ROBUST LIKELIHOOD INFERENCE FOR LATENT VARIABLE MODELS USING MARKOV CHAIN MONTE CARLO

Michael R. Kosorok

department of statistics

university of wisconsin at madison

Jeffrey A. Douglas

department of statistics

university of illinois

Jimmy de la Torre

department of psychology

university of illinois

March 5, 2003

This research was partially funded by U.S. National Cancer Institute grant CA75142.

Correspondence should be sent to Jeff Douglas, 101 Illini Hall, 725 S. Wright St, Champaign, IL 61820. E-Mail: jeffdoug@stat.uiuc.edu.

Abstract

Latent variable models are routinely used in psychometrics to characterize dependence in hierarchical, longitudinal, or multivariate data. Parametric models in which the complete data likelihood can be specified provide useful and convenient approaches to handling the complexities which arise. However, likelihood inference with latent variables can be quite challenging because of the intractability of the associated marginal likelihood. This problem can be further exacerbated when the number of parameters involved is large and when robust estimation of the variance is desired. We propose using Markov chain Monte Carlo to obtain the posterior mean estimator and the Fisher information matrix, both computed under the presumed likelihood, and using Monte Carlo quadrature to obtain the remaining components of the robust asymptotic variance. Evaluation of the marginal likelihood is not needed. We demonstrate asymptotic normality of the mean estimator and consistency of the robust variance estimator when the number of data clusters is large but the likelihood may be incorrectly specified. An analysis of mulivariate binary survey data is given as an example.

Key words: Convergence of posterior distribution; Maximum likelihood; Metropolis-Hastings algorithm; Missing data; Random effects; Latent variables.

1. Introduction

A standard approach to modeling in the presence of latent variables has been to use parametric models in which the complete data likelihood can be specified. However, likelihood inference often requires the evaluation of numerically intractable marginal likelihood integrals. In contrast, Bayesian analysis based on examination of the posterior distribution through the Metropolis-Hastings algorithm (MHA) only requires evaluation of the unconditional likelihood and the prior distribution (see Besag and Green, 1993; Gilks et al, 1993; and Smith and Roberts, 1993). For this reason we have seen the MHA and other Markov chain Monte Carlo techniques become increasingly popular for fitting complicated latent variable models in psychometrics. MCMC algorithms have been developed for factor analysis (Ansari and Jedidi, 200; Lee and Zhu, 2002; and Meng and Schilling, 1996), item response theory (Albert, 1992; and Patz and Junker, 1999a and 1999b), and latent class modeling (Garret and Zeger, 2000; Hoijting and Molenaar, 1997; and Lenk and DeSarbo, 2000). Because the output of an MCMC analysis leads directly to approximation of the mean of the marginal posterior distribution, rather than approximation of the posterior mode or maximum likelihood estimator, MCMC is generally considered a method for Bayesian analysis. However, when the likelihood model is correct, and when the number of data clusters is large, it is well known that the posterior distribution and the distribution of the maximum likelihood estimator coincide (see, for example, Ghosal, Ghosh, and Samanta, 1995; and Johnson, 1967 and 1970).

When fitting complicated parametric models, goodness-of-fit is always an issue. In many cases formal diagnostics may reveal that the model being used does not fit, but first-order residual analyses or scientific judgement may indicate that it is still capable of providing meaningful parameter estimates. Latent variable models in psychometrics are rarely viewed as precisely correct. In fact, doing so would require reification of the latent variable, which often represents a psychological construct. However, subject to fitting adequately, the utility of these models is appreciated when highly multivariate data can be summarized by a much smaller set of interpretable latent variables. As part of an extensive discussion of the utility of misspecified models in psychometrics given in his presidential address, Jan de Leeuw (1987) points out that indices based on the AIC or cross validation seem more appropriate than significance tests of model fit.

For situations in which we have found a useful model but acknowledge that it only approximates the true distribution, Huber (1967) demonstrated that likelihood inference can still be performed, provided an adjustment is made to the variance estimator so that it is consistent for $I_0^{-1} I_0^* I_0^{-1}$, where I_0 is the expectation of the negative second derivative of the misspecified log-likelihood and I_0^* is the expectation of the outer product of the cluster-level score component. The likelihood in this context has been referred to as a "working likelihood" (see Zeger, Liang, and Self, 1985)—because it is used for inference but not necessarily held to be the true probability distribution—and its maximizer is consistent for the parameter which locally minimizes the Kullback-Leibler distance between the presumed likelihood and the true probability distribution. This approach is particularly useful when the potentially misspecified model addresses the questions of interest but it is unclear how to embed the chosen model into a richer class for further improving goodness-of-fit. When using likelihood inference it is rarely absolutely clear that the given likelihood is correct, and robust likelihood inference is therefore just likelihood inference with the acknowledged necessity of computing a variance estimator which is robust against possible misspecification of the likelihood.

Hence the goal of this paper is to describe a method for conducting frequency based robust likelihood inference with latent variables which does not require integration of the conditional likelihood. This method utilizes MCMC to first obtain both the posterior mean estimator of the likelihood maximizer and a consistent estimator of the Fisher information matrix I_0 and, second, uses Monte Carlo quadrature to obtain a consistent estimator for I_0^*. In section 2, we present a formal basis for determining the meaningfulness of a possibly misspecified working likelihood. In section 3, we describe the setting for which the proposed methods are to be applied. We detail the proposed algorithm and provide large sample theory in section 4. Section 5 presents an analysis of real data using an item response model. The paper closes with a discussion in section 6.

