Technical Report #112

Further Details on Robust Likelihood Inference with Missing Data Using Markov Chain Monte Carlo

Michael R. Kosorok, Ph.D.
Jeffrey A. Douglas, Ph.D.
Daode Huang, Ph.D.
Further Details on Robust Likelihood Inference with Missing Data Using Markov Chain Monte Carlo

Department of Biostatistics Technical Report 112
University of Wisconsin-Madison

Michael R. KOSOROK, Jeffrey A. DOUGLAS, and Daode HUANG

Maximum likelihood inference in the presence of missing data 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 first obtain both the maximum likelihood estimator and the working Fisher information matrix and, second, using Monte Carlo quadrature to obtain the remaining components of the robust asymptotic variance. Evaluation of the marginal likelihood is not needed. We demonstrate consistency and asymptotic normality when the number of independent and identically distributed data clusters is large but the likelihood may be incorrectly specified. An analysis of longitudinal ordinal data is given for an example.

KEY WORDS: Convergence of posterior distributions, Maximum likelihood, Metropolis-Hastings algorithm, Missing data, Random effects, Latent variables.

1. INTRODUCTION

Frequency based maximum likelihood (ML) inference involving missing data 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; Besag et al., 1995; Chib and Greenberg, 1995;
Gilks et al., 1993; and Smith and Roberts, 1993). When the likelihood model is correct,
and when the number of independent and identically distributed 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; Johnson, 1967 and
1970; and Johnson and Ladalla, 1979).

When the proposed likelihood is not necessarily correct but capable of providing meaning-
ful parameter estimates, Huber (1967) demonstrated that maximum likelihood inference can
still be carried out, provided an appropriate 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 deriva-
tive of the log-likelihood and $I_0^*$ is the expectation of the outer product of the cluster-level
score component. The likelihood in this context is referred to as a "working likelihood"
(see Zeger, Liang, and Self, 1985, for example)—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 working
likelihood and the true probability distribution. It is important to point out that whenever
maximum likelihood is used, it is usually not absolutely clear that the given likelihood is
correct: maximum working likelihood (MWL) inference is therefore just maximum likeli-
hood inference with the acknowledged necessity of computing a variance estimator which is
robust against possible misspecification of the likelihood.

An alternative approach to estimation which also acknowledges that the true probability
distribution cannot be known is the method of estimating functions (Godambe, 1991). The
quasi-score estimating function is a familiar example, and only requires knowing the mean
function and variance structure of the data (Wedderburn, 1974). Li and McCullagh (1994) pointed out that because a quasi-score function, unlike the likelihood score function, is generally not a gradient of any potential function, comparison of parameter values can be difficult and the problem is compounded when the score function has multiple roots. They address this issue by defining a class of conservative score functions which are the projection of some true score function. Hanfelt and Liang (1995) consider a similar problem and develop methods for constructing approximate likelihood ratios that require only the knowledge of the moment structure specified by the estimating function. Maximum working likelihood inference is quite distinct from these approaches in that one begins by specifying a complete working likelihood function and initially proceeds as with ordinary maximum likelihood inference.

In this paper, we describe a method for obtaining frequency based maximum working likelihood inference in the presence of missing data which does not require integration of the conditional likelihood. This method utilizes Markov chain Monte Carlo (MCMC) to first obtain both the MWL estimator and a consistent estimator of the working Fisher information matrix \( I_0 \) and, second, uses Monte Carlo quadrature to obtain a consistent estimator for \( I_0^* \). In section 2, we more formally describe the setting for which the proposed methods are to be applied and introduce the notation which will be used throughout the paper. We then detail the proposed algorithm and provide large sample theory in section 3. Section 4 presents a worked example demonstrating analytically how the true variance (based on MWL) can be arbitrarily larger than the usual ML variance estimator even in a very simple and commonly encountered setting. Section 5 then presents an analysis of longitudinal arthritis data which utilizes the proposed new methods. The paper then closes with a brief discussion in section 6.
2. THE SETTING AND NOTATION

We begin by assuming that the data consist of \( n \) independent and identically distributed clusters. Let \( \beta \) be the unknown fixed parameter vector of interest; \( \theta_i \) be the missing (or latent) random variables (possibly vectors) for each cluster with sample space \( \Theta \), \( i = 1 \ldots n \); and let \( \ell_i^*(\beta|\theta_i) \) be the working log-likelihood for each cluster conditional on \( \theta_i \), with conditional score and information

\[
s_i^*(\beta|\theta_i) = \frac{\partial \ell_i^*(\beta|\theta_i)}{\partial \beta} \quad \text{and} \quad I_i^*(\beta|\theta_i) = -\frac{\partial s_i^*(\beta|\theta_i)}{\partial \beta},
\]

respectively, \( i = 1 \ldots n \). Also denote \( g(\theta_i; \beta) \) to be the working density for \( \theta_i \) and \( \mu(\cdot) \) to be the appropriate measure for \( \theta_i \) so that the working marginal log-likelihood component

\[
\ell_i(\beta) \equiv \log \left( \int_{\Theta} \exp \{ \ell_i^*(\beta|\theta) \} g(\theta; \beta) \mu(d\theta) \right)
\]

is correctly defined \( (i = 1 \ldots n) \), and let

\[
r^{(1)}(\theta; \beta) = \frac{\partial \log [g(\theta; \beta)]}{\partial \beta} \quad \text{and} \quad R^{(2)}(\theta; \beta) = \frac{\partial r^{(1)}(\theta; \beta)}{\partial \beta}.
\]

Provided that the appropriate derivatives with respect to \( \beta \) can be taken through the integral sign in (1), we also define

\[
s_i(\beta) = \frac{\partial \ell_i(\beta)}{\partial \beta} = \frac{\int_{\Theta} \frac{\partial}{\partial \beta} \left\{ \exp \{ \ell_i^*(\beta|\theta) \} g(\theta; \beta) \right\} \mu(d\theta)}{\int_{\Theta} \exp \{ \ell_i^*(\beta|\theta) \} g(\theta; \beta) \mu(d\theta)}
\]

\[
= \frac{\int_{\Theta} \left\{ s_i^*(\beta|\theta) + r^{(1)}(\theta; \beta) \right\} \exp \{ \ell_i^*(\beta|\theta) \} g(\theta; \beta) \mu(d\theta)}{\int_{\Theta} \exp \{ \ell_i^*(\beta|\theta) \} g(\theta; \beta) \mu(d\theta)}
\]

