vec}(\mathbf{R}) \partial \text{vec}(\mathbf{R})'}. \tag{A.28}$$

As an example of how to obtain the desired quantity, let $i = 1$. Then,

$$\frac{\partial^2 r_1^{-1}}{\partial \mathbf{r} \partial \mathbf{r}'} = \mathbf{E}\mathbf{G}_{11}\mathbf{E}'. \tag{A.29}$$

The cross-partial derivative, with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, is

$$\frac{\partial^2 x_k}{\partial \boldsymbol{\mu} \partial \boldsymbol{\sigma}'} = \frac{\partial^2 x_k}{\partial \boldsymbol{\mu} \partial (\mathbf{r}^{-1})'} \frac{\partial r^{-1}}{\partial \boldsymbol{\sigma}}. \tag{A.30}$$

In Equation (A.30), $\frac{\partial^2 x_k}{\partial \boldsymbol{\mu} \partial (\mathbf{r}^{-1})'}$ may be obtained from

$$\mathbf{H} = -(\mathbf{R} \otimes \mathbf{I})\mathbf{K}\mathbf{E}'. \tag{A.31}$$

The $p^2 \times p^*$ matrix $\mathbf{H}$ is structured as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_p \end{bmatrix}, \tag{A.32}$$

where

$$\mathbf{H}_k = \frac{\partial^2 x_k}{\partial \boldsymbol{\mu} \partial (\mathbf{r}^{-1})'}. \tag{A.33}$$

A.3 Full Derivatives for the SNP Log-Likelihood

A.3.1 First Derivatives of l_1

Using the chain rule and

$$\frac{\partial l_1}{\partial \mathbf{x}} = \frac{2}{P(\mathbf{x})} \frac{\partial P(\mathbf{x})}{\partial \mathbf{x}}, \tag{A.34}$$

where

$$\frac{\partial P(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{x}} \right)' \mathbf{a}, \tag{A.35}$$

the first derivatives of l_1 with respect to $\boldsymbol{\psi}$ are given by

$$\frac{\partial l_1}{\partial \boldsymbol{\mu}} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)' \frac{\partial l_1}{\partial \mathbf{x}} \tag{A.36}$$

and

$$\frac{\partial l_1}{\partial \boldsymbol{\sigma}} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}\right)' \frac{\partial l_1}{\partial \mathbf{x}}. \tag{A.37}$$

Finally,

$$\frac{\partial l_1}{\partial \boldsymbol{\nu}} = \frac{2}{P(\mathbf{x})} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial P(\mathbf{x})}{\partial \mathbf{c}}, \tag{A.38}$$

where

$$\frac{\partial P(\mathbf{x})}{\partial \mathbf{c}} = (\mathbf{B}^{-1})' \boldsymbol{\eta} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{c}}\right)' \mathbf{a}. \tag{A.39}$$

A.3.2 Second Derivatives of l_1

The second derivatives of l_1 are as follows:

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)' \boldsymbol{\Gamma} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}, \tag{A.40}$$

where

$$\boldsymbol{\Gamma} = -\frac{2}{P(\mathbf{x})^2} \frac{\partial P(\mathbf{x})}{\partial \mathbf{x}} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{x}}\right)' + \sum_{k=1}^{t} a_k \frac{\partial^2 \eta_k}{\partial \mathbf{x} \partial \mathbf{x}'}, \tag{A.41}$$

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}\right)' \boldsymbol{\Gamma} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}} + \frac{2}{P(\mathbf{x})} \sum_{i=1}^{p} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{x}}\right)_i \frac{\partial^2 x_i}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}, \tag{A.42}$$

and

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'} = \frac{2}{P(\mathbf{x})} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \boldsymbol{\Xi}^{(1)} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}} + \sum_{k=1}^{t} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{c}}\right)_k \frac{\partial^2 c_k}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'}, \tag{A.43}$$

where

$$\boldsymbol{\Xi}^{(1)} = -\frac{1}{P(\mathbf{x})} \frac{\partial P(\mathbf{x})}{\partial \mathbf{c}} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{c}}\right)' + 2 \left(\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{c}}\right)' \mathbf{B}^{-1} + \sum_{k=1}^{t} a_k \frac{\partial^2 \eta_k}{\partial \mathbf{c} \partial \mathbf{c}'}. \tag{A.44}$$

The cross-partial derivatives are given by:

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\mu} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)' \boldsymbol{\Gamma} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}} + \frac{2}{P(\mathbf{x})} \sum_{i=1}^{p} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{x}}\right)_i \frac{\partial^2 x_i}{\partial \boldsymbol{\mu} \partial \boldsymbol{\sigma}'}, \tag{A.45}$$

and using the chain rule with

$$\begin{aligned} \frac{\partial^2 l_1}{\partial \boldsymbol{\mu} \partial \mathbf{x}'} &= -\frac{2}{P(\mathbf{x})^2} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial P(\mathbf{x})}{\partial \mathbf{x}} \left(\frac{\partial P(\mathbf{x})}{\partial \mathbf{x}}\right)' \\ &\quad -\frac{2}{P(\mathbf{x})} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \left[\left(\mathbf{B}^{-1}\right)' \frac{\partial \boldsymbol{\eta}}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{x}}{\partial \mathbf{c}}\right)' \frac{\partial^2 P(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right], \end{aligned} \tag{A.46}$$

where

$$\frac{\partial^2 P(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} = \sum_{k=1}^{t} a_k \frac{\partial^2 \eta_k}{\partial \mathbf{x} \partial \mathbf{x}'}, \tag{A.47}$$

then

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'} = \frac{\partial^2 l_1}{\partial \boldsymbol{\nu} \partial \mathbf{x}'} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}} \tag{A.48}$$

and

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\nu} \partial \boldsymbol{\sigma}'} = \frac{\partial^2 l_1}{\partial \boldsymbol{\nu} \partial \mathbf{x}'} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}. \tag{A.49}$$

A.3.3 First Derivatives of l_2

The first derivatives are

$$\frac{\partial l_2}{\partial \boldsymbol{\mu}} = -\mathbf{x} \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)', \tag{A.50}$$

$$\frac{\partial l_2}{\partial \boldsymbol{\sigma}} = -\mathbf{x} \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}\right)', \tag{A.51}$$

and

$$\frac{\partial l_2}{\partial \boldsymbol{\nu}} = \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial l_2}{\partial \mathbf{c}}, \tag{A.52}$$

where

$$\frac{\partial l_2}{\partial \mathbf{c}} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{c}}\right)' \mathbf{x}. \tag{A.53}$$

A.3.4 Second Derivatives of l_2

The second derivatives are of l_2 are as follows:

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = -\left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)' \frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}, \tag{A.54}$$

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = -\left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}\right)' \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}} - \sum_{i=1}^{p} x_i \frac{\partial^2 x_i}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}, \tag{A.55}$$

and

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'} = \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \boldsymbol{\Xi}^{(2)} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}} + \sum_{k=1}^{t} \left(\frac{\partial l_2}{\partial \mathbf{c}}\right)_k \frac{\partial^2 c_k}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'}, \tag{A.56}$$

where

$$\boldsymbol{\Xi}^{(2)} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{c}}\right)' \frac{\partial \mathbf{x}}{\partial \mathbf{c}} - \sum_{i=1}^{p} x_i \frac{\partial^2 x_i}{\partial \mathbf{c} \partial \mathbf{c}'}. \tag{A.57}$$

The cross-partial derivatives are given by:

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\mu} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}\right)' \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}} - \sum_{i=1}^{p} x_i \frac{\partial^2 x_i}{\partial \boldsymbol{\mu} \partial \boldsymbol{\sigma}'}, \tag{A.58}$$

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\mu}'} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}}, \tag{A.59}$$

and

$$\frac{\partial^2 l_2}{\partial \boldsymbol{\nu} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial \mathbf{x}}{\partial \boldsymbol{\sigma}}. \tag{A.60}$$

A.3.5 First Derivatives of l_3

l_3 is only a function of $\boldsymbol{\nu}$. So, all derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are 0. The first derivatives of l_3 are

$$\frac{\partial l_3}{\partial \boldsymbol{\nu}} = \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \frac{\partial l_3}{\partial \mathbf{c}}, \tag{A.61}$$

where

$$\frac{\partial l_3}{\partial \mathbf{c}} = \left(\frac{\partial \mathbf{s}}{\partial \mathbf{c}}\right)' \mathbf{s}^{-1}. \tag{A.62}$$