2. Neighborhoods of Misspecified Likelihoods

Let $\mathcal{P}_*$ be a class of distributions which constitutes the working model for the data at hand. Typically, these models are indexed by some parameter set. Let P_0 be the true distribution of the data, and let $\mathcal{P} \supset \mathcal{P}_*$ be a sufficiently large model so that $P_0 \in \mathcal{P}$. We assume that the distributions in $\mathcal{P}$ have a dominating measure ρ. We say that the working model $\mathcal{P}_*$ is misspecified if $P_0 \notin \mathcal{P}_*$. For each $P \in \mathcal{P}$, let $P^* = \arg\max_{Q \in \mathcal{P}_*} \int \log(dQ/d\rho)\, dP$. We assume P^* is unique. Then P^* is the closest element of $\mathcal{P}_*$ to P in terms of the Kullback-Leibler discrepancy. Let $\phi : \mathcal{P} \mapsto \mathcal{B}$ be the function on the model which extracts the parameter of interest, where $\mathcal{B}$ is the parameter space. We say that the working model is meaningful if $e\left[\phi(P_0^*), \phi(P_0)\right] \leq \delta$, where $e : \mathcal{B} \times \mathcal{B} \mapsto [0, \infty)$ is a discrepancy measure and δ is some tolerance threshold.

In the setting we consider, $\mathcal{P}_*$ is a class of distributions indexed by a parameter vector β of constant dimension. Thus, the penalty term of the AIC (Akaike, 1987) is constant, and the only remaining issue is finding the particular value of β that maximizes fidelity to the data, as measured by the working likelihood function. A preliminary step might be using the AIC to arrive at the class of distributions $\mathcal{P}_*$, when several nonnested classes are under consideration. Akaike (1987) considers the AIC for use in factor analysis, and Bozdogan (1987) studies the relationship between Kullback-Liebler distance and the AIC, and proposes extensions of the AIC which penalize overparametrization more heavily.

2.1. Example 1: Confirmatory Factor Analysis

A familiar special case of the framework for defining acceptable neighborhoods of misspecified models given above is seen in covariance structure analysis, where goodness-of-fit indices are so commonly used. Confirmatory factor analysis is often applied to investigate the goodness-of-fit of an independent clusters factor pattern, for the purpose of validating separate unidimensional subscores on either educational or psychological assessments. Let P_0 denote the true probability distribution of a response vector Y with population covariance given by $\phi(P_0) = \Sigma = \Lambda_0\Lambda_0' + \Psi_0$, where Λ_0 and Ψ can be viewed as the factor pattern and specific variance matrix of an orthogonal factor model. In particular, suppose that Σ has diagonal elements equal to 1 and

$$\Lambda_0 = \begin{pmatrix} 0.85 & 0.15 \\ 0.85 & 0.15 \\ 0.85 & 0.15 \\ 0.15 & 0.85 \\ 0.15 & 0.85 \\ 0.15 & 0.85 \end{pmatrix}.$$

In a typical application, the first 3 elements of Y might be designed to measure one psychological construct, with the final three elements intended to measure another. Restricting the factors to be uncorrelated, and using an independent clusters structure we find the covariance matrix $\phi(P_0^*) = \Lambda_*\Lambda_*' + \Psi_*$ that minimizes the Kullback-Liebler discrepancy by fitting the covariance matrix Σ_0 with the maximum likelihood criterion under the constraints of an independent clusters structure with uncorrelated factors. This results in the factor pattern

$$\Lambda_* = \begin{pmatrix} 0.863 & 0.0 \\ 0.863 & 0.0 \\ 0.863 & 0.0 \\ 0.0 & 0.863 \\ 0.0 & 0.863 \\ 0.0 & 0.863 \end{pmatrix}.$$

For the discrepancy function $e\,[\phi(P_0^*), \phi(P_0)]$ we choose the well-known Goodness-of-Fit Index (GFI),

$$e\,[\phi(P_0^*), \phi(P_0)] = 1 - GFI = \frac{tr\,[(\Sigma_*^{-1}\Sigma_0 - I)^2]}{tr\,[(\Sigma_*\Sigma_0)^2]}.$$

In this example we have $e\,[\phi(P_0^*), \phi(P_0)] = 0.0305$ or equivalently $GFI = 0.9695$. Thus, the restricted model is obviously misspecified, but the GFI indicates that the independent clusters structure results in an adequate fit, and would be seen as a useful reduction of the data.

2.2. *Example 2: A Mixed Linear Model*

Suppose we wish to estimate the parameters of a simple linear regression model for clusters consisting of m observations in each cluster, where each observation has real valued outcome Y_j and predictor X_j measurements, $j = 1 \ldots m$. We will use the following working random intercept model for $Y_1 \ldots Y_m$ given $X \equiv (X_1 \ldots X_m)'$ and θ:
$Y_j = a + bX_j + \gamma\theta + \epsilon_j$, where $\theta \sim N(0,1)$ and $\epsilon_j \sim N(0, \sigma^2)$ are independent and identically distributed, $j = 1 \ldots m$, $\gamma \geq 0$, and $\sigma \geq 0$. We also assume that X_j, $j = 1 \ldots m$, are i.i.d. $N(0, \xi^2)$, and denote the resulting working model $\mathcal{P}_*$.

Now consider the following enlarged model $\mathcal{P}$: $Y_1, \ldots, Y_m$ given $X_1, \ldots, X_m$ and θ has the form $Y_j = a + UbX_j + \gamma\theta + \epsilon_j$, where U is a mean 1 variance τ^2 random variable, and the remaining aspects of the model are the same as $\mathcal{P}_*$. Hence, $\mathcal{P}_*$ is the special case of $\mathcal{P}$

where $\tau = 0$. Assume the parameters of interest are $\phi(P) = (a, b, \gamma)$, and that the true distribution of the data is $P_0 \in \mathcal{P}$ with $P_0 \notin \mathcal{P}_*$. Also let e be the uniform distance and set $\delta = 0$. We have the following result, the proof of which is given in the appendix:

Proposition 1. In this situation, $\phi(P_0^*) = \phi(P_0)$. However, I_0 and I_0^* can be very different. Specifically, let $\hat{b}$ denote the maximum working likelihood estimator for b_*, and let ζ_* denote the naive limiting variance of $\sqrt{n}(\hat{b} - b_*)$ based on I_0^{-1}. The correct variance based on $I_0^{-1} I_0^* I_0^{-1}$ has the form $\zeta_*(1 + c_m m)$, where $c_m \geq 1/2$ for all $m \geq 2$.