(2)
and

\[
I_i(\beta) = -\frac{\partial s_i(\beta)}{\partial \beta} =
\]

\[
= \frac{\int_\Theta \left\{ I_i(\beta|\theta) + R^{(3)}(\theta; \beta) \right\} \exp \left[ I_i^*(\beta|\theta) \right] g(\theta; \beta) \mu(d\theta)}{\int_\Theta \exp \left[ I_i^*(\beta|\theta) \right] g(\theta; \beta) \mu(d\theta)}
\]

\[
- \frac{\int_\Theta \left\{ s_i^*(\beta|\theta) + r^{(1)}(\theta; \beta) - s_i(\beta) \right\} \cdot \left\{ s_i^*(\beta|\theta) + r^{(1)}(\theta; \beta) - s_i(\beta) \right\}^T \exp \left[ I_i^*(\beta|\theta) \right] g(\theta; \beta) \mu(d\theta)}{\int_\Theta \exp \left[ I_i^*(\beta|\theta) \right] g(\theta; \beta) \mu(d\theta)}
\]

\[
(3)
\]

where superscript $T$ denotes transpose, for $i = 1 \ldots n$. Denote the full working log-likelihood

for $\beta$ and $\theta \equiv \{ \theta_1, \theta_2, \ldots, \theta_n \}$ as

\[
L^*_n(\beta, \theta) \equiv \sum_{i=1}^{n} \left\{ I_i^*(\beta|\theta_i) + \log \left[ g(\theta_i; \beta_i) \right] \right\}
\]

and denote the full working marginal log-likelihood as

\[
L_n(\beta) \equiv \sum_{i=1}^{n} \ell_i(\beta).
\]

We will introduce additional notation in the remainder of this paper as the need arises.

3. The Estimation Algorithm and Large Sample Theory

3.1 The Algorithm

The algorithm we propose consists of two major steps.

3.1.1 The First Step. The first step is to use an MHA to generate a Markov chain in $\beta$ and $\theta, \{ \beta_j, \theta^j \}, j = 1 \ldots m$, with an equilibrium distribution which equals the posterior

\[
C \exp \left[ L^*_n(\beta, \theta) \right] \pi(\beta),
\]

\[
(4)
\]
where $C$ is the normalizing constant, for a suitable non-negative “working” 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 of (4) will just be the normalized likelihood function. When the sample size is sufficiently large that the prior distribution will have little influence on parameter estimation, one may attempt to define $\pi(\cdot)$ in a way that allows draws to be obtained from the full conditional distributions in order to perform Gibbs sampling (Smith and Roberts, 1993), a special case of the MHA. 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. However, McCulloch (1997) cautions that for models with random effects using diffuse priors, the posterior distribution may not exist, and suggests alternative algorithms for maximum likelihood estimation of generalized linear mixed models. 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 4. Let

$$\bar{\beta}_n^m = m^{-1} \sum_{j=1}^{m} \beta^j$$

be an estimator of the posterior mean of $\beta$; let $\bar{\beta}_n^m$ be an estimator of the posterior mode; and let

$$W_n^m = nm^{-1} \sum_{j=1}^{m} \left\{ \beta^j - \bar{\beta}_n^m \right\} \left\{ \beta^j - \bar{\beta}_n^m \right\}^T$$
be an estimator of the posterior variance of $\beta$ (ignoring the $\theta$ components).

The next important question is how to determine whether $m$ is large enough for a specific data set. To accomplish this, we will estimate the Monte Carlo error and then take $m$ large enough so that the resulting Monte Carlo error is less than some predetermined $\epsilon$ times the estimated MWL component of the variance. The resulting "convergence criterion" is essentially equivalent to the "scale reduction factor" criterion described in Tierney (1992).

Our approach to estimating the Monte Carlo error for this purpose will be described in Section 3.3. For an excellent overview of MCMC convergence diagnostics, see Cowles and Carlin (1996).

3.1.2 The Second Step. The second step is to generate a sequence of random deviates $\theta^1, \theta^2, \ldots, \theta^\nu$ which form a Markov chain, or are simply independent, with equilibrium distribution $g(\theta; \beta^m)$, so that we can obtain estimators

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

and

$$
\tilde{D}_i(\beta) = \frac{\sum_{j=1}^{\nu} \left\{ I_i^*(\beta|\theta^j) + R^{(2)}(\theta^j; \beta) \right\} \exp \left[ \ell^*_i(\beta|\theta^j) \right]}{\sum_{j=1}^{\nu} \exp \left[ \ell^*_i(\beta|\theta^j) \right]}
$$

$$
= \frac{\sum_{j=1}^{\nu} \left\{ s_i^*(\beta|\theta^j) + r^{(1)}(\theta^j; \beta) \right\} \left\{ s_i^*(\beta|\theta^j) + r^{(1)}(\theta^j; \beta) \right\}^T \exp \left[ \ell^*_i(\beta|\theta^j) \right]}{\sum_{j=1}^{\nu} \exp \left[ \ell^*_i(\beta|\theta^j) \right]}
$$

for a chosen value of $\beta$, which are jointly consistent—conditional on the $n$ data clusters—for all $s_i(\beta)$ and $D_i(\beta) \equiv I_i(\beta) - s_i(\beta)\tilde{s}_i^T(\beta), i = 1 \ldots n$. We then compute