A.3.6 Second Derivatives of l_3

The second derivatives of l_3 are

$$\frac{\partial^2 l_3}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'} = \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}}\right)' \boldsymbol{\Xi}^{(3)} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\nu}} + \sum_{k=1}^{t} \left(\frac{\partial l_3}{\partial \mathbf{c}}\right)_k \frac{\partial^2 c_k}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}'}, \tag{A.63}$$

where

$$\boldsymbol{\Xi}^{(3)} = -\left(\frac{\partial \mathbf{s}}{\partial \mathbf{c}}\right)' \mathbf{E} \left(\mathbf{K}((\mathbf{S}^{-1})' \otimes \mathbf{S}^{-1})\right) \mathbf{E}' \frac{\partial \mathbf{s}}{\partial \mathbf{c}} - \sum_{j=1}^{p^*} s_j^{-1} \frac{\partial^2 s_j}{\partial \mathbf{c} \partial \mathbf{c}'}. \tag{A.64}$$

A.3.7 First Derivatives of l_4

l_4 is only a function of $\boldsymbol{\sigma}$. So, all derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are 0. The first derivatives of l_4 are

$$\frac{\partial l_4}{\partial \boldsymbol{\sigma}} = \left(\frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}}\right)' \frac{\partial l_4}{\partial \mathbf{r}}, \tag{A.65}$$

where

$$\frac{\partial l_4}{\partial \mathbf{c}} = \mathbf{E}\mathbf{K}\text{vec}(\mathbf{R}). \tag{A.66}$$

A.3.8 Second Derivatives of l_4

The second derivatives of l_4 are

$$\frac{\partial^2 l_4}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = \left(\frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}}\right)' \frac{\partial^2 l_4}{\partial \mathbf{r} \partial \mathbf{r}'} \frac{\partial \mathbf{r}}{\partial \boldsymbol{\sigma}} + \sum_{i=1}^{p} \left(\frac{\partial l_4}{\partial \mathbf{r}}\right)_i \frac{\partial^2 r_i}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}, \tag{A.67}$$

where

$$\frac{\partial^2 l_4}{\partial \mathbf{r} \partial \mathbf{r}'} = -\mathbf{E}\mathbf{K} \left(\mathbf{R}' \otimes \mathbf{R}\right) \mathbf{E}'. \tag{A.68}$$

APPENDIX B

Derivatives for the Observed Information Estimator

This Appendix presents dervatives needed for the observed information estimator $\mathcal{F}_{CJ}$ presented in Chapter 6. In particular, the derivative $\nabla_{\boldsymbol{\omega}} \log \boldsymbol{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega})$ is needed, for an arbitrary response pattern $\mathbf{y}$.

Using Equation (6.7),

$$\begin{aligned} \nabla_{\boldsymbol{\omega}} \log \boldsymbol{\Pi}(\boldsymbol{\theta}^*|\mathbf{y}, \boldsymbol{\omega}) &= \nabla_{\boldsymbol{\omega}} \log \bar{u} - \nabla_{\boldsymbol{\omega}} \log \bar{v} \\ &= \frac{1}{\bar{u}} \nabla_{\boldsymbol{\omega}} \bar{u} - \frac{1}{\bar{v}} \nabla_{\boldsymbol{\omega}} \bar{v}. \end{aligned} \tag{B.1}$$

Then, let $\nabla_{\boldsymbol{\zeta}}$ return the $d_1 \times 1$ vector of partial derivatives with respect to the item parameters $\boldsymbol{\zeta}$ and let $\nabla_{\boldsymbol{\xi}}$ return the $d_2 \times 1$ vector of partial derivatives with respect to the group parameters $\boldsymbol{\xi}$. Therefore, $\nabla_{\boldsymbol{\omega}} \bar{u} = ((\nabla_{\boldsymbol{\zeta}} \bar{u})', (\nabla_{\boldsymbol{\xi}} \bar{u})')'$, and $\nabla_{\boldsymbol{\omega}} \bar{v} = ((\nabla_{\boldsymbol{\zeta}} \bar{v})', (\nabla_{\boldsymbol{\xi}} \bar{v})')'$. Below, each of these 4 quantities will be defined.