Hence the magnitude of the bias in the variance can be made arbitrarily large as the cluster size increases.

2.3. Assessing Usefulness

As we have shown, choosing $\delta = 0$ is possible, but other values of δ can also be useful, as in the GFI. The mixed linear model example also demonstrates that whether or not a working model is meaningful is unrelated to whether or not $I_0 = I_0^*$ for Euclidean parameters. Hence testing whether $I_0 = I_0^*$ will say nothing about the usefulness of the working likelihood. On the other hand, such a test could potentially help determine whether a robust estimator of the variance is needed. While it is not difficult to extend the approach of White (1982) to develop such a test in the latent variables setting, we have found from our own experience that this test has too little power to be generally helpful. Another point is that the chosen examples are analytically tractable so that the usefulness of the parameters can be directly assessed. For most applications, it will generally be very difficult—if not impossible—to analytically evaluate the parameters in this way. Thus, at some level, deciding whether a working model is useful is a matter of scientific judgement.

3. The Data Setting and Notation

We assume hereafter that the data consist of n independent and identically distributed clusters. In psychometrics, a cluster would generally be comprised of the multiple responses from a fixed person. Let β be the unknown fixed parameter vector of interest; θ_i be the latent random variables (possibly vectors) for each cluster with sample space T, $i = 1 \ldots n$; and let $\ell_i^*(\beta|\theta_i)$ be the presumed log-likelihood for each cluster conditional on θ_i, with conditional score and information $s_i^*(\beta|\theta_i) \equiv \partial \ell_i^*(\beta|\theta_i)/(\partial\beta)$ and $I_i^*(\beta|\theta_i) \equiv -\partial s_i^*(\beta|\theta_i)/(\partial\beta)$, respectively, $i = 1 \ldots n$. Also denote $g(\theta_i;\beta)$ to be the presumed density for θ_i and $\mu(\cdot)$ to be the appropriate measure for θ_i, generally counting measure when θ is discrete and Lebesgue measure when θ is continuous, so that the marginal log-likelihood integral $\ell_i(\beta) \equiv \log(\int_T \exp[\ell_i^*(\beta|\theta)]g(\theta;\beta)\mu(d\theta))$ is well defined ($i = 1 \ldots n$), and let $r(\theta;\beta) \equiv \partial \log\left[g(\theta;\beta)\right]/(\partial\beta)$ and $R(\theta;\beta) \equiv -\partial r(\theta;\beta)/(\partial\beta)$. The form of ℓ_i^* together with g specify the working model $\mathcal{P}_*$

Provided that the appropriate derivatives with respect to β can be taken through the integral sign in the marginal log-likelihood integral, we also define

$$
\begin{aligned}
s_i(\beta) \equiv \frac{\partial \ell_i(\beta)}{\partial\beta} &= \frac{\int_T \frac{\partial}{\partial\beta}\left\{\exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\right\}\mu(d\theta)}{\int_T \exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)} \\
&= \frac{\int_T \left\{s_i^*(\beta|\theta) + r(\theta;\beta)\right\}\exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)}{\int_T \exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)}
\end{aligned}
\tag{1}
$$

and

$$
\begin{aligned}
I_i(\beta) \equiv -\frac{\partial s_i(\beta)}{\partial\beta} &= \frac{\int_T \left\{I_i^*(\beta|\theta) + R(\theta;\beta)\right\}\exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)}{\int_T \exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)} \\
&- \frac{\int_T \left\{s_i^*(\beta|\theta) + r(\theta;\beta) - s_i(\beta)\right\}\left\{s_i^*(\beta|\theta) + r(\theta;\beta) - s_i(\beta)\right\}' \exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)}{\int_T \exp\left[\ell_i^*(\beta|\theta)\right] g(\theta;\beta)\mu(d\theta)},
\end{aligned}
\tag{2}
$$

for $i = 1 \ldots n$. Denote the full working log-joint density of the clusters and $\Theta \equiv \{\theta_1, \ldots, \theta_n\}$ as $L_n^*(\beta,\Theta) \equiv \sum_{i=1}^n \left\{\ell_i^*(\beta|\theta_i) + \log\left[g(\theta_i;\beta)\right]\right\}$; and denote the full marginal log-likelihood as $L_n(\beta) \equiv \sum_{i=1}^n \ell_i(\beta)$.

4. The Estimation Algorithm and Large Sample Theory

4.1. The Algorithm

The algorithm we propose consists of two major steps.

The First Step: The first step is to use an MHA to generate a Markov chain in β and Θ of length m, $\{\beta^j, \Theta^j\}$, $j = 1 \ldots m$, with an equilibrium distribution which equals the posterior $C \exp\left[L_n^*(\beta, \Theta)\right] \pi(\beta)$ where C is the normalizing constant, for a suitable non-negative prior measure $\pi(\cdot)$. The reason we call this a working prior is that we are not doing Bayesian inference but are taking advantage of numerical methods used by Bayesians to compute quantities which can then be used for frequentist inference. For most frequentist applications $\pi(\cdot)$ will be constant on an appropriate compact set and zero outside of this set. In such cases the normalized posterior will just be the normalized likelihood function. A benefit of using the MHA for frequentist applications with missing data or random effects is that Bayesian computational methods can be used to avoid computation of the marginal likelihood, yet with flat priors the stationary distribution remains proportional to the likelihood function. We will assume throughout that the "burn in" stage for all Markov chains has already been omitted and that the chain is in equilibrium. Details of an MHA used for a specific data analysis will be given later in section 5. Let $\bar{\beta}_n^m \equiv m^{-1} \sum_{j=1}^m \beta^j$ be an estimator of the posterior mean of β, and let $W_n^m \equiv nm^{-1} \sum_{j=1}^m \left\{\beta^j - \bar{\beta}_n^m\right\} \left\{\beta^j - \bar{\beta}_n^m\right\}'$ be an estimator of the posterior variance of $\sqrt{n}(\beta - \beta_0)$. The chain length m is chosen so that the Monte Carlo error is less than some predetermined ϵ times the estimated variance of the posterior mean, using multivariate diagnostics of Kosorok (2000).

