A Metropolis-Hastings Robbins-Monro Algorithm for Maximum Likelihood Nonlinear Latent Structure Analysis with a Comprehensive Measurement Model

Li Cai

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2008

Approved by

Robert C. MacCallum, Ph.D.

David M. Thissen, Ph.D.

Daniel J. Bauer, Ph.D.

Stephen H. C. du Toit, Ph.D.

Margaret R. Burchinal, Ph.D.

ABSTRACT

Li Cai: A Metropolis-Hastings Robbins-Monro Algorithm for Maximum Likelihood Nonlinear Latent Structure Analysis with a Comprehensive Measurement Model
(Under the direction of Robert C. MacCallum and David M. Thissen)

A Metropolis-Hastings Robbins-Monro (MH-RM) algorithm is proposed for maximum likelihood estimation in a general nonlinear latent structure model. The MH-RM algorithm represents a synthesis of the Markov chain Monte Carlo method, widely adopted in Bayesian statistics, and the Robbins-Monro stochastic approximation algorithm, well known in the optimization literature. The general latent structure model not only encompasses linear structural equations among latent variables, but also includes provisions for nonlinear latent regressions. Based on item response theory, a comprehensive measurement model provides the link between the latent structure and the observed variables. The MH-RM algorithm is shown to converge to a local maximum of the likelihood surface with probability one. Its significant advantages in terms of flexibility and efficiency over existing algorithms are illustrated with applications to real and simulated data. Implementation issues are discussed in detail. In addition, this dissertation integrates research on the parametrization and estimation of complex nonlinear latent variable models and furthers the understanding of the relationship between latent trait models and incomplete data estimation.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors, Drs. Robert MacCallum and David Thissen, for their unwavering support and guidance over the years, and I thank members of my committee for the many stimulating discussions and suggestions that led to a much improved research project.

I thank my friends and colleagues at the Thurstone Psychometric Lab. I also thank Dr. Michael Edwards for kindly sharing his MultiNorm program and data sets. I am indebted to the many mathematicians and statisticians that worked on the theories of stochastic processes and optimization, and to computer scientists who helped create TEX, LATEX, and C++. I would not have come this far intellectually without your insights.

Generous financial support from the Educational Testing Service (in the form of a Harold Gulliksen Psychometric Research Fellowship) and the National Science Foundation (Grant # SES-0717941) is gratefully acknowledged.

Finally, I thank my family, especially my wife, to whom this work is dedicated, for being there for me.

To Wenjing

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In the present research, *latent structure model* refers to a class of parametric statistical models that specify linear or nonlinear relations among a set of continuous latent variables. The term is admittedly influenced by Lazarsfeld's (1950) chapter on latent structure analysis, although only continuous latent traits will be considered in the sequel. The observed variables become indicators of the latent variables via a *comprehensive measurement model* such that an arbitrary mixture of metric and non-metric variables is permitted at the manifest level, including e.g., dichotomous, ordered polytomous, and nominal responses.

For a variety of technical reasons, it is a preferable practice that all parameters in the latent structure model be jointly estimated using full-information maximum likelihood (FIML). However, the likelihood function of the latent structure model involves intractable high dimensional integrals that present serious numerical challenges. For standard Gaussian quadrature based methods, the amount of computation increases exponentially as a function of the number of latent variables. This is known in the literature as the "curse of dimensionality." Existing Monte Carlo based methods also become cumbersome when both the number of latent variables and the number of observed variables are large.

The primary objective of the present research is to outline a newly proposed estimation algorithm known as Metropolis-Hastings Robbins-Monro (MH-RM; Cai,

2006) and apply it to solve the parameter estimation problem in maximum likelihood latent structure modelling. The MH-RM algorithm combines the Metropolis-Hastings (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) algorithm and the Robbins-Monro (RM; Robbins & Monro, 1951) stochastic approximation (SA; see e.g., Kushner & Yin, 1997) algorithm. The MH-RM algorithm was initially proposed by Cai (2006) for exploratory item factor analysis. A general convergence proof has been worked out and MH-RM performed satisfactorily in preliminary comparisons against leading item factor analysis software packages. MH-RM can handle large-scale analysis with many items, many factors, and thousands of respondents. It is flexible enough to seamlessly incorporate the mixing of different item response models, missing data, and multiple groups. It is well-suited to general computer programming for confirmatory analysis with arbitrary user-defined constraints. It is efficient in the use of Monte Carlo and unlike the EM algorithm it also produces an estimate of the parameter information matrix as an automatic by-product.

MH-RM has the potential of becoming a general and self-adaptive algorithm for arbitrarily high dimensional latent trait analysis. While the use of MH-RM is novel in its own right, a significant by-product of the present research is the integration of research on the parametrization and estimation of complex nonlinear latent variable models. To that end, a review of relevant background information is in order.

1.1 Background

Latent structure models are of considerable historical, theoretical, and practical value. To the research psychometrician, latent variable models of such a general kind encompass a large selection of psychometric models and techniques developed during the past six decades, including, e.g., exploratory factor analysis (e.g., Thurstone, 1947; see also Cudeck & MacCallum, 2007 and the references therein), confirmatory factor analysis (Jöreskog, 1969), covariance structure analysis (Bock & Bargmann,

1966; Bollen, 1989; Jöreskog, 1970), and item response theory (Lord & Novick, 1968; Thissen & Wainer, 2001). To the applied statistician, latent structure models are described in a distinctly modern statistical language, using the general framework of hierarchical models (see e.g., Bartholomew & Knott, 1999). These models offer fertile new ground for interesting applications of modern statistical and computational theory (see e.g. Dunson, 2000 for a statistician's view on latent variable models). To the statistical software programmers, the advent of a general modelling framework permits the development of general software packages that combine features of existing software such as Lisrel (Jöreskog & Sörbom, 2001), Testfact (Bock et al., 2003), and Multilog (Thissen, 2003). To administrators of testing programs and ultimately test users, latent structure models provide essential tools for item analysis, test assembly, and score reporting. For instance, as noted by von Davier and Sinharray (2004), the reporting methods used in NAEP rely on a special multidimensional item response model with covariate effects.

Latent structure models have been invented (and reinvented) under many different names. However, three main currents of research can be identified:

1. the extension of factor analysis and structural equation modelling to categorical indicators,

2. the growth of multidimensional item response theory (IRT), especially full-information item factor analysis (FIFA), and

3. the infusion of statistical concepts such as hierarchical models, mixture models, and mixed-effects models into psychometrics.

Main topics from these three areas shall be reviewed in turn from section 1.1.1, to section 1.1.3, in rough chronological order.

1.1.1 Limited-information Categorical Factor Analysis

Limited-information methods have a long tradition in psychometrics. This line of work began with the now classical treatment of factor analysis of categorical data by Christoffersson (1975) and Muthén (1978). It was soon realized that Muthén's (1978) approach could handle far richer structural models than the common factor model. Indeed, Muthén (1984) showed later that Jöreskog's (1970) linear structural model could be generalized to the case of mixed continuous and ordinal outcomes.

Building upon the equivalence of the "underlying response" formulation and the IRT formulation of categorical factor analysis (Takane & de Leeuw, 1987), several multi-stage estimators based on univariate and bivariate (hence limited) information have been proposed (e.g., Lee, Poon, & Bentler, 1992, 1995a, 1995b; Muthén & Muthén, 2007), and they compared favorably in empirical research against optimal but more computationally demanding estimators such as full information maximum likelihood (see e.g., Bolt, 2005; Knol & Berger, 1991). These estimators share the common feature that estimates of the category thresholds and polychoric correlations are obtained in the first one or two stages, as well as the asymptotic covariance matrix of these estimates. In the final stage, the remaining structural parameters are estimated using generalized least squares (GLS).

As far as statistical theory is concerned, these GLS-based estimators are grounded on sound principles. Furthermore, they have important ties to Browne's (1984) asymptotically distribution free method for moment structures, also known as the method of estimating equations in the statistical literature (Godambe, 1960) and the generalized method of moments in the econometric literature (Hansen, 1982). The concept of limited-information has also motivated recent development of goodness-of-fit indices for categorical data models (e.g., Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005).

However, currently popular GLS methods rely on pair-wise estimation of the

polychoric/polyserial correlations, so the resulting polychoric correlation matrix may not be positive definite. It is interesting to note that the dependence on pair-wise estimation is also a consequence of the "curse of dimensionality." Obtaining a full polychoric/polyserial correlation matrix by maximum likelihood requires as many dimensions of numerical integration as the number of observed variables, which is typically large. In addition, if the sample size is not too large, the asymptotic covariance matrix of the polychoric correlations cannot be determined accurately, which may adversely affect the GLS estimation of structural parameters. Although remedies such as the diagonally weighted least squares estimator appear to work in practice (see e.g., Flora & Curran, 2004 for recent simulation results), current implementations of these estimators often require *ad hoc* corrections for zero cell-counts in the marginal contingency tables. Furthermore, computation can become intense if the number of observed variables is large. To a psychometrician, the main drawback of GLS-based methods is the complete lack of support of other (more interesting) types of measurement models, such as the nominal model (Bock, 1972), not to mention the awkwardness when missing responses are present. Finally, due to the multi-stage nature of the estimation procedure, a fully Bayesian analysis is cumbersome. A notable exception to the above criticism is the Monte Carlo EM method for estimating polychoric correlations due to Song and Lee (2003), but this method is based on FIML so that one might as well estimate the structural parameters directly, with even lower dimensional integrals to solve.

A promising alternative that uses limited information but does not rely on multi-stage estimation of thresholds and polychorics is Jöreskog and Moustaki's (2001) Underlying Bivariate Normal (UBN) estimator (also known as the bivariate composite likelihood estimator). However, joint estimation of all parameters using a Newton-type algorithm can lead to large optimization problems. For instance, if one were to conduct a factor analysis of 100 Likert items in three independent samples, there

would be well over 1000 parameters to be jointly estimated. As will be shown, joint optimization of all parameters is unnecessary, but the current UBN formulation does not shed light on how the problem may be justifiably reduced to lower dimensions.

It should be noted that McDonald's (1967) treatise on nonlinear factor analysis and the associated NOHARM software (Fraser & McDonald, 1988) for parameter estimation can also be classified as using limited-information. However, the NOHARM method is based on ordinary least squares, and it is not even efficient among the class of limited-information estimators. This method has also seen limited practical use.

General latent structure models have taken preliminary shape within the limited-information approach. The GLS-based estimation method has enjoyed a high degree of popularity among practitioners over the past decades partly because of the existence of successful software programs such as Lisrel or Mplus. Many interesting applications ensued, but it can be concluded from examining the trend of recent research that traditional GLS-based estimation methods have failed to keep up with the ever-increasing complexity of measurement and structural models.

1.1.2 Full-information Item Factor Analysis

Occurring around the same period as Muthén (1978) proposed the GLS estimator, Bock and colleagues popularized the maximum marginal likelihood (MML) estimator in the field of IRT (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Thissen, 1982). This estimator uses full-information and circumvents many problems associated with the limited-information approach. It eventually led to active research on FIFA (Bock, Gibbons, & Muraki, 1988; Gibbons et al., 2007; Meng & Schilling, 1996; Mislevy, 1986; Muraki & Carlson, 1995; Schilling & Bock, 2005). Wirth and Edwards (2007) provide a recent overview of FIFA.

Full-information item factor analysis, as the name suggests, is factor analysis of categorical item-level data, using a full-information estimator such as MML. In educational and psychological testing, FIFA is a state-of-the-art method for item analysis.

It provides a wealth of information about scale dimensionality and item appropriateness, and it has also been used widely as a tool for modelling local dependence, e.g., testlets (Wainer & Kiely, 1987).

While standard IRT models are *unidimensional* in the sense that only one factor is in the model, recent applications of IRT in domains such as personality and health outcomes (see e.g., Bjorner, Chang, Thissen, & Reeve, 2007; Reeve et al., 2007; Reise, Ainsworth, & Haviland, 2003; Thissen, Reeve, Bjorner, & Chang, 2007) have prompted the increasing use of FIFA. However, the MML method leads to analytically intractable high dimensional integrals in the likelihood function, which is already nonlinear. The difficult nonlinear optimization problem is further complicated by the fact that there are typically many items and many respondents in applications of FIFA. Thus the use of standard Newton-type algorithms for maximizing the FIFA log-likelihood, e.g., Bock and Lieberman (1970), does not generalize well to real psychological and educational testing situations.

A break-through was made when Bock and Aitkin (1981) proposed a quadrature based EM algorithm. The main idea of Bock and Aitkin (1981) is remarkably simple: in step one, "make up" artificial data by conditioning on provisional estimates and observed data; step two, estimate parameters and go back to step one and repeat until parameters do not change further. This paper not only made important contributions to the computational methods of the day and established the *de facto* standard of parameter estimation method in the IRT field for the next two and half decades, but it was also profoundly influential in helping shape psychometricians' view on latent variable models. Evidence of its lasting impact is provided by over 400 citations since its publication, in fields ranging from education and psychology to biostatistics and medicine.

Bock and Aitkin (1981) were also among the first to investigate the possibility of further structural modelling within the computational framework of EM. For

instance, their empirical characterization of the latent ability distribution is at the crossroads of latent trait models and latent class models. The model for structured item parameters is effectively a Multiple-Indicator Multiple-Cause (MIMIC; see e.g., Bollen, 1989) model, widely known in the structural equation modelling community.

Because of the IRT orientation of FIFA, the kinds of structural models in widespread use are usually not as rich as those seen in the categorical factor analysis and structural equation modelling domain. The difference can be attributed to the difficulty of numerically evaluating high dimensional integrals in the EM algorithm for FIFA. The so-called "curse of dimensionality" is partially alleviated with recent development in statistical computing, especially Markov chain Monte Carlo (MCMC; see e.g., Tierney, 1994).

A clear trend that began with Albert's (1992) Bayesian analysis of the two-parameter normal ogive IRT model has been the wider acceptance of MCMC-based estimation algorithms in FIFA (Béguin & Glas, 2001; Edwards, 2005; Meng & Schilling, 1996; Patz & Junker, 1999a, 1999b; Segall, 1998; Shi & Lee, 1998). These developments blurred the boundaries between traditional domains of psychometrics such as IRT and structural equation modelling. Equipped with powerful modern computational methods and converging theories on latent variables models, general latent structure models have finally come to fruition.

1.1.3 Random-Effects, Mixtures, and Latent Variables

During the 1990's, statisticians turned their interest to general latent variable models. Early contributors include Bartholomew and Knott (1999), Skrondal and Rabe-Hesketh (2004), and Moustaki and her coworkers (see e.g., Moustaki, 2000, 2003, 2007; Moustaki & Knott, 2000). Important theoretical and computational results were also obtained by S.-Y. Lee and his associates (see e.g., Lee, Song, & Lee, 2003; Lee & Zhu, 2000; Song & Lee, 2001; Zhu & Lee, 2002). At the same time, complex extensions to standard IRT started to surface within the psychometric literature. For

instance, in a series of papers, Fox and colleagues have developed a multilevel IRT model (Fox, 2003, 2005; Fox & Glas, 2001) that is a synthesis of IRT and random effects regression models.

The literature has reached a consensus that latent variables are synonymous with random coefficients, factors, random effects, missing data, counterfactuals, unobserved heterogeneities, hidden volatilities, disturbances, errors, etc. The fact that they are known by different names is mostly a disciplinary terminology issue rather than real difference in their nature. In addition, it is not useful to classify models based on the type of observed variables. For a vector of observed variables, say, $\mathbf{y} = (y_1, \ldots, y_n)'$ that possesses density $f(\mathbf{y})$, all latent variable models can be expressed as (cf. Bartholomew & Knott, 1999):

$$f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{x})h(\mathbf{x})d\mathbf{x}, \tag{1.1}$$

where $\mathbf{x} = (x_1, \ldots, x_p)'$ is a p-dimensional vector of latent variables, $f(\mathbf{y}|\mathbf{x})$ the conditional density of observed variables given latent variables, and $h(\mathbf{x})$ the density of the latent variables. The y's need not be all continuous, and neither must all the x's be so. The p-fold integral over $\mathbf{x}$ can also be a mixture of integration and summation, as dictated by the type of random variables. Note that the term density is used generically to refer to either the probability density function of an absolutely continuous random variable or the probability mass function of a discrete random variable. One need not distinguish between the two because conceptually they are both Radon-Nikodym derivatives of probability measures.

Equation (1.1) is of fundamental importance so comments are in order. First, the latent variable model is not uniquely determined. Any arbitrary transformation of $\mathbf{x}$ can preserve the same $f(\mathbf{y})$ by simply making a change of variables in the integral. This indeterminacy cannot be resolved based on statistical theory alone. Instead, strong parametric assumptions on either $f(\mathbf{y}|\mathbf{x})$ or $h(\mathbf{x})$ are necessary, and they must come from substantive theory. The position taken here is to let $f(\mathbf{y}|\mathbf{x})$ be a member

of the exponential family while restricting $\mathbf{x}$ to have an absolutely continuous multivariate distribution. This specification affords enough flexibility (as far as the role of latent traits in mental test theory is concerned) and translates into easily interpretable parameters. Second, $f(\mathbf{y})$ is of the form of a mixture distribution, where the conditional density $f(\mathbf{y}|\mathbf{x})$ is mixed over $h(\mathbf{x})$. This implies that unless $f(\mathbf{y}|\mathbf{x})$ and $h(\mathbf{x})$ form conjugate pairs, the resulting integral does not have closed-form solution and must be approximated numerically. Third, Equation (1.1) has a hierarchical interpretation, wherein $h(\mathbf{x})$ may be conceived of as a "prior" density that completes the specification of a Bayesian two-level model. In a similar vein, the x's can also be thought of as nuisance parameters or random effects that must be integrated out to arrive at a genuine likelihood. This is the statistical basis for MML estimation. Finally, as the third point suggests, all inferences about $\mathbf{x}$ should be based on the posterior distribution $f(\mathbf{x}|\mathbf{y})$. For instance, the best mean square predictor of $\mathbf{x}$ is the posterior mean, a fact well-utilized in normal theory linear mixed models (Harville, 1977) and IRT scoring (Thissen & Wainer, 2001).

The historical circle that began with Thurstone (1947), Lazarsfeld (1950), and Lord and Novick (1968) is finally complete. Now that Equation (1.1) provides a general setup for latent variable models and the role of maximum likelihood estimation, the immediate central issue becomes one of computation, especially methods for evaluating multidimensional integrals.

1.2 Numerical Integration in FIML

To facilitate discussions, it is henceforth assumed that there are N independent respondents, each measured on n observed variables or items, and the total number of latent variables is p. Equation (1.1) makes it clear that the biggest obstacle in general latent structure analysis stems from the need to evaluate high dimensional integrals. Depending on how the integrals are approximated, existing algorithms for maximum likelihood estimation in latent variable models can be grouped roughly

into five classes, with a gradation from deterministic to stochastic.

1.2.1 Laplace Approximation

This class of methods is characterized by the use of Laplace approximation (Tierney & Kadane, 1986). Psychometric applications of this method can be found in Kass and Steffey (1989) and Thomas (1993). The Laplace method is fast (see e.g., Raudenbush, Yang, & Yosef, 2000, in a slightly different application), but a notable feature of this method is that the error of approximation decreases only as the number of observed variables increases. When few items are administered to each examinee, such as in an adaptive test design, or when there are relatively few items loading on a factor, such as in the presence of testlets (Wainer & Kiely, 1987), the degree of imprecision in approximation can be substantial and may lead to biased parameter estimates. Raudenbush et al. (2000) argue for the use of higher-order Laplace approximation, but the complexity of software implementation grows dramatically as the order of approximation increases. In addition, the truncation point in the asymptotic series expansion (6th-degree in their paper) of the integrand function is essentially arbitrary. Furthermore, from the perspective of random effects models, it is well known that if the error of approximation depends on cluster size (which is the same as the number of observed variables in this context), the standard Laplace method cannot be applied to models with crossed random effects (e.g., Kuk, 1999).

1.2.2 Adaptive Quadrature

This class is a direct generalization of Bock and Aitkin's (1981) quadrature-based approach. By replacing the original fixed-point quadrature with adaptive Gaussian quadrature (e.g., Naylor & Smith, 1982; Schilling & Bock, 2005), approximations to the high dimensional integrals impose significantly less computational burden. Adapting the quadrature nodes also stabilizes likelihood computations, because when n is large the likelihood becomes so concentrated that standard Gaussian

quadrature rules do not accurately capture its mass. With care in implementation, pointwise convergence of the estimates to a local maximum of the likelihood function can be obtained when an efficient quadrature rule is used in conjunction with either a Newton-type algorithm or the EM algorithm. Because of over two decades of success of Gaussian quadrature in IRT, it is often considered a gold standard against which other methods are compared. However, quadrature-based algorithms, e.g., those implemented in Testfact (Bock et al., 2003) or Gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2004), are still quite limited in the number of factors that they can handle simply because the number of quadrature points must grow exponentially as the dimensionality of the latent traits increases. In addition, because the EM algorithm does not provide information on sampling variability upon convergence, Testfact does not provide standard errors.

1.2.3 Monte Carlo EM

This class of methods is intimately related to Wei and Tanner's (1990) MCEM algorithm, wherein Monte Carlo integration replaces numerical quadrature in the E-step (e.g., Meng & Schilling, 1996; Song & Lee, 2005). The latent variables are treated as missing data, and their plausible values are multiply imputed from the posterior predictive distribution $f(\mathbf{x}|\mathbf{y})$ of the missing data given the observed data and provisional estimates. As it will become clear in section 3, the connection between MCEM and likelihood-based approaches to missing data (e.g., Little & Rubin, 1987; Schafer, 1997) provides a strong motivation for the MH-RM algorithm.

To achieve pointwise convergence, simulation size (the number of random draws) must increase as the estimates move closer to the maximum so that Monte Carlo error in the E-step does not overwhelm real changes in the M-step. Adaptive MCEM algorithms have been devised (e.g., Booth & Hobert, 1999), but the number of random draws in the final iterations of these automated algorithms can be as high as tens of thousands (Jank, 2004), dramatically slowing its convergence.

As far as estimation is concerned, MCEM is inefficient in the use of simulated data because a new set of random draws are generated at each E-step, discarding all previous draws. Experience with the design of IRT estimation software that implements the EM algorithm such as Multilog (Thissen, 2003) suggests that the E-step is usually much more time consuming than the M-step due to expensive exponential function calls and nested loop operations over both N and n. It is therefore not surprising that the E-step simulations in MCEM should take much more time than any other step. Discarding random draws appears to waste much needed computational resources and is clearly undesirable.

1.2.4 Simulated Maximum Likelihood and Variants

While quadrature-based EM and MCEM work on the log of the marginal likelihood function, Geyer and Thompson's (1992) simulated maximum likelihood approach seeks a direct Monte Carlo approximation to the marginal likelihood using importance sampling. The simulated likelihood is then optimized using standard numerical techniques such as Newton-Raphson. The appeal of simulated maximum likelihood is that in theory simulation is done only once, at the beginning of the estimation algorithm. However, the resulting estimates become sensitive to the initial approximation to the likelihood, especially the choice of the importance sampling distribution. This leads to alternative, doubly-iterative procedures in which parameters of the importance sampling distribution are updated after each optimization step. McCulloch and Searle (2001) advise caution on the implementation of simulated maximum likelihood, and Jank (2004) show that simulated maximum likelihood tends to be less efficient than MCEM. The details are too intricate and beyond the scope of the present research but briefly, one must pay close attention to the balance of simulation size and the updating of the importance sampling distribution to ensure convergence (Cappé, Douc, Moulines, & Robert, 2002).

Closely relate to simulated maximum likelihood is Qian and Shapiro's (2006)

Sample Average Approximation (SAA) method. SAA is not doubly-iterative. Returning to the log of the likelihood function, SAA exploits the structures of some latent variable models and uses Monte Carlo or quasi-Monte Carlo sampling from $h(\mathbf{x})$ to directly approximate the log-likelihood. For instance, Qian and Shapiro (2006) considered exploratory FIFA of dichotomous items. SAA is relatively new and holds promise for a class of relatively simple latent variable models wherein the mixing density $h(\mathbf{x})$ does not contain parameters. However, for more general models the performance of SAA is unknown. It is also subject to the same criticism as the UBN approach (Jöreskog & Moustaki, 2001) in the sense that the size of the optimization problem in SAA can become unnecessarily large (if n is large) because SAA fails to exploit the conditional independence structure often encountered in models arising out of test theory. Furthermore, SAA (in its original form in Qian & Shapiro, 2006) obscures the important connection between missing data imputation and latent variable modelling as the integration is taken with respect to $h(\mathbf{x})$ instead of $f(\mathbf{x}|\mathbf{y})$.

1.2.5 Fully Bayesian MCMC

This class of algorithms is purely stochastic and may at best be regarded as approximations to maximum likelihood because the algorithms usually do not involve optimization at all. A defining characteristic is the use of fully Bayesian sampling-based estimation methods such as Markov chain Monte Carlo. Within the Bayesian estimation framework, maximum likelihood can be approximated by choosing a non-informative prior distribution. One then constructs an ergodic Markov chain whose unique invariant measure is the posterior distribution of the parameters, and then after a certain "burn-in" period, samples from the chain may be regarded as random draws from the posterior, from which any functional of the posterior distribution can be estimated. While the basic principle is easy to state, the implementations vary to a wide extent (e.g., Albert, 1992; Diebolt & Ip, 1996; Patz & Junker, 1999a; Segall, 1998; Shi & Lee, 1998), and the relative algorithmic efficiency of the existing implementa-

tions have not been entirely settled (see e.g., Edwards, 2005). In addition, great care and experience are needed to handle the numerical results because the chains only converge *weakly* and the use of convergence diagnostics can be cumbersome with models having both large n and p.

1.3 Stochastic Approximation Algorithms

From the preceding discussion, it seems clear that a flexible and efficient algorithm that converges pointwise to the the maximum likelihood estimate is much desired for high dimensional latent structural analysis. Indeed, in the research reported here, the MH-RM algorithm is suggested to address most of the afore-mentioned difficulties. The MH-RM algorithm was initially proposed by Cai (2006) for exploratory FIFA and it compared favorably against leading IRT parameter estimation software packages in preliminary investigations.

The MH-RM algorithm is well-suited to general computer programming for high dimensional analysis with large n, p, and N. It is efficient in the use of Monte Carlo because the simulation size is fixed and usually small throughout the iterations. In addition, it also produces an estimate of the parameter information matrix as a by-product that can be used subsequently for standard error estimation and goodness-of-fit testing (e.g., Cai et al., 2006).

In brief, MH-RM is a data augmented Robbins-Monro type stochastic approximation algorithm driven by the random imputations produced by a Metropolis-Hastings sampler. The MH-RM algorithm is motivated by Titterington's (1984) recursive algorithm for incomplete data estimation, and is a close relative of Gu and Kong's (1998) SA algorithm. It can also be conceived of as a natural extension of the Stochastic Approximation EM algorithm (SAEM; Celeux & Diebolt, 1991; Celeux, Chauveau, & Diebolt, 1995; Delyon, Lavielle, & Moulines, 1999). Probability one convergence of the sequence of estimates to a local maximum of the likelihood surface will be established along essentially the same line as Gu and Kong's (1998) theorem 1.