At this point, it is convient to define

$$\varrho(\boldsymbol{\theta}, \check{\boldsymbol{\theta}}) = \frac{f(\mathbf{y}|\check{\boldsymbol{\theta}}, \boldsymbol{\zeta}) \varphi(\check{\boldsymbol{\theta}}; \boldsymbol{\xi})}{f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\zeta}) \varphi(\boldsymbol{\theta}; \boldsymbol{\xi})}. \tag{B.2}$$

Now, consider $\nabla_{\boldsymbol{\zeta}} \bar{u}$:

$$\nabla_{\boldsymbol{\zeta}}\bar{u} = C^{-1}\sum_{c=1}^{C} q(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*})\nabla_{\boldsymbol{\zeta}}\alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}), \tag{B.3}$$

where

$$\nabla_{\boldsymbol{\zeta}}\alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}) = \begin{cases} 0 & \text{if } \alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}) \geq 1 \\ \varrho(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*})\nabla_{\boldsymbol{\zeta}} \log\left(\dfrac{f(\mathbf{y}|\boldsymbol{\theta}^{*}, \boldsymbol{\zeta})}{f(\mathbf{y}|\boldsymbol{\theta}^{(c)}, \boldsymbol{\zeta})}\right) & \text{otherwise.} \end{cases} \tag{B.4}$$

Note that the second case in Equation (B.4) involves a difference of complete data log-likelihood derivatives for items, which are used in MH-RM.

Next, consider $\nabla_{\boldsymbol{\zeta}}\bar{v}$:

$$\nabla_{\boldsymbol{\zeta}}\bar{v} = J^{-1}\sum_{j=1}^{J} \nabla_{\boldsymbol{\zeta}}\alpha(\boldsymbol{\theta}^{*}, \boldsymbol{\theta}^{(j)}), \tag{B.5}$$

where $\nabla_{\boldsymbol{\zeta}}\alpha(\boldsymbol{\theta}^{*}, \boldsymbol{\theta}^{(j)})$ may be calculated by Equation (B.4), substituting in appropriate arguments.

Next, consider $\nabla_{\boldsymbol{\xi}}\bar{u}$:

$$\nabla_{\boldsymbol{\zeta}}\bar{u} = C^{-1}\sum_{c=1}^{C} q(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*})\nabla_{\boldsymbol{\xi}}\alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}), \tag{B.6}$$

where

$$\nabla_{\boldsymbol{\xi}}\alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}) = \begin{cases} 0 & \text{if } \alpha(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*}) \geq 1 \\ \varrho(\boldsymbol{\theta}^{(c)}, \boldsymbol{\theta}^{*})\nabla_{\boldsymbol{\xi}} \log\left(\dfrac{\varphi(\boldsymbol{\theta}^{*}; \boldsymbol{\xi})}{\varphi(\boldsymbol{\theta}; \boldsymbol{\xi})}\right) & \text{otherwise.} \end{cases} \tag{B.7}$$

Note that the second case in Equation (B.7) involves a difference of complete data

log-likelihood derivatives for density models, which are used in MH-RM.

Finally, consider $\nabla_{\boldsymbol{\xi}}\bar{v}$:

$$\nabla_{\boldsymbol{\xi}}\bar{v} = J^{-1}\sum_{j=1}^{J}\nabla_{\boldsymbol{\xi}}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}), \tag{B.8}$$

where $\nabla_{\boldsymbol{\xi}}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)})$ may be calculated by Equation (B.7), substituting in appropriate arguments.

Bibliography

Adams, R. (Ed.). (2005). *PISA 2003 technical report* (Tech. Rep.). Paris, France: Organisation for Economic Co-operation and Development.

Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. (2005). Skew-normal linear mixed models. *Journal of Data Science*, *3*, 415-438.

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

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2014). SGP: An r package for the calculation and visualization of student growth percentiles & percentile growth trajectories. [Computer software manual]. Retrieved from `http://centerforassessment.github.com/SGP/` (R package version 1.2-0.0)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (chap. 17-20). Reading, MA: Addison-Wesley.

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

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society - Series B*, *61*, 265-285.

Cai, L. (2008a). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model.* Department of Psychology, University of North Carolina,

Chapel Hill. (Unpublished Doctoral Dissertation)

Cai, L. (2008b). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309-329.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L. (in press-a). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*.

Cai, L. (in press-b). Two-tier item factor analysis modeling. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (Second ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software Internatonal, Inc.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 23-28.

Chen, J., Zhang, D., & Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution.

Biostatistics, *3*(3), 347-360.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, 1-38.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: method and application. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (p. 259-273). London: Chapman and Hall.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological Bulletin*, *136*, 103-127.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700-725.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363-390.

Gallant, A. R., & Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Assset pricing applications. *Econometrica*, *57*(5), 1091-1120.

Ghidey, W., Lesaffre, E., & Verbeke, G. (2008). A comparison of methods for estimating the random efects distribution of a linear mixed model. *Statistical Methods in Medical Research*, *19*, 575–600.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Herman, J. L., & Linn, R. L. (2013). *On the road to assessing deeper learning: The status of smarter balanced and parcc assessment consortia* (Tech. Rep. No. 823). Los Angeles, CA: University of California: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494-495.