The Second Step: The second step is to generate a sequence of random deviates $\theta^1, \ldots, \theta^\nu$ which form a Markov chain, or are simply independent, with equilibrium distribution $g(\theta; \bar{\beta}_n^m)$, so that we can obtain the estimator

$$\hat{s}_i^\nu(\hat{\beta}) \equiv \frac{\sum_{j=1}^\nu \left\{s_i^*(\hat{\beta}|\theta^j) + r(\theta^j; \hat{\beta})\right\} \exp\left[\ell_i^*(\hat{\beta}|\theta^j)\right]}{\sum_{j=1}^\nu \exp\left[\ell_i^*(\hat{\beta}|\theta^j)\right]},$$

where $\hat{\beta}$ is consistent for β_0, and where β_0 is the parameter value which locally minimizes the Kullback-Leibler distance between the presumed marginal log-likelihood and the true probability distribution. We then compute $V_n^\nu(\hat{\beta}) \equiv n^{-1} \sum_{i=1}^n \hat{s}_i^\nu(\hat{\beta}) \{\hat{s}_i^\nu\}' (\hat{\beta})$.

As $n \to \infty$, $m/n \to \infty$, $\nu/n \to \infty$ and subject to regularity conditions, $n^{1/2} \left(\bar{\beta}_n^m - \beta_0\right)$ is asymptotically normal with mean zero and covariance which can be consistently estimated by $W_n^m V_n^\nu(\bar{\beta}_n^m) W_n^m$. We will present these asymptotic results in the next section.

4.2. Large Sample Theory

Regularity Conditions. The standard conditions for maximum working likelihood estimation will be needed (seed Huber, 1967; or Zeger, Liang, and Self, 1985), which include that the data clusters are independent and identically distributed, that there is a unique maximizer β_0 of the expected likelihood, that the marginal likelihood is three times differentiable for all β in a convex neighborhood $B \subset \mathbf{R}^p$ of β_0, that both $I_0^* \equiv \mathbf{E}_0\left[s_i(\beta_0) s_i'(\beta_0)\right]$ and $I_0 \equiv \mathbf{E}_0\left[I_i(\beta_0)\right]$ are finite, and that I_0 is non-singular. In addition to these likelihood conditions for the maximizer $\hat{\beta}_n$, we need the following conditions on integrability and on the prior measure π:

(A) There exists an $\epsilon > 0$ such that, almost surely,

$$\limsup_{n\to\infty} \sqrt{n} \int_{\beta \in B} \left|\sqrt{n}(\beta - \hat{\beta}_n)\right|^{2+\epsilon} \exp\left[L_n(\beta) - L_n(\hat{\beta}_n)\right] d\beta < \infty.$$

(B) $\pi(\beta_0) > 0$, $\sup_{\beta \in B} \pi(\beta)/\pi(\beta_0) < \infty$, and $\pi(\beta)$ is continuous at β_0.

The integrability condition (A) is needed to establish consistency of moment estimates for the Markov chain in β. Although this condition is difficult to check analytically, it can be evaluated empirically by plotting the quantiles of $\left[n \left\{\beta^j - \bar{\beta}_n^m\right\}' (W_n^m)^{-1} \left\{\beta^j - \bar{\beta}_n^m\right\}\right]^{3/2}$, $j = 1 \ldots m$, against the quantiles of a sample of m independent chi-squared variates with p degrees of freedom and raised to the $3/2$ power. These two sets of quantiles should approximately agree.

Main Results. The following theorem yields consistency and asymptotic normality of the working likelihood maximizer. The proof of this result is essentially classical and will be omitted.

Theorem 1. Under the likelihood regularity conditions for $\hat{\beta}_n$, $\lim_{n\to\infty} \hat{\beta}_n = \beta_0$ in probability and $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges in distribution to Z_0, as $n \to \infty$, where $Z_0 \sim \mathrm{N}\left(0, I_0^{-1} I_0^* I_0^{-1}\right)$.

The following two theorems establish consistency of the estimators obtained through the algorithm in section 4.1. Proofs are given in the appendix.

Theorem 2. Under the aforementioned likelihood regularity conditions and conditions (A) and (B), there exists a sequence m_n such that $m_n < \infty$ for every n and

(i) $\lim_{n\to\infty} \sqrt{n}\left(\bar{\beta}_n^{m_n} - \hat{\beta}_n\right) = 0$ in probability;
(ii) $\sqrt{n}\left(\bar{\beta}_n^{m_n} - \beta_0\right)$ converges in distribution to the same Z_0 given in theorem 1, as $n \to \infty$; and
(iii) $\lim_{n\to\infty} W_n^{m_n} = I_0^{-1}$ in probability.

Theorem 3. Under the conditions of theorem 2, and provided we can generate a Markov chain with equilibrium distribution $g(\theta; \beta)$, for any $\beta \in B$, there exists a sequence $\{\nu_n\}$ such that $\nu_n < \infty$ for every $n \geq 0$ and $\lim_{n\to\infty} V_n^{\nu_n}(\bar{\beta}_n^{m_n}) = I_0^*$ in probability.

Taken together, the results of this section establish the asymptotic validity of the estimation algorithm given in section 4.1, and also that $W_n^m V_n^\nu(\bar{\beta}_n^m) W_n^m$ converges to $I_0^{-1} I_0^* I_0^{-1}$.