SA algorithms have been well studied in the fields of systems engineering, adaptive control, and signal processing (see e.g., Kushner & Yin, 1997; Spall, 1999) since the pioneering work of Robbins and Monro (1951). Until recently, statistical applications of SA algorithms have remained predominantly in the area of generalized and nonlinear mixed-effects modeling (Gu & Kong, 1998; Gu & Zhu, 2001; Gu, Sun, & Huang, 2004; Gueorguieva & Agresti, 2001; Kuhn & Lavielle, 2005; Makowski & Lavielle, 2006; Zhu & Lee, 2002). The present research represents one of the first applications of SA algorithms to solve parameter estimation problems in latent structure analysis. However, before further discussion of the MH-RM algorithm, it is worthwhile to flesh out the details of a nonlinear latent structure model that will become a context for the application of the MH-RM algorithm.

CHAPTER 2

A Latent Structure Model

Consistent with the review of historical background in Chapter 1, the latent structure model considered here can be regarded either as a generalization of categorical structural equation models or as an extension of multidimensional IRT models. As a structural equation model, it not only incorporates a full Lisrel-type latent regression model, but also permits nonlinear latent regressions such that polynomial and interaction effects between latent variables can be assessed. However, it differs from Arminger and Muthén's (1998) and Lee and Zhu's (2000) nonlinear models in that the measurement part is developed directly from multidimensional IRT. For instance, the current framework includes a dichotomous IRT model with respondent guessing effect, as well as a multidimensional nominal model (Thissen, Cai, & Bock, 2006). Both models are rarely discussed in the categorical factor analysis and structural equation modelling literature.

2.1 Latent Structural Models

Recall that there are N independent respondents, n observed variables or items, and p latent variables. Let the vector of latent variables for respondent i be denoted as $\mathbf{x}_i$. For convenience, the p-dimensional vector $\mathbf{x}_i$ is further partitioned into two sub-vectors: $\boldsymbol{\xi}_i$ $(p_1 \times 1)$ and $\boldsymbol{\eta}_i$ $(p_2 \times 1)$, such that $\mathbf{x}_i = (\boldsymbol{\xi}_i', \boldsymbol{\eta}_i')'$ and $p = p_1 + p_2$. As a notational convention adopted from this point on, boldface capital letters denote matrices and boldface lower case letters are vectors.

2.1.1 Linear Structural Model

A linear structure is assumed for $\boldsymbol{\xi}_i$:

$$\boldsymbol{\xi}_i = \boldsymbol{\tau} + \boldsymbol{\Delta}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\boldsymbol{\tau}$ is a $p_1 \times 1$ vector of latent variable means, $\boldsymbol{\Delta}$ is a $p_1 \times p_1$ matrix of regression coefficients, and $\boldsymbol{\varepsilon}_i$ is a $p_1 \times 1$ vector of multivariate normally distributed error terms with zero means and covariance matrix $\boldsymbol{\Phi}$. Equation (2.1) represents a standard system of linear equations among $\boldsymbol{\xi}_i$. It is further assumed that $(\mathbf{I}_{p_1} - \boldsymbol{\Delta})$ is nonsingular so that $\boldsymbol{\xi}_i$ can be expressed as

$$\boldsymbol{\xi}_i = (\mathbf{I}_{p_1} - \boldsymbol{\Delta})^{-1}(\boldsymbol{\tau} + \boldsymbol{\varepsilon}_i). \tag{2.2}$$

Upon defining $\mathbf{A} = (\mathbf{I}_{p_1} - \boldsymbol{\Delta})^{-1}$, Equation (2.2) implies that the distribution of $\boldsymbol{\xi}_i$ is p_1-variate normal with mean $\mathbf{A}\boldsymbol{\tau}$ and covariance matrix $\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'$:

$$\boldsymbol{\xi}_i \sim \mathcal{N}_{p_1}(\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') \tag{2.3}$$

2.1.2 Nonlinear Structural Model

The latent variables in $\boldsymbol{\eta}_i$ are endogenous and are assumed to follow a nonlinear regression equation:

$$\boldsymbol{\eta}_i = \boldsymbol{\tau}_* + \boldsymbol{\Delta}_* g(\boldsymbol{\xi}_i) + \boldsymbol{\zeta}_i, \tag{2.4}$$

where $g(\cdot)$ is a continuous vector-valued nonlinear function that maps the p_1-dimensional vector $\boldsymbol{\xi}_i$ into a q-dimensional vector $g(\boldsymbol{\xi}_i)$, and $\boldsymbol{\tau}_*$ is a $p_2 \times 1$ vector of intercepts, $\boldsymbol{\Delta}_*$ a $p_2 \times q$ matrix of regression coefficients. Finally $\boldsymbol{\zeta}_i$ is a $p_2 \times 1$ vector of normally distributed error terms that is uncorrelated with $\boldsymbol{\varepsilon}_i$ and has zero means and covariance matrix $\boldsymbol{\Psi}$:

$$\boldsymbol{\zeta}_i \sim \mathcal{N}_{p_2}(\mathbf{0}, \boldsymbol{\Psi}). \tag{2.5}$$

This nonlinear model is quite flexible. For example, elements in $g(\boldsymbol{\xi}_i)$ may be polynomials of elements in $\boldsymbol{\xi}_i$ so that it can model latent variable interactions (Kenny & Judd, 1984).

For computational reasons, it is convenient to define

$$\mathbf{\Pi} = \begin{bmatrix} \boldsymbol{\tau}_*' \\ \boldsymbol{\Delta}_*' \end{bmatrix}, \quad \boldsymbol{\xi}_{*i} = \begin{bmatrix} 1 \\ g(\boldsymbol{\xi}_i) \end{bmatrix} \tag{2.6}$$

and rewrite Equation (2.4) as:

$$\boldsymbol{\eta}_i = \mathbf{\Pi}' \boldsymbol{\xi}_{*i} + \boldsymbol{\zeta}_i. \tag{2.7}$$

Equation (2.7) implies that conditional on $\boldsymbol{\xi}_{*i}$, the distribution of $\boldsymbol{\eta}_i$ is p_2-variate normal with mean $\mathbf{\Pi}'\boldsymbol{\xi}_{*i}$ and covariance matrix $\mathbf{\Psi}$:

$$\boldsymbol{\eta}_i | \boldsymbol{\xi}_{*i} \sim \mathcal{N}_{p_2}(\mathbf{\Pi}'\boldsymbol{\xi}_{*i}, \mathbf{\Psi}). \tag{2.8}$$

2.2 Measurement Models

The measurement models are developed as multidimensional IRT models. The basic principle of conditional independence (Lord & Novick, 1968) is assumed. Denote the ith respondent's vector of responses to the set of n observed variables as $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in})'$. The conditional independence assumption states that conditional on the respondent's latent trait level $\mathbf{x}_i$, the y_{ij}'s are mutually independent. Therefore, it is sufficient in the sequel to consider measurement models for a single response y_{ij} to a generic item j. Before embarking on model development, it is also useful to define an indicator function for categorical response models

$$\chi_k(y) = \begin{cases} 1, & \text{if } y = k, \\ 0, & \text{otherwise,} \end{cases} \tag{2.9}$$

for nonnegative integer $k \in \{0, 1, 2, \ldots\}$.

2.2.1 Dichotomous Response with Guessing Effect

This model is a generalization of the so-called 3-parameter logistic model (3PL) in unidimensional IRT. Given $\mathbf{x}_i$, the conditional probability of observing $y_{ij} = 1$ is

$$P(y_{ij} = 1 | \mathbf{x}_i, \boldsymbol{\theta}_j) = c(\kappa_j) + \frac{1 - c(\kappa_j)}{1 + \exp[-(\gamma_j + \boldsymbol{\beta}_j'\mathbf{x}_i)]}, \tag{2.10}$$

where $\boldsymbol{\theta}_j = (\gamma_j, \boldsymbol{\beta}_j', \kappa_j)'$ is a $(p+2) \times 1$ vector of parameters with γ_j being the intercept, $\boldsymbol{\beta}_j$ the $p \times 1$ vector of slopes, and $c(\kappa_j)$ the so-called guessing parameter, where κ_j is the logit of guessing:

$$c(\kappa_j) = \frac{1}{1 + \exp(-\kappa_j)}. \tag{2.11}$$

The logit reparametrization transforms a bounded parameter space to an unbounded one, and is customarily done in standard IRT software packages, e.g., Multilog (Thissen, 2003). Given Equation (2.10), the conditional probability of observing $y_{ij} = 0$ is equal to

$$P(y_{ij} = 0|\mathbf{x}_i, \boldsymbol{\theta}_j) = 1 - P(y_{ij} = 1|\mathbf{x}_i; \boldsymbol{\theta}_j). \tag{2.12}$$

The conditional density for y_{ij} is that of a Bernoulli variable:

$$f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) = P(y_{ij} = 1|\mathbf{x}_i, \boldsymbol{\theta}_j)^{y_{ij}} P(y_{ij} = 0|\mathbf{x}_i, \boldsymbol{\theta}_j)^{1-y_{ij}}. \tag{2.13}$$

2.2.2 Graded Response

This model is the multidimensional counterpart of Samejima's (1969) graded response model. Let $y_{ij} \in \{0, 1, 2, \ldots, K_j - 1\}$ be the response from respondent i to item j in K_j ordered categories. The development starts from defining the following logistic conditional cumulative response probabilities for each category:

$$\begin{aligned}
P(y_{ij} \geq 0|\mathbf{x}_i, \boldsymbol{\theta}_j) &= 1, \\
P(y_{ij} \geq 1|\mathbf{x}_i, \boldsymbol{\theta}_j) &= \frac{1}{1 + \exp[-(\gamma_{1,j} + \boldsymbol{\beta}_j'\mathbf{x}_i)]}, \\
P(y_{ij} \geq 2|\mathbf{x}_i, \boldsymbol{\theta}_j) &= \frac{1}{1 + \exp[-(\gamma_{2,j} + \boldsymbol{\beta}_j'\mathbf{x}_i)]}, \\
&\vdots \\
P(y_{ij} \geq K_j - 1|\mathbf{x}_i, \boldsymbol{\theta}_j) &= \frac{1}{1 + \exp[-(\gamma_{K_j-1,j} + \boldsymbol{\beta}_j'\mathbf{x}_i)]},
\end{aligned} \tag{2.14}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\gamma}_j', \boldsymbol{\beta}_j')'$ is a $(p + K_j - 1) \times 1$ vector of parameters, and $\boldsymbol{\gamma}_j = (\gamma_{1,j}, \ldots, \gamma_{K_j-1,j})'$ contains $K_j - 1$ intercepts that are strictly ordered. Equation (2.14)

implies that the category response probability is the difference between two adjacent cumulative probabilities:

$$P(y_{ij} = k|\mathbf{x}_i, \boldsymbol{\theta}_j) = P(y_{ij} \geq k|\mathbf{x}_i, \boldsymbol{\theta}_j) - P(y_{ij} \geq k+1|\mathbf{x}_i, \boldsymbol{\theta}_j), \tag{2.15}$$

where $P(y_{ij} \geq K_j|\mathbf{x}_i, \boldsymbol{\theta}_j)$ is identically equal to zero to ensure Equation (2.15) is well-defined for $k = 0, 1, 2, \ldots, K_j - 1$. With the indicator function defined in Equation (2.9), the conditional density for y_{ij} is a multinomial with trial size 1 in K_j categories:

$$f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) = \prod_{k=0}^{K_j-1} P(y_{ij} = k|\mathbf{x}_i, \boldsymbol{\theta}_j)^{\chi_k(y_{ij})}. \tag{2.16}$$

2.2.3 Nominal Response

This model is a recent reparametrization of Bock's (1972) original nominal model (Thissen et al., 2006). Let $y_{ij} \in \{0, 1, 2, \ldots, K_j - 1\}$ be the response from respondent i to item j in K_j nominal (unordered) categories. Conditional on $\mathbf{x}_i$, the category response probability for category k is defined as

$$P(y_{ij} = k|\mathbf{x}_i, \boldsymbol{\theta}_j) = \frac{\exp[a_k(\boldsymbol{\alpha}_j)\boldsymbol{\beta}_j'\mathbf{x}_i + c_k(\boldsymbol{\gamma}_j)]}{\sum_{m=0}^{K_j-1} \exp[a_m(\boldsymbol{\alpha}_j)\boldsymbol{\beta}_j'\mathbf{x}_i + c_m(\boldsymbol{\gamma}_j)]}, \tag{2.17}$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j', \boldsymbol{\beta}_j', \boldsymbol{\gamma}_j')'$ is a vector of parameters with $\boldsymbol{\beta}_j$ being a p-vector of slopes, $\boldsymbol{\alpha}_j$ a $(K_j - 2) \times 1$ vector that defines the "ordering" of categories, and $\boldsymbol{\gamma}_j$ a $(K_j - 1) \times 1$ vector of intercepts. The scalar parameters $a_k(\boldsymbol{\alpha}_j)$ and $c_k(\boldsymbol{\gamma}_j)$ in Equation (2.17) are the elements of K_j-dimensional vectors $\mathbf{a}(\boldsymbol{\alpha}_j)$ and $\mathbf{c}(\boldsymbol{\gamma}_j)$, respectively, who are themselves linear functions of $\boldsymbol{\alpha}_j$ and $\boldsymbol{\gamma}_j$, as defined below:

$$\mathbf{a}(\boldsymbol{\alpha}_j) = \begin{bmatrix} a_0(\boldsymbol{\alpha}_j) \\ \vdots \\ a_{K_j-1}(\boldsymbol{\alpha}_j) \end{bmatrix} = \mathbf{F}(K_j) \begin{bmatrix} 1 \\ \boldsymbol{\alpha}_j \end{bmatrix}, \tag{2.18}$$

and

$$\mathbf{c}(\boldsymbol{\gamma}_j) = \begin{bmatrix} c_0(\boldsymbol{\alpha}_j) \\ \vdots \\ c_{K_j-1}(\boldsymbol{\alpha}_j) \end{bmatrix} = \mathbf{F}(K_j)\boldsymbol{\gamma}_j, \tag{2.19}$$

where $\mathbf{F}(K_j)$ is a $K_j \times (K_j - 1)$ linear-Fourier basis matrix:

$$\mathbf{F}(K_j) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & f_{2,2} & \cdots & f_{2,(K_j-1)} \\ 2 & f_{3,2} & \cdots & f_{3,(K_j-1)} \\ \vdots & \vdots & & \vdots \\ K_j - 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{2.20}$$

and a typical element $f_{k,m}$ for $k = 1, 2, \ldots, K_j$ and $m = 1, 2, \ldots, K_j - 1$ takes its value from a Fourier sine-series:

$$f_{k,m} = \sin\left\{\frac{\pi(k-1)(m-1)}{K_j - 1}\right\}.$$

Again making use of the indicator function defined in Equation (2.9), the conditional density for y_{ij} under the nominal response model is

$$f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) = \prod_{k=0}^{K_j-1} P(y_{ij} = k|\mathbf{x}_i, \boldsymbol{\theta}_j)^{\chi_k(y_{ij})}. \tag{2.21}$$

Note that Equation (2.16) and Equation (2.21) are identical in form.

The seemingly complicated reparametrization achieves several goals. First, the category response probability, defined in Equation (2.17) using the multinominal logit, is invariant under arbitrary affine transformation of the logits. Therefore, the following restrictions must be in place for identification (see Thissen et al., 2006):

$$a_0(\boldsymbol{\alpha}_j) = 0, \quad a_{K_j-1}(\boldsymbol{\alpha}_j) = K_j - 1, \quad c_0(\boldsymbol{\gamma}_j) = 0.$$

It is clear that the linear-Fourier basis matrix implements these restrictions. Second, the linear-Fourier matrix provide (partially) orthogonal bases that essentially serve to "smooth" category boundaries or define partial ordering of the categories. If one fixes some of elements of $\boldsymbol{\alpha}_j$ and/or $\boldsymbol{\gamma}_j$ to zero, one can effectively create models that are between the original nominal model and constrained models such as the generalized partial credit model (Muraki, 1992).

2.2.4 Continuous Response

This model corresponds to the conditionally normal model assumed in common factor analysis. Specifically, y_{ij} is no longer an integer, but rather a number on the real line that has a conditional normal density:

$$f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(y_{ij} - \alpha_j - \boldsymbol{\beta}_j'\mathbf{x}_i)^2}{2\sigma_j^2} \right\}, \tag{2.22}$$

where $\boldsymbol{\theta}_j = (\alpha_j, \boldsymbol{\beta}_j', \sigma_j)$ is a $(p+2)$-dimensional vector of parameters, with α_j being the measurement intercept, $\boldsymbol{\beta}_j$ the slopes and σ_j the unique variance.

2.3 Observed and Complete Data Likelihoods

2.3.1 Observed Data Likelihood

Invoking the conditional independence assumption, let the conditional density for the observed vector of responses $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})'$ be

$$f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \prod_{j=1}^{n} f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) \tag{2.23}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_n')'$ is a vector of measurement model parameters. By definition (see section 2.1 and Equation 2.7),

$$\mathbf{x}_i = \begin{bmatrix} \boldsymbol{\xi}_i \\ \boldsymbol{\eta}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}_i \\ \boldsymbol{\Pi}'\boldsymbol{\xi}_{*i} + \boldsymbol{\zeta}_i \end{bmatrix},$$

so one can rewrite Equation (2.23) as

$$f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = f(\mathbf{y}_i|\boldsymbol{\xi}_i, \boldsymbol{\Pi}'\boldsymbol{\xi}_{*i} + \boldsymbol{\zeta}_i, \boldsymbol{\theta}). \tag{2.24}$$

Let $H_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the p-variate normal distribution function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Equation (2.24) implies that marginal density of $\mathbf{y}_i$ is

$$\begin{aligned} f(\mathbf{y}_i|\boldsymbol{\omega}) &= f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\Delta}, \boldsymbol{\Phi}, \boldsymbol{\Pi}, \boldsymbol{\Psi}) \\ &= \int\int f(\mathbf{y}_i|\boldsymbol{\xi}, \boldsymbol{\Pi}'\boldsymbol{\xi}_{*i} + \boldsymbol{\zeta}, \boldsymbol{\theta}) H_{p_1}(d\boldsymbol{\xi}|\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') H_{p_2}(d\boldsymbol{\zeta}|\mathbf{0}, \boldsymbol{\Psi}), \end{aligned} \tag{2.25}$$

where the integrals above are Lebesgue-Stieltjes integrals over $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$ with respect to the distribution functions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, respectively (see Equations 2.3 and 2.5), and $\boldsymbol{\omega} \in \boldsymbol{\Omega} \subset \mathbb{R}^d$ is defined as a d-dimensional vector containing all free parameters in $\boldsymbol{\theta}$, $\boldsymbol{\tau}$, $\boldsymbol{\Delta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Psi}$. Recall that $\mathbf{A} = (\mathbf{I}_{p_1} - \boldsymbol{\Delta})^{-1}$ (see Equation 2.2). Let $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_N')'$ be an $N \times n$ matrix of observed responses. The observed data likelihood is simply

$$L(\boldsymbol{\omega}|\mathbf{Y}) = \prod_{i=1}^{N} f(\mathbf{y}_i|\boldsymbol{\omega}). \tag{2.26}$$

Note that $L(\boldsymbol{\omega}|\mathbf{Y})$ contains N integrals of p dimensions, which makes its direct optimization extremely challenging.

2.3.2 Complete Data Likelihood

It is clear from the treatment in section 2.3.1 that the $\boldsymbol{\xi}$'s and $\boldsymbol{\zeta}$'s are treated as missing data that are integrated out to arrive at the marginal likelihood. Had they been observed, the complete data likelihood would simplify considerably. Equivalently stated, if an imputation scheme produces values of the $\mathbf{x}_i$'s, the optimization of the complete data likelihood would become easy, because $\mathbf{x}_i$ is completely determined by $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$. Let $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_N')'$ be an $N \times p$ matrix of missing data, so that complete data may be written as $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$.

Let $h_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the density of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It follows from Equation (2.8) that the complete data likelihood is:

$$L(\boldsymbol{\omega}|\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) h_{p_1}(\boldsymbol{\xi}_i|\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') h_{p_2}(\boldsymbol{\eta}_i|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i}, \boldsymbol{\Psi}). \tag{2.27}$$

The complete data likelihood is of a factored form. To further simplify the analysis, a restriction is placed on $\boldsymbol{\omega}$ to partition it into three independent sub-vectors $\boldsymbol{\omega} = (\boldsymbol{\omega}_1', \boldsymbol{\omega}_2', \boldsymbol{\omega}_3')'$, where:

1. $\boldsymbol{\omega}_1 = \boldsymbol{\theta}$ contains measurement parameters,

2. $\boldsymbol{\omega}_2$ contains linear structural parameters in $\boldsymbol{\tau}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Phi}$, and

3. $\boldsymbol{\omega}_3$ contains nonlinear structural parameters in $\boldsymbol{\Pi}$ and $\boldsymbol{\Psi}$.

It is assumed that $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ are not linked via parameter space restrictions and/or hyperparameters.

Let $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_N')'$ be an $N \times p_1$ matrix and $\mathbf{H} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_N')'$ be an $N \times p_2$ matrix such that $\mathbf{X} = (\boldsymbol{\Xi}, \mathbf{H})$. Rearrangement of the individual terms in Equation (2.27) leads to the following alternative expression of the complete data likelihood as the product of three independent parts:

$$L(\boldsymbol{\omega}|\mathbf{Z}) = L(\boldsymbol{\omega}_1|\mathbf{Z})L(\boldsymbol{\omega}_2|\boldsymbol{\Xi})L(\boldsymbol{\omega}_3|\mathbf{X}), \tag{2.28}$$

where

$$L(\boldsymbol{\omega}_1|\mathbf{Z}) = \prod_{i=1}^{n}\prod_{j=1}^{N} f(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) \tag{2.29}$$

is the measurement model complete data likelihood, and

$$L(\boldsymbol{\omega}_2|\boldsymbol{\Xi}) = \prod_{j=1}^{N} h_{p_1}(\boldsymbol{\xi}_i|\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') \tag{2.30}$$

is the linear structural model complete data likelihood, and

$$L(\boldsymbol{\omega}_3|\mathbf{X}) = \prod_{j=1}^{N} h_{p_2}(\boldsymbol{\eta}_i|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i}, \boldsymbol{\Psi}) \tag{2.31}$$

is the nonlinear structural model complete data likelihood. Note that Equation (2.29) corresponds to n (possibly nonlinear) regression models; Equation (2.30) corresponds to a normal theory moment structure model; and Equation (2.31) corresponds to a normal theory multivariate regression model.

CHAPTER 3

A Metropolis-Hastings Robbins-Monro Algorithm

3.1 The EM Algorithm and Fisher's Identity

Using the notation of Chapter 2, where $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, and the complete data likelihood is $L(\boldsymbol{\omega}|\mathbf{Z})$ for a d-dimensional parameter vector $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, and suppose $\mathbf{X} \in \mathcal{E}$, where $\mathcal{E}$ is some sample space. The task is to compute the MLE $\hat{\boldsymbol{\omega}}$ based on the observed data likelihood $L(\boldsymbol{\omega}|\mathbf{Y})$.

Let $l(\boldsymbol{\omega}|\mathbf{Y}) = \log L(\boldsymbol{\omega}|\mathbf{Y})$ and $l(\boldsymbol{\omega}|\mathbf{Z}) = \log L(\boldsymbol{\omega}|\mathbf{Z})$. Instead of maximizing $l(\boldsymbol{\omega}|\mathbf{Y})$ directly, Dempster, Laird, and Rubin (1977) transformed the observed data estimation problem into a sequence of complete data estimation problems by iteratively maximizing the conditional expectation of $l(\boldsymbol{\omega}|\mathbf{Z})$ over $F(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$, where $F(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$ denotes the conditional distribution of missing data given observed data. Let the current estimate be $\boldsymbol{\omega}^*$. One iteration of the EM algorithm consists of: (a) the E(xpectation) step, in which the expected complete-data log-likelihood

$$\mathcal{Q}(\boldsymbol{\omega}|\boldsymbol{\omega}^*) = \int_{\mathcal{E}} l(\boldsymbol{\omega}|\mathbf{Z}) F(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega}^*), \tag{3.1}$$

is computed, and (b) the M(aximization)-step, in which $\mathcal{Q}(\boldsymbol{\omega}|\boldsymbol{\omega}^*)$ is maximized to yield an updated estimate. Let

$$\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z}) = \nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Z}) \tag{3.2}$$

be the gradient of the complete data log-likelihood, where $\nabla_{\boldsymbol{\omega}}$ denotes the gradient operator that returns a vector of all partial derivatives with respect to $\boldsymbol{\omega}$ for a scalar-

valued function. By Fisher's identity (Fisher, 1925), the conditional expectation of $\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})$ equals the gradient of $l(\boldsymbol{\omega}|\mathbf{Y})$:

$$\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y}) = \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\omega}|\mathbf{Z}) F(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega}). \tag{3.3}$$

Equation (3.3) is the key to the entire EM machinery, and the MH-RM algorithm is strongly motivated by this equality.

3.2 MH-RM as a Generalized RM Algorithm

Robbins and Monro's (1951) algorithm is a root-finding algorithm for noise-corrupted regression functions. In the simplest case, let $\rho(\cdot)$ be a real-valued function of a real variable θ. If $\rho(\cdot)$ were known and continuously differentiable, one can use Newton's procedure

$$\theta_{k+1} = \theta_k + [-\nabla_\theta \rho(\theta_k)]^{-1} \rho(\theta_k)$$

to find the root. Alternatively, if differentiability cannot be assumed, one can use the following successive approximation

$$\theta_{k+1} = \theta_k + \epsilon \rho(\theta_k)$$

in a neighborhood of the root if ϵ is sufficiently small. If $\rho(\cdot)$ is unknown, but noisy observations can be taken at levels of θ at one's discretion, one can use the following RM recursive filter

$$\theta_{k+1} = \theta_k + \epsilon_k r_k, \tag{3.4}$$

where r_k is a noisy estimate of $\rho(\theta_k)$ and $\{\epsilon_k; k \geq 1\}$ is a sequence of decaying *gain constants* such that:

$$\epsilon_k \in (0, 1], \ \sum_{k=1}^{\infty} \epsilon_k = \infty, \text{ and } \sum_{k=1}^{\infty} \epsilon_k^2 < \infty. \tag{3.5}$$

Taken together, the three conditions above ensures that the gain constants decrease *slowly* to zero. The intuitive appeal of this algorithm is that r_k does not have to be highly accurate. This can be understood from the following: if θ_k is still far away

from the root, taking a large number of observations to compute a good estimate of $\rho(\theta_k)$ is inefficient because r_k is useful only insofar as it provides the right direction for the next move. The decaying gain constants eventually eliminate the noise effect so that the mean path of the sequence of estimates converges to the root.