$$
V_n^\nu(\beta) \equiv n^{-1} \sum_{i=1}^{n} \tilde{s}_i(\beta)\tilde{s}_i^T(\beta) \quad \text{and} \quad C_n^\nu(\beta) \equiv n^{-1} \sum_{i=1}^{n} \tilde{D}_i(\beta).
$$
If we let $\beta_0$ be the parameter value which locally minimizes the Kullback-Leibler distance between the working marginal log-likelihood and the true probability distribution, both $n^{1/2} \left( \hat{\beta}_n^m - \beta_0 \right)$ and $n^{1/2} \left( \tilde{\beta}_n^m - \beta_0 \right)$ provided $n \to \infty$, $m/n \to \infty$, and subject to several regularity conditions—are asymptotically normal with mean zero and covariance which can be consistently estimated by either $W_n^m V_n^\nu(\hat{\beta}) W_n^m$ (no matrix inversions are needed) or by

$$
\left[ V_n^\nu(\hat{\beta}) + C_n^\nu(\hat{\beta}) \right]^{-1} V_n^\nu(\hat{\beta}) \left[ V_n^\nu(\hat{\beta}) + C_n^\nu(\hat{\beta}) \right]^{-1},
$$

where $\hat{\beta}$ is either $\hat{\beta}_n^m$ or $\tilde{\beta}_n^m$ or some other consistent estimator of $\beta_0$. Thus only one of either $W_n^m(\hat{\beta})$ or $C_n(\hat{\beta})$ needs to be computed. We will present these asymptotic results in the next section.

### 3.2 Large Sample Theory

We will first present the required regularity conditions and then present the main theoretical results. All proofs are given in the Appendix.

#### 3.2.1 Regularity Conditions

Conditions (L1) through (L6) below apply to the marginal working likelihood, the true probability distribution, and the working likelihood maximizer $\hat{\beta}_n$. These conditions are essentially the standard ones for MWL theory (see Huber, 1967; or Zeger, Liang, and Self, 1985, for a more applied example) but are presented here in a non-standard level of generality which we will need. Condition (L5') relates to the posterior mode $\tilde{\beta}_n$ while Conditions (B1) through (B3) and (B3') apply to the prior measure $\pi$. Conditions (U') and (M'), as well as Lemma 1, are additional restrictions on the working likelihood and true probability distribution which are needed for the asymptotic consistancy of the proposed MCMC estimators.
(L1) The data consists of independent and identically distributed random pairs \( \{X_i, Y_i\} \),
\( i = 1 \ldots n \), on a probability space \( \Omega, \mathcal{F}, \Pi_0 \), with joint density \( \Pi_0(x, y) = P_0(y|x)q(x) \), with respect to a \( \sigma \)-finite measure \( \xi(x, y) \), where \( P_0(y|x) \) is the conditional density of \( Y \) given \( X \). For \( \mathcal{F} \)-measurable functions \( f \), let
\[
\mathbb{E}_0[f(X, Y)] = \int_{\Omega} f(x, y) P_0(y|x)q(x)\xi(dx, dy).
\]
Assume also that \( L_0 \equiv \mathbb{E}_0[\log P_0(Y|X)] > -\infty \). Note that \( \Pi_0 \) is the true probability distribution of the data.

(L2) Denote \( P_\beta(y|x) \) as the working conditional density parameterized by \( \beta \in \mathcal{R}^p \), i.e.
\[
P_\beta(y_i|x_i) = \exp \{ \ell_i(\beta) \}, \ i = 1 \ldots n, \text{ were } p \text{ is finite; and let the Kullback-Leibler distance be } K_0(\beta) = L_0 - L(\beta), \text{ where } L(\beta) = \mathbb{E}_0[\log P_\beta(Y_i|X_i)]. \text{ Assume that } K_0(\beta)
\]
has a unique minimizer over the convex set \( B \subset \mathcal{R}^p \) at \( \beta_0 \in B, \text{ where } K_0(\beta_0) < \infty, \text{ and } \mathcal{R}^p \text{ is endowed with the Euclidean norm } | \cdot |. \text{ Since this is a working likelihood, it is not necessary that } P_\beta(y|x) = P_0(y|x) \text{ for any } \beta \in B.

(L3) For all \( \beta \in B, \ell_i(\beta) \) is three times differentiable with respect to \( \beta \) and both \( I_0^* \equiv \mathbb{E}_0[s_i(\beta_0)s_i^T(\beta_0)] \) and \( I_0 \equiv \mathbb{E}_0[I_i(\beta_0)] \) are finite.

(L4) For every compact \( T \subset B \), there exists an \( \mathcal{F} \)-measurable real valued function \( h \) such that \( \mathbb{E}_0[h(X_i, Y_i)] < \infty \) and, for all \( \beta \in T, \)

- (a) \( |s_i(\beta)| \leq h(X_i, Y_i), \)
- (b) \( \|I_i(\beta)\| \leq h(X_i, Y_i), \)
- (c) \( \left\| \frac{\partial I_i(\beta)}{\partial \beta} \right\| \leq h(X_i, Y_i), \) and
- (d) \( |s_i(\beta)| \times \|I_i(\beta)\| \leq h(X_i, Y_i), \)
where for the matrix or tensor $A$, $\|A\|$ is the maximum in absolute value over all its elements.

(L5) Let $\tilde{\beta}_n$ be a sequence of maximizers of $L_n(\beta)$ over $B$ such that $\forall \epsilon > 0$, $\exists n_0 < \infty$ and a convex compact $B_\epsilon \subset B$ where $P \left[ \tilde{\beta}_n \in B_\epsilon \right] \geq 1 - \epsilon$, $\forall n \geq n_0$.

(L6) $I_0$ is non-singular.

The following condition replaces tightness of the likelihood maximizer as specified in (L5) with tightness of the posterior maximizer, permitting interchangeability of the two methods of estimation:

(L5') Let $\tilde{\beta}_n$ be a sequence of maximizers of $\exp \{ L_n(\beta) \} \pi(\beta)$ over $B$ such that for all $\epsilon > 0$, there exists an $n_0 < \infty$ and a convex compact $B_\epsilon \subset B$ where $P \left[ \tilde{\beta}_n \in B_\epsilon \right] \geq 1 - \epsilon$, $\forall n \geq n_0$.

The following conditions on the prior measure $\pi$ will also be needed. Condition (B3') below is a stronger version of (B3) which will sometimes prove useful.