5. An Example in Item Response Modeling

Here we consider data on "social life feelings" listed and analyzed in Bartholomew and Knott (1999), and originally reported in Schuessler (1982). In particular, we are interested in one scale designed to measure a construct labeled "economic self-determination". Data from an intercultural version of this scaled were obtained from 1490 Germans. The data consist of yes or no reponses to the following five questions that are listed in Bartholomew and Knott (1990): 1. Anyone can raise his standard of living if he is willing to work at it. 2. Our country has too many poor people who can do little to raise their standard of living. 3. Individuals are poor because of lack of effort on their part. 4. Poor people could improve their lot if they tried. 5. Most people have a good deal of freedom in deciding how to live.

We are interested in seeing if the dependence in these five responses can be adequately explained be a unidimensional latent variable θ. The model we consider is referred to as the logit/normit model by Bartholomew and Knott, and it is mathematically equivalent to a two-parameter logistic item response model with a normal distribution for the latent trait. Let Y_{ij} denote the response of the ith person on the jth variable, and let θ_i denote the value of the latent trait for the ith person. We assume that $\theta \sim N(0,1)$, and the conditional distribution of Y_{ij} is given by

$$log\left(\frac{P[Y_{ij}=1|\theta_i]}{P[Y_{ij}=0|\theta_i]}\right) = \beta_{j0} + \beta_{j1}\theta_i.$$

Let $P_j(\theta_i)$ denote $P[Y_{ij}=1|\theta_i]$. Assuming conditional independence of the items responses given θ, the conditional working likelihood function is

$$\prod_{i=1}^{n}\prod_{j=1}^{5}[P_j(\theta_i)]^{y_{ij}}[1-P_j(\theta_i)]^{1-y_{ij}}.$$

Bartholomew and Knott (1999) also assume a $N(0,1)$ distribution for θ, and they use the EM algorithm to arrive at the marginal maximum likelihood estimate of β. They measure goodness-of-fit by comparing expected and observed frequencies for the 32 possible response patterns. This results in a chi-square statistic of 32.84 on 16 degrees of freedom, with a corresponding p-value less than 0.01. However, Bartholomew and Knott argue that all marginal, two-way, and three-way residuals are very small and seem to provide sufficent grounds for justifying the use of the unidimensional logit/normit model with these data. Based on this reasoning that the model is likely misspecified but adequate, we fit the same model, only using a frequentist MCMC approach, and include robust estimates of the standard errors.

5.1. Implementation of the MHA

The initial values of β_{j0} were set to 0 for $j = 1, 2, ..., 5$, and the initial values of β_{j1} were set to 0 for $j = 1, 2, ..., 5$. Initial values for θ were obtained by summing the five items, ranking the sums, and transforming the ranks to the corresponding quantiles of a standard normal distribution. The parameters were updated one-by-one in the MHA, rather than rejecting or accepting proposals in blocks of multiple parameters. Proposal distributions for the β_{j0}'s were independent normal distributions with means equal to the current values and standard deviations tuned to accept roughly one in four proposals. The β_{j1}'s were proposed from independent log-normal distributions with location parameters equal to the logarithm of the current values and scale parameters tuned to accept one in four. Finally, the proposal distributions for the θ_i's were independent normal distributions with means equal to the current values and having a common standard deviation calibrated to accept roughly one in four proposals over the whole vector of θ for the 1490 subjects.

Our aim was to perform a frequentist analysis, which was achieved by setting the marginal posterior distribution equal to the normalized marginal likelihood function. Accordingly, Lebesgue measure was used as a prior for all item parameters, with β_{11} restricted to be positive to preserve identifiability. The prior distribution for the latent trait was $N(0, 1)$.

A chain of length 10,000 was constructed in this manner and the first 2,000 observations were discarded. Posterior mean estimators using the remaining 8,000 observations were calculated. The two separate covariance matrices described in section 3 were computed, using $\nu = 100$, and are denoted $\hat{\Sigma}_1 \equiv \frac{1}{n}\hat{W}$ and the robust $\hat{\Sigma}_2 \equiv \frac{1}{n}\hat{W}\hat{V}\hat{W}$.

5.2. Results

The resulting parameter estimates and standard errors are given in Table 1 below. The posterior mean estimates using flat prior distributions are nearly identical to the marginal maximum likelihood estimates listed in Table 4.3 of Bartholomew and Knott (1999). The largest absolute difference between the estimates using these two methods was 0.07 for parameter β_{41}. Note that for all parameters $\hat{\sigma}_1$ is less than the robust standard error estimate $\hat{\sigma}_2$. In many cases the absolute difference is small. However, the robust variance estimator ranges from 1 percent larger to 32 percent larger. This small but positive difference between the naive and robust variance estimates may reflect failure to condition on a second but less important latent trait, that might be required for the assumption of conditional independence to strictly hold.

Table 1: Parameter estimates and standard errors for logit/normit fit of economic self-determination data.

parameter	estimate	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_2^2/\hat{\sigma}_1^2$
β_{10}	-2.383	0.146	0.168	1.324
β_{20}	0.797	0.064	0.066	1.073
β_{30}	0.993	0.095	0.101	1.139
β_{40}	-0.683	0.135	0.144	1.149
β_{50}	-1.103	0.077	0.082	1.138
β_{11}	1.233	0.157	0.177	1.261
β_{21}	0.716	0.095	0.100	1.134
β_{31}	1.535	0.167	0.168	1.018
β_{41}	2.620	0.426	0.440	1.065
β_{51}	0.928	0.106	0.113	1.143

6. Discussion

The latent variable models used in psychometrics can often be quite intricate, and may be difficult to inspect for goodness-of-fit. Even when formal goodness-of-fit tests do indicate some model misspecification, analysis of predicted versus observed means or frequencies, higher-order residuals, or even scientific judgement might suggest that the model is still capable of providing considerable insight. However, in such situations the usual estimator of the parameter estimator's sampling variance is not correct. The worked examples based on the simple linear model with random effects/latent variables demonstrates that the difference between the naive (non-robust) and robust variance estimators can be quite large. Given this, it may be prudent in large studies to use robust methods as a general rule, at least for diagnostic purposes, especially considering that the variance estimators have the same limit when the model is correctly specified.