The MH-RM algorithm is an extension of the basic algorithm in Equation (3.4) to multi-parameter problems that involve stochastic augmentation of missing data. Let

$$\mathbf{J}(\boldsymbol{\omega}|\mathbf{Z}) = -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Z})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} \tag{3.6}$$

be the complete data information matrix, and let $\mathcal{K}(\cdot, A|\boldsymbol{\omega}, \mathbf{Y})$ be a Markov transition kernel such that for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ and any measurable set $A \in \mathcal{E}$, it generates a uniformly ergodic chain satisfying

$$\int_A F(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega}) = \int_{\mathcal{E}} F(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})\mathcal{K}(\mathbf{X}, A|\boldsymbol{\omega}, \mathbf{Y}). \tag{3.7}$$

In practice, it is often useful to exploit the relation $F(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega}) \propto L(\mathbf{Z}|\boldsymbol{\omega})$ and construct a Metropolis-Hastings sampler that has the desired target distribution.

Let initial values be $(\boldsymbol{\omega}^{(0)}, \boldsymbol{\Gamma}_0)$, where $\boldsymbol{\Gamma}_0$ is a $d \times d$ symmetric positive definite matrix. Let $\boldsymbol{\omega}^{(k)}$ be the parameter estimate at the end of iteration k. The $(k+1)$th iteration of the MH-RM algorithm consists of:

1. *Stochastic Imputation*:

   Draw m_k sets of missing data $\left\{\mathbf{X}_j^{(k)}; j = 1, \ldots, m_k\right\}$ from the transition kernel $\mathcal{K}(\cdot, A|\boldsymbol{\omega}^{(k)}, \mathbf{Y})$ to form m_k sets of complete data

   $$\left\{\mathbf{Z}_j^{(k)} = (\mathbf{Y}, \mathbf{X}_j^{(k)}); j = 1, \ldots, m_k\right\}. \tag{3.8}$$

2. *Stochastic Approximation*: Using the relation in Equation (3.3), compute a Monte Carlo approximation to $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y})$ as

   $$\tilde{\mathbf{s}}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} \mathbf{s}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_j^{(k)}), \tag{3.9}$$

and a recursive stochastic approximation of the conditional expectation of the information matrix of the complete data log-likelihood as:

$$\mathbf{\Gamma}_k = \mathbf{\Gamma}_{k-1} + \epsilon_k \left\{ \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{J}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_j^{(k)}) - \mathbf{\Gamma}_{k-1} \right\}. \tag{3.10}$$

3. *Robbins-Monro Update*: Set the new parameter estimate to:

$$\boldsymbol{\omega}^{(k+1)} = \boldsymbol{\omega}^{(k)} + \epsilon_k \mathbf{\Gamma}_k^{-1} \tilde{\mathbf{s}}_k. \tag{3.11}$$

The iterations are terminated when the estimates converge (see section 4.3.3 for details on convergence analysis). In practice, ϵ_k may be taken as $1/k$, in which case the choice of $\mathbf{\Gamma}_0$ becomes arbitrary. Though the simulation size m_k is allowed to depend on the iteration number k, it is by no means required. In fact, the algorithm converges with a fixed and relatively small simulation size.

The MH-RM for maximum likelihood estimation is not too different from the engineering application of the RM algorithm for the identification and control of a dynamical system with observational noise. Finding the MLE amounts to finding the root of $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y})$, but because of missing data, $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y})$ is difficult to evaluate directly. In contrast, the gradient of the complete data log-likelihood $\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})$ is much simpler. Making use of Fisher's identity in Equation (3.3), the conditional expectation of $\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})$ is equal to $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y})$, so if one can augment missing data by sampling from a Markov chain having $F(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$ as its target, $\nabla_{\boldsymbol{\omega}} l(\boldsymbol{\omega}|\mathbf{Y})$ can be approximated via Monte Carlo integration. This is the logic behind Equation (3.9), and the key to understanding the asymptotic (in time) behavior of MH-RM.

As to the matrix $\mathbf{\Gamma}_k$, it is an approximation to the conditional expectation of $\mathbf{J}(\boldsymbol{\omega}|\mathbf{Z})$ over $F(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$. In multi-parameter optimization problems, use of curvature information often speeds up convergence. The complete data information matrix is easy to compute, and the recursive filter in Equation (3.10) helps stabilize the Monte Carlo noise. The term $\mathbf{\Gamma}_k^{-1} \tilde{\mathbf{s}}_k$ serves precisely the same role as r_k in Equation (3.4).

In Equation (3.11), MH-RM proceeds by using the same recursive filter as Equation (3.4) to average out the effect of the simulation noise on parameter estimates, so that the sequence of estimates converges to the root of $\nabla_{\omega} l(\boldsymbol{\omega}|\mathbf{Y})$.

3.3 Relation of MH-RM to Existing Algorithms

Cai (2006) showed that when the complete data log-likelihood corresponds to that of the generalized linear model for exponential family outcomes, the MH-RM algorithm can be derived as an extension of the SAEM algorithm by the same linearization argument that leads to the iteratively reweighted least squares algorithm (IRLS; McCullagh & Nelder, 1989) for maximum likelihood estimation in generalized linear models. Cai's (2006) result also implies that if the complete data model is ordinary multiple linear regression for Gaussian outcomes, the SAEM algorithm and the MH-RM algorithm are numerically equivalent. In other cases when this finite-time numeric equivalence does not hold, Delyon et al. (1999) showed that the SAEM algorithm has the same asymptotic (in time) behavior as the stochastic gradient algorithm. Equation (3.11) makes it clear that the MH-RM algorithm is an elaborated stochastic gradient algorithm that takes second derivative information into account. This implies that the MH-RM algorithm and the SAEM algorithm share the same asymptotic dynamics.

MH-RM is closely related to Titterington's (1984) algorithm in that both algorithms use the conditional expectation of the information matrix of the complete data log-likelihood. It becomes Gu and Kong's (1998) stochastic approximation Newton-Raphson algorithm if $\boldsymbol{\Gamma}_k$ is replaced by an estimate of the information matrix of the observed data log-likelihood. By the missing information principle (Orchard & Woodbury, 1972), the step size of the proposed MH-RM algorithm cannot be larger than that of Gu and Kong's (1998) algorithm. However, the MH-RM algorithm is much easier to implement and more stable than Gu and Kong's (1998) algorithm whenever the complete data likelihood is of a factored form. This will subsequently

be important because the latent structure model considered here has a factored complete data likelihood (see Equation 2.27).

If one sets ϵ_k to be identically equal to unity throughout the iterations and m_k to some relatively large number, the MH-RM algorithm becomes a Monte Carlo Newton-Raphson algorithm (MCNR; McCulloch & Searle, 2001). Unlike MCEM, there is no explicit maximization step in the MH-RM algorithm, so the MH-RM is not transparently related with MCEM. However, if $\epsilon_k \equiv 1$, the Robbins-Monro Update step can be thought of as a single iteration of maximization, in the same spirit as Lange's (1995) algorithm with a single iteration of Newton-Raphson in the M-step, which is locally equivalent to the EM.

In addition to ϵ_k being unity, if the number of random draws is also equal to one, i.e., $m_k \equiv 1$ for all k, the MH-RM algorithm becomes a close relative of Diebolt and Ip's (1996) stochastic EM (SEM) algorithm. The sequence of estimates produced by the SEM algorithm forms a time-homogeneous stationary Markov chain. The mean of its invariant distribution is close to the MLE, and the variance reflects loss of information due to missing data. In psychometric models similar to FIFA, the SEM algorithm is found to converge quickly to a close vicinity of the MLE (see e.g., Fox, 2003). Thus, the version of MH-RM similar to the SEM algorithm leads to a simple and effective method for computing start values for the subsequent MH-RM iterations with decreasing gain constants.

3.4 The Convergence of MH-RM

Recall that $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. Reference to $\mathbf{Y}$ will be suppressed in this section because it is fixed once observed. To avoid intricate notation, it is sufficient to consider $m_k = 1$ for all k. First define the following expectations:

$$\bar{\mathbf{J}}(\boldsymbol{\omega}) = \int_{\mathcal{E}} \mathbf{J}(\boldsymbol{\omega}|\mathbf{Z}) F(d\mathbf{X}|\boldsymbol{\omega}), \text{ and } \bar{\mathbf{s}}(\boldsymbol{\omega}) = \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\omega}|\mathbf{Z}) F(d\mathbf{X}|\boldsymbol{\omega}).$$

Due to the similarity of Equation (3.11) and Gu and Kong's (1998) Equation (5), it can be verified that the following ordinary differential equation (ODE) governs the asymptotic (in time) behavior of MH-RM:

$$\begin{bmatrix} \dot{\boldsymbol{\omega}}(t) \\ \dot{\boldsymbol{\Gamma}}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}(t)^{-1}\bar{\mathbf{s}}(\boldsymbol{\omega}(t)) \\ \bar{\mathbf{J}}(\boldsymbol{\omega}(t)) - \boldsymbol{\Gamma}(t) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\omega}(0) \\ \boldsymbol{\Gamma}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{\Gamma} \end{bmatrix}, \tag{3.12}$$

where $\dot{\boldsymbol{\omega}}(t)$ and $\dot{\boldsymbol{\Gamma}}(t)$ use Newton's notation for time derivatives. Consider the solution $(\boldsymbol{\omega}(t), \boldsymbol{\Gamma}(t))$, for $t \geq 0$. A point $(\boldsymbol{\omega}^*, \boldsymbol{\Gamma}^*)$ is a stability point if the above ODE admits the only solution $\boldsymbol{\omega}(t) = \boldsymbol{\omega}^*$, $\bar{\mathbf{J}}(\boldsymbol{\omega}(t)) = \boldsymbol{\Gamma}^*$, $t \geq 0$ when $\boldsymbol{\omega}(0) = \boldsymbol{\omega}^*$, $\boldsymbol{\Gamma}(0) = \boldsymbol{\Gamma}^*$. A set $\mathcal{D}$ is called a domain of attraction of a stability point $(\boldsymbol{\omega}^*, \boldsymbol{\Gamma}^*)$ if the solution of the above ODE with $(\boldsymbol{\omega}(0), \boldsymbol{\Gamma}(0)) \in \mathcal{D}$ remains in $\mathcal{D}$ indefinitely and converges to the stability point $(\boldsymbol{\omega}^*, \boldsymbol{\Gamma}^*)$. Clearly, for MLE $\hat{\boldsymbol{\omega}}$, the point $(\hat{\boldsymbol{\omega}}, \bar{\mathbf{J}}(\hat{\boldsymbol{\omega}}))$ is a stability point.

The same regularity conditions as Gu and Kong's (1998) theorem 1 are assumed to hold. These conditions guarantee (a) the integrability, convergence, and continuity of the Markov transition kernel, (b) the continuity and the existence of sufficient moments for functions $\mathbf{J}(\boldsymbol{\omega}|\mathbf{Z})$ and $\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})$. If the process $\{(\boldsymbol{\omega}^{(k)}, \boldsymbol{\Gamma}_k), k \geq 1\}$ as defined by Equation (3.11) is a bounded sequence, then

$$\boldsymbol{\omega}^{(k)} \to \hat{\boldsymbol{\omega}}, \text{ almost surely as } k \to \infty, \tag{3.13}$$

provided that the following recurrence condition also holds: $\{(\boldsymbol{\omega}^{(k)}, \boldsymbol{\Gamma}_k), k \geq 1\}$ belongs to a compact subset of the domain of attraction $\mathcal{D}$ of the stability point $(\hat{\boldsymbol{\omega}}, \bar{\mathbf{J}}(\hat{\boldsymbol{\omega}}))$. This result is a direct consequence of Gu and Kong's (1998) theorem 1, which is in turn based on results in Benveniste, Métivier, and Priouret (1990).

3.5 Approximating the Information Matrix

Fisher's identity in Equation (3.3) suggests the following procedure to recursively approximate the score vector:

$$\hat{\mathbf{s}}_k = \hat{\mathbf{s}}_{k-1} + \epsilon_k \left\{ \tilde{\mathbf{s}}_k - \hat{\mathbf{s}}_{k-1} \right\},$$

where $\tilde{\mathbf{s}}_k$ is a Monte Carlo estimate of the observed data score function as defined in Equation (3.9).

Following Louis's (1982) derivations, the information matrix of the observed data log-likelihood is

$$\begin{aligned} -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} &= \int_{\mathcal{E}} \left[\mathbf{J}(\boldsymbol{\omega}|\mathbf{Z}) - \mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})[\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})]'\right] F(d\mathbf{X}|\mathbf{Y},\boldsymbol{\omega}) \\ &+ \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\omega}|\mathbf{Z}) F(d\mathbf{X}|\mathbf{Y},\boldsymbol{\omega}) \int_{\mathcal{E}} [\mathbf{s}(\boldsymbol{\omega}|\mathbf{Z})]' F(d\mathbf{X}|\mathbf{Y},\boldsymbol{\omega}). \end{aligned} \tag{3.14}$$

This is a direct consequence of Orchard and Woodbury's (1972) missing information principle. Let

$$\tilde{\mathbf{G}}_k = \frac{1}{m_k}\sum_{j=1}^{m_k}\left[\mathbf{J}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_j^{(k)}) - \mathbf{s}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_j^{(k)})[\mathbf{s}(\boldsymbol{\omega}^{(k)}|\mathbf{Z}_j^{(k)})]'\right].$$

be a Monte Carlo estimate of the first conditional expectation in Equation (3.14). This estimate is too noisy, so a better recursive SA estimate is

$$\hat{\mathbf{G}}_k = \hat{\mathbf{G}}_{k-1} + \epsilon_k\left\{\tilde{\mathbf{G}}_k - \hat{\mathbf{G}}_{k-1}\right\}.$$

Putting the pieces together, the observed data information matrix can be approximated as

$$\mathcal{I}_k = \hat{\mathbf{G}}_k + \hat{\mathbf{s}}_k\hat{\mathbf{s}}_k'. \tag{3.15}$$

Provided that the log-likelihood is smooth, the almost sure convergence of $\boldsymbol{\omega}^{(k)} \to \hat{\boldsymbol{\omega}}$ in Equation (3.13) implies that

$$\mathcal{I}_k \to -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}, \text{ almost surely as } k \to \infty.$$

Upon convergence of the MH-RM algorithm, the inverse of $\mathcal{I}_k$ is the large-sample covariance matrix of parameter estimates.

CHAPTER 4

Implementation of MH-RM

This chapter focuses on the details of implementing the MH-RM algorithm for the latent structural model. For the majority of the applications described in Chapter 5, the random-walk Metropolis sampler developed in section 4.1 is the most convenient choice under full generality. Section 5.6 will describe an alternative special-purpose formulation of the MH-RM algorithm for full-information maximum likelihood estimation of the tetrachoric correlation matrix that relies on a data augmented Gibbs sampler – a special case of the Metropolis-Hastings algorithm with 100 percent acceptance rate (see e.g., Chib & Greenberg, 1995 for a discussion of the connection between the Gibbs sampler and the Metropolis-Hastings algorithm). The alternative sampler, however, requires switching the measurement part of the latent structure model to multidimensional normal ogive models, which do not readily support nominal responses. That restriction, in addition to the numerical complexities of the normal cumulative distribution functions significantly undermines the purpose of the present research, i.e., to develop a fully general framework for modelling and parameter estimation. Thus, the special-purpose MH-RM algorithm will be relegated to section 5.6.

4.1 A Metropolis-Hastings Sampler

The MCMC imputation procedure can be derived in a similar way as in Patz and Junker (1999a) from a Metropolis-within-Gibbs calculation (Chib & Greenberg, 1995).

Let $f(\mathbf{x}_i|\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_N,\mathbf{Y},\boldsymbol{\omega})$ be the full conditional density for $\mathbf{x}_i$, and let $\mathbf{x}_i^l$ be the value of $\mathbf{x}_i$ in the lth iteration of a Gibbs sampler with the following steps:

$$\begin{aligned}
\text{Draw } \mathbf{x}_1^l &\sim f(\mathbf{x}_1|\mathbf{x}_2^{l-1},\ldots,\mathbf{x}_N^{l-1},\mathbf{Y},\boldsymbol{\omega}) \\
\text{Draw } \mathbf{x}_2^l &\sim f(\mathbf{x}_2|\mathbf{x}_1^{l},\mathbf{x}_3^{l-1},\ldots,\mathbf{x}_N^{l-1},\mathbf{Y},\boldsymbol{\omega}) \\
&\vdots \\
\text{Draw } \mathbf{x}_i^l &\sim f(\mathbf{x}_i|\mathbf{x}_1^{l},\ldots,\mathbf{x}_{i-1}^{l},\mathbf{x}_{i+1}^{l-1},\ldots,\mathbf{x}_N^{l-1},\mathbf{Y},\boldsymbol{\omega}) \\
&\vdots \\
\text{Draw } \mathbf{x}_N^l &\sim f(\mathbf{x}_N|\mathbf{x}_1^{l},\ldots,\mathbf{x}_{N-1}^{l},\mathbf{Y},\boldsymbol{\omega}) \qquad (4.1)
\end{aligned}$$

Let the transition kernel defined by this Gibbs sampler be $\mathcal{K}(\mathbf{X},A|\boldsymbol{\theta},\mathbf{Y})$. Standard results (e.g., Gelfand & Smith, 1990; Geman & Geman, 1984) ensure that it satisfies the invariance condition in Equation (3.7). Hence if $\mathbf{X}_l=\{\mathbf{x}_i^l;i=1,\ldots,N\}$, the sequence $\{\mathbf{X}_l;l\geq 0\}$ converges in distribution to $F(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})$. It is also easy to recognize that the full conditional densities on the right hand side of (4.1) do not depend on past updates. Thus the coordinates can be sampled independently of each other.

The full conditionals are still difficult to sample directly from, but they are specified up to a proportionality constant, i.e.,

$$\begin{aligned}
& f(\mathbf{x}_i|\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_N,\mathbf{Y},\boldsymbol{\omega}) \propto L(\boldsymbol{\omega}|\mathbf{Z}) \\
\propto\ & f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})h_{p_1}(\boldsymbol{\xi}_i|\mathbf{A}\boldsymbol{\tau},\mathbf{A}\boldsymbol{\Phi}\mathbf{A}')h_{p_2}(\boldsymbol{\eta}_i|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i},\boldsymbol{\Psi}).
\end{aligned}$$

This suggests coupling the Gibbs sampler with the MH algorithm. Let

$$\begin{aligned}
& \alpha(\mathbf{x}_i,\mathbf{x}_i^*|\boldsymbol{\omega},\mathbf{y}_i) \\
=\ & \min\left\{\frac{f(\mathbf{y}_i|\mathbf{x}_i^*,\boldsymbol{\theta})h_{p_1}(\boldsymbol{\xi}_i^*|\mathbf{A}\boldsymbol{\tau},\mathbf{A}\boldsymbol{\Phi}\mathbf{A}')h_{p_2}(\boldsymbol{\eta}_i^*|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i}^*,\boldsymbol{\Psi})q(\mathbf{x}_i^*,\mathbf{x}_i)}{f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})h_{p_1}(\boldsymbol{\xi}_i|\mathbf{A}\boldsymbol{\tau},\mathbf{A}\boldsymbol{\Phi}\mathbf{A}')h_{p_2}(\boldsymbol{\eta}_i|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i},\boldsymbol{\Psi})q(\mathbf{x}_i,\mathbf{x}_i^*)},1\right\} \qquad (4.2)
\end{aligned}$$

be the acceptance probability of moving from state $\mathbf{x}_i$ to $\mathbf{x}_i^*$, where $\boldsymbol{\xi}_{*i}^*=(1,g(\boldsymbol{\xi}_i^*)')'$ takes the same form as Equation (2.6). To draw each $\mathbf{x}_i$, the following MH transition kernel is used:

$$\mathcal{K}(\mathbf{x}_i,d\mathbf{x}_i^*|\boldsymbol{\theta},\mathbf{y}_i)=q(\mathbf{x}_i,\mathbf{x}_i^*)\alpha(\mathbf{x}_i,\mathbf{x}_i^*|\boldsymbol{\omega},\mathbf{y}_i)d\mathbf{x}_i^* \qquad (4.3)$$

for $\mathbf{x}_i^* \neq \mathbf{x}_i$ and $\mathcal{K}(\mathbf{x}_i, \{\mathbf{x}_i\}|\boldsymbol{\theta}, \mathbf{y}_i) = 1 - \int_{\mathbf{x}_i^* \neq \mathbf{x}_i} \mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\boldsymbol{\theta}, \mathbf{y}_i)$, where $q(\mathbf{x}_i, \mathbf{x}_i^*)$ is any aperiodic recurrent transition density. Piecing the Gibbs part and the MH part together, the transition kernel for generating the stochastic imputations is

$$\mathcal{K}(\mathbf{X}, d\mathbf{X}^*|\boldsymbol{\theta}, \mathbf{Y}) = \prod_{i=1}^{N} \mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\boldsymbol{\theta}, \mathbf{y}_i). \tag{4.4}$$

In the sequel, a simple random walk chain $\mathbf{x}_i^* = \mathbf{x}_i + \mathbf{e}_i$ is used to generate the proposal draws, where the increment density is that of a scaled standard multivariate normal distribution in p dimensions, i.e., $\mathbf{e}_i \sim \mathcal{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$. The scalar parameter c adjusts the dispersion of the increments, so one can change its value to tune the acceptance ratio of the MH chain. Simple calculation shows that $q(\mathbf{x}_i, \mathbf{x}_i^*) = |2\pi c^2 \mathbf{I}_p| \exp\{-(\mathbf{x}_i^* - \mathbf{x}_i)'(\mathbf{x}_i^* - \mathbf{x}_i)/(2c^2)\}$ for this increment density. Because of the symmetry of the increment density $q(\mathbf{x}_i, \mathbf{x}_i^*) = q(\mathbf{x}_i^*, \mathbf{x}_i)$, Equation (4.2) can be further reduced to

$$\alpha(\mathbf{x}_i, \mathbf{x}_i^*|\boldsymbol{\omega}, \mathbf{y}_i) = \min\left\{ \frac{f(\mathbf{y}_i|\mathbf{x}_i^*, \boldsymbol{\theta}) h_{p_1}(\boldsymbol{\xi}_i^*|\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') h_{p_2}(\boldsymbol{\eta}_i^*|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i}^*, \boldsymbol{\Psi})}{f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) h_{p_1}(\boldsymbol{\xi}_i|\mathbf{A}\boldsymbol{\tau}, \mathbf{A}\boldsymbol{\Phi}\mathbf{A}') h_{p_2}(\boldsymbol{\eta}_i|\boldsymbol{\Pi}'\boldsymbol{\xi}_{*i}, \boldsymbol{\Psi})}, 1 \right\}. \tag{4.5}$$

The kernel in Equation (4.4) represents a remarkably simple sampling plan because all the conditioning kernels $\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\boldsymbol{\theta}, \mathbf{y}_i)$ on the right hand side can be evaluated independently of each other. This means that the N Gibbs updates in Equation (4.1) can be finished in parallel, if a matrix-oriented programming language such as GAUSS (Aptech Systems, Inc., 2003) is used. In brief, one first generates an $N \times p$ matrix $\mathbf{E}$, whose ith row is $\mathbf{e}_i'$, from a matrix normal distribution (Mardia, Kent, & Bibby, 1979) with independent rows each distributed as $\mathcal{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$, and compute the proposals as $\mathbf{X}^* = \mathbf{X} + \mathbf{E}$. Then for all rows, one evaluates the acceptance probabilities in Equation (4.5) as a "dot" division of the numerator and the denominator.

Since $\mathbf{Y}$ is fixed, let $\mathcal{K}_k(\cdot, A) = \mathcal{K}(\cdot, A|\boldsymbol{\omega}^{(k)}, \mathbf{Y})$ be the transition kernel in the $(k+1)$th iteration of MH-RM. From initial state $\mathbf{X}_0^{(k)}$, a sequence $\{\mathbf{X}_l^{(k)}; l \geq 0\}$ is generated by iterating $\mathcal{K}_k(\mathbf{X}, A)$, i.e.,

$$\Pr(\mathbf{X}_l^{(k)} \in A|\mathbf{X}_0^{(k)}) = \mathcal{K}_k^l(\mathbf{X}_0^{(k)}, A),$$

where $\mathcal{K}_k^l(\mathbf{X}_0^{(k)}, A)$ denotes the lth iterate of the kernel. The sequence of random imputations $\{\mathbf{X}_j^{(k)}; j = 1, \ldots, m_k\}$ can be chosen from $\{\mathbf{X}_l^{(k)}; l \geq 0\}$ as a subsequence, using standard "burn-in" and/or "thinning" methods. The initial state can be chosen as the last element of $\{\mathbf{X}_j^{(k-1)}; j = 1, \ldots, m_{k-1}\}$, i.e., $\mathbf{X}_0^{(k)} = \mathbf{X}_{m_{k-1}}^{(k-1)}$. Experience with this MCMC sampling procedure suggests that the only parameter that has to be tweaked on a case-by-case basis is the scalar dispersion parameter c in proposal generation. For high-dimensional problems, c generally needs to be smaller than 1, and the right choice can be made by monitoring the rejection rates of the MH chain for a brief period of time. It is also worthwhile to point out that standard subsampling methods have little impact on the asymptotic behavior of the MH-RM algorithm because the convergence result (3.13) does not require uncorrelated imputations. For any value of k, discarding a large number of initial iterates of the chain before generating the first imputation $\mathbf{X}_1^{(k)}$ is not useful for speeding up convergence. In a similar way, "thinning" off many iterates between $\mathbf{X}_j^{(k)}$ and $\mathbf{X}_{j+1}^{(k)}$ to reduce autocorrelation is not helpful either. If the starting values are sufficiently close to the MLE, one may even take $m_k \equiv 1$ for all k and set the number of "burn-in" iterates as small as 5.

4.2 Complete Data Models and Derivatives

Having described the MH sampler that produces the random imputations, the complete data log-likelihood and derivatives are needed to complete the specification of the MH-RM algorithm for latent structure analysis. The first and second order derivatives of the complete data models with respect to the *unrestricted* parameters will be given. These results conveniently allow the implementation of the MH-RM algorithm under general *linear* restrictions on the parameters. The linear constraint capabilities can be implemented with a parameter segmenting technique described by Thissen (1982). Due to conditional independence, it is sufficient to consider one respondent and one item at a time for the measurement models.