Irincheeva, I., Cantoni, E., & Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, 1-18.

Jara, A., Quintana, F., & San Martin, E. (2008). Linear mixed models with skew-elliptical distributions: a Bayesian approach. *Computational Statistics and Data Analysis*, *52*(11), 5033-5045.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, vol. 1.* New York, NY: John Wiley and Sons.

Knott, M., & Tzamourani, P. (2007). Bootstrapping the estimated latent distribution of the two-parameter latent trait model. *British Journal of Mathematical and Statistical Psychology*, *60*, 175-191.

Koenker, R. (2005). *Quantile regression.* New York, NY: Cambridge University Press.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *3*, *46*(1), 33-50.

Lin, T. I., & Lee, J. C. (2008). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Statistics in Medicine*, *27*, 1490-1507.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society - Series B*, *44*(2), 226-233.

Magder, L. S., & Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association*, *91*(435), 1141-1151.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E.

(1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087-1092.

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

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve IRT model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343-369.

Montanari, A., & Viroli, C. (2010a). Heteroscedastic factor mixture analysis. *Statistical Modelling*, *10*(4), 441-460.

Montanari, A., & Viroli, C. (2010b). A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, *37*(3), 473-487.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: an R package for global optimization by differential evolution. *Journal of Statistical Software*, *40*(6), 1-26.

Papageorgiou, G., & Hinde, J. (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Statistics and Computing*, *22*, 79-92.

Patz, R. A., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal of control and optimization*, *30*(4), 838-855.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (p. 45-57). London: Chapman and Hall.

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items scale scores for patterns of summed scores. In *Test scoring* (p. 253-292). Mahwah, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, *17*.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: an application to student growth percentiles. *Journal of Educational Measurement*, *49*(4), 446-465.

Smith, S. P. (1995). Differentiation of the Cholesky algorithm. *Journal of Comutational and Graphical Statistics*, *4*(2), 134-147.

Spall, J. C. (1997, December). Accelerated second-order stochastic optimization using only function measurements. In *Proceedings of the IEEE conference on decision and control* (p. 1417-1424). San Diego, CA.

Spall, J. C. (1998). Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans. Aerospace and Electronic Systems*, *34*, 817-823.

Spall, J. C. (2003). *Introduction to stochastic search and optimization.* Hoboken, NJ: Wiley-Interscience.

Storm, R., & Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*, 341-359.

Thissen, D., Cai, L., & Bock, R. D. (2011). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (p. 43-75). New York, NY: Taylor & Francis.

Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2014, July). Extending the use of multidimensional IRT calibration as projection: many-to-one linking

and linear computation of projected scores. Paper presented at the 2014 International Meeting of the Psychometric Society, Madison, WI.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL^{TM}3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, *20*(9), 1497-1505.

Tian, W., Cai, L., Thissen, D., & Xin, T. (2012). Numerical differentiation methods for computing error covariance matrices in item response theory modeling : An evaluation and a new proposal. *Educational and Psychological Measurement*, *73*(3), 412-439.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217-221.

Vitoratou, S., Ntzoufras, I., & Moustaki, I. (2014). Marginal likelihood estimation from the Metropolis output: tips and tricks for efficient implementation in generalized linear latent variable models. *Journal of Statistical Computation and Simulation*, *84*(10), 2091-2105.

Vock, D. M., Davidian, M., & Tsiatis, A. A. (2012). Mixed model analysis of censored longitudinal data with flexible random-effects density. *Biostatistics*, *13*(1), 61-73.

Vock, D. M., Davidian, M., & Tsiatis, A. A. (2014). SNP_NLMM: a SAS macro to implement a flexible random-effects density for generalized linear and nonlinear mixed models. *Journal of Statistical Software*, *56*, 1-21.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the

EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, *85*, 699-704.

Woods, C. M. (2007). Ramsay curve IRT for Likert-type data. *Applied Psychological Measurement*, *31*(3), 195-212.

Woods, C. M. (2008). Ramsay curve item response theory for the three-parameter item response theory model. *Applied Psychological Measurement*, *36*(6), 447-465.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Yamamoto, K., & Muraki, E. (1991, April). Non-linear transformation of IRT scale to account for the effect of nonnormal ability distribution on the item parameter estimation. Paper Presented at the 1991 annual meeting of the American Educational Research Association, Chicago, IL.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*(3), 795-802.