The most important accomplishment of the methods proposed in this paper is that the power of Markov chain Monte Carlo methods can now be utilized for frequentist inference which is robust to likelihood misspecification. Another important point is that all of the given methods will work not only when latent variables are involved but also in simpler working likelihood settings not involving latent variables, or when the latent variables represent missing manifest variables rather than psychological constructs.

Appendix

Proof of proposition 1. Let $\Phi(z)$ be the standard normal cummulative distribution function, ℓ_i the working log-likelihood, and $\tilde{\beta} = (a, b, \gamma, \sigma)^T$. Then

$$\begin{aligned}
\ell_i(\tilde{\beta}) &= \log\left[\int_{\Re} \exp\left(-m\log\sigma - \frac{\sum_{j=1}^{m}\left\{Y_{ij} - a - bX_{ij} - \gamma\theta\right\}^2}{2\sigma^2}\right) d\Phi(\theta)\right] \\
&= -m\log\sigma - \frac{\sum_{j=1}^{m}(Y_{ij} - a - bX_{ij})^2}{2\sigma^2} \\
&\quad + \frac{m^2\gamma^2\left(1 + \frac{m\gamma^2}{\sigma^2}\right)^{-1}\left(\overline{Y}_i - a - b\overline{X}_i\right)^2}{2\sigma^4} - \frac{1}{2}\log\left(1 + \frac{m\gamma^2}{\sigma^2}\right), \qquad (3)
\end{aligned}$$

where $\overline{Y}_i = m^{-1}\sum_{j=1}^{m} Y_{ij}$ and $\overline{X}_i = m^{-1}\sum_{j=1}^{m} X_{ij}$. If we set $r = m\gamma^2\left(\sigma^2 + m\gamma^2\right)^{-1}$, (3) becomes

$$\ell_i(\beta) = -m\log\sigma - \frac{\sum_{j=1}^{m}(Y_{ij} - a - bX_{ij})^2}{2\sigma^2} + \frac{mr\left(\overline{Y}_i - a - b\overline{X}_i\right)^2}{2\sigma^2} + \frac{1}{2}\log\left(1 - r\right), \qquad (4)$$

where we have reparameterized $\tilde{\beta}$ to become $\beta = (a, b, r, \sigma)^T$.

The score for (4) now becomes

$$s_i(\beta) = \begin{bmatrix} \dfrac{-mr\begin{pmatrix}1\\ \overline{X}_i\end{pmatrix}\left(\overline{Y}_i - a - b\overline{X}_i\right) + \sum_{j=1}^m \begin{pmatrix}1\\ X_{ij}\end{pmatrix}(Y_{ij} - a - bX_{ij})}{\sigma^2} \\[2em] \dfrac{m\left(\overline{Y}_i - a - b\overline{X}_i\right)^2}{2\sigma^2} - \dfrac{1}{2(1-r)} \\[2em] -\dfrac{m}{\sigma} + \dfrac{-mr\left(\overline{Y}_i - a - b\overline{X}_i\right)^2 + \sum_{j=1}^m (Y_{ij} - a - bX_{ij})^2}{\sigma^3} \end{bmatrix};$$

and the information becomes $I_i(\beta)$, with the first two columns being

$$\sigma^{-2} \begin{bmatrix} -mr\begin{pmatrix}1 & \overline{X}_i\\ \overline{X}_i & \overline{X}_i^2\end{pmatrix} + \sum_{j=1}^m \begin{pmatrix}1 & X_{ij}\\ X_{ij} & X_{ij}^2\end{pmatrix} \\[2em] m\left(\overline{Y}_i - a - b\overline{X}_i\right)\left(1,\ \overline{X}_i\right) \\[2em] 2\sigma^{-1}\left\{-mr\left(\overline{Y}_i - a - b\overline{X}_i\right)\left(1,\ \overline{X}_i\right) + \sum_{j=1}^m (Y_{ij} - a - bX_{ij})(1,\ X_{ij})\right\} \end{bmatrix},$$

the lower-right 2×2 block being

$$\begin{bmatrix} \dfrac{1}{2(1-r)^2}, & \dfrac{m\left(\overline{Y}_i - a - b\overline{X}_i\right)^2}{\sigma^3} \\[2em] \dfrac{m\left(\overline{Y}_i - a - b\overline{X}_i\right)^2}{\sigma^3}, & -\dfrac{m}{\sigma^2} + \dfrac{3\left\{-mr\left(\overline{Y}_i - a - b\overline{X}_i\right)^2 + \sum_{j=1}^m (Y_{ij} - a - bX_{ij})^2\right\}}{\sigma^4} \end{bmatrix},$$

and the remaining components defined by symmetry. For the Kullback-Leibler minimizer β_*, $\mathbf{E}s_i(\beta_*) = 0$; thus, if we let $E_1 = m^{-1}\sum_{j=1}^m \mathbf{E}\left(Y_{ij} - a_* - b_*X_{ij}\right)^2$ and $E_2 = \mathbf{E}\left(\overline{Y}_i - a_* - b_*\overline{X}_i\right)^2$, then by solving the score equation we obtain $\sigma_*^2 = (1 - 1/m)^{-1}(E_1 - E_2)$ and $r_* = 1 - \sigma_*^2/(mE_2)$.