4.2.1 Linear Latent Structure

Using notation developed in Sections 2.1.1 and 2.3.2, let $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\tau}$ and $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}'$ be the mean vector and covariance matrix of $\boldsymbol{\xi}_i$. Equation (2.30) implies that the complete data log-likelihood can be written as

$$\log L(\boldsymbol{\omega}_2|\boldsymbol{\Xi}) \propto -\frac{N}{2}\left[\log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\xi}) + (\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})\right] \tag{4.6}$$

(see Mardia et al., 1979, p. 97), where $\bar{\boldsymbol{\xi}} = N^{-1}\boldsymbol{\Xi}'\mathbf{1}_N$ is the mean vector of data matrix $\boldsymbol{\Xi}$, and $\mathbf{S}_{\xi} = N^{-1}\boldsymbol{\Xi}'\boldsymbol{\Xi} - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}'$ is the covariance matrix. Let ω_i and ω_j be the ith and jth element of $\boldsymbol{\omega}_2$. Newton's notation for derivatives will be used. For instance, let the first derivatives of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with respect to ω_j be denoted $\dot{\boldsymbol{\mu}}_j$ and $\dot{\boldsymbol{\Sigma}}_j$, respectively. Similarly, let the second derivatives of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with respect to ω_i and ω_j be denoted $\ddot{\boldsymbol{\mu}}_{ij}$ and $\ddot{\boldsymbol{\Sigma}}_{ij}$, respectively. The first derivative of $L(\boldsymbol{\omega}_2|\boldsymbol{\Xi})$ with respect to ω_j is

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\omega}_2|\boldsymbol{\Xi})}{\partial \omega_j} &= -\frac{N}{2}\Big\{\text{tr}[\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{S}_{\xi})] - 2\dot{\boldsymbol{\mu}}_j'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}) \\ &\quad -(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})\Big\},\end{aligned} \tag{4.7}$$

and the second derivative of $L(\boldsymbol{\omega}_2|\boldsymbol{\Xi})$ with respect to ω_i and ω_j is

$$\begin{aligned}&\frac{\partial^2 \log L(\boldsymbol{\omega}_2|\boldsymbol{\Xi})}{\partial \omega_i \partial \omega_j} = \\ &-\frac{N}{2}\Big\{\text{tr}[(\boldsymbol{\Sigma}^{-1}\ddot{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\Sigma}^{-1} - 2\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma} - \mathbf{S}_{\xi}) + \boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i] \\ &-2[\ddot{\boldsymbol{\mu}}_{ij}'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}) - \dot{\boldsymbol{\mu}}_j'\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}) - \dot{\boldsymbol{\mu}}_j'\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\mu}}_i] \\ &-(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})'(\boldsymbol{\Sigma}^{-1}\ddot{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\Sigma}^{-1} - 2\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}^{-1})(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu})\Big\},\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}\dot{\boldsymbol{\mu}}_j &= \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\boldsymbol{\tau} + \mathbf{A}\dot{\boldsymbol{\tau}}_j, \\ \ddot{\boldsymbol{\mu}}_{ij} &= \mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\boldsymbol{\tau} + \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\boldsymbol{\tau} + \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\dot{\boldsymbol{\tau}}_i + \mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\dot{\boldsymbol{\tau}}_j,\end{aligned}$$

$$
\begin{aligned}
\dot{\boldsymbol{\Sigma}}_j &= \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Phi}}_j\mathbf{A}' + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\dot{\boldsymbol{\Delta}}_j'\mathbf{A}', \\
\ddot{\boldsymbol{\Sigma}}_{ij} &= \mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\dot{\boldsymbol{\Phi}}_i\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Delta}}_j\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\dot{\boldsymbol{\Delta}}_i'\mathbf{A}' + \\
&\quad \mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\dot{\boldsymbol{\Phi}}_j\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Phi}}_j\mathbf{A}'\dot{\boldsymbol{\Delta}}_i\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Delta}}_i\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\dot{\boldsymbol{\Delta}}_j'\mathbf{A}' + \mathbf{A}\dot{\boldsymbol{\Phi}}_i\mathbf{A}'\dot{\boldsymbol{\Delta}}_j'\mathbf{A}' + \\
&\quad \mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\dot{\boldsymbol{\Delta}}_i'\mathbf{A}'\dot{\boldsymbol{\Delta}}_j'\mathbf{A}' + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\dot{\boldsymbol{\Delta}}_j'\mathbf{A}'\dot{\boldsymbol{\Delta}}_i'\mathbf{A}'.
\end{aligned}
$$

4.2.2 Nonlinear Latent Structure

The notation here follows from Sections 2.1.2 and 2.3.2. Recall that $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_N')'$ is an $N \times p_1$ matrix and $\mathbf{H} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_N')'$ is an $N \times p_2$ matrix such that $\mathbf{X} = (\boldsymbol{\Xi}, \mathbf{H})$. Let $\boldsymbol{\Xi}_* = (\boldsymbol{\xi}_{*1}', \ldots, \boldsymbol{\xi}_{*N}')'$ be an $N \times (q+1)$ matrix consisting of $\boldsymbol{\xi}_{*i}$ (see Equation 2.6) as its ith row. Equation (2.31) implies that the complete data model corresponds to a multivariate regression model that regresses $\mathbf{H}$ on $\boldsymbol{\Xi}_*$, with $\boldsymbol{\Pi}$ being the regression coefficient matrix and $\boldsymbol{\Psi}$ the error covariance matrix. The complete data log-likelihood is:

$$
\log L(\boldsymbol{\omega}_3|\mathbf{X}) \propto -\frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathrm{tr}[(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'] \tag{4.9}
$$

(see Mardia et al., 1979, p. 158). Let ω_i and ω_j be the ith and jth element of $\boldsymbol{\omega}_3$, respectively. The first derivative of the complete data log-likelihood with respect to ω_j is

$$
\begin{aligned}
\frac{\partial \log L(\boldsymbol{\omega}_3|\mathbf{X})}{\partial \omega_j} &= -\frac{N}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) + \frac{1}{2}\mathrm{tr}\Big\{2\dot{\boldsymbol{\Pi}}_j\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'\boldsymbol{\Xi}_* \\
&\quad + \boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})\Big\}.
\end{aligned} \tag{4.10}
$$

The second derivative of $\log L(\boldsymbol{\omega}_3|\mathbf{X})$ with respect to ω_i and ω_j is

$$
\begin{aligned}
&\frac{\partial^2 \log L(\boldsymbol{\omega}_3|\mathbf{X})}{\partial \omega_i \partial \omega_j} = \frac{N}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j) \\
&-\mathrm{tr}\Big\{\dot{\boldsymbol{\Pi}}_j\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'\boldsymbol{\Xi}_* + \dot{\boldsymbol{\Pi}}_j\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Pi}}_i'\boldsymbol{\Xi}_*'\boldsymbol{\Xi}_* \\
&+\dot{\boldsymbol{\Pi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'\boldsymbol{\Xi}_* \\
&+\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_i\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_j\boldsymbol{\Psi}^{-1}(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})'(\mathbf{H} - \boldsymbol{\Xi}_*\boldsymbol{\Pi})\Big\}.
\end{aligned} \tag{4.11}
$$

4.2.3 Dichotomous Response with Guessing Effect

Suppressing subscripts i and j used in section 2.2.1 define

$$T = \frac{1}{1+\exp[-(\gamma+\boldsymbol{\beta}'\mathbf{x})]},\ P = c(\kappa) + [1-c(\kappa)]T,$$

where $c(\kappa)$ is defined in Equation (2.11). The individual contribution to the complete data log-likelihood can be written as

$$l = y\log P + (1-y)\log(1-P). \tag{4.12}$$

Using the chain-rule of differentiation, the first derivatives of (4.12) are

$$\frac{\partial l}{\partial \gamma} = \frac{\partial l}{\partial P}\frac{\partial P}{\partial \gamma},\ \frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial P}\frac{\partial P}{\partial \boldsymbol{\beta}},\ \frac{\partial l}{\partial \kappa} = \frac{\partial l}{\partial P}\frac{\partial P}{\partial \kappa},$$

where

$$\begin{aligned}
\frac{\partial l}{\partial P} &= \left(\frac{y}{P} - \frac{1-y}{1-P}\right),\\
\frac{\partial P}{\partial \gamma} &= [1-c(\kappa)]T(1-T),\\
\frac{\partial P}{\partial \boldsymbol{\beta}} &= [1-c(\kappa)]T(1-T)\mathbf{x},\\
\frac{\partial P}{\partial \kappa} &= (1-T)c(\kappa)[1-c(\kappa)]
\end{aligned}$$

Using the product rule, the second derivatives of (4.12) are

$$\begin{aligned}
\frac{\partial^2 l}{\partial\gamma\partial\gamma} &= \left(\frac{\partial}{\partial\gamma}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\gamma} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\gamma}\frac{\partial P}{\partial\gamma}\right),\\
\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} &= \left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\boldsymbol{\beta}'} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\boldsymbol{\beta}'},\right)\\
\frac{\partial^2 l}{\partial\kappa\partial\kappa} &= \left(\frac{\partial}{\partial\kappa}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\kappa} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\kappa}\frac{\partial P}{\partial\kappa}\right),\\
\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\gamma} &= \left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\gamma} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\gamma}\right),\\
\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\kappa} &= \left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\kappa} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\kappa}\right),\\
\frac{\partial^2 l}{\partial\gamma\partial\kappa} &= \left(\frac{\partial}{\partial\gamma}\frac{\partial l}{\partial P}\right)\frac{\partial P}{\partial\kappa} + \frac{\partial l}{\partial P}\left(\frac{\partial}{\partial\gamma}\frac{\partial P}{\partial\kappa}\right),
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial\gamma}\frac{\partial l}{\partial P} &= \left(-\frac{y}{P^2}-\frac{1-y}{(1-P)^2}\right)\frac{\partial P}{\partial\gamma}', \\
\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial l}{\partial P} &= \left(-\frac{y}{P^2}-\frac{1-y}{(1-P)^2}\right)\frac{\partial P}{\partial\boldsymbol{\beta}}', \\
\frac{\partial}{\partial\kappa}\frac{\partial l}{\partial P} &= \left(-\frac{y}{P^2}-\frac{1-y}{(1-P)^2}\right)\frac{\partial P}{\partial\kappa}', \\
\frac{\partial}{\partial\gamma}\frac{\partial P}{\partial\gamma} &= [1-c(\kappa)]T(1-T)(1-2T), \\
\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\boldsymbol{\beta}'} &= [1-c(\kappa)]T(1-T)(1-2T)\mathbf{x}\mathbf{x}', \\
\frac{\partial}{\partial\kappa}\frac{\partial P}{\partial\kappa} &= (1-T)c(\kappa)[1-c(\kappa)][1-2c(\kappa)], \\
\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\gamma} &= [1-c(\kappa)]T(1-T)(1-2T)\mathbf{x}, \\
\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial P}{\partial\kappa} &= -c(\kappa)[1-c(\kappa)]T(1-T)\mathbf{x}, \\
\frac{\partial}{\partial\gamma}\frac{\partial P}{\partial\kappa} &= -c(\kappa)[1-c(\kappa)]T(1-T).
\end{aligned}$$

As is often done in practice (Thissen, 2003), a normal prior with mean μ and variance σ^2 can be placed on κ. While the resulting solution is technically no longer the maximum likelihood estimate, the complete data model can be understood as having a penalized log-likelihood:

$$\bar{l} \propto l - \frac{(\kappa-\mu)^2}{2N\sigma^2}.$$

This implies that the first derivative of $\bar{l}$ with respect to κ becomes

$$\frac{\partial\bar{l}}{\partial\kappa} = \frac{\partial l}{\partial\kappa} - \frac{\kappa-\mu}{N\sigma^2}.$$

The prior also leads to the addition of a ridge term to the second derivative of $\bar{l}$ with respect to κ:

$$\frac{\partial^2\bar{l}}{\partial\kappa\partial\kappa} = \frac{\partial^2 l}{\partial\kappa\partial\kappa} - \frac{1}{N\sigma^2}.$$

The other derivatives remain unchanged.

4.2.4 Graded Response

Subscripts i and j used in section 2.2.2 will be suppressed. Let $y \in \{0, 1, \ldots, K-1\}$ be the response to a graded item in K categories. Let

$$\begin{aligned} T_0 &= 1, \\ T_1 &= \frac{1}{1+\exp[-(\gamma_1 + \boldsymbol{\beta}'\mathbf{x})]}, \\ T_2 &= \frac{1}{1+\exp[-(\gamma_2 + \boldsymbol{\beta}'\mathbf{x})]}, \\ &\vdots \\ T_{K-1} &= \frac{1}{1+\exp[-(\gamma_{K-1} + \boldsymbol{\beta}'\mathbf{x})]}, \\ T_K &= 0 \end{aligned}$$

be the cumulative response probabilities as defined in Equation (2.14) such that

$$P_k = T_k - T_{k+1}$$

is the category response probability for $k \in \{0, 1, \ldots, K-1\}$. Using the indicator function defined in Equation (2.9), the log-likelihood for the complete data model is

$$l = \sum_{k=0}^{K-1} \chi_k(y) \log P_k = \sum_{k=0}^{K-1} \chi_k(y) \log(T_k - T_{k+1}). \tag{4.13}$$

The first derivatives of (4.13) are

$$\begin{aligned} \frac{\partial l}{\partial \gamma_k} &= \frac{\partial}{\partial \gamma_k}\Big(\chi_{k-1}(y)\log(T_{k-1} - T_k) + \chi_k(y)\log(T_k - T_{k+1})\Big) \\ &= -\left(\frac{\chi_{k-1}(y)}{T_{k-1} - T_k} - \frac{\chi_k(y)}{T_k - T_{k+1}}\right)\frac{\partial T_k}{\partial \gamma_k} \\ \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{k=0}^{K-1} \frac{\chi_k(y)}{T_k - T_{k+1}}\left(\frac{\partial T_k}{\partial \boldsymbol{\beta}} - \frac{\partial T_{k+1}}{\partial \boldsymbol{\beta}}\right), \end{aligned}$$

where

$$\frac{\partial T_k}{\partial \gamma_k} = T_k(1 - T_k), \; \frac{\partial T_k}{\partial \boldsymbol{\beta}} = T_k(1 - T_k)\mathbf{x}.$$

The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial \gamma_k^2} &= -\left(\frac{\chi_{k-1}(y)}{(T_{k-1}-T_k)^2}+\frac{\chi_k(y)}{(T_k-T_{k+1})^2}\right)\left(\frac{\partial T_k}{\partial \gamma_k}\right)^2 \\
&\quad -\left(\frac{\chi_{k-1}(y)}{T_{k-1}-T_k}-\frac{\chi_k(y)}{T_k-T_{k+1}}\right)\left(\frac{\partial}{\partial \gamma_k}\frac{\partial T_k}{\partial \gamma_k}\right) \\
\frac{\partial^2 l}{\partial \gamma_{k-1}\partial \gamma_k} &= \frac{\chi_{k-1}(y)}{(T_{k-1}-T_k)^2}\left(\frac{\partial T_{k-1}}{\partial \gamma_{k-1}}\right)\left(\frac{\partial T_k}{\partial \gamma_k}\right) \\
\frac{\partial^2 l}{\partial \gamma_{k+1}\partial \gamma_k} &= \frac{\chi_k(y)}{(T_k-T_{k+1})^2}\left(\frac{\partial T_{k+1}}{\partial \gamma_{k+1}}\right)\left(\frac{\partial T_k}{\partial \gamma_k}\right) \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \gamma_k} &= -\frac{\chi_k(y)}{(T_k-T_{k+1})^2}\left(\frac{\partial T_k}{\partial \gamma_k}\right)\left(\frac{\partial T_k}{\partial \boldsymbol{\beta}}-\frac{\partial T_{k+1}}{\partial \boldsymbol{\beta}}\right) \\
&\quad +\frac{\chi_{k-1}(y)}{(T_{k-1}-T_k)^2}\left(\frac{\partial T_k}{\partial \gamma_k}\right)\left(\frac{\partial T_{k-1}}{\partial \boldsymbol{\beta}}-\frac{\partial T_k}{\partial \boldsymbol{\beta}}\right) \\
&\quad -\left(\frac{\chi_{k-1}(y)}{T_{k-1}-T_k}-\frac{\chi_k(y)}{T_k-T_{k+1}}\right)\left(\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial T_k}{\partial \gamma_k}\right) \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= \sum_{k=0}^{K-1}\left\{-\frac{\chi_k(y)}{(T_k-T_{k+1})^2}\left(\frac{\partial T_k}{\partial \boldsymbol{\beta}}-\frac{\partial T_{k+1}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial T_k}{\partial \boldsymbol{\beta}'}-\frac{\partial T_{k+1}}{\partial \boldsymbol{\beta}'}\right)\right. \\
&\quad \left.+\frac{\chi_k(y)}{T_k-T_{k+1}}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial T_k}{\partial \boldsymbol{\beta}'}-\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial T_{k+1}}{\partial \boldsymbol{\beta}'}\right)\right\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \gamma_k}\frac{\partial T_k}{\partial \gamma_k} &= T_k(1-T_k)(1-2T_k) \\
\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial T_k}{\partial \gamma_k} &= T_k(1-T_k)(1-2T_k)\mathbf{x} \\
\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial T_k}{\partial \boldsymbol{\beta}'} &= T_k(1-T_k)(1-2T_k)\mathbf{x}\mathbf{x}'.
\end{aligned}$$

4.2.5 Nominal Response

Subscripts i and j used in section 2.2.3 will be suppressed. Let $y \in \{0, 1, \ldots, K-1\}$ be the response to an item in K nominal categories. Let

$$P_k = \frac{\exp[a_k(\boldsymbol{\alpha})\boldsymbol{\beta}'\mathbf{x} + c_k(\boldsymbol{\gamma})]}{\sum_{m=0}^{K-1}\exp[a_m(\boldsymbol{\alpha})\boldsymbol{\beta}'\mathbf{x} + c_m(\boldsymbol{\gamma})]},$$

be the category response probability. It follows from Equations (2.18) and (2.19), that for $k \in \{0, 1, \ldots, K-1\}$, $a_k(\boldsymbol{\alpha}) = \mathbf{f}'_{k+1}\boldsymbol{\alpha}$, and $c_k(\boldsymbol{\gamma}) = \mathbf{f}'_{k+1}\boldsymbol{\gamma}$, where $\mathbf{f}'_k$ is the $(k+1)$th

row of $\mathbf{F}(K)$ (see Equation 2.20). The log-likelihood is

$$l = \sum_{k=0}^{K} \chi_k(y) \log P_k. \tag{4.14}$$

Because

$$\begin{aligned}
\frac{\partial P_k}{\partial \boldsymbol{\alpha}} &= P_k \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m) \right\} (\boldsymbol{\beta}'\mathbf{x}) \\
\frac{\partial P_k}{\partial \boldsymbol{\beta}} &= P_k \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m)'\boldsymbol{\alpha} \right\} \mathbf{x} \\
\frac{\partial P_k}{\partial \boldsymbol{\gamma}} &= P_k \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m) \right\},
\end{aligned}$$

it follows that the first derivatives of (4.14) are

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\alpha}} &= \sum_{k=0}^{K} \frac{\chi_k(y)}{P_k} \frac{\partial P_k}{\partial \boldsymbol{\alpha}} = \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m) \right\} (\boldsymbol{\beta}'\mathbf{x}), \\
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{k=0}^{K} \frac{\chi_k(y)}{P_k} \frac{\partial P_k}{\partial \boldsymbol{\beta}} = \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m)'\boldsymbol{\alpha} \right\} \mathbf{x}, \\
\frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{k=0}^{K} \frac{\chi_k(y)}{P_k} \frac{\partial P_k}{\partial \boldsymbol{\gamma}} = \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} P_m(\mathbf{f}_k - \mathbf{f}_m) \right\}.
\end{aligned}$$

The second derivatives of (4.14) are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\alpha}} (\mathbf{f}_k - \mathbf{f}_m)' \right\} (\boldsymbol{\beta}'\mathbf{x}), \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\beta}} (\mathbf{f}_k - \mathbf{f}_m)' (\boldsymbol{\beta}'\mathbf{x}) + P_m \mathbf{x} (\mathbf{f}_k - \mathbf{f}_m)' \right\}, \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\alpha}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\gamma}} (\mathbf{f}_k - \mathbf{f}_m)' \right\} (\boldsymbol{\beta}'\mathbf{x}), \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\beta}} (\mathbf{f}_k - \mathbf{f}_m)'\boldsymbol{\alpha} \right\} \mathbf{x}', \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\gamma}} (\mathbf{f}_k - \mathbf{f}_m)'\boldsymbol{\alpha} \right\} \mathbf{x}', \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} &= \sum_{k=0}^{K} \chi_k(y) \left\{ \sum_{m=0}^{K-1} \frac{\partial P_m}{\partial \boldsymbol{\gamma}} (\mathbf{f}_k - \mathbf{f}_m)' \right\}.
\end{aligned}$$

4.2.6 Continuous Response

Subscripts i and j used in section 2.2.4 will be suppressed. It is also convenient to define $\mathbf{x}_* = (1, \mathbf{x}')'$ and $\boldsymbol{\beta}_* = (\alpha, \boldsymbol{\beta}')'$ so that the individual contribution to the complete data log-likelihood can be written as

$$l = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{(y - \mathbf{x}_*'\boldsymbol{\beta}_*)^2}{\sigma^2}. \tag{4.15}$$

The first derivatives of (4.15) are

$$\frac{\partial l}{\partial \boldsymbol{\beta}_*} = \frac{(y - \mathbf{x}_*'\boldsymbol{\beta}_*)\mathbf{x}_*}{\sigma^2}, \frac{\partial l}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(y - \mathbf{x}_*'\boldsymbol{\beta}_*)^2}{2\sigma^4}.$$

The second derivatives are

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*'} = -\frac{\mathbf{x}_*\mathbf{x}_*'}{\sigma^2}, \frac{\partial^2 l}{\partial \boldsymbol{\beta}_* \partial \sigma^2} = -\frac{(y - \mathbf{x}_*'\boldsymbol{\beta}_*)\mathbf{x}_*}{\sigma^4}, \frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2\sigma^4} - \frac{(y - \mathbf{x}_*'\boldsymbol{\beta}_*)^2}{2\sigma^6}.$$

4.3 Acceleration and Convergence

The asymptotic convergence result of MH-RM says nothing of its finite-time behavior. Experience suggests that it is crucial to ensure that the algorithm does not get "stuck" in locations far from the MLE during the initial stage of iterations. This is clear because the sequence of gain constants is deterministic. With finite-precision floating point operations, they eventually go to zero and the sequence of estimates $\boldsymbol{\omega}^{(k)}$ is bound to converge to some limit, though it may not be the MLE. The problem of premature "convergence" can be effective solved with a combination of the following strategies.

4.3.1 Adaptive Gain Constants

Schilling and Bock (2005) reported success with the parameter expansion method (PX-EM; Liu, Rubin, & Wu, 1998) for speeding up convergence of a quadrature based EM algorithm in the context of exploratory item factor analysis. However, PX-EM does not easily generalize to confirmatory analysis with arbitrary constraints.

On the other hand, it is well-known from the work on the rate of convergence of the EM algorithm (e.g., Meng & Rubin, 1991) that components of $\boldsymbol{\omega}^{(k)}$ converge at different rates due to difference in the *fraction of missing information* (Orchard & Woodbury, 1972). Thus it seems "unfair" for those parameters having lower fractions of missing information to share the same sequence of gain constants with other parameters having higher fractions of missing information. This suggests using vector valued gain constants that take into account the differential rates of convergence.

Recall that $\boldsymbol{\omega}$ is a d-dimensional vector. Let $\boldsymbol{\epsilon}_k = (\epsilon_{1k}, \ldots, \epsilon_{dk})'$, and write $\{\boldsymbol{\epsilon}_k; k \geq 1\}$ as the sequence of vector-valued gain constants. Then for parameter ω_ℓ, the sequence of gain constants is defined by $\{\epsilon_{\ell k}; k \geq 1\}$. Kesten (1958) finds that frequent sign changes in the the successive differences between adjacent estimates is often an indication that the estimate is close to the MLE. Therefore, the gain constants should be made "larger" when the sign change is infrequent. One way to accomplish this is to decrease the gain constant only if two successive changes are of opposite sign. To formalize this ideal, let $\triangle_\ell^{(k)} = \omega_\ell^{(k)} - \omega_\ell^{(k-1)}$. Then choose $\epsilon_{\ell(k+1)}$ such that $\epsilon_{\ell(k+1)} < \epsilon_{\ell k}$ if and only $\triangle_\ell^{(k)}$ and $\triangle_\ell^{(k-1)}$ have different signs. This method allows slower moving parameters to have larger step sizes.

4.3.2 Multi-stage Gain Constants

The adaptive gain constants can be further augmented with a three-stage procedure. First, from some starting values, run M_1 MH-RM iterations, wherein both the gain constants and the m_k's are set to 1 for $k = 1, \ldots, M_1$. At the end of iteration M_1, run another M_2 iterations and at the end of iteration $M_1 + M_2$, the sequence of parameter estimates obtained from the last M_2 iterations are averaged and used as the start value for the subsequent MH-RM iterations with decreasing gain constants.

This multi-stage procedure is motivated by a simple example. Suppose one is

faced with the task of solving for the root of the following equation

$$\frac{1}{1+\exp(-0.5\theta)} - 0.5 = 0.$$

Of course, one does not need advanced root-finding algorithms to see that the root is equal to 0. However, if one insists, Newton's algorithm can be used. Now suppose one is faced with a stochastic version of the root-finding problem

$$\frac{1}{1+\exp(-0.5\theta)} - 0.5 + u = 0,$$

where $u \sim \mathcal{N}(0,1)$. Newton's method cannot be used, but the Robbins-Monro method is still applicable. Let the starting value of θ be 2. Figure 4.1 compares the rate of convergence of three implementations of the Robbins-Monro method: with strictly decreasing gains, with constant gains, and with a two-stage (constant and then decreasing) gain sequence. The iteration history of Newton's method for the deterministic problem is also plotted to serve as a baseline. It is apparent from the figure that though the Robbins-Monro method with decreasing gain constants eventually converged to the root, it took well over 100 cycles. The constant gain method pushed θ to a close neighborhood of the root, but started oscillating around 0 as a stochastic process. The two stage method used constant gains up to cycle 20, and then switched over to decreasing gains. It converged to 0 in less than half of the cycle count of the decreasing gain method. Applying this technique to the MH-RM algorithm leads to the three-stage gain procedure that performs well in practice.

4.3.3 Convergence Check

Convergence of the MH-RM can be monitored by computing a window of successive differences, say $\{\max_\ell \triangle_\ell^{(k)}, \max_\ell \triangle_\ell^{(k-1)}, \ldots, \max_\ell \triangle_\ell^{(k-W)}\}$, where $\max_\ell$ stands for taking the maximum over $\ell = 1, \ldots, d$ successive differences, and W is a predetermined window size. The iterations are terminated if and only if all differences in the window are less than some small number. This method prevents premature stop due to random variation. In practice, $W = 3$ seems a reasonable choice.