(B1) $\pi(\beta_0) > 0$,  
(B2) $\pi(\beta)/\pi(\beta_0) \leq M < \infty$ for all $\beta \in B$, and  
(B3) $\pi(\beta)$ is continuous at $\beta_0$.  
(B3') There exists an $M < \infty$ such that

$$\left| s_\pi(\beta) \equiv \frac{\partial \log \pi(\beta)}{\partial \beta} \right| \leq M, \quad \left| I_\pi(\beta) \equiv -\frac{\partial s_\pi(\beta)}{\partial \beta} \right| \leq M, \quad \text{and} \quad \left| \frac{\partial I_\pi(\beta)}{\partial \beta} \right| \leq M$$

for all $\beta \in B$.

The following integrability and moment conditions, for $d > 0$, will allow us to utilize a local approximation to the (working) posterior distribution in the estimation of posterior
moments. Condition (M^d) implies condition (U^d) but can be at least partially evaluated by examining the posterior central moments. 

(U^d) 

$$\lim_{c \to \infty} \limsup_{n \to \infty} \sqrt{n} \int_{B \in B, ||\sqrt{n}(\beta - \hat{\beta}_n)|| > c} \left| \sqrt{n}(\beta - \hat{\beta}_n) \right|^d \exp \left[ L_n(\beta) - L_n(\hat{\beta}_n) \right] d\beta = 0$$

(M^d) There exists an $c > 0$ such that 

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

The following lemma provides a condition implying (M^d) and (U^d) but which is substantially easier to validate in some settings:

**Lemma 1.** If $B$ is compact and there exists a positive definite matrix $C_0$ such that $I_0(\beta) - C_0$ has nonnegative eigenvalues for all $\beta \in B$, then conditions (M^d) and (U^d) are satisfied for all finite $d > 0$.

The details of how these conditions are satisfied for a specific setting are illustrated for a simple working likelihood model in section 5. The establishment of these conditions will frequently not be easy, analytically, but they are likely to hold for most reasonable settings.

This is particularly true of Conditions (L1), (L2), and (L3) which involve making sure that the given model is a regression model of some sort, believing that there is a maximum likelihood solution, and then checking for the existence of derivatives. Conditions (B1) through (B3) are in the control of the researcher and thus will be easy to establish. Condition (L4) is not easy to verify, but will generally work out if $B$ is compact (away from extreme points), $\beta_0$ is in the interior of $B$, $X_i$ is bounded, and enough moments of $Y_i$ exist (sometimes $Y_i$ will need to have a moment generating function). If $W^n$ is positive definite, then $I_0(\beta_0)$ is most likely positive definite and thus Condition (L6) will follow from Condition (L5).
tion (L4). Condition (L4) and (L6) now imply there exists a convex compact neighborhood $B_0$ of $\beta_0$ such that $E_0(I_0(\beta))$ is non-singular for all $\beta_0 \in B_0$. If we then assume that $B$ is essentially restricted to $B_0$, then Conditions (L5) will automatically be satisfied. Lemma 1 will thus apply and Conditions (U$^d$) and (M$^d$) will be satisfied. An alternative method of checking Condition (M$^2$) is by examining the marginal distribution of

$$\left[ n \left\{ \beta^j - \bar{\beta}^m_n \right\}^T (W^m_n)^{-1} \left\{ \beta^j - \bar{\beta}^m_n \right\} \right]^{3/2}$$

and comparing it with the distribution of a chi-squared random variable with $p$ degrees of freedom and raised to the $3/2$ power (these two distributions should approximately agree).

3.2.2 Main Results The following theorem yields consistency and asymptotic normality of the working likelihood maximizer. Although condition (L6) serves a purpose similar to Lemma 2 of Huber (1967), Huber’s approach otherwise differs substantially from ours and is not as general. The theorem below is also similar to Theorem 1 of Kosorok and Chao (1995) but focuses on the likelihood maximizer rather than on the zero of the score.

**Theorem 1.** Under conditions (L1)-(L6),

(i) $\lim_{n \to \infty} \widehat{\beta}_n \to \beta_0$ in probability and

(ii) $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to $Z$, as $n \to \infty$, where $Z \sim N(0, I_0^{-1} I_0^{-1})$.

The following corollary allows us to replace the likelihood maximizer $\widehat{\beta}_n$ with the posterior mode $\widetilde{\beta}_n$:

**Corollary 1.** Suppose conditions (L1)-(L4), (L5'), (L6), (B1), (B2), and (B3') obtain. If $\widetilde{\beta}_n^{m_n}$ is an estimator of the posterior mode (possibly obtained from a Markov chain of length $m_n$) such that $\lim_{n \to \infty} \sqrt{n}(\widetilde{\beta}_n^{m_n} - \widetilde{\beta}_n) = 0$ in probability, then Theorem 1 implies that
\( \sqrt{n}(\tilde{\beta}_n^{m_n} - \beta_0) \) converges in distribution to the same \( Z \) given in result (ii) of Theorem 1, as \( n \to \infty \).

The following theorem verifies that the MCMC algorithm given in section 3.1 for the posterior distribution of \( \beta \) is a consistent method of estimating certain MWL parameters:

**Theorem 2.** Under conditions (L1)–(L6) and (B1)–(B3), there exists a sequence \( m_n \) such that \( m_n < \infty \) for every \( n \) and

(i) under condition (U_1), \( \lim_{n \to \infty} \sqrt{n} \left( \tilde{\beta}_n^{m_n} - \tilde{\beta}_n \right) = 0 \) in probability;

(ii) under condition (U_1), \( \sqrt{n}(\tilde{\beta}_n^{m_n} - \beta_0) \) converges in distribution to the same \( Z \) given in result (ii) of Theorem 1, as \( n \to \infty \); and

(iii) under condition (U_2), \( \lim_{n \to \infty} W_n^{m_n} = I_0^{-1} \) in probability.