Now, $\mathbf{E}s_i(\beta_*) = 0$ implies that

$$\mathbf{E}\left[\left\{-mr_*\begin{pmatrix}1\\ \overline{X}_i\end{pmatrix}\left(a_0 + U_i b_0 \overline{X}_i - a_* - b_*\overline{X}_i + \gamma_0\theta_i + \overline{\eta}_i\right)\right.\right.$$

$$+\sum_{j=1}^{m}\begin{pmatrix}1\\X_{ij}\end{pmatrix}(a_0+U_ib_0X_{ij}-a_*-b_*X_{ij}+\gamma_0\theta_i+\eta_{ij})\Bigg|X_i\Bigg\}\Bigg]=0, \tag{5}$$

where $\overline{\eta}_i=m^{-1}\sum_{j=1}^{m}\eta_{ij}$. But (5) implies

$$\mathbf{E}\left[-mr_*\begin{pmatrix}1&\overline{X}_i\\\overline{X}_i&\overline{X}_i^2\end{pmatrix}+\sum_{j=1}^{m}\begin{pmatrix}1&X_{ij}\\X_{ij}&X_{ij}^2\end{pmatrix}\right]\begin{pmatrix}a_0-a_*\\b_0-b_*\end{pmatrix}=0,$$

which then implies $a_*=a_0$ and $b_*=b_0$ by positive definiteness of the left-hand side. Consequently,

$$E_1=\mathbf{E}\left[\left(m^{-1}\sum_{j=1}^{m}\{(U_i-1)b_*X_{ij}+\gamma_0\theta_i+\eta_{ij}\}^2\Bigg|X_i\right)\right]=b_*^2\tau^2\xi^2+\gamma_0^2+\sigma_0^2$$

and $E_2=\mathbf{E}\left[\left(\left\{(U_i-1)b_*\overline{X}_i+\gamma_0\theta_i+\overline{\eta}_i\right\}^2\Big|X_i\right)\right]=b_*^2\tau^2\xi^2/m+\gamma_0^2$. Thus $\sigma_*^2=\sigma_0^2+b_*^2\tau^2\xi^2$ and $r_*=m\gamma_0^2/\left(\sigma_*^2+m\gamma_0^2\right)$, which implies $\gamma_*^2=\gamma_0^2$.

It is clear at this juncture that the working maximum likelihood estimators for a_*, b_*, and γ_* are consistent for a_0, b_0, and γ_0 but that the estimator for σ_*^2 will not be consistent for σ_0^2. If we let $[A]_{kl}$ denote the k,l'th element of the matrix A, then, after some derivation, the naive variance (based on the working likelihood information) of $\sqrt{n}(\hat{b}-b_*)$ is $\left[\{\mathbf{E}I_i(\beta_*)\}^{-1}\right]_{22}=\sigma_*^2/\left[(m-r_*)\xi^2\right]\equiv\zeta_*$. Note that the off-diagonals $\left[\{\mathbf{E}I_i(\beta_*)\}^{-1}\right]_{2k}$, for $k\neq 2$, are zero.

Thus to get the robust (correct) variance, we need only to determine $\mathbf{E}\left[s_i(\beta_*)\right]_2^2$, where $[a]_k$ is the k'th element of the vector a:

$$\begin{aligned}\sigma_*^4\mathbf{E}\left[s_i(\beta_*)\right]_2^2&=\mathbf{E}\left[-mr_*\left(\overline{Y}_i-a_*-b_*\overline{X}_i\right)\overline{X}_i+\sum_{j=1}^{m}\left(Y_{ij}-a_*-b_*X_{ij}\right)X_{ij}\right]^2\\&=\left(m^2-1+3(1-r_*)^2\right)b_*^2\tau^2\xi^4+(m-r_*)\xi^2\sigma_0^2+r_*(1-r_*)\xi^2(\sigma_*^2-\sigma_0^2),\end{aligned}$$

which implies

$$\left[\{\mathbf{E}I_i(\beta_*)\}^{-1}\right]_{22}\mathbf{E}\left[s_i(\beta_*)\right]_2^2\left[\{\mathbf{E}I_i(\beta_*)\}^{-1}\right]_{22}=\frac{\left[m^2+(2r_*-1)(r_*-2)\right]b_*^2\tau^2\xi^2+\sigma_0^2}{(m-r_*)^2\xi^2}. \tag{6}$$

If we let $\rho = b_*^2\tau^2\xi^2/\sigma_*^2$, then $0 \le \rho < 1$ and (6) can be rewritten as

$$\frac{[m^2 + (2r_* - 1)(r_* - 2)]\,\sigma_*^2\rho}{(m - r_*)^2\xi^2} + \frac{(1-\rho)\sigma_*^2}{(m - r_*)\xi^2} = \frac{\sigma_*^2}{(m - r_*)\xi^2}\left[1 + \rho \times \frac{m^2 - m + 2(1 - r_*)^2}{m - r_*}\right]$$
$$= \zeta_*\left[1 + \rho \times \frac{m^2 - m + 2(1 - r_*)^2}{m - r_*}\right],$$

and the desired result follows.

Proof of theorem 2. First observe that, for any $d : 0 \le d \le 2$, (A) combined with (B) implies

$$\lim_{c\to\infty}\limsup_{n\to\infty}\sqrt{n}\int_{\beta\in B:\left|\sqrt{n}(\beta-\hat\beta_n)\right|>c}\left|\sqrt{n}(\beta - \hat\beta_n)\right|^d \exp\left\{L_n(\beta) - L_n(\hat\beta)\right\}\frac{\pi(\beta)}{\pi(\beta_0)}d\beta = 0. \tag{7}$$