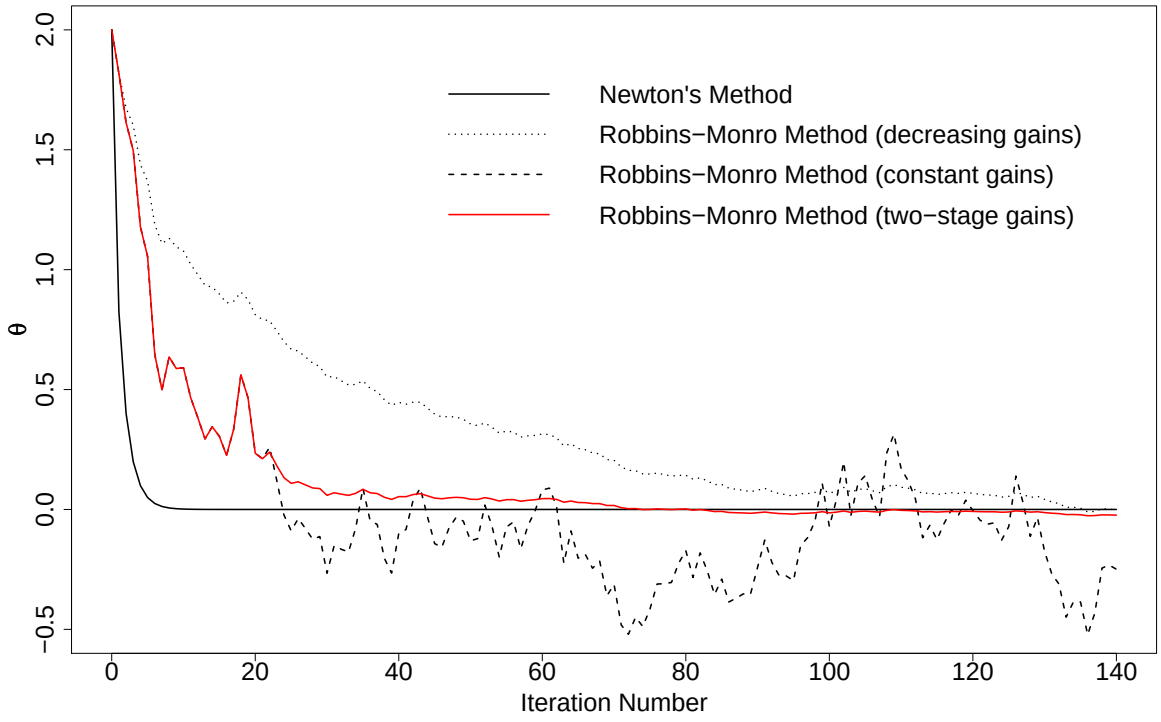

Figure 4.1: The Effect of Gain Constants on the Robbins-Monro Iterations

CHAPTER 5

Applications of MH-RM

In this chapter, the MH-RM algorithm will be applied to fit some latent structure models of realistic complexity to both real and simulated data sets. Whenever possible, estimates from the MH-RM algorithm will be compared with an alternative algorithm. Data analysis with MH-RM is conducted using a C++ program that implements the model and the algorithm as described in Chapters 2, 3, and 4. The computer used throughout this chapter is a laptop equipped with Intel Core^{TM}2 Duo processor at 2.0GHz with 2GB RAM and running Windows XP. The goal here is to illustrate the flexibility and efficiency of MH-RM as a general estimation algorithm for the kinds of latent variable models that frequently arise in psychological research.

5.1 One-Parameter Logistic IRT Model for LSAT6 Data

The Law School Admission Test section 6 (LSAT6) data set is a well-known real data set analyzed by a number of authors including Bock and Lieberman (1970), Bock and Aitkin (1981), and Thissen (1982), among others. This data set consists of 1000 responses to 5 dichotomous items. The unidimensional 1-parameter logistic IRT model is known to fit the data well. Table 5.1 contains parameter estimates and standard errors for item intercepts and a slope parameter that is constrained to be equal across items. As a comparison, the same model is estimated in Mplus (Muthén & Muthén, 2007) using the EM algorithm with adaptive Gaussian quadrature (20 points). Mplus converged in 21 cycles and took 1 second. The two sets of estimates

are almost identical.

Due to the nature of the unidimensional model, it is not realistic to expect a Monte Carlo based estimation method to outperform a specialized estimation algorithm, i.e., EM with Gaussian quadrature, in terms of CPU time. However, even in this unfavorable comparison, MH-RM converged in just 12 seconds at cycle 46 with 2 imputations per cycle. This clearly demonstrates that MH-RM can, in principle, be used in unidimensional IRT estimation.

5.2 Three-Parameter Logistic IRT Model for LSAT6 Data

MH-RM allows the imposition of univariate logit-normal priors on the guessing parameters in the three-parameter IRT model (see section 2.2.1). To demonstrate this capability, and to verify results under arguably the most popular IRT model for educational tests, the 3PL model is fitted to the LSAT6 data set using both the C++ implementation of MH-RM and Multilog (Thissen, 2003). Prior research suggest that very little guessing is involved in these items, and that priors on the lower asymptote parameters are crucial for stable convergence to the MLE. The prior in the present analysis is chosen to be logit-normal with mean -1.4 and standard deviation 0.5. This corresponds to a multiple-choice test with 5 response alternatives as the prior mean -1.4 is roughly equal to $\log(0.2/0.8)$.

Table 5.2 presents estimates and standard errors from both programs. Multilog used 19 equally-spaced fixed quadrature nodes to perform the numerical integration in the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981). It converged at cycle 21 in less than 1 second. MH-RM used 90 seconds in 123 cycles with 5 imputations per cycle. It is clear that the two programs produced almost identical point estimates. The standard errors are slightly different. Multilog standard errors are known to be quite inaccurate (see Cai, in press), which appears to be the case here. Note in the last column that it produced standard errors that are *larger* than the prior standard deviation. Taken together, the performance of MH-RM is entirely acceptable.

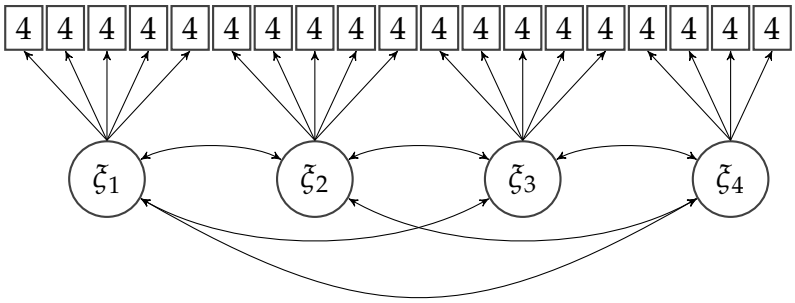

Figure 5.1: Path Diagram for Confirmatory Item Factor Analysis

5.3 Four-Dimensional Confirmatory Item Factor Analysis

Confirmatory item factor models (Edwards, 2005) for ordinal responses can be obtained as a special case of the multidimensional graded IRT model presented in section 2.2.2 by placing restrictions on the item slopes. In his dissertation Edwards (2005) used Bayesian MCMC estimation methods to fit a four-dimensional item factor analysis model to a simulated data set having 19 items and 2000 respondents. The factor pattern is that of a perfect-cluster simple structure, i.e., each item loads on one and only one factor with no cross-loadings. The latent variables are constrained to have zero means and unit variances to identify the model. The correlations among the factors are freely estimated. Figure 5.1 depicts the model as a path diagram. The number 4 in the rectangles indicates that observed variables are scored in 4 categories.

Edwards (2005) implemented a version of data augmented Gibbs sampling algorithm in a C++ software package (MultiNorm). Because Edwards (2005) used the normal ogive parametrization, whereas the present MH-RM implementation is based on the logistic, all generating item parameter values and estimates in Table 5.4 are converted to logistic metric. Edwards (2005) reported that the generating parameter values are consistent with estimates that one frequently encounters in practice.[1]

[1]Data set kindly supplied by Dr. Mike Edwards, The Ohio State University.

Edwards (2005) reported that the MCMC estimation ran for 60,000 cycles and required 3 1/2 hours to finish. A thinning interval of 50 was used, resulting in a final MCMC sample of 1200, of which the first 200 were discarded as burn-in. In comparison, MH-RM converged in 3 minutes and 54 seconds on a similarly equipped computer before convergence at cycle 484 with 1 imputation per cycle. The dispersion constant for the MH proposal density is tuned to .45, which produced a chain that accepted about 40 percent of the imputations.

Tables 5.3 and 5.4 give generating parameter values and estimates from both algorithms. The MultiNorm estimates are posterior means, which is different from the MLE by definition. Despite this difference, and the difference in normal vs. logistic parametrization, the two methods yielded essentially identical estimates. However, MH-RM is several orders of magnitude faster than MCMC.

5.4 Latent Variable Interaction Analysis

The modelling of two-way interactions in latent variables has received considerable attention from methodologists since the seminal work of Kenny and Judd (1984). See Marsh, Wen, and Hau (2004) for a recent review. Klein and Moosbrugger (2000) showed that full-information maximum likelihood estimation for such polynomial structural equation models is feasible. Indeed, latent variable models with interaction effects can be specified as nonlinear mixed models and SAS PROC NLMIXED can be used to estimate their parameters with adaptive Gaussian quadrature.

As pointed out in section 2.1.2, latent variable interactions can be obtained as a special case of the general nonlinear latent structure model. MH-RM can therefore be used to estimate the parameters. Using notation of Chapter 2, let

$$\eta = \tau_* + \mathbf{\Delta}_* g(\boldsymbol{\xi}) + \zeta = \tau_* + \begin{bmatrix} \delta_{*1} & \delta_{*2} & \delta_{*3} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_1\xi_2 \end{bmatrix} + \zeta, \tag{5.1}$$

be the structural equation for latent variable interaction analysis, where

$$g\left(\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}\right) = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_1\xi_2 \end{bmatrix}$$

is the nonlinear function that produces the polynomial effects in $\boldsymbol{\xi}$. The structural model has 10 free parameters: 4 regression coefficients (τ_*, δ_{*1}, δ_{*2}, δ_{*3}), two means for ξ_1 and ξ_2 (τ_1, τ_2), three dispersion components for ξ_1 and ξ_2 (ϕ_{11}, ϕ_{21}, ϕ_{22}), and one variance component (ψ_{11}) for ζ.

Following Marsh et al. (2004), let there be three continuous indicators per factor. In matrix notation, the measurement model can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{bmatrix} + \begin{bmatrix} \beta_1 & 0 & 0 \\ \beta_2 & 0 & 0 \\ \beta_3 & 0 & 0 \\ 0 & \beta_4 & 0 \\ 0 & \beta_5 & 0 \\ 0 & \beta_6 & 0 \\ 0 & 0 & \beta_7 \\ 0 & 0 & \beta_8 \\ 0 & 0 & \beta_9 \end{bmatrix} \begin{bmatrix} \eta \\ \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{bmatrix}, \tag{5.2}$$

where u_j is the uniqueness term with mean 0 and variance σ_j^2. To identify the model, let y_1, y_4, and y_7 be scaling indicators by fixing β_1, β_4, and β_7 to one and the corresponding α's to zero. Thus the measurement model has 21 parameters: 6 intercepts, 6 slopes (loadings), and 9 unique variances. Figure 5.2 shows the path diagram for this latent variable interaction model.

One data set was simulated from the model described above with $N = 1000$ and generating parameter values similar to those reported in Marsh et al. (2004).[2] The

[2]Data set kindly supplied by Ms. Wenjing Huang, University of North Carolina – Chapel Hill.

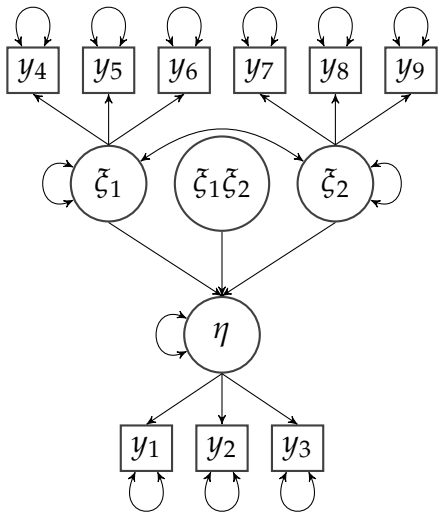

Figure 5.2: Path Diagram for Latent Variable Interaction

C++ program implementing MH-RM and SAS PROC NLMIXED (SAS Institute Inc., 2004) were both used to fit the model to the simulated data set. Table 5.5 presents generating parameter values and estimates from both programs for the structural model. Table 5.6 shows the results for the measurement model. The two sets of point estimates are virtually identical and the standard errors are also quite close.

In NLMIXED, a quasi-Newton algorithm in conjunction with 6-point adaptive Gaussian quadrature was specified; computation time was 3 hours 45 minutes. However, several gradient elements are still larger than 0.001 at the NLMIXED solution. In contrast, MH-RM required 258 cycles, and 1 minute 42 seconds, to converge, with the number of imputations per MH-RM cycle set to 1. The dispersion constant for the MH proposal density was tuned to .8, which produced a chain that accepted about 33 percent of the imputations.

5.5 Latent Mediated Regression with Dichotomous Indicators

Mediation, or indirect effect, is a topic of much interest among psychologists (Baron & Kenny, 1986). Figure 5.3 shows a path diagram (measurement part omitted)

for a latent variable mediation model. In the diagram, ξ_1 to ξ_4 are used to predict ξ_5, which in turn predicts ξ_6. If observed variables are all continuous and normally distributed, estimating and testing such a latent mediated regression model is not a difficult task. Now suppose each latent variable is measured by a distinct set of 10 dichotomous observed variables (not shown in the path diagram). In that case, maximum likelihood estimation requires 6-dimensional numerical integration. If one were to use a 5-point quadrature rule to evaluate the integral, $5^6 = 15625$ function evaluations are needed for each integral.

It is clear that this model can be obtained as a special case of the latent structure model (see section 2.1.1). The measurement model is the graded response model presented in section 2.2.2 for two categories. The factor pattern is again that of a perfect cluster simple structure so that each observed variable has exactly two measurement parameters. For the latent structural model, $\mathbf{\Delta}$ contains 5 free parameters:

$$\mathbf{\Delta} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \delta_{51} & \delta_{52} & \delta_{53} & \delta_{54} & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_{65} & 0 \end{bmatrix}.$$

The equation disturbance covariance matrix $\mathbf{\Phi}$ takes the following form:

$$\begin{bmatrix} 1 & & & & & \\ \phi_{21} & 1 & & & & \\ \phi_{31} & \phi_{32} & 1 & & & \\ \phi_{41} & \phi_{42} & \phi_{43} & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

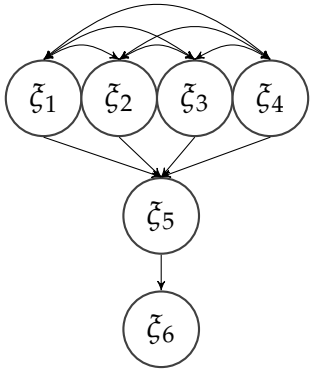

Figure 5.3: Path Diagram for Latent Mediated Regression

The 6 free elements correspond to the saturated correlations among the predictors. To identify the model, $\boldsymbol{\tau}$ is constrained to be a vector of zeros.

Taken together, this model has 120 parameters in the measurement part, and 11 parameters in the structural part. One data set was generated from this model with $N = 500$. Tables 5.7 and 5.8 present generating values for the measurement intercepts and slopes, respectively. Consistent with standard practice (e.g., Chen & Thissen, 1997), the 60 slopes were sampled from a log-normal distribution with (normal) mean 0 and (normal) standard deviation 0.5. To generate the intercepts, thresholds are first sampled from a normal distribution with mean 0 and standard deviation 1.5. Intercepts were then obtained by taking the negative of the product of each threshold and the corresponding slope.

Both Mplus (Muthén & Muthén, 2007) and MH-RM were used to estimate the parameters of this model from the simulated data set. In Mplus, an EM algorithm with adaptive Gaussian quadrature was employed. The number of quadrature points per dimension was set at 5. In MH-RM, 1 imputation per cycle was taken, and the dispersion constant for the MH proposal density was tuned to 0.55, which produced a chain that accepted about 30 percent of the imputations.

Figure 5.4 plots MH-RM estimates of measurement intercepts against Mplus es-

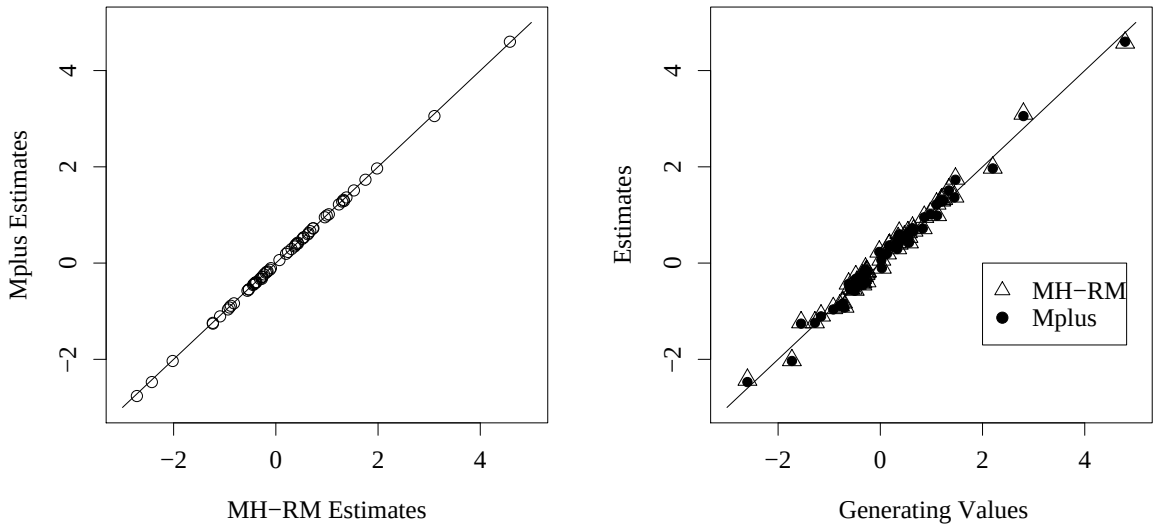

Figure 5.4: Latent Mediated Regression: Intercept Estimates

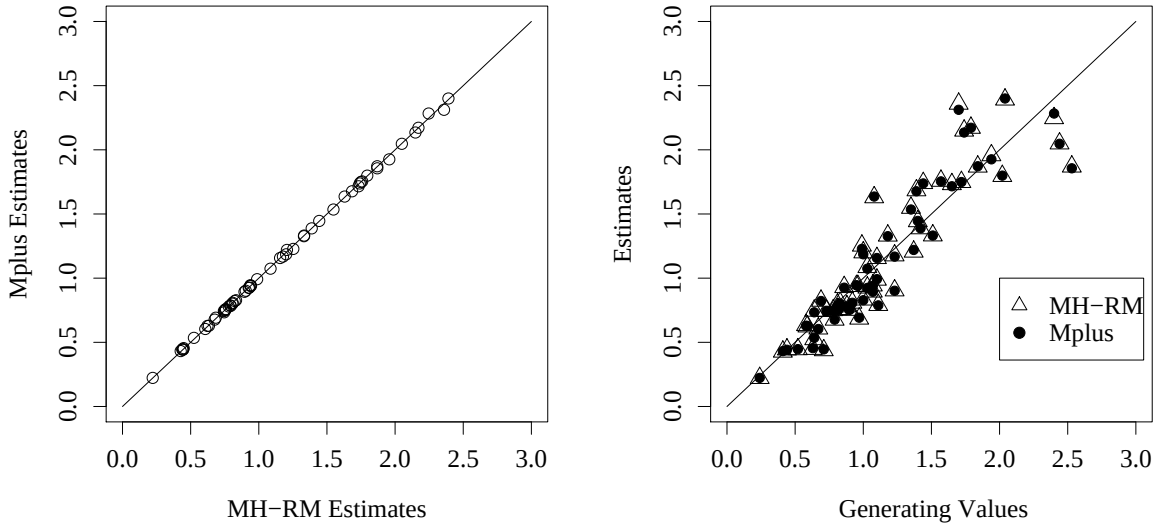

Figure 5.5: Latent Mediated Regression: Slope Estimates

timates (left panel), and both sets of estimates against true generating values (right panel). Figure 5.5 presents similar information for the measurement slopes. As one can easily tell from both figures, the points closely follow the 45 degree reference line in the two panels on the left, indicating the close agreement of MH-RM and Mplus estimates. Given the relatively small N and the fact that this is a one-replication sampling experiment, the degree of variability of the estimated parameters shown in the two panels on the right, when plotted against generating values, is entirely expected.

For the structural parameters, Table 5.9 presents results from both MH-RM and Mplus. The two algorithms produced almost identical point estimates for the structural parameters of interest. Mplus required 2 hours 53 minutes and 38 cycles, whereas MH-RM converged at cycle 598 in 4 minutes 43 seconds. MH-RM is clearly more efficient than quadrature-based EM for problems of this type.

5.6 Full-information Estimation of Tetrachoric Correlations

As mentioned in section 1.1.1, standard estimators in categorical structural equation modelling (e.g., weighted least squares) are multi-stage estimators. A tetrachoric/polychoric correlation matrix is estimated from bivariate marginal frequency tables in the first stage and in the second stage, the structural parameters (e.g., factor loadings and correlations) are estimated. Song and Lee (2003) pointed out that pairwise estimation of the tetrachoric/polychoric correlations from bivariate marginal tables does not lead to maximum likelihood estimates for the full correlation matrix. In practice, the estimated correlation matrix often turns out to be not positive definite, due to the lack of explicit restrictions in the pairwise estimation procedure that ensures the positive definiteness of the full matrix (see Rousseeuw & Molenberghs, 1994, for a nontechnical discussion of the shape of correlation matrices). The pairwise procedure also does not handle missing data in a statistically justified manner. Full maximum likelihood estimation is ideal, but to compute an MLE of a full $n \times n$ tetrachoric correlation matrix n-fold numerical integration is necessary; this is still an

insurmountable difficulty.

With no loss of generality, this section's focus is on full-information maximum likelihood estimation of the tetrachoric correlations; the extension to the polychoric case is straightforward. The approach developed here is similar in principle to Song and Lee's (2003) work, but with important differences in both the modelling framework and the estimation algorithm. Specifically, two competing implementations of the MH-RM algorithm are given. The first is based on further augmenting the model with a set of underlying response variates (Albert, 1992). The other one is based on the latent structure model as it is described in Chapter 2, placing special restrictions on the parameters.

One can use a Gibbs sampler to produce the multiple imputations in the first approach. In that case the complete data model is linear, permitting explicit closed-form estimates of the correlation parameters. The second approach enjoys the advantage of having a uniformly smaller fraction of missing information because no underlying response variates need to be introduced in addition to the factor scores, which leads to a potentially more rapidly converging Robbins-Monro sequence. Less missing information means less Monte Carlo variability; less variability means that there is less error to be filtered out by the RM recursion in 3.11. The additional benefit of the second approach is that the existing C++ program can be used directly.

5.6.1 An Approach Using Underlying Response Variates

The underlying response variates for each dichotomous observed variable can be understood either as a mechanical consequence of the probit link function (Albert, 1992; Chib & Greenberg, 1998), or as a necessity of the psychometric theory behind IRT (Lord & Novick, 1968; Thissen & Wainer, 2001). Recall that there are n dichotomous observed variables and N respondents. For respondent i, let the number of latent variables also be n so that $\boldsymbol{\xi}_i$ is an $n \times 1$ vector normal random variable with mean $\boldsymbol{\tau}$ and covariance matrix $\boldsymbol{\Phi}$. As an identification restriction, $\boldsymbol{\Phi}$ has unit diagonal

elements.

Let $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{in}^*)$ be an $n \times 1$ vector of underlying response random variables that are conditionally independent given $\boldsymbol{\xi}_i$. From a Thurstonian factor analysis tradition, one can start with the following:

$$\mathbf{y}_i^* = \lambda \mathbf{I}_n \boldsymbol{\xi}_i + \mathbf{u}_i, \tag{5.3}$$

where λ can be understood as the factor loading, and $\mathbf{u}_i$'s represent the unique factor terms that are normally distributed with mean zero and covariance matrix $(1 - \lambda^2)\mathbf{I}_n$ such that $\mathbf{y}_i^*$ is conventionally scaled. The observed variables are connected to the underlying response variates in the following manner: if y_{ij}^* is larger than 0, the observed response y_{ij} is 1, and $y_{ij} = 0$ if $y_{ij}^* \leq 0$. This is equivalent to stating that the item response function for item j is the following normal probability integral

$$P(y_{ij} = 1 | \xi_{ij}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{B\xi_{ij}} \exp\left(-\frac{1}{2}t^2\right) dt, \tag{5.4}$$

where

$$B = \frac{\lambda}{\sqrt{1 - \lambda^2}}$$

can be understood as a slope parameter.

Recall that by definition, the tetrachoric correlation matrix of the observed variables is the Pearson correlation matrix of the underlying response variables. If the diagonal elements of the unique factor covariance matrix $(1 - \lambda^2)\mathbf{I}_n$ can be made arbitrarily small, the factor correlation matrix $\boldsymbol{\Phi}$ becomes equivalent to the tetrachoric correlation matrix. This is the guiding insight from Song and Lee (2003) and they proceeded with the factor analysis formulation to arrive at a Monte Carlo EM algorithm for estimating the tetrachoric correlations by FIML.

It is easy to show that $\lambda \to 1$ as $B \to \infty$, so making the unique variance arbitrarily small amounts to picking a sufficiently large B. Equation (5.3) suggests that the following parametrization of the underlying response variables leads to the same

item response function as in Equation (5.4)

$$\mathbf{y}_i^* = B\mathbf{I}_n\boldsymbol{\xi}_i + \mathbf{u}_i^*, \tag{5.5}$$

where $\mathbf{u}_i^*$ is standard multivariate normally distributed. The implication of the present formulation is that each y_{ij}^* is required to have mean $B\xi_{ij}$ for a sufficiently big positive constant B, say, 10, and standard deviation 1. Setting B to 10 is equivalent to setting the unique variance to 0.005, which means that y_{ij}^* is practically a perfect surrogate of ξ_{ij}. From the IRT perspective, the high slope B ensures that the response function is practically a step-function with nearly perfect discrimination. Equation (5.5) is the parametrization used in this section because it makes the approach in section 5.6.2 a conceptual continuation of the present formulation.

If one can produce random imputations of both $\mathbf{y}_i^*$ and $\boldsymbol{\xi}_i$ for $i = 1, \ldots, N$, the threshold parameters as well as the tetrachoric correlation matrix can be estimated in closed form as the mean and correlation matrix of the sampled $\boldsymbol{\xi}$'s. In other words, $\boldsymbol{\tau}$ contains the negative of the threshold parameters whereas the lower triangular elements of $\boldsymbol{\Phi}$ are the tetrachoric correlations.

Some bias is incurred because the unique variance cannot be entirely eliminated, though it can be made arbitrarily small by fixing B to a sufficiently large value. The small bias is a price that one pays in exchange for a dramatically simpler sampling scheme. In particular, the n observed variables (and the underlying response variate for each) are guaranteed to be conditionally independent with the introduction of n factors in $\boldsymbol{\xi}$. This conditional independence reduces the multivariate sampling of $\mathbf{y}^*$ to univariate sampling. If one does not introduce the $\boldsymbol{\xi}$'s, the distribution of $\mathbf{y}^*$ given $\mathbf{y}$ will be multivariate truncated normal, and the development of effective sampling schemes for the multivariate truncated normal distribution is still an open arena for research (see Geweke, 1991 for one algorithm). On the other hand, univariate truncated normal sampling is comparatively a much easier task. A two-step Gibbs sampling scheme that produces $\mathbf{y}_i^*$ and $\boldsymbol{\xi}_i$ is as follows.