The following theorem establishes consistency for the remaining parameter estimators:

**Theorem 3.** Under condition (L1)–(L6), and provided we can generate a Markov chain with equilibrium distribution \( g(\theta; \beta) \), for any \( \beta \in \mathcal{B} \), then there exists a sequence of lengths of this chain \( \{\nu_n\} \) such that \( \nu_n < \infty \) for every \( n \geq 0 \) and

(i) \( \lim_{n \to \infty} V_n^{\nu_n}(\tilde{\beta}) = I_0^* \) in probability and

(ii) \( \lim_{n \to \infty} C_n^{\nu_n}(\tilde{\beta}) = I_0 - I_0^* \) in probability,

where \( \tilde{\beta} \) is any consistent estimator of \( \beta_0 \) (such as \( \tilde{\beta}_n^{m_n}, \tilde{\beta}_n^{\nu_n}, \) or \( \tilde{\beta}_n \)).

Taken together, the results of this section establish the asymptotic validity of the MWL estimation algorithm given in section 3.1 and also that \( W_n^{m_n} V_n^{\nu_n} W_n^{m_n} \) converges to \( I_0^{-1} I_0^* I_0^{-1} \).

### 3.3 Monte Carlo Error

We begin by examining the asymptotic behavior of \( n^{1/2} \left( \tilde{\beta}_n^{m} - \beta_0 \right) \), which can be expressed
as
\[ n^{1/2} \left( \bar{\beta}_n^m - \beta_0 \right) = n^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) + n^{1/2} \left( \bar{\beta}_n - \bar{\beta}_n \right) + n^{1/2} \left( \bar{\beta}_n - \beta_0 \right), \]

where \( \bar{\beta}_n \) is the (working) posterior mean for the given sample. Results given in the previous section guarantee that the second term converges to zero in probability if condition \((U^1)\) holds. Theorem 1 gives the asymptotic behavior of the third term. When the sum of the autocovariances of the Markov chain converges, we have that conditional on the data and for large \( m \), \( m^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) \) is asymptotically normal with finite covariance \( \mathbf{H}_0 \), by Theorem 2.1 of Geyer (1992), since our MHA is time-reversible, stationary, and ergodic by construction.

Based on the foregoing arguments, we can construct a sequence \( \{m_n, n \geq 1\} \) such that each \( m_n \) is finite and
\[ \limsup_{n \to \infty} \left| n^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) \right| = \limsup_{n \to \infty} \left| (n/m_n)^{1/2} m_n^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) \right| = 0 \]
in probability. The main question we want to address in this section boils down to determining \( m_n \) for a specific data set. To accomplish this, we will utilize the results of Kosorok (1998) to estimate the variance of \( m^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) \), the Monte Carlo error, for large \( m \) and holding \( n \) fixed, and then take \( m = m_n \) large enough so that
\[ (n/m_n) \left\| \text{var} \left[ m_n^{1/2} \left( \bar{\beta}_n^m - \bar{\beta}_n \right) \right] \right\| < \epsilon \]
for some prespecified \( \epsilon \). Essentially, Kosorok's (1992) results involve generalizing the approach of Geyer (1992) from the univariate to the multivariate Markov chain setting. This approach is how we will implement Tierney's (1992) scale reduction factor convergence criteria which was mentioned earlier in Section 3.1.1.
We now need several definitions. For a stationary multivariate Markov chain \( \{ \beta^i, i \geq 1 \} \), let

\[
G_k \equiv \frac{E \left[ \{ \beta^1 - \mu \} \{ \beta^{1+k} - \mu \}^T \right] + E \left[ \{ \beta^{1+k} - \mu \} \{ \beta^1 - \mu \}^T \right]}{2}
\]

be the symmetrized \( k \)'th autocovariance, where \( \mu = E [ \beta^1 ] \) and \( k \geq 0 \); and define \( \Gamma_k \equiv G_{2k} + G_{2k+1} \), also for \( k \geq 0 \). Also, for a subchain of length \( m \), define the estimators

\[
\tilde{G}_k \equiv \frac{\sum_{i=1}^{m-k} \left\{ \beta^i - \hat{\mu} \right\} \left\{ \beta^{i+k} - \hat{\mu} \right\}^T + \left\{ \beta^{i+k} - \hat{\mu} \right\} \left\{ \beta^i - \hat{\mu} \right\}^T}{2m},
\]

where \( \hat{\mu} = m^{-1} \sum_{i=1}^{m} \beta^i \); and \( \tilde{\Gamma}_k = \tilde{G}_{2k} + \tilde{G}_{2k+1} \). For a real symmetric matrix \( A \), let \( \lambda_{\text{min}} (A) \) denote the smallest eigenvalue of \( A \).

The Monte Carlo error estimator we will use, the first of three estimators given in Kosorok (1998), has the form

\[
\tilde{H}_1 \equiv -\tilde{G}_0 + 2 \sum_{i=0}^{m^*_1-1} \tilde{\Gamma}_i,
\]

where

\[
m^*_1 \equiv \inf \left\{ k \geq 1 : \lambda_{\text{min}} (\tilde{\Gamma}_k) \leq 0 \right\}.
\]

Kosorok (1998) points out that this estimator is asymptotically conservative for \( H_0 \), in the sense that for any real vector \( v \) of the same dimension as the number of columns of \( H_0 \),

\[
\liminf_{n \to \infty} v^T \tilde{H}_1 v \geq v^T H_0 v.
\]

As Geyer (1992) points out, the motivation for using a (possibly) upward biased variance estimator is that in many applications, overestimates are less of a problem than underestimates. This is certainly true in our setting since we are using the variance estimator to
help us determine when our chain is sufficiently long. Geyer also provides a comparison of these variance estimators with other estimators (e.g., window or batched means estimators) in the univariate Markov chain setting.

4. A WORKED EXAMPLE

We will now present a worked example to illustrate the potential problems resulting from not using the proposed robust variance estimator when using maximum likelihood. This example will also serve to demonstrate how the assumptions needed for the asymptotic theory to work out can hold in practice. Suppose we wish to estimate the parameters of a simple linear regression model for data consisting of \( n \) clusters with \( m \) observations in each cluster, where each observation has real valued outcome \( Y_{ij} \) and predictor \( X_{ij} \) measurements, \( j = 1 \ldots m \), \( i = 1 \ldots n \). We will use the following random effects model for estimation of \( Y_{i1} \ldots Y_{im} \) given \( X_i \equiv (X_{i1} \ldots X_{im})^T \):

\[
Y_{ij} | X_i = a + bX_{ij} + \gamma \theta_i + \epsilon_{ij},
\]

where \( \theta_i \sim N(0, 1) \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \) are independent and identically distributed, \( j = 1 \ldots m \), \( i = 1 \ldots n \); \( \gamma \geq 0 \); and \( \sigma \geq 0 \). Thus we are modelling the cluster effect with a random intercept term.