Note that the density associated with our Markov chain is

$$\frac{\exp\left\{L_n^*(\beta;\theta)\right\}\pi(\beta)}{\int_B\int_T \exp\left\{L_n^*(b;t)\right\}\pi(b)\left\{\prod_{i=1}^n \mu(dt_i)\right\}db},$$

where $t \equiv \{t_1, \ldots, t_n\}'$. This density admits the marginal density

$$\frac{\exp\left\{L_n(\beta)\right\}\pi(\beta)}{\int_B \exp\left\{L_n(b)\right\}\pi(b)db} = \frac{\sqrt{n}\exp\left\{L_n(\beta) - L_n(\hat\beta_n)\right\}\pi(\beta)/\pi(\beta_0)}{\sqrt{n}\int_B \exp\left\{L_n(b) - L_n(\hat\beta_n)\right\}\left\{\pi(b)/\pi(\beta_0)\right\}db}.$$

By a Taylor's expansion,

$$\begin{aligned}&\sqrt{n}\exp\left\{L_n(\beta) - L_n(\hat\beta_n)\right\}\pi(\beta)/\pi(\beta_0)\\ &= \sqrt{n}\exp\left\{S_n'(\hat\beta_n)(\beta - \hat\beta_n) - \frac{1}{2}\sqrt{n}(\beta - \hat\beta_n)'\left[\frac{I_n(\beta^*)}{n}\right]\sqrt{n}(\beta - \hat\beta_n)\right\}\frac{\pi(\beta)}{\pi(\beta_0)}\end{aligned} \tag{8}$$

for some β^* on the line segment between β and $\hat\beta_n$, and thus we have that $\forall c < \infty$,

$$\sqrt{n}\int_{\beta\in B:\left|\sqrt{n}(\beta-\hat\beta_n)\right|\le c}\left|\sqrt{n}(\beta - \hat\beta_n)\right|^d \exp\left\{L_n(\beta) - L_n(\hat\beta)\right\}\frac{\pi(\beta)}{\pi(\beta_0)}d\beta$$

is asymptotically equivalent to

$$\sqrt{n}\int_{\beta\in B:\left|\sqrt{n}(\beta-\hat\beta_n)\right|\le c}\left|\sqrt{n}(\beta - \hat\beta_n)\right|^d \exp\left\{-\frac{1}{2}\sqrt{n}(\beta - \hat\beta_n)'I_0\sqrt{n}(\beta - \hat\beta_n)\right\}\frac{\pi(\beta)}{\pi(\beta_0)}d\beta$$

which, by the likelihood regularity conditions and (B), is asymptotically equivalent to $\int_{u \leq c} |u|^d \exp\{-u' I_0 u/2\}\, du$, for finite $d > 0$. Since c is arbitrary, we then have by (7) that

$$\lim_{n\to\infty} \frac{\int_B \sqrt{n}(\beta - \hat{\beta}) \exp\{L_n(\beta)\}\, \pi(\beta) d\beta}{\int_B \exp\{L_n(\beta)\}\, \pi(\beta) d\beta} = 0$$

and

$$\lim_{n\to\infty} \frac{\int_B n(\beta - \hat{\beta})(\beta - \hat{\beta})' \exp\{L_n(\beta)\}\, \pi(\beta) d\beta}{\int_B \exp\{L_n(\beta)\}\, \pi(\beta) d\beta} = I_0,$$

in probability. The result now follows by the convergence of the Markov chain conditional on the data, by the ergodic theorem, and by theorem 1.

Proof of theorem 3. For each $n \geq 1$, and conditional on the data, we have by the likelihood regularity conditions, and by the ergodic theorem, that $\forall \epsilon > 0$, there exists a $\nu_n < \infty$ such that

$$\mathbf{P}\left[\left| n^{-1} \sum_{i=1}^{n} \hat{s}_i^{\nu_n}(\hat{\beta}) \left\{\hat{s}_i^{\nu_n}(\hat{\beta})\right\}' - n^{-1} \sum_{i=1}^{n} s_i(\hat{\beta}) s_i'(\hat{\beta}) \right| > \epsilon \right] = 0$$

for all $m \geq m_n$, where the probability is with respect to the Markov chain conditional on the data. By standard arguments, the consistency of $\hat{\beta}$ now yields the desired results.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 1987*, 317–332.

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

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary data. *Psychometrika, 65*, 475–496.

Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis.* London: Arnold.

Besag, J., & Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Ser. B, 55*, 25–37.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions.

de Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika, 53*, 437–454.

Garret, E., & Zeger, S. (2000). Latent class model diagnosis. *Biometrics, 56*, 1055–1067.

Ghosal, S., Ghosh, J. K., & Samanta, T. (1995), On convergence of posterior distributions. *Annals of Statistics, 23*, 2145–2152.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., & Kirby, A. J. (1993), Modeling complexity: Applications of Gibbs Sampling in medicine. *Journal of the Royal Statistical Society, Ser. B, 55*, 39–52.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171–189.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. in *Proceedings of the 5th Berkeley Symposium,* Vol. 1, Berkeley: University of California Press, pp. 221–234.

Johnson, R. A. (1967). An asymptotic expansion for posterior distributions. *Annals of Mathematical Statistics, 38*, 1899–1906.

Johnson, R. A. (1970). Asymptotic expansion associated with posterior distributions. *Annals of Mathematical Statistics, 41*, 851–864.

Kosorok, M. R. (2000). Monte Carlo error estimation for multivariate Markov chains. *Statistics & Probability Letters, 46*, 85–93.

Lenk, P. J.,& DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika, 65*, 93–119.

Lee, S. Y., & Zhu, H. T. (2002). Maximum likelihood estimation of nonlinear structural equation model. *Psychometrika, 67*, 189–210.

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.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response theory. *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.

Schuessler, K. F. (1982). *Measuring Social Life Feelings.* Jossey-Bass Social and Behavioral Science Series. San Francisco: Jossey-Bass.

Smith, A. F. M, & Roberts, G. O. (1993). Bayesian computation via the Gibbs Sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Ser. B, 55*, 3–23.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.

Zeger, S. L., Liang, K.-Y., & Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika, 72*, 31–38.