First, the distribution of y_{ij}^*, conditional on $\mathbf{y}_i$, $\boldsymbol{\xi}_i$, τ, and $\boldsymbol{\Phi}$ is

$$y_{ij}^* | \mathbf{y}_i, \boldsymbol{\xi}_i, \tau \sim \begin{cases} \mathcal{N}(B\xi_{ij}, 1) \text{ truncated on the left by 0 if } y_{ij} = 1 \\ \mathcal{N}(B\xi_{ij}, 1) \text{ truncated on the right by 0 if } y_{ij} = 0. \end{cases}$$

This is easy to generate by inverting the standard normal cumulative distribution function (see e.g., Albert, 1992).

Second, the distribution of $\boldsymbol{\xi}_i$, conditional on $\mathbf{y}_i$, $\mathbf{y}_i^*$, $\boldsymbol{\tau}$, and $\boldsymbol{\Phi}$ is normal

$$\mathcal{N}_n \left((B^2\mathbf{I}_n + \boldsymbol{\Phi}^{-1})^{-1}(B\mathbf{I}_n\mathbf{y} + \boldsymbol{\Phi}^{-1}\boldsymbol{\tau}), (B^2\mathbf{I}_n + \boldsymbol{\Phi}^{-1})^{-1} \right).$$

This follows directly from Equation (5.5) and elementary Bayesian results on the normal theory linear model.

Once the imputations are obtained, one can follow the second and third steps of the MH-RM algorithm as described in section 3.2 to obtain updated estimates of $\boldsymbol{\tau}$ and $\boldsymbol{\Phi}$, and iterate the process until convergence. Because the complete data model is linear, a result in Cai (2006) can be utilized to show that the MH-RM iteration is formally equivalent to a direct application of the SAEM algorithm.

5.6.2 An Approach Using Logistic Approximation

As is made clear by Equations (5.4) and (5.5), one can use the existing latent structure model and software to obtain an alternative estimation method for the tetrachoric correlation matrix. In this approach, n factors are specified and for item j, its vector of slopes takes the following form

$$\boldsymbol{\beta}_j' = (0, \ldots 0, 1.702 \times B, 0, \ldots, 0)',$$

where the nonzero entry is the jth component. The item intercepts are set to zero, and as in the last section the factor means $\boldsymbol{\tau}$ and $\boldsymbol{\Phi}$ are estimated with the diagonal elements of $\boldsymbol{\Phi}$ restricted to unity for identification. The Metropolis random-walk sampler is then used to draw $\boldsymbol{\xi}$'s directly.

The validity of this alternative approach rests on the similarity of the normal ogive and logistic item response functions. The scaling constant 1.702 is used to maximize this similarity. However, it must be pointed out that the two distributions are not identical, especially in the tails. Also, even if the underlying response variables are not explicitly present in this method, the use of the logistic IRT model implicitly assumes the existence of underlying response variables that are distributed as multivariate logistic variables. Therefore, one cannot expect that the two approaches will yield exactly identical correlations.

5.6.3 An Example

To demonstrate the feasibility of FIML estimation of the tetrachoric correlation matrix using MH-RM, a proof-of-concept simulated example consisting of 5 dichotomous variables is shown. Three methods will be used to estimate the item thresholds and the correlations: 1) the underlying response variables method as in section 5.6.1, 2) the logistic approximation method as in section 5.6.2, and 3) as a benchmark the currently popular pair-wise estimation method. For the last method, Mplus (Muthén & Muthén, 2007) is used.

Table 5.10 lists the generating parameter values. The following procedure was used to generate the data set. First, one multivariate normal sample with correlations equal to the generating correlations and means equal to the generating thresholds parameters was obtained for $N = 200$. Next, the variables were dichotomized to 0-1 according to the thresholds. If a sampled value is larger than the threshold, the result was coded as 1, and 0 otherwise.

Table 5.11 shows the means and Pearson correlations for the obtained multivariate sample. Table 5.12 presents estimates from the dichotomized variables. The threshold estimates from all three methods are close. Some elements in the tetrachoric correlation matrix show slight discrepancy when compared across methods, but the overall pattern is in clear agreement. With essentially one replication at a

small N, little can be said about parameter recovery. However, this small example shows what MH-RM promises when the number of variables becomes large and when there are missing data.

The underlying response variables method was implemented in GAUSS (Aptech Systems, Inc., 2003), which is an interpreted matrix algebra package. However, because the GAUSS programming were vectorized, i.e., no explicit looping was used, there is little interpretation overhead. Thus the efficiency of the GAUSS program should be treated as comparable to the compiled C++ program that implements the second approach using logistic approximation. The GAUSS program required 251 seconds. Using similar options (in terms of the number of imputations, number of thinning cycles, etc.), the C++ program required 115 seconds.

Table 5.1: LSAT6 One-Parameter Logistic Model Estimates

		Item Intercepts (*SE*)				
	Slope (*SE*)	1	2	3	4	5
MH-RM	0.75 (.07)	2.73 (.12)	1.00 (.08)	0.24 (.07)	1.31 (.08)	2.10 (.10)
Mplus	0.76 (.07)	2.73 (.13)	1.00 (.08)	0.25 (.07)	1.31 (.09)	2.10 (.11)

Table 5.2: LSAT6 Three-Parameter Logistic Model Estimates

	Intercept (*SE*)		Slope (*SE*)		Logit(c) (*SE*)	
Item	MH-RM	Multilog	MH-RM	Multilog	MH-RM	Multilog
1	2.52 (.23)	2.53 (.25)	0.82 (.26)	0.84 (.19)	-1.40 (.50)	-1.40 (.51)
2	0.65 (.19)	0.65 (.28)	0.83 (.28)	0.85 (.15)	-1.40 (.49)	-1.41 (.59)
3	-0.28 (.31)	-0.27 (.27)	1.27 (.40)	1.21 (.21)	-1.43 (.48)	-1.45 (.45)
4	0.98 (.18)	0.98 (.22)	0.75 (.22)	0.77 (.14)	-1.40 (.50)	-1.41 (.57)
5	1.80 (.18)	1.80 (.21)	0.71 (.23)	0.71 (.15)	-1.40 (.50)	-1.40 (.56)

Table 5.3: Four-Dimensional Item Factor Analysis: Factor Correlation Estimates

	Factor Correlations					
	(ξ_2, ξ_1)	(ξ_3, ξ_1)	(ξ_3, ξ_2)	(ξ_4, ξ_1)	(ξ_4, ξ_2)	(ξ_4, ξ_3)
True	*0.75*	*0.70*	*0.75*	*0.60*	*0.50*	*0.80*
MH	0.74	0.69	0.74	0.63	0.52	0.81
MC	0.74	0.69	0.73	0.63	0.51	0.80

Note. True = Generating values; MH = MH-RM estimates; MC = MCMC estimates.

Table 5.4: Four-Dimensional Item Factor Analysis: Item Parameter Estimates

	Slope			Intercept 1			Intercept 2			Intercept 3		
Item	True	MH	MC	True	MH	MC	True	MH	MC	True	MH	MC
1	*2.43*	2.43	2.43	*2.45*	2.58	2.43	*0.56*	0.79	0.68	*-3.13*	-2.82	-2.83
2	*2.66*	2.70	2.66	*2.76*	3.11	2.91	*0.51*	0.75	0.63	*-3.00*	-2.82	-2.81
3	*2.09*	2.19	2.16	*3.17*	3.56	3.39	*1.36*	1.66	1.55	*-1.74*	-1.58	-1.60
4	*1.63*	1.73	1.75	*1.06*	1.18	1.11	*-0.44*	-0.45	-0.51	*-2.31*	-2.28	-2.31
5	*2.08*	1.96	2.01	*2.03*	2.17	2.09	*0.56*	0.74	0.68	*-1.36*	-1.23	-1.28
6	*2.31*	2.20	2.21	*0.53*	0.73	0.63	*-1.12*	-0.88	-0.97	*-3.57*	-3.43	-3.42
7	*1.84*	1.77	1.75	*1.40*	1.58	1.50	*0.00*	0.18	0.12	*-2.52*	-2.42	-2.45
8	*1.46*	1.46	1.48	*0.07*	0.18	0.12	*-1.00*	-0.97	-1.04	*-2.54*	-2.51	-2.55
9	*2.08*	1.91	1.91	*1.92*	2.05	1.96	*0.27*	0.43	0.36	*-2.48*	-2.35	-2.37
10	*1.99*	1.92	1.96	*0.95*	1.13	1.06	*-0.68*	-0.59	-0.66	*-2.86*	-2.83	-2.86
11	*2.81*	2.88	2.88	*1.75*	2.13	1.96	*-0.36*	-0.15	-0.26	*-2.91*	-2.85	-2.88
12	*3.13*	3.11	3.10	*1.96*	2.28	2.11	*-0.41*	-0.15	-0.26	*-2.84*	-2.76	-2.77
13	*2.30*	2.14	2.16	*2.16*	2.26	2.14	*0.56*	0.67	0.58	*-1.97*	-1.83	-1.87
14	*1.87*	1.82	1.86	*0.63*	0.82	0.75	*-1.06*	-0.86	-0.92	*-2.84*	-2.65	-2.66
15	*2.14*	2.10	2.09	*0.39*	0.70	0.61	*-1.19*	-0.98	-1.04	*-3.47*	-3.39	-3.37
16	*2.09*	2.09	2.08	*2.62*	2.95	2.81	*0.83*	1.10	1.00	*-1.65*	-1.53	-1.57
17	*2.13*	2.10	2.09	*3.47*	3.77	3.59	*1.36*	1.65	1.55	*-1.28*	-1.11	-1.16
18	*1.60*	1.45	1.48	*1.77*	1.83	1.79	*-0.10*	0.03	-0.02	*-2.11*	-2.01	-2.04
19	*1.38*	1.28	1.31	*1.11*	1.15	1.12	*-0.61*	-0.45	-0.49	*-3.44*	-3.25	-3.23

Note. True = Generating values; MH = MH-RM estimates; MC = MCMC estimates.

Table 5.5: Latent Variable Interaction: Structural Model Estimates

Parameter		Generating Value	Estimate (*SE*)	
			MH-RM	NLMIXED
Intercept	τ_*	*-1.50*	-1.61 (.15)	-1.60 (.16)
Slope for ξ_1	δ_{*1}	*0.70*	0.80 (.06)	0.80 (.07)
Slope for ξ_2	δ_{*2}	*0.50*	0.51 (.10)	0.51 (.11)
Slope for $\xi_1\xi_2$	δ_{*3}	*0.25*	0.20 (.05)	0.20 (.05)
Residual Variance for ζ	ψ_{11}	*0.60*	0.66 (.09)	0.67 (.09)
Factor Mean for ξ_1	τ_1	*2.00*	1.96 (.04)	1.96 (.05)
Factor Mean for ξ_2	τ_2	*1.00*	0.99 (.03)	0.99 (.04)
Variance for ξ_1	ϕ_{11}	*1.50*	1.42 (.08)	1.45 (.11)
Cov(ξ_1,ξ_2)	ϕ_{21}	*0.00*	-0.05 (.05)	-0.05 (.05)
Variance for ξ_2	ϕ_{22}	*0.70*	0.76 (.05)	0.77 (.08)

Table 5.6: Latent Variable Interaction: Measurement Model Estimates

	Intercept (*SE*)			Slope (*SE*)			Uniqueness (*SE*)		
Item	True	MH	SAS	True	MH	SAS	True	MH	SAS
1	*0.0*	–	–	*1.0*	–	–	*.7*	.72(.06)	.71(.07)
2	*1.0*	1.02(.04)	1.03(.04)	*.8*	.80(.02)	.80(.03)	*.8*	.84(.05)	.84(.06)
3	*1.4*	1.35(.05)	1.36(.05)	*.5*	.52(.03)	.51(.03)	*1.5*	1.52(.07)	1.52(.07)
1	*0.0*	–	–	*1.0*	–	–	*.7*	.72(.05)	.70(.06)
2	*1.0*	1.03(.06)	1.05(.08)	*.8*	.80(.03)	.79(.04)	*.8*	.78(.04)	.79(.05)
3	*1.4*	1.39(.08)	1.40(.09)	*.5*	.52(.04)	.52(.04)	*1.5*	1.57(.07)	1.56(.07)
1	*0.0*	–	–	*1.0*	–	–	*.7*	.76(.05)	.75(.07)
2	*1.0*	0.97(.06)	0.98(.08)	*.8*	.80(.05)	.80(.07)	*.8*	.78(.05)	.78(.05)
3	*1.4*	1.49(.07)	1.50(.07)	*.5*	.46(.05)	.45(.06)	*.5*	1.58(.07)	1.58(.07)

Note. True = Generating values; MH = MH-RM estimates; SAS = SAS PROC NLMIXED estimates.

Table 5.7: Latent Mediated Regression: Measurement Intercept Generating Values

	Distinct Observed Variables for Each Latent Variable									
	1	2	3	4	5	6	7	8	9	10
ξ_1	-0.26	1.20	0.13	0.54	-0.32	-0.79	-0.27	0.17	-0.38	0.83
ξ_2	2.20	-0.46	1.11	-0.59	0.36	-0.27	-0.70	1.45	0.56	-0.29
ξ_3	0.98	0.55	1.19	0.63	1.34	0.86	-0.54	-0.35	0.03	-0.62
ξ_4	-0.28	0.35	4.79	0.02	2.80	-0.35	-1.28	-0.46	-0.73	0.65
ξ_5	-0.36	0.32	-0.02	-0.92	0.33	0.47	0.18	-1.73	0.37	-2.60
ξ_6	-1.16	-0.29	-0.50	-1.55	-3.66	1.10	-0.35	1.23	1.47	-0.48

Table 5.8: Latent Mediated Regression: Measurement Slope Generating Values

	Distinct Observed Variables for Each Latent Variable									
	1	2	3	4	5	6	7	8	9	10
ξ_1	1.07	1.03	0.59	1.42	1.57	0.83	0.78	2.44	1.23	0.41
ξ_2	1.40	0.88	1.37	1.44	0.71	1.08	1.74	1.72	0.64	0.52
ξ_3	2.04	0.58	1.35	0.64	1.84	0.95	2.02	0.96	0.92	0.82
ξ_4	0.63	1.51	2.40	0.79	1.94	0.90	1.07	0.97	1.70	1.10
ξ_5	1.11	1.03	0.24	1.00	0.67	0.73	0.44	0.99	1.18	1.39
ξ_6	0.81	0.86	0.69	1.23	2.53	0.81	1.00	1.65	1.10	1.79

Table 5.9: Latent Mediated Regression: Structural Model Estimates

		Estimate (*SE*)	
Parameter	Generating Value	MH-RM	Mplus
δ_{51}	*0.60*	0.51 (.11)	0.51 (.10)
δ_{52}	*0.30*	0.26 (.14)	0.27 (.12)
δ_{53}	*0.40*	0.49 (.16)	0.49 (.13)
δ_{54}	*0.50*	0.33 (.09)	0.34 (.10)
δ_{65}	*0.50*	0.51 (.05)	0.51 (.07)
ϕ_{21}	*0.30*	0.24 (.06)	0.25 (.06)
ϕ_{31}	*0.40*	0.38 (.09)	0.38 (.06)
ϕ_{32}	*0.60*	0.61 (.10)	0.61 (.05)
ϕ_{41}	*0.20*	0.22 (.06)	0.22 (.07)
ϕ_{42}	*0.50*	0.42 (.07)	0.42 (.06)
ϕ_{43}	*0.30*	0.31 (.06)	0.31 (.06)

Table 5.10: Generating Tetrachoric Correlations and Thresholds

	Variable				
Variable	1	2	3	4	5
1	-1.00				
2	0.72	-0.50			
3	0.48	0.56	0.00		
4	0.48	0.40	0.20	0.50	
5	0.72	0.68	0.32	0.64	1.00

Note. Thresholds are on the diagonal.

Table 5.11: Means and Pearson Correlations of the Underlying Response Variables

	Variable				
Variable	1	2	3	4	5
1	-0.94				
2	0.67	-0.46			
3	0.43	0.55	0.04		
4	0.41	0.29	0.13	0.51	
5	0.66	0.63	0.32	0.62	1.05

Note. Means are on the diagonal.

Table 5.12: Comparison of Three Estimation Methods for Tetrachoric Correlations

1	2	3	4	5
-1.15,-1.17,-1.15				
.72, .68, .70	-.51,-.50,-.51			
.46, .50, .51	.68, .67, .66	.05, .04, .05		
.50, .55, .59	.28, .23, .28	.03, .01, .03	.50, .51, .51	
.52, .53, .45	.76, .73, .60	.07, .11, .10	.48, .53, .53	1.13, 1.11, 1.13

Note. Thresholds are on the diagonal. The estimates in left-to-right order are obtained from: 1) the underlying response variables method as in section 5.6.1, 2) the logistic approximation method as in section 5.6.2, and 3) the Mplus implementation of the pair-wise method, in that order.

CHAPTER 6

Preliminary Sampling Experiments with MH-RM

While the applications to real and simulated data shown in the last chapter tell convincing success stories about the MH-RM algorithm in a wide variety of modelling contexts, questions such as parameter recovery and accuracy of standard error estimates are best answered by simulation studies. To that end, some results from three small scale sampling experiments are reported in this chapter. The C++ program implementing MH-RM is used throughout for both random data generation and model fitting. Only a relatively small number of replications is attempted under each condition because of the preliminary nature of these sampling experiments.

6.1 A Unidimensional Model

The data generation model is a standard 2-parameter logistic IRT model with $n = 10$ items. The reason for including this relatively simple model in the simulation is to investigate the accuracy of the estimated standard errors produced as a by-product of the MH-RM iterations. While results on the convergence of the approximation to the observed data information matrix as given in section 3.5 hold under fairly general conditions, they are asymptotic results that may be difficult to verify given the finite sample size and finite computing time in practice.

Theory on multiple imputation (Little & Rubin, 1987) suggests that while taking the number of imputations per MH-RM cycle m_k to be identically equal to 1 is enough to ensure the convergence of the sequence of point estimates to the MLE, it may lead

to downward bias in the estimated standard errors. In a slightly different context, taking $m_k = 1$ is analogous to conducting a single round of imputation for survey nonresponse, a practice that tends to produce biased standard errors. Multiple imputation theory also suggests that even when the fraction of missing information is large, 5 to 10 imputations is usually good enough. An infinite number of imputations guarantees full efficiency, but a diminishing marginal returns effect is clearly at work. Limited preliminary work with the MH-RM algorithm shows that 5 to 10 imputations per cycle may be more than sufficient for high quality standard errors. This conjecture is investigated in the present simulation.

Three sample size conditions are considered: 200, 1000, and 3000, corresponding to small, large, and very large N, respectively, for a 10-item test. As in section 5.5, the generating slope parameters are sampled from a log-normal (0,0.5) distribution, and the thresholds are sampled from a normal (0,1.5) distribution. Intercepts are then obtained from the negative product of thresholds and slopes. The latent trait variable is scaled as standard normal. Table 6.2 lists the generating item parameters in the first column, where α_j is the intercept and β_j is the slope for item j.

The number of Monte Carlo replications is set to 200. Though this is smaller than the typical number of replications seen in the literature, it is sufficient to detect clear trends and to verify the accuracy of both the point estimates and the standard errors. After all, the aim of this simulation study is not about Type I errors or making accurate power tables. The number of imputations for the MH-RM algorithm is set to 5, and the dispersion of the proposal density in the Metropolis random-walk sampler is set to 2.4.

Timing information is listed in Table 6.1. It is clear that as sample size increases, the average number of MH-RM cycles decreases steadily. This is natural because MH-RM has the same asymptotic (in time) behavior as the SAEM algorithm, whose rate of convergence is inversely related to the fraction of missing information, which

decreases as N increases. On the other hand, the amount of computation per cycle increases linearly in N. The two processes interact such that the average CPU time per replication for $N = 3000$ is only about 2 times the average CPU time for $N = 200$, as opposed to a factor of 15 when predicted from N alone.

Table 6.2, 6.3, and 6.4 present summaries of simulation results for $N = 200$, $N = 1000$, and $N = 3000$, respectively. Let θ denote a generic item parameter, and let $\hat{\theta}$ be its MLE. A variety of statistics are computed to examine parameter recovery, including the mean of the point estimates $E(\hat{\theta})$, absolute bias $(\hat{\theta} - \theta)$, and the mean and standard deviation of a univariate z statistic, computed as

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

for each parameter, where $se(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$ obtained from Equation (3.15). The z statistic should have a mean of 0 and standard deviation of 1, if the point estimate is correctly centered and the standard error estimate is accurate. The tables show that as N increases, the quality of the point estimates and standard errors becomes better, as indicated by the diminishing bias and the closer agreement between the Monte Carlo standard deviation of the point estimates $SD(\hat{\theta})$ and the mean of estimated standard errors $E\{se(\hat{\theta})\}$. The mean and standard deviation of the univariate z's also approach 0 and 1, respectively.

To further aid interpretation, a series of plots are made using information from the tables. For $N = 200$ and the item intercept parameters, the left panel in Figure 6.1 plots $E(\hat{\theta})$ against θ, and the right panel plots $\log E\{se(\hat{\theta})\}$ against $\log SD(\hat{\theta})$. While the point estimates are aligned correctly against the true generating values, the standard errors are generally underestimated. Figure 6.2 shows a similar pattern for the slopes. When N is increased to 1000, the downward bias of the standard error estimates largely disappears, as shown in Figures 6.3 and 6.4. Figures 6.5 and 6.6 show results for $N = 3000$, which is even better than the $N = 1000$ condition.