Let \( \phi(z) \) be the standard normal density and denote \( \tilde{\beta} = (a, b, \gamma, \sigma)^T \). Then

\[
\ell_i(\tilde{\beta}) = \log \left[ \int_{\mathbb{R}} \exp \left( -m \log \sigma - \sum_{j=1}^{m} \left( \frac{Y_{ij} - a - bX_{ij} - \gamma \theta_j}{2\sigma^2} \right)^2 \right) \phi(\theta) d\theta \right]
\]

\[
= -m \log \sigma - \sum_{j=1}^{m} \frac{(Y_{ij} - a - bX_{ij})^2}{2\sigma^2}
\]

\[
+ \frac{m^2 \gamma^2}{2\sigma^4} \left( 1 + \frac{m \gamma^2}{\sigma^2} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - b\bar{X}_i)^2 \right) - \frac{1}{2} \log \left( 1 + \frac{m \gamma^2}{\sigma^2} \right), \tag{5}
\]
where $\bar{Y}_i = m^{-1} \sum_{j=1}^{m} Y_{ij}$ and $\bar{X}_i = m^{-1} \sum_{j=1}^{m} X_{ij}$. If we set $r = m \gamma^2 (\sigma^2 + m \gamma^2)^{-1}$, (5) becomes

$$
\ell_i(\beta) = -m \log \sigma - \frac{\sum_{j=1}^{m} (Y_{ij} - a - bX_{ij})^2}{2\sigma^2} + \frac{mr (Y_i - b\bar{X}_i)^2}{2\sigma^2} - \frac{1}{2} \log (1 - r),
$$

(6)

where we have reparameterized $\tilde{\beta}$ to become $\beta = (a, b, r, \sigma)^T$. The above likelihood is the "working likelihood" simply because we will be using it for inference.

The score for (6) now becomes

$$
s_i(\beta) = \begin{bmatrix}
-\frac{mr}{\sigma^2} \left( \frac{1}{\bar{X}_i} \right) (Y_i - a - b\bar{X}_i) + \sum_{j=1}^{m} \left( \frac{1}{X_{ij}} \right) (Y_{ij} - a - bX_{ij}) \\
\frac{m (Y_i - a - b\bar{X}_i)^2}{2\sigma^2} - \frac{1}{2(1 - r)} \sum_{j=1}^{m} (Y_{ij} - a - bX_{ij})^2
\end{bmatrix},
$$

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

$$
\sigma^{-2} \begin{bmatrix}
-\frac{mr}{\bar{X}_i} + \sum_{j=1}^{m} \left( \frac{1}{X_{ij}} \right) X_{ij} \\
m (Y_i - a - b\bar{X}_i) (1, \bar{X}_i) \\
2\sigma^{-1} \left\{ -mr (Y_i - a - b\bar{X}_i) (1, \bar{X}_i) + \sum_{j=1}^{m} (Y_{ij} - a - bX_{ij}) (1, X_{ij}) \right\}
\end{bmatrix},
$$

the lower-right $2 \times 2$ block being

$$
\begin{bmatrix}
\frac{1}{2(1 - r)^2} & \frac{m (Y_i - a - b\bar{X}_i)^2}{\sigma^2} \\
\frac{m (Y_i - a - b\bar{X}_i)^2}{\sigma^2} - \frac{mr (Y_i - a - b\bar{X}_i)^2}{\sigma^2} + \sum_{j=1}^{m} (Y_{ij} - a - bX_{ij})^2
\end{bmatrix}.
$$
and the remaining components defined by symmetry. For the Kulback-Leibler minimizer $\beta_0$, $E_s_i(\beta_0) = 0$; thus, if we let $E_1 = m^{-1} \sum_{j=1}^m (Y_{ij} - a_0 - b_0 X_{ij})^2$ and $E_2 = E (\bar{Y}_i - a_0 - b_0 \bar{X}_i)^2$, then by solving the score equation we obtain $\sigma_0^2 = (1 - 1/m)^{-1} (E_1 - E_2)$ and $r_0 = 1 - \sigma_0^2 / (mE_2)$.

Although we are using the forgoing likelihood for inference, suppose that the data is actually generated as follows: $X_{ij}$ are i.i.d. $N(0, \phi^2)$ and $Y_{i1}, \ldots, Y_{im}$ given $X_{i1}, \ldots, X_{im}$ has the form

$$Y_{ij} = a_* + U_i b_* X_{ij} + \gamma_* \theta_i + \eta_{ij},$$

where $U_i$ are i.i.d. mean 1 variance $\tau^2$ random variables, $\theta_i$ are i.i.d. $N(0, 1)$, and $\eta_{ij}$ are i.i.d. $N(0, \sigma_i^2)$, $j = 1 \ldots m$, $i = 1 \ldots n$. We will now explore the behavior of the working likelihood estimators of $\beta$.

To begin with, $E_s_i(\beta_0) = 0$ implies that

$$E \left[ \left\{ -mr \left( \frac{1}{X_i} \right) (\bar{Y}_i - a_0 - b_0 \bar{X}_i) + \sum_{j=1}^m \left( \frac{1}{X_{ij}} \right) (Y_{ij} - a_0 - b_0 X_{ij}) \bigg| X_i \right\} \right] = 0$$

which implies that

$$E \left[ \left\{ -mr_0 \left( \frac{1}{X_i} \right) (a_* + U_i b_* \bar{X}_i - a_0 - b_0 \bar{X}_i + \gamma_* \theta_i + \bar{\eta}_i) \right. \right.$$

$$+ \sum_{j=1}^m \left( \frac{1}{X_{ij}} \right) (a_* + U_i b_* X_{ij} - a_0 - b_0 X_{ij} + \gamma_* \theta_i + \eta_{ij}) \bigg| X_i \right\} \right] = 0, \quad (7)$$

where $\bar{\eta}_i = m^{-1} \sum_{j=1}^m \eta_{ij}$. But (7) implies

$$E \left[ -mr_0 \left( \frac{1}{X_i} \frac{\bar{X}_i}{X_i^2} \right) + \sum_{j=1}^m \left( \frac{1}{X_{ij}} \frac{X_{ij}}{X_{ij}^2} \right) \right] \left( \frac{a_* - a_0}{b_* - b_0} \right) = 0.$$
which then implies \( a_0 = a_* \) and \( b_0 = b_* \) by positive definiteness of the left-hand side.

Consequently,

\[
E_3 = E \left[ \left( m^{-1} \sum_{j=1}^{m} ((U_i - 1)b_0X_{ij} + \gamma_i\theta_i + \eta_i)^2 \right| X_i \right)
\]

\[
= b_0^2 \tau^2 \phi^2 + \gamma_*^2 + \sigma_*^2
\]

and

\[
E_2 = E \left[ \left( \left( \left( U_i - 1 \right) b_0 \bar{X}_i + \gamma_i \theta_i + \bar{\eta}_i \right)^2 \right| X_i \right)
\]

\[
= \frac{b_0^2 \tau^2 \phi^2}{m} + \gamma_*^2.
\]

Thus \( \sigma_0^2 = \sigma_*^2 + b_0^2 \tau^2 \phi^2 \) and

\[
r_0 = \frac{m \gamma_*^2}{\sigma_0^2 + m \gamma_*^2},
\]

which implies \( \gamma_0^2 = \gamma_*^2 \).

We also obtain

\[
EI_t(\beta_0) = \sigma_0^{-2} \begin{bmatrix}
m(1 - r_0) & 0 & 0 & 0 \\
0 & (m - r_0) \phi^2 & 0 & 0 \\
0 & 0 & \frac{1}{2(1 - r_0)} & \frac{1}{\sigma_0(1 - r_0)} \\
0 & 0 & \frac{1}{\sigma_0(1 - r_0)} & \frac{2(m - r_0)}{\sigma_0^2}
\end{bmatrix},
\]

since

\[
- \frac{m}{\sigma_0^2} + 3 \times \frac{mE_1 - m r_0 E_2}{\sigma_0^4} = - \frac{m}{\sigma_*^2} + 3 \times \frac{(m - r_0) \sigma_*^2 + (1 - r_0) m \gamma_*^2}{\sigma_*^4}
\]

\[
= \frac{2(m - r_0)}{\sigma_*^2}.
\]
Clearly, $EI_i(\beta_0)$ is bounded and positive definite whenever $0 < \phi^2 < \infty$, $|b_*| < \infty$, $0 \leq \tau^2 < \infty$, $0 < \sigma_*^2 < \infty$, $0 < \gamma_*^2 < \infty$ (implying that $0 < r_0 < 1$), and $m \geq 2$. It is also clear that $EI_i(\beta)$ is continuous in a convex compact neighborhood $B$ of $\beta_0$ and therefore that condition (L6) is satisfied. Clearly, conditions (L1) and (L2) are satisfied, and it is not hard to show that the first three derivatives of $\ell_i(\beta)$ exist on $B$ and hence that conditions (L3) and (L4) are readily satisfied. If we restrict $\beta$ to $B$, (L5) is also satisfied. We are thus tacitly using a uniform prior restricted to $B$, and hence conditions (B1), (B2), and (B3) are satisfied along with (B3'). By Lemma 1, condition (U2) is also satisfied and thus Theorems 1, 2, and 3, as well as Corollary 1 all follow.

It is clear at this juncture that the working maximum likelihood estimators for $a_0$, $b_0$, and $\gamma_0$ are consistent for $a_*$, $b_*$, and $\gamma_*$ but that the estimator for $\sigma_*^2$ will not be consistent for $\sigma_*^2$. We will now turn our attention to estimating the variance of $\sqrt{n}(\hat{b} - b_0)$ to explore the relationship between the naive working information based variance and the proposed robust variance. If we let $[A]_{kl}$ denote the $k,l$'th element of the matrix $A$, then the naive variance (based on the working likelihood information) of $\sqrt{n}(\hat{b} - b_0)$ is

$$\left[\{EI_i(\beta_0)\}^{-1}\right]_{22} = \frac{\sigma_0^2}{(m - r_0)\phi^2} \equiv \zeta_0.$$

To get the robust (correct) variance, we need to determine $E[s_i(\beta_0)]_2^2$, where $[a]_k$ is the $k$'th element of the vector $a$:

$$\sigma_0^4 E[s_i(\beta_0)]_2^2$$

$$= E \left[ -m r_0 \left( \bar{Y}_i - a_0 - b_0 \bar{X}_i \right) \bar{X}_i + \sum_{j=1}^m (Y_{ij} - a_0 - b_0 X_{ij}) X_{ij} \right]^2$$
\[
E \left[ E \left\{ \left[ -mr_0 \left( (u_i - 1)b_0 \bar{X}_i + \gamma_0 \theta_i + \eta_i \right) \bar{X}_i + \sum_{j=1}^m (u_i - 1)b_0 X_{ij} \gamma_0 \theta_{ij} + \eta_{ij} X_{ij} \right]^2 \right\} \mid X_i \right] \right]
\]

\[
= E \left[ m^2 (1 - r_0)^2 \gamma_0^2 \bar{X}_i^2 + b_0^2 \tau^2 \left\{ m(1 - r_0) \bar{X}_i^2 + \sum_{i=1}^m \left( X_{ij} - \bar{X}_i \right)^2 \right\}^2 + \sum_{j=1}^m \left( X_{ij} - r_0 \bar{X}_i \right)^2 \sigma^2_\psi \right]
\]

\[
= b_0^2 \tau^2 \left( m^2 - 1 + 3(1 - r_0)^2 \right) \phi^4 + m(1 - r_0)^2 \gamma_0^2 \phi^2 + [m - 1 + (1 - r_0)^2] \sigma^2_\psi \phi^2
\]

\[
= (m^2 - 1 + 3(1 - r_0)^2) b_0^2 \tau^2 \phi^4 + r_0(1 - r_0) \phi^2 \sigma^2_\psi + (1 - r_0)^2 \phi^2 \sigma^2_\psi + (m - 1) \phi^2 \sigma^2_\psi
\]

\[
= (m^2 - 1 + 3(1 - r_0)^2) b_0^2 \tau^2 \phi^4 + (m - r_0) \phi^2 \sigma^2_\psi + r_0(1 - r_0) \phi^2 (\sigma^2_\psi - \sigma^2_\psi),
\]

which implies

\[
\left[ \{ EI_i(\beta_0) \}^{-1} \right]_{22} E [ s_i(\beta_0) ]^2 \left[ \{ EI_i(\beta_0) \}^{-1} \right]_{22}
\]

\[
= \frac{[m^2 - 1 + 3(1 - r_0)^2] b_0^2 \tau^2}{(m - r_0)^2} + \frac{(m - r_0) \sigma^2_\psi}{(m - r_0)^2 \phi^2} + \frac{r_0(1 - r_0)(\sigma^2_\psi - \sigma^2_\psi)}{(m - r_0)^2 \phi^2}
\]

\[
= \frac{[m^2 + (2r_0 - 1)(r_0 - 2)] b_0^2 \tau^2}{(m - r_0)^2} + \frac{\sigma^2_\psi}{(m - r_0)^2 \phi^2}. \tag{8}
\]

If we let \( \delta = b_0^2 \tau^2 \phi^2 / \sigma^2_\psi \), then \( 0 \leq \delta < 1 \) and (8) can be rewritten as

\[
\psi_0 = \frac{[m^2 + (2r_0 - 1)(r_0 - 2)] \sigma^2_\psi \delta}{(m - r_0)^2 \phi^2} + \frac{(1 - \delta) \sigma^2_\psi}{(m - r_0)^2 \phi^2}
\]

\[
= \frac{\sigma^2_\psi}{(m - r_0)^2 \phi^2} \left[ 1 + \delta \times \frac{m^2 - m + 2(1 - r_0)^2}{m - r_0} \right]
\]

\[
= \zeta_0 \left[ 1 + \delta \times \frac{m^2 - m + 2(1 - r_0)^2}{m - r_0} \right].
\]

Thus \( \psi_0 \) can be made to be arbitrarily greater than the naive likelihood estimator \( \zeta_0 \) by allowing the cluster size to increase. This happens inspite of the fact that the working likelihood estimator is consistent for \( a_* \), \( b_* \), and \( \gamma_* \) (although it is not consistent for \( \sigma^2_\psi \)).
5. AN ANALYSIS OF LONGITUDINAL DISEASE STATUS DATA

An example is given involving longitudinal data for an ordinal variable describing global disease status for patients with arthritis. The data were obtained from the study *Pharmaceutical Care Outcomes: The Patient's Role* (PCOPR) described in Chewning et al (1998).

Our analysis includes data from 576 men and women with either osteoarthritis or rheumatoid arthritis. Clinical and quality of life variables were measured at six-month intervals for a maximum follow-up period of three years. The outcome variable of interest in the analysis is a scale of global arthritis severity with three ordered categories which can be thought of as 1=severe, 2=moderate and 3=mild. These longitudinally collected measurements of global disease severity will be modeled as depending on the covariates: time since being diagnosed with arthritis, sex, type of arthritis, and an index of the total number of swollen and painful joints.

5.1 Description of the Model

At six-month intervals, patients rated the severity of their arthritis on a Likert scale. Let $Y_{ij}$ denote the rating given by the $i$th subject at the $j$th follow-up appointment. Time-independent covariates included in the model are sex (0=male, 1=female), and type of arthritis (0=rheumatoid arthritis, 1=osteoarthritis). Time-dependent covariates are the number of years since being diagnosed with arthritis, scaled in units of decades, and the Ritchie articular index for the total number of swollen and painful joints (Ritchie, et al, 1968). This index is an enumeration of tender joints and has been rescaled to have a theoretical minimum of 0 and a theoretical maximum of 1.

A proportional odds model is used for the ordinal observations, with a random intercept to account for the statistical dependence of repeated measurements taken on a single patient.
over time. Specifically, the logit of the probability of responding at or below a particular response category is given by

\[
\log\left[\frac{P(Y_{ij} \leq k)}{1 - P(Y_{ij} \leq k)}\right] = \alpha_k + \gamma \theta_i + \omega_1 \text{sex}_i + \omega_2 \text{type}_i + \omega_3 \text{time}_{ij} + \omega_4 \text{joint}_{ij}
\]

for \(1 \leq k \leq 3\), where \(\gamma > 0\) and \(\alpha_1 < \alpha_2 < \alpha_3 = \infty\). The latent random variable \(\theta\) is assumed to have a standard normal distribution, and the standard deviation of the random intercept is parametrized by \(\gamma\). The prior measure for the structural parameters of the model, \(\pi(\cdot)\), is set equal to Lebesgue measure on the set where the intercept parameters are properly ordered, and is equal to 0 outside of this set.

5.2 Implementation of the Metropolis-Hastings Algorithm

The latent variables and the structural parameters of the model were divided into four blocks; \(\theta\), \(\alpha\), \(\omega\), and \(\gamma\). The initial value \(\theta^0\) was obtained by drawing \(n\) independent standard normal random variates. The initial value \(\omega^0\) was set equal to 0 and \(\gamma^0\) was set equal to 1. The intercept parameters, \(\alpha_1^0\) and \(\alpha_3^0\), were chosen to fit the marginal response probabilities for the three options, given the values of the other parameters.

Let \(\theta^j\), \(\alpha^j\), \(\omega^j\), and \(\gamma^j\) denote the elements of the Markov chain after the \(j\)th iteration. These were updated in four separate blocks. The proposed value of \(\omega^{j+1}\) was obtained by adding independent Gaussian noise to each component of the current value with standard deviations tuned to yield an acceptance rate