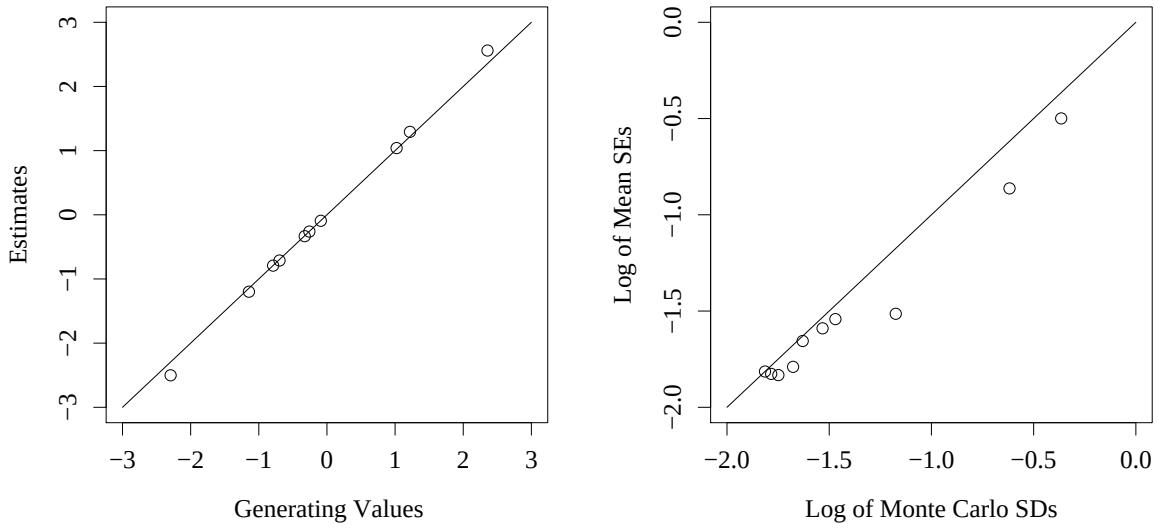

Figure 6.1: Unidimensional IRT Model ($N = 200$): Intercepts

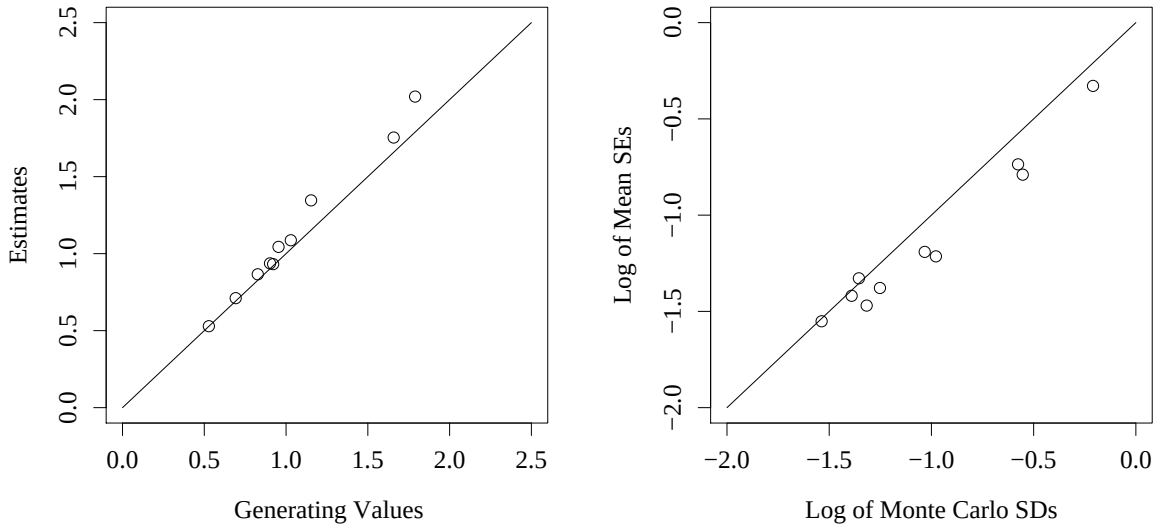

Figure 6.2: Unidimensional IRT Model ($N = 200$): Slopes

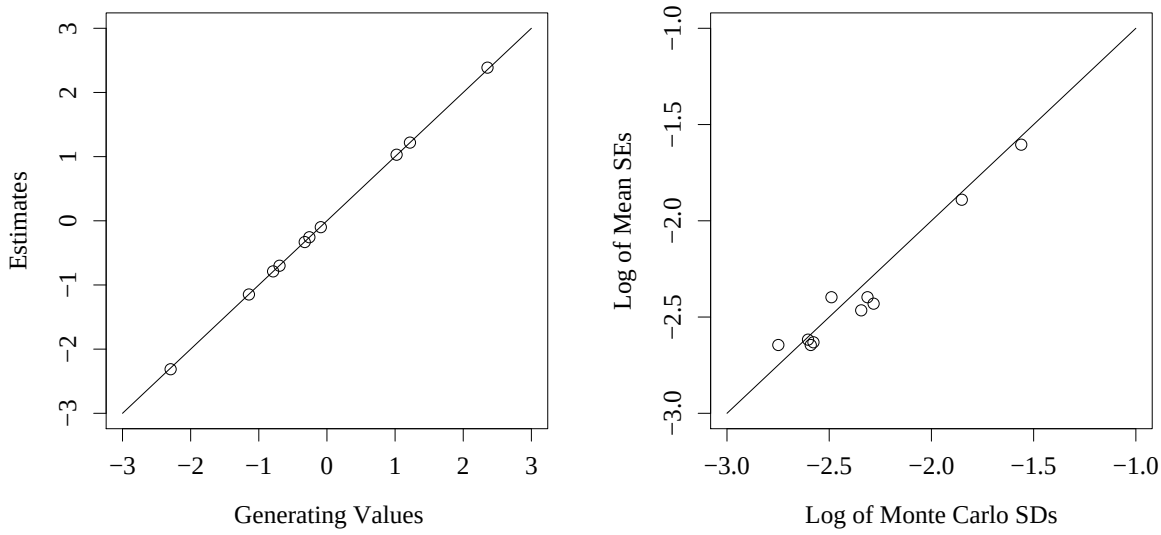

Figure 6.3: Unidimensional IRT Model ($N = 1000$): Intercepts

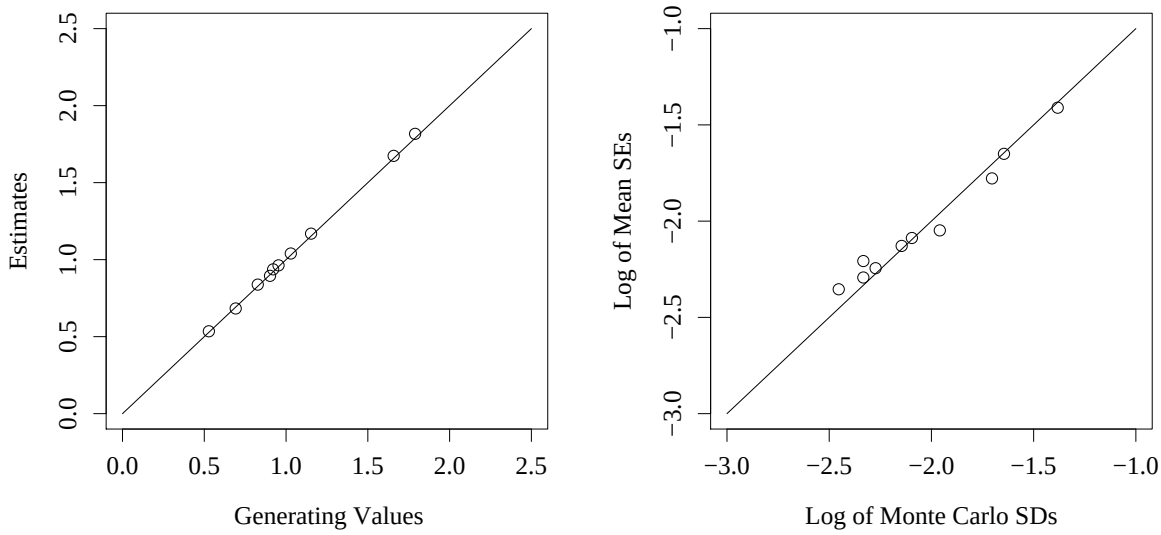

Figure 6.4: Unidimensional IRT Model ($N = 1000$): Slopes

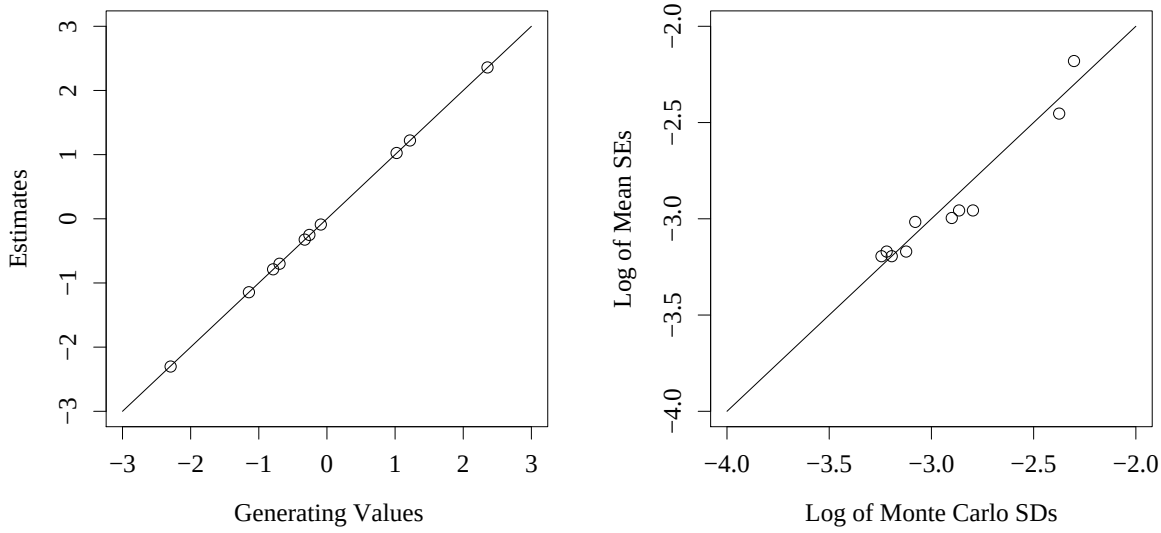

Figure 6.5: Unidimensional IRT Model ($N = 3000$): Intercepts

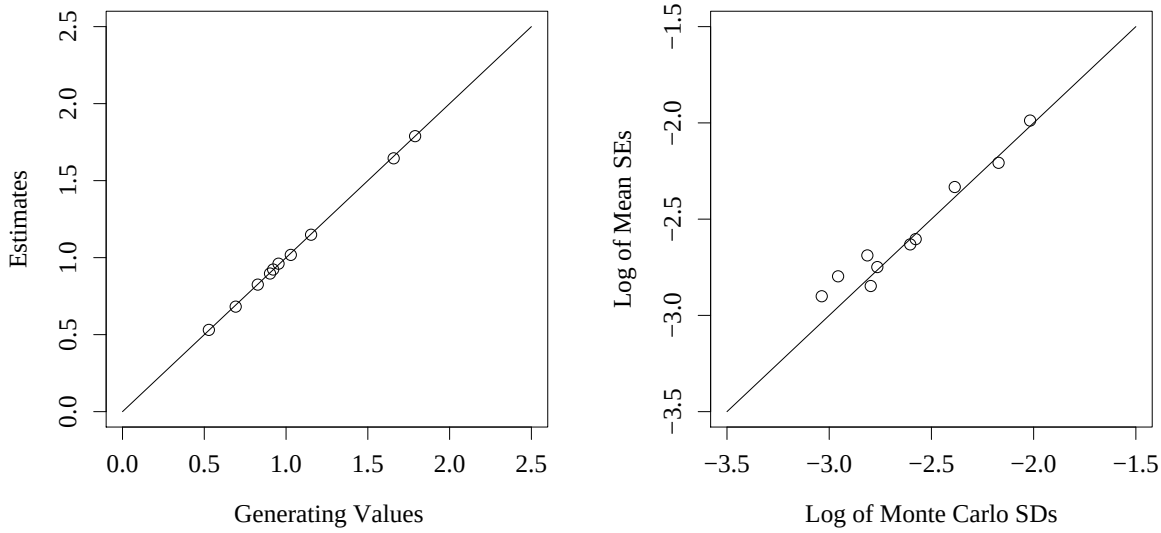

Figure 6.6: Unidimensional IRT Model ($N = 3000$): Slopes

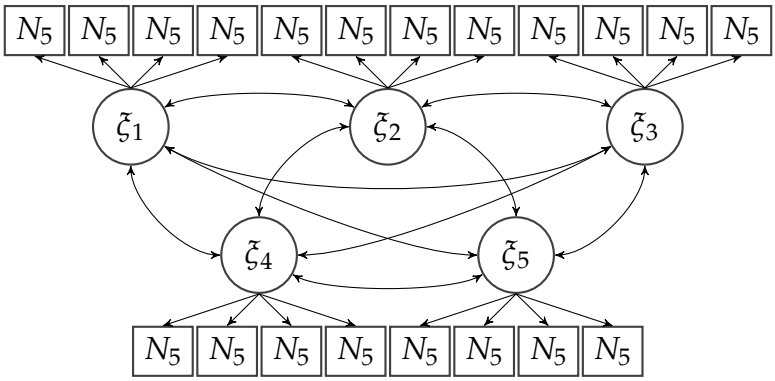

Figure 6.7: Path Diagram for A Constrained Multidimensional Nominal Model

6.2 A Constrained Multidimensional Nominal Model

This simulation study may be the first empirical verification of the multidimensional nominal model, so the data generating model is kept relatively simple. A nominal categories confirmatory item factor model with five correlated factors and perfect simple structure factor pattern is used. Figure 6.7 shows a path diagram for this model, where the symbol N_5 in each of the rectangles represents nominal observed variables in 5 categories. There are $n = 20$ items, $N = 500$ respondents, and each factor is measured by a distinct set of four items possessing the same $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ parameters (see section 2.2.3). The goal of this study is on parameter recovery of the point estimates with a relatively small sample size. Hence the number of imputations per MH-RM cycle is set to $m_k = 1$ to reduce the run time.

For 5 nominal categories, the linear-Fourier basis matrix (see Equation 2.20) is given by

$$\mathbf{F}(5) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0.707 & 1 & 0.707 \\ 2 & 1 & 0 & -1 \\ 3 & 0.707 & -1 & 0.707 \\ 4 & 0 & 0 & 0 \end{bmatrix}.$$

Let $\boldsymbol{\alpha}_j = (1, 2, 2)'$ and $\boldsymbol{\gamma}_j = (1, 0.5, 0, 0)'$ for all $j = 1, \ldots, n$. The nominal category parameters are

$$\mathbf{a}(\boldsymbol{\alpha}_j) = \mathbf{F}(5) \begin{bmatrix} 1 \\ \boldsymbol{\alpha}_j \end{bmatrix} = \begin{bmatrix} 0 & 5.12 & 1 & 3.12 & 4 \end{bmatrix}',$$

and

$$\mathbf{c}(\boldsymbol{\gamma}_j) = \mathbf{F}(5)\boldsymbol{\gamma}_j = \begin{bmatrix} 0 & 1.354 & 2.500 & 3.354 & 4 \end{bmatrix}'.$$

Because the last two components of $\boldsymbol{\gamma}_j$ are equal to zero, they were constrained to be zero during model fitting using the linear restrictions capability of the C++ program.

With 100 replications, bias and variability of the parameter estimates were assessed. On average it required MH-RM 378 iterations in 1253 seconds to complete each replication. The lowest number of iterations was 297 (in 960 seconds) and the highest number was 427 (in 1639 seconds). The Metropolis proposal dispersion was set at 0.3.

Due to the perfect cluster factor pattern, each item loads on one and only one factor. Table 6.5 shows the generating values and estimates of the item slope parameters side by side. Overall, the bias in point estimates is relatively small. The most interesting feature is that the bias are all in the negative direction. It may be attributable to the small sample size $N = 500$, relative to the number of item parameters $d = 130$.[1] The Monte Carlo standard deviations of the estimated slopes and the mean of the estimated standard errors are also shown. Given that m_k is set to 1, the estimated standard errors for the slopes tend to be negatively biased. This is to be expected based on results in section 6.1. Table 6.6 presents a similar set of results for the factor correlation matrix.

Table 6.7 shows point estimates and bias information for the nominal categories parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. The bias tends to be smaller in proportion than that of the

[1] In a subsequent sampling experiment for $N = 1000$ that is not reported here, the negative bias decreased almost by half.

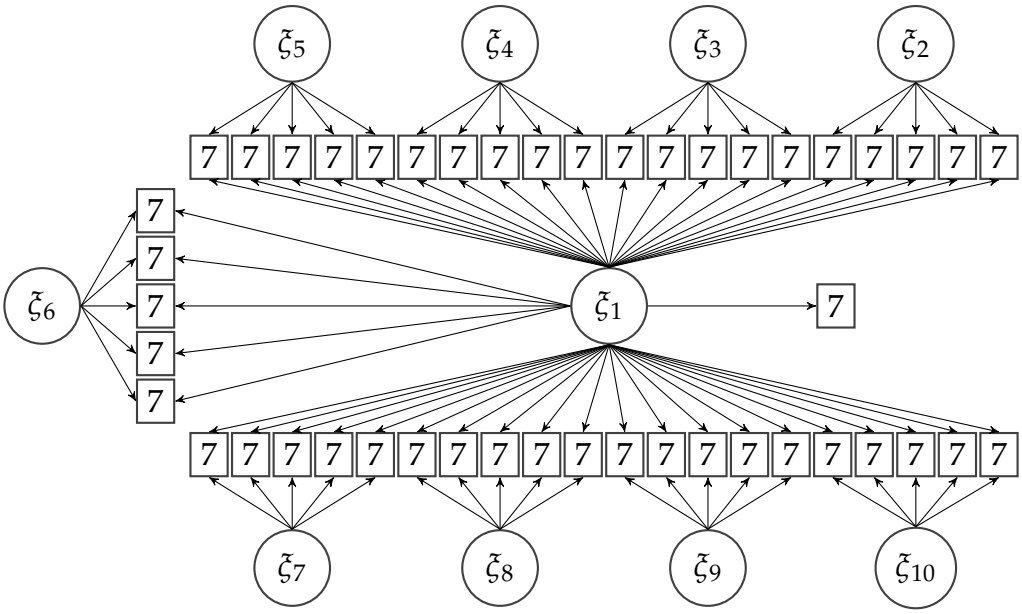

Figure 6.8: Path Diagram for a Bifactor Type Model for Graded Responses

slopes. Finally table 6.8 compares Monte Carlo standard deviations of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ with the average of the estimated standard errors. Again, the estimated standard errors are smaller, but not appreciably so. This seems to suggest that even when m_k is small, the standard error estimates are not entirely useless as indicators of sampling variability for the parameters in confirmatory item factor analysis.

6.3 A Bifactor Type Model for Graded Responses

The bifactor model (Holzinger & Swineford, 1937) is an important special case of confirmatory factor analysis. At the item level, item bifactor analysis (Gibbons & Hedeker, 1992; Gibbons et al., 2007) is also extensively used for modelling local dependence. Much of the focus in the methodological literature on item bifactor analysis deals with special purpose algorithms for reducing the dimension of integration so that numerical quadrature can be applied. Analyses of such special cases can be useful at times, but the generality of the latent structure model and the MH-RM algorithm has enabled psychometricians, for the first time, to specify full information

item bifactor models (or closely related variants that would break the special case algorithm due to Gibbons & Hedeker, 1992) as a standard confirmatory item factor analysis model by simply placing restrictions on parameters, and fit the model by maximum likelihood. While the generality afforded by this approach is attractive (e.g., item bifactor models that mix item types are now possible), its performance in practice is best checked with a simulation study.

Figure 6.8 shows the path diagram of the data generation model. This model is motivated by the bifactor analysis reported in Gibbons et al. (2007) on Lehman's (1988) classical quality of life scale. One can think of ξ_1 as the primary dimension underlying the 46 observed variables, each scored on a 7-point ordinal scale (hence the 7 in the rectangles). The items also fall into 9 clusters that have 5 item in each cluster. The 9 additional factors that are orthogonal to the primary dimension can be thought of as nuisance dimensions. In the context of quality of life scales, these are often termed domain-specific factors. Note that there is a singleton item that loads on the primary dimension only. The existence of such items is not unusual in practice. In fact, in Lehman's (1988) original scale, there is a global life satisfaction item that does not appear to belong to any specific domain. The presence of such an item that belongs only to the primary dimension is a departure from the standard bifactor model in which each item is required to load on two and only two factors. As a confirmatory item factor analysis model, the dimensionality of integration for maximum likelihood fitting of the model in Figure 6.8 is equal to 10, which is the number of dimensions that MH-RM attempts to cover in this simulation study.

The generating parameter values are given in Tables 6.9 and 6.10. These parameter values are chosen to be similar to the bifactor item parameters reported in Gibbons et al. (2007). The sample size is $N = 2000$. The Metropolis random walk proposal dispersion is set at .25, and $m_k = 1$.

A total of 100 replications were conducted, with average run time of 1191 seconds

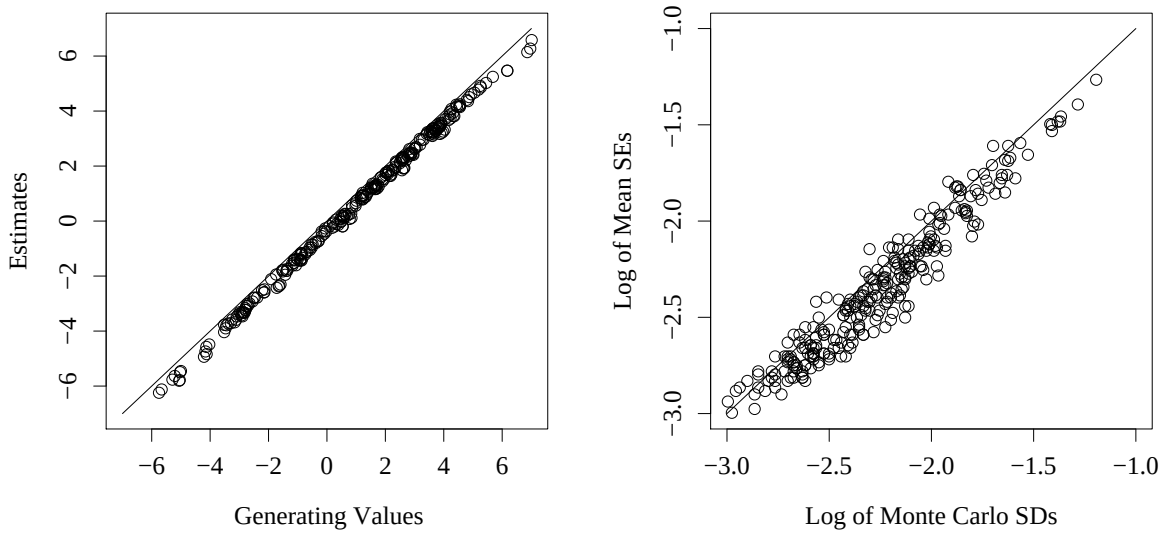

Figure 6.9: Bifactor Type Model: Intercepts

(in 847 cycles) per replication. The minimum cycle count was 719, and the maximum 968. The minimum run time was 803 seconds, and the maximum was 1878 seconds.

The left panel in Figure 6.9 compares the average of the point estimates of item intercepts against the generating values. The panel on the right plots the log of the average of the estimated standard errors against the log of the Monte Carlo standard deviations of the intercept estimates. The intercept point estimates show a slight downward bias, and so are the standard error estimates. Figure 6.10 shows a similar set of plots for the primary dimension item slopes. These slopes are slightly upwardly biased, but their standard errors are still underestimated. Surprisingly, the slopes for the specific dimensions are recovered very well, including their standard error estimates (see Figure 6.11). Note that this model contains 367 parameters, so even with a sample size of 2000, the ratio of respondents to parameters is still not too large. Further simulation work is needed to gather conclusive evidence about the quality of parameter estimates, but the current results show clear promise for MH-RM.

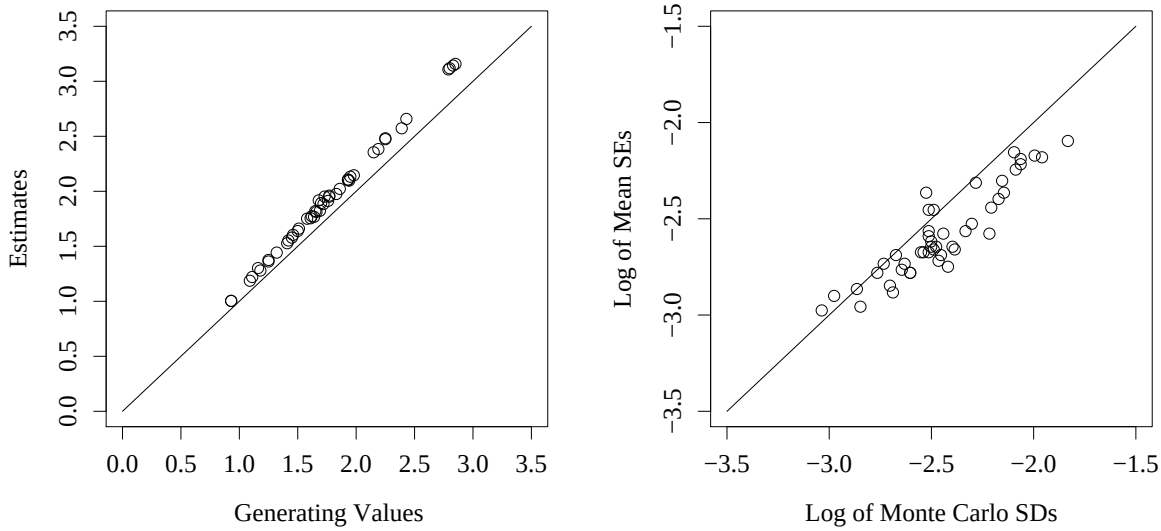

Figure 6.10: Bifactor Type Model: Slopes for the Primary Dimension

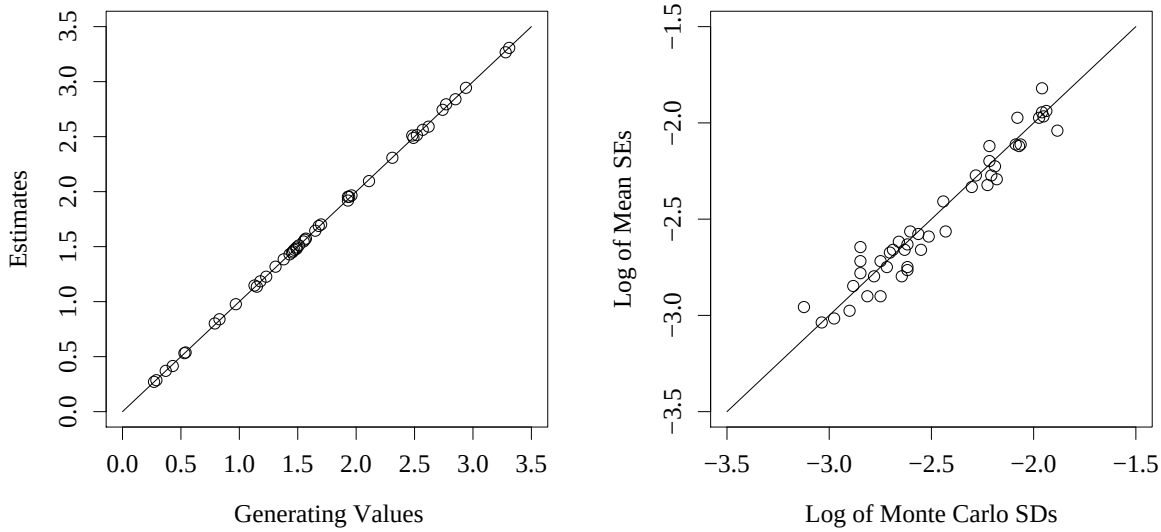

Figure 6.11: Bifactor Type Model: Slopes for Specific Dimensions

Table 6.1: Timing the MH-RM for Unidimensional IRT Simulation

N		Mean	*SD*	Min	Max
200	# cycles	1625	301	925	2761
	# seconds	352	73	128	654
1000	# cycles	515	69	335	710
	# seconds	530	97	234	783
3000	# cycles	257	37	122	337
	# seconds	708	230	391	1186

Table 6.2: Unidimensional IRT Model ($N = 200$)

	θ	$E(\hat{\theta})$	$E(\hat{\theta}-\theta)$	$SD(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$	$SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$
α_1	-0.79	-0.79	0.00	0.16	0.16	0.03	0.98
β_1	0.53	0.53	0.00	0.22	0.21	-0.08	0.99
α_2	1.02	1.04	0.02	0.20	0.19	0.02	1.00
β_2	0.92	0.93	0.01	0.26	0.27	-0.08	0.95
α_3	-1.15	-1.20	-0.05	0.23	0.21	-0.11	0.97
β_3	1.03	1.09	0.06	0.36	0.30	-0.02	1.00
α_4	2.36	2.56	0.20	0.69	0.61	-0.10	1.08
β_4	1.79	2.02	0.23	0.81	0.72	-0.10	1.08
α_5	-0.33	-0.33	-0.01	0.22	0.20	0.02	1.05
β_5	1.66	1.75	0.10	0.56	0.48	-0.14	0.99
α_6	-0.26	-0.26	0.00	0.17	0.16	0.01	1.05
β_6	0.83	0.87	0.04	0.25	0.24	0.03	0.98
α_7	-2.29	-2.50	-0.21	0.54	0.42	-0.16	0.99
β_7	1.15	1.35	0.19	0.57	0.45	0.14	0.96
α_8	1.22	1.29	0.07	0.31	0.22	0.13	1.05
β_8	0.95	1.04	0.09	0.38	0.30	0.09	0.96
α_9	-0.09	-0.09	0.00	0.17	0.16	-0.02	1.03
β_9	0.90	0.94	0.04	0.29	0.25	-0.03	1.08
α_{10}	-0.70	-0.71	-0.02	0.19	0.17	-0.04	1.11
β_{10}	0.69	0.71	0.02	0.27	0.23	-0.06	1.09

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E(\hat{\theta}-\theta)$ = absolute bias; $SD(\hat{\theta})$ = *SD* of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated *SE*s; $E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = mean of univariate z statistics using estimated *SE*s; $SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = *SD* of univariate z statistics using estimated *SE*s.

Table 6.3: Unidimensional IRT Model ($N = 1000$)

	θ	$E(\hat{\theta})$	$E(\hat{\theta} - \theta)$	$SD(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$	$SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$
α_1	-0.79	-0.79	0.00	0.08	0.07	0.05	1.05
β_1	0.53	0.53	0.01	0.09	0.10	0.04	0.90
α_2	1.02	1.03	0.01	0.10	0.08	0.05	1.13
β_2	0.92	0.94	0.02	0.12	0.12	0.07	0.98
α_3	-1.15	-1.15	0.00	0.10	0.09	0.02	1.08
β_3	1.03	1.04	0.01	0.14	0.13	0.00	1.10
α_4	2.36	2.39	0.03	0.21	0.20	-0.01	1.03
β_4	1.79	1.82	0.03	0.25	0.24	-0.04	1.04
α_5	-0.33	-0.33	0.00	0.10	0.09	-0.03	1.16
β_5	1.66	1.67	0.02	0.19	0.19	-0.04	1.01
α_6	-0.26	-0.26	0.00	0.06	0.07	0.02	0.91
β_6	0.83	0.84	0.01	0.10	0.11	0.05	0.97
α_7	-2.29	-2.31	-0.02	0.16	0.15	-0.02	1.03
β_7	1.15	1.17	0.02	0.18	0.17	-0.01	1.06
α_8	1.22	1.22	0.00	0.08	0.09	-0.02	0.92
β_8	0.95	0.96	0.01	0.12	0.12	0.01	0.98
α_9	-0.09	-0.10	-0.01	0.07	0.07	-0.13	1.06
β_9	0.90	0.90	-0.01	0.10	0.11	-0.11	0.90
α_{10}	-0.70	-0.70	0.00	0.07	0.07	-0.03	1.01
β_{10}	0.69	0.68	-0.01	0.10	0.10	-0.14	0.96

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E(\hat{\theta} - \theta)$ = absolute bias; $SD(\hat{\theta})$ = *SD* of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated *SE*s; $E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = mean of univariate z statistics using estimated *SE*s; $SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = *SD* of univariate z statistics using estimated *SE*s.

Table 6.4: Unidimensional IRT Model ($N = 3000$)

	θ	$E(\hat{\theta})$	$E(\hat{\theta}-\theta)$	$SD(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$	$SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$
α_1	-0.79	-0.79	0.00	0.04	0.04	0.08	1.05
β_1	0.53	0.53	0.00	0.05	0.05	0.03	0.88
α_2	1.02	1.03	0.00	0.05	0.05	0.06	0.95
β_2	0.92	0.92	0.00	0.06	0.07	0.00	0.89
α_3	-1.15	-1.14	0.00	0.06	0.05	0.05	1.18
β_3	1.03	1.02	-0.01	0.08	0.07	-0.20	1.01
α_4	2.36	2.36	0.00	0.10	0.11	-0.04	0.90
β_4	1.79	1.79	0.00	0.13	0.14	-0.07	0.99
α_5	-0.33	-0.32	0.00	0.06	0.05	0.06	1.10
β_5	1.66	1.65	-0.01	0.11	0.11	-0.20	1.04
α_6	-0.26	-0.25	0.01	0.04	0.04	0.16	0.95
β_6	0.83	0.82	0.00	0.05	0.06	-0.06	0.85
α_7	-2.29	-2.30	-0.01	0.09	0.09	-0.02	1.05
β_7	1.15	1.15	0.00	0.09	0.10	-0.08	0.93
α_8	1.22	1.22	0.00	0.06	0.05	0.02	1.09
β_8	0.95	0.96	0.01	0.07	0.07	0.06	1.01
α_9	-0.09	-0.09	0.00	0.04	0.04	0.04	1.00
β_9	0.90	0.90	-0.01	0.06	0.06	-0.12	1.00
α_{10}	-0.70	-0.70	0.00	0.04	0.04	-0.09	0.95
β_{10}	0.69	0.68	-0.01	0.06	0.06	-0.20	1.04

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E(\hat{\theta}-\theta)$ = absolute bias; $SD(\hat{\theta})$ = *SD* of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated *SE*s; $E\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = mean of univariate z statistics using estimated *SE*s; $SD\left\{\frac{\hat{\theta}-\theta}{se(\hat{\theta})}\right\}$ = *SD* of univariate z statistics using estimated *SE*s.

Table 6.5: Multidimensional Nominal Model: Slopes

Item	β	$E(\hat{\beta})$	$E(\hat{\beta}-\beta)$	$SD(\hat{\beta})$	$E\{se(\hat{\beta})\}$	$\frac{E\{se(\hat{\beta})\}}{SD(\hat{\beta})}$
1	1.00	1.04	0.04	0.12	0.10	0.85
2	1.50	1.56	0.06	0.17	0.16	0.96
3	1.20	1.29	0.09	0.13	0.13	1.00
4	0.90	0.95	0.05	0.10	0.10	0.96
5	1.00	1.05	0.05	0.12	0.11	0.89
6	1.50	1.59	0.09	0.19	0.17	0.94
7	1.20	1.25	0.05	0.15	0.13	0.87
8	0.90	0.92	0.02	0.11	0.10	0.90
9	1.00	1.07	0.07	0.12	0.11	0.88
10	1.50	1.59	0.09	0.19	0.18	0.94
11	1.20	1.28	0.08	0.14	0.13	0.92
12	0.90	0.96	0.06	0.11	0.10	0.93
13	1.00	1.02	0.02	0.10	0.11	1.02
14	1.50	1.58	0.08	0.18	0.18	1.02
15	1.20	1.25	0.05	0.14	0.13	0.95
16	0.90	0.94	0.04	0.12	0.10	0.79
17	1.00	1.05	0.05	0.12	0.10	0.86
18	1.50	1.59	0.09	0.17	0.17	1.02
19	1.20	1.28	0.08	0.13	0.13	0.96
20	0.90	0.94	0.04	0.10	0.10	0.97

Note. β = Generating item slope; $E(\hat{\beta})$ = mean of point estimates; $E(\hat{\beta}-\beta)$ = absolute bias; $SD(\hat{\beta})$ = *SD* of point estimates; $E\{se(\hat{\beta})\}$ = mean of estimated *SE*s.

Table 6.6: Multidimensional Nominal Model: Factor Correlations

	ϕ_{ij}	$E(\hat{\phi}_{ij})$	$E(\hat{\phi}_{ij} - \phi_{ij})$	$SD(\hat{\phi}_{ij})$	$E\{se(\hat{\phi}_{ij})\}$	$\frac{E\{se(\hat{\phi}_{ij})\}}{SD(\hat{\phi}_{ij})}$
ϕ_{21}	0.70	0.71	0.01	0.03	0.04	1.10
ϕ_{31}	0.60	0.61	0.01	0.04	0.04	1.03
ϕ_{32}	0.70	0.71	0.01	0.03	0.04	1.14
ϕ_{41}	0.60	0.61	0.01	0.04	0.04	0.98
ϕ_{42}	0.50	0.51	0.01	0.04	0.05	1.05
ϕ_{43}	0.50	0.51	0.01	0.04	0.05	1.09
ϕ_{51}	0.80	0.81	0.01	0.03	0.03	1.37
ϕ_{52}	0.60	0.62	0.02	0.04	0.04	1.01
ϕ_{53}	0.60	0.61	0.01	0.03	0.04	1.22
ϕ_{54}	0.50	0.51	0.01	0.05	0.05	1.02

Note. ϕ_{ij} = Generating correlation; $E(\hat{\phi}_{ij})$ = mean of point estimates; $E(\hat{\phi}_{ij} - \phi_{ij})$ = absolute bias; $SD(\hat{\phi}_{ij})$ = SD of point estimates; $E\{se(\hat{\phi}_{ij})\}$ = mean of estimated SEs.

Table 6.7: Multidimensional Nominal Model: $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ Estimate and Bias

True Values	$\alpha_1 = 1$	$\alpha_2 = 2$	$\alpha_3 = 2$	$\gamma_1 = 1$	$\gamma_2 = 1$
Item 1	(0.98,-.02)	(1.99,-.01)	(2.02, .02)	(1.00, .00)	(.51, .01)
Item 2	(0.97,-.03)	(1.98,-.02)	(1.98,-.02)	(0.99,-.01)	(.52, .02)
Item 3	(1.04, .04)	(1.97,-.03)	(2.01, .01)	(1.03, .03)	(.52, .02)
Item 4	(1.03, .03)	(1.98,-.02)	(2.00, .00)	(1.01, .01)	(.53, .03)
Item 5	(1.02, .02)	(1.99,-.01)	(2.02, .02)	(1.03, .03)	(.55, .05)
Item 6	(0.98,-.02)	(1.99,-.01)	(2.00, .00)	(1.01, .01)	(.51, .01)
Item 7	(1.00, .00)	(1.97,-.03)	(2.02, .02)	(1.00, .00)	(.50, .00)
Item 8	(0.99,-.01)	(1.99,-.01)	(2.01, .01)	(1.00, .00)	(.51, .01)
Item 9	(1.04, .04)	(1.98,-.02)	(2.00, .00)	(1.03, .04)	(.58, .08)
Item 10	(0.98,-.02)	(1.99,-.01)	(1.99,-.01)	(1.01, .01)	(.52, .02)
Item 11	(1.02, .02)	(1.96,-.04)	(2.00,-.01)	(1.01, .01)	(.51, .01)
Item 12	(1.02, .02)	(1.95,-.05)	(1.96,-.04)	(1.03, .03)	(.57, .07)
Item 13	(0.98,-.02)	(2.01, .01)	(2.00, .00)	(1.00,-.01)	(.51, .01)
Item 14	(0.98,-.02)	(2.00,-.01)	(1.99,-.01)	(1.01, .01)	(.53, .03)
Item 15	(1.02, .02)	(1.99,-.01)	(2.03, .03)	(1.01, .01)	(.48,-.01)
Item 16	(1.00, .00)	(1.99,-.01)	(2.00, .00)	(1.01, .01)	(.53, .03)
Item 17	(0.99,-.01)	(1.96,-.04)	(1.99,-.01)	(1.01, .01)	(.53, .03)
Item 18	(0.97,-.03)	(1.98,-.02)	(2.00, .00)	(1.01, .01)	(.50, .00)
Item 19	(1.00, .00)	(1.97,-.03)	(2.00, .00)	(1.02, .02)	(.53, .03)
Item 20	(0.98,-.01)	(1.99,-.01)	(2.01, .01)	(1.00, .00)	(.50, .00)

Note. The entries in the parentheses are (mean of point estimates, bias), where bias is defined as the mean of point estimates minus true generating value.

Table 6.8: Multidimensional Nominal Model: $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ Standard Errors

Item	α_1	α_2	α_3	γ_1	γ_2
1	(.17, .18, 1.04)	(.14, .13, 0.89)	(.15, .15, 1.01)	(.10, .09, 0.90)	(.24, .22, 0.91)
2	(.19, .16, 0.88)	(.14, .12, 0.83)	(.12, .13, 1.09)	(.10, .10, 0.96)	(.23, .24, 1.03)
3	(.16, .17, 1.02)	(.11, .12, 1.03)	(.14, .13, 0.94)	(.10, .09, 0.95)	(.23, .23, 0.99)
4	(.16, .19, 1.18)	(.13, .13, 1.01)	(.14, .16, 1.09)	(.08, .09, 1.03)	(.20, .22, 1.08)
5	(.20, .19, 0.94)	(.14, .13, 0.90)	(.16, .15, 0.93)	(.11, .09, 0.85)	(.26, .23, 0.86)
6	(.16, .17, 1.04)	(.14, .13, 0.87)	(.13, .13, 1.02)	(.10, .10, 1.03)	(.27, .25, 0.90)
7	(.16, .17, 1.10)	(.13, .12, 0.98)	(.14, .14, 0.97)	(.10, .09, 0.90)	(.22, .23, 1.05)
8	(.23, .20, 0.87)	(.14, .14, 0.98)	(.17, .16, 0.98)	(.09, .09, 1.00)	(.23, .22, 0.94)
9	(.20, .18, 0.93)	(.13, .13, 0.97)	(.16, .15, 0.91)	(.10, .09, 0.90)	(.23, .23, 0.98)
10	(.18, .17, 0.97)	(.14, .13, 0.93)	(.13, .13, 1.02)	(.11, .11, 1.00)	(.27, .25, 0.93)
11	(.20, .17, 0.84)	(.13, .12, 0.91)	(.13, .14, 1.05)	(.11, .10, 0.84)	(.26, .23, 0.91)
12	(.19, .19, 1.01)	(.14, .13, 0.90)	(.16, .15, 0.98)	(.10, .09, 0.94)	(.20, .22, 1.13)
13	(.20, .19, 0.95)	(.15, .13, 0.90)	(.16, .15, 0.92)	(.09, .09, 1.03)	(.23, .22, 0.96)
14	(.19, .18, 0.95)	(.13, .14, 1.05)	(.14, .14, 0.99)	(.11, .11, 0.97)	(.25, .25, 1.02)
15	(.19, .18, 0.93)	(.12, .13, 1.02)	(.14, .14, 1.04)	(.10, .10, 0.99)	(.22, .23, 1.09)
16	(.22, .20, 0.92)	(.13, .14, 1.03)	(.18, .16, 0.88)	(.10, .09, 0.88)	(.23, .22, 0.97)
17	(.19, .18, 0.97)	(.13, .13, 0.95)	(.15, .15, 0.96)	(.10, .09, 0.89)	(.23, .22, 0.98)
18	(.17, .17, 0.98)	(.13, .13, 0.99)	(.14, .13, 0.96)	(.09, .10, 1.10)	(.25, .24, 0.98)
19	(.16, .17, 1.07)	(.14, .12, 0.89)	(.13, .14, 1.01)	(.10, .09, 0.94)	(.25, .23, 0.92)
20	(.19, .20, 1.04)	(.12, .13, 1.09)	(.17, .16, 0.93)	(.08, .09, 1.11)	(.20, .22, 1.04)

Note. The entries in the parentheses are (Monte Carlo *SD* of point estimates, mean of estimated standard errors, the ratio of the latter over the former).

Table 6.9: Generating Parameter Values for the Bifactor Type Model: Items 1–23

		Intercepts						Slopes	
Item		1	2	3	4	5	6	ξ_1	$\xi_2 - \xi_6$
	1	4.33	2.95	2.15	0.75	-0.91	-2.93	2.39	–
ξ_2	2	5.19	3.62	2.89	1.24	-0.31	-2.16	1.70	1.93
	3	4.16	2.65	1.62	0.64	-1.00	-2.91	1.58	1.45
	4	6.86	3.98	2.57	0.53	-1.72	-5.06	2.79	3.28
	5	6.17	3.87	2.62	0.78	-1.63	-4.21	2.83	2.85
	6	3.81	1.90	0.57	-0.58	-2.86	-5.29	1.73	2.49
ξ_3	7	3.54	2.03	1.17	0.21	-1.53	-3.52	1.09	1.38
	8	3.61	1.72	0.39	-0.92	-3.16	-5.75	2.15	2.74
	9	3.74	1.64	0.26	-0.74	-2.85	-5.03	1.95	2.52
	10	3.63	2.55	2.02	1.19	-0.89	-2.51	1.25	0.37
	11	4.87	3.86	3.00	1.66	-0.18	-2.79	1.61	1.55
ξ_4	12	4.45	3.25	2.13	1.15	-0.92	-3.24	1.41	1.49
	13	5.25	4.05	2.80	1.65	-0.49	-3.09	1.66	1.68
	14	4.53	3.00	2.09	1.17	-0.39	-2.55	1.69	0.54
	15	3.71	2.50	1.61	0.56	-1.13	-2.95	1.83	0.27
	16	3.83	2.34	1.29	0.18	-1.39	-3.47	1.94	0.83
ξ_5	17	3.84	2.59	1.51	0.51	-1.12	-2.98	1.45	0.97
	18	4.32	2.82	1.90	0.87	-0.59	-2.39	1.72	1.13
	19	4.83	2.87	1.66	0.49	-1.43	-3.52	2.25	1.93
	20	3.68	2.51	1.53	0.61	-1.03	-3.03	1.50	1.23
	21	4.26	3.18	2.36	1.22	-0.18	-1.74	0.93	0.29
ξ_6	22	3.91	2.85	2.01	0.99	-0.86	-2.65	1.46	1.46
	23	4.21	3.42	2.58	1.86	0.31	-1.49	1.11	1.15

Table 6.10: Generating Parameter Values for the Bifactor Type Model: Items 24–46

		Intercepts						Slopes	
Item		1	2	3	4	5	6	ξ_1	$\xi_6 - \xi_{10}$
ξ_6	24	4.22	3.31	2.33	1.42	-0.02	-2.14	1.51	2.11
	25	5.05	3.86	3.07	2.24	0.46	-1.36	1.65	2.31
	26	2.62	1.51	0.86	-0.08	-1.10	-2.22	1.25	1.43
ξ_7	27	4.55	3.19	2.26	0.98	-1.36	-4.03	1.93	1.96
	28	4.79	3.77	2.87	1.74	-0.38	-2.88	1.86	1.70
	29	2.38	1.84	1.01	0.09	-1.13	-2.72	1.18	0.79
	30	3.60	2.60	1.57	0.56	-1.20	-3.38	1.64	1.49
	31	4.56	3.50	2.57	1.37	-0.73	-3.29	1.76	1.31
ξ_8	32	7.01	5.68	4.16	2.21	-1.06	-4.16	2.25	2.48
	33	5.45	3.88	2.78	1.22	-1.18	-4.13	1.93	1.56
	34	4.55	3.54	2.60	1.23	-0.53	-2.76	1.42	1.18
	35	4.43	3.44	2.82	1.56	-0.06	-1.91	0.93	0.53
	36	4.36	2.96	2.18	0.83	-0.87	-2.89	2.43	1.94
ξ_9	37	5.27	3.66	2.90	1.33	-0.28	-2.14	1.77	1.51
	38	4.23	2.70	1.66	0.68	-0.90	-2.82	1.68	3.31
	39	6.96	4.03	2.63	0.54	-1.65	-5.04	2.80	2.94
	40	6.18	3.93	2.65	0.84	-1.59	-4.12	2.85	2.57
	41	3.88	1.96	0.58	-0.52	-2.79	-5.21	1.77	1.47
ξ_{10}	42	3.60	2.11	1.23	0.21	-1.51	-3.44	1.16	2.77
	43	3.66	1.77	0.48	-0.84	-3.11	-5.66	2.19	2.62
	44	3.82	1.69	0.33	-0.68	-2.79	-5.00	1.98	0.43
	45	3.68	2.60	2.08	1.24	-0.87	-2.47	1.32	1.65
	46	4.94	3.91	3.02	1.69	-0.09	-2.71	1.62	1.57

CHAPTER 7

Discussions and Future Directions

7.1 Discussions

The present research extends work in Cai (2006) on exploratory item factor analysis using the MH-RM algorithm to the general confirmatory setting within the modelling context of a flexible nonlinear latent structure model. The latent structure model synthesizes many existing psychometric models and provides a rich framework for the development of new models. The MH-RM algorithm combines elements of MCMC with Stochastic Approximation and makes important connections between the missing data formulation, the EM algorithm, and maximum marginal likelihood estimation. A C++ program has been written that implements both the model and the algorithm, permitting general confirmatory analysis.

Theoretical results on the MH-RM algorithm have been extended, and details on the practical implementation of the MH-RM algorithm are discussed. Issues that have been encountered and resolved include: 1) algorithmic acceleration with multistage and adaptive gain constants, 2) estimation under user-defined linear restrictions using a parameter space segmentation technique, 3) exploratory studies on an alternative Gibbs sampler for the normal ogive parametrization.

A variety of real and simulated data sets are used to compare MH-RM with currently available software packages for latent variable modelling. The problems range from small to large, and from simple to complex. Without exception, MH-RM

compared favorably against alternatives in terms of efficiency, while maintaining the same degree of accuracy.

Some Monte Carlo sampling experiments were conducted to examine different aspects of this new estimation algorithm, including parameter recovery, standard error estimation, and CPU time. While firm conclusions cannot be drawn solely on the current limited set of simulations, the results are promising. It is safe to conclude that MH-RM may become the first self-adaptive algorithm for high-dimensional maximum likelihood latent variable modelling that is not subject to the "curse of dimensionality."

7.2 Future Directions

There are two avenues for extensions of the present research. The first, and most obvious one, is to extend the MH-RM algorithm to modelling frameworks that are not covered by the current latent structure model. For instance, theoretical and empirical results on MH-RM suggest that it may also work well for generalized and nonlinear mixed effects models. Another possibility is to investigate the applicability of MH-RM for multilevel latent structure modelling, where the respondents are themselves nested within some larger unit or cluster. This includes multilevel structural equation modelling as a special case. Yet another possibility is mixture modelling. MH-RM resembles, in some ways, the SAEM algorithm, which was originally proposed to solve finite mixture problems. Therefore it is reasonable to expect MH-RM to be applicable in that context as well. In general, MH-RM is most likely to excel in dealing with high-dimensional problems where both the number of latent variables and the number of observed variables are large.

The second avenue is extending the algorithm itself. Theory on the Robbins-Monro method suggests that further improvement of MH-RM's efficiency and rate of convergence is possible. For example, Polyak and Juditski (1992) showed that if the gain constants converge to zero at a rate less than $O(N^{-1})$, the "off-line" averaged

estimate

$$\tilde{\boldsymbol{\omega}}^{(k)} = \frac{1}{K} \sum_{j=k-K(k)+1}^{k} \boldsymbol{\omega}^{(j)}$$

converges at an optimal rate, where $K(k)$ is a "window of averaging" that is allowed to depend on k. Kushner and Yin (1997) give detailed analysis and justification for this procedure based on the time-scale separation argument. The effect of Polyak averaging on the finite-time behavior of MH-RM deserves further study.

In general, more simulation work using MH-RM is necessary. A feature of the present simulations is that the data generating model and the fitted model are identical, which means that there is no model error (in the sense of MacCallum, 2003). This is strikingly unrealistic. Better external validity can be achieved by incorporating model error into the simulations and only then can one be somewhat optimistic about the usefulness of MH-RM for latent structural modelling in real psychological research.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269.

Aptech Systems, Inc. (2003). GAUSS (Version 6.08) [Computer software]. Maple Valley, WA: Author

Arminger, G., & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, *63*, 271–300.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.

Benveniste, A., Métivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*. Berlin: Springer-Verlag.

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive asessment. *Quality of Life Research*, *16*, 95–108.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

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

Bock, R. D., & Bargmann, R. (1966). Analysis of covariance structures. *Psychometrika*, *46*, 443–449.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4 user's guide*. Chicago, IL: SSI International.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179–197.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bolt, D. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (p. 27-71). Mahwah, NJ: Earlbaum.

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.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Cai, L. (2006). *Full-information item factor analysis by Markov chain Monte Carlo stochastic approximation*. Unpublished master's thesis, Department of Statistics, University of North Carolina at Chapel Hill.

Cai, L. (in press). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.

Cappé, O., Douc, R., Moulines, E., & Robert, C. (2002). On the convergence of the Monte Carlo maximum likelihood method for latent variable models. *Scandinavian Journal of Statistics*, *29*, 615–635.

Celeux, G., Chauveau, D., & Diebolt, J. (1995). *On stochastic versions of the EM algorithm* (Tech. Rep. No. 2514). The French National Institute for Research in Computer Science and Control.

Celeux, G., & Diebolt, J. (1991). *A stochastic approximation type EM algorithm for the mixture problem* (Tech. Rep. No. 1383). The French National Institute for Research in Computer Science and Control.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*, 347-361.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: Historical developments and new directions*. Mahwah, NJ: Laurence Erlbaum Associates.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, *27*, 94–128.

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

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

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society – Series B*, *62*, 355–366.

Edwards, M. C. (2005). *A Markov chain Monte Carlo approach to confirmatory item factor analysis*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

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

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Fox, J.-P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology*, *56*, 65–81.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*, 145-172.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267-269.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In *Computing science and statistics: Proceedings of the twenty-third symposium on the interface* (p. 571-578.).

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society – Series B*, *54*, 657–699.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423–436.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, *31*, 1208–1211.

Gu, M. G., & Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *The Proceedings of the National Academy of Sciences*, *95*, 7270–7274.

Gu, M. G., Sun, L., & Huang, C. (2004). A universal procedure for parametric frailty models. *Journal of Statistical Computation and Simulation*, *74*, 1–13.

Gu, M. G., & Zhu, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society – Series B*, *63*, 339–355.

Gueorguieva, R. V., & Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, *96*, 1102–1112.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–338.

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Jank, W. (2004). Quasi-Monte Carlo sampling to improve the effciency of Monte Carlo EM. *Computational Statistics and Data Analysis*, *48*, 685–701.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*, 239–251.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL user's guide*. Chicago, IL: Scientific Software International.

Kass, R., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *84*, 717–726.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201-210.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, *29*, 41-59.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457-474.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response theory models. *Multivariate Behavioral Research*, *26*, 457-477.

Kuhn, E., & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, *49*, 1020–1038.

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, *63*, 143–158.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. New York: Springer.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society – Series B*, *57*, 425-437.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Wiley.

Lee, S.-Y., Poon, W. Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*, 89–105.

Lee, S.-Y., Poon, W. Y., & Bentler, P. M. (1995a). A three stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, *55*, 45–51.

Lee, S.-Y., Poon, W. Y., & Bentler, P. M. (1995b). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–358.

Lee, S.-Y., Song, X.-Y., & Lee, J. C. K. (2003). Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics*, *28*, 111–134.

Lee, S.-Y., & Zhu, H.-T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *53*, 209–232.

Lehman, A. F. (1988). A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning*, *11*, 51–62.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Liu, C., Rubin, D. B., & Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85*, 755–770.

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*, 226–233.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113-139.

Makowski, D., & Lavielle, M. (2006). Using SAEM to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*, 45–60.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*, 275-300.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, *15*.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.

Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*, 1254–1267.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3–31.

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, *24*, 211–223.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, *56*, 337-357.

Moustaki, I. (2007). Factor analysis and latent structure of categorical and metric data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions.* Mahwah, NJ: Laurence Erlbaum Associates.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391-411.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.

Muthén, B. (1978). Contributions of factor analysis to dichotomous variables. *Psychometrika*, *43*, 551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén & Muthén. (2007). Mplus (Version 4.21) [Computer software]. Los Angeles, CA: Author.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society – Series C*, *31*, 214–225.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and application. In L. M. Lecam, J. Neyman, & E. L. Scott (Eds.), *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability* (pp. 697–715). Berkeley, CA: University of California Press.

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

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Polyak, B. T., & Juditski, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, *30*, 838-855.

Qian, Z., & Shapiro, A. (2006). Simulation-based approach to estimation of latent variable models. *Computational Statistics and Data Analysis*, *51*, 1243–1259.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual. *U.C. Berkeley Division of Biostatistics Working Paper Series*, *160*.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22–31.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2003). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*, 95–101.

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

Rousseeuw, P. J., & Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, *48*, 276–279.

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

SAS Institute Inc. (2004). *SAS/STAT® 9.1 users guide*. Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Segall, D. O. (1998). IFACT computer program Version 1.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [Computer software]. Seaside, CA: Defense Manpower Data Center.

Shi, J.-Q., & Lee, S.-Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *51*, 233–252.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Song, X.-Y., & Lee, S.-Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology*, *54*, 237–263.

Song, X.-Y., & Lee, S.-Y. (2003). Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivariate Behavioral Research*, *38*, 57–79.

Song, X.-Y., & Lee, S.-Y. (2005). A multivariate probit latent variable model for analyzing dichotomous responses. *Statistica Sinica*, *15*, 645–664.

Spall, J. C. (1999). Stochastic optimization: Stochastic approximation and simulated annealing. In J. G. Webster (Ed.), *Encyclopedia of electrical and electronics engineering* (Vol. 20, p. 529-542). New York: Wiley.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

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

Thissen, D. (2003). *MULTILOG 7 user's guide*. Chicago, IL: SSI International.

Thissen, D., Cai, L., & Bock, R. D. (2006). A new parameterization of the nominal item response model. *Paper presented at the 71st Annual Meeting of the Psychometric Society, Montreal, Canada*.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, *16*, 109–116.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309–322.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: The University of Chicago Press.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, *22*, 1701–1762.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society – Series B*, *46*, 257-267.

von Davier, M., & Sinharray, S. (2004). Application of the stochastic EM method to latent regression models. *ETS Research Report*, *RR-04-34*.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–202.

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.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Zhu, H.-T., & Lee, S.-Y. (2002). Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method. *Statistics and Computing*, *12*, 175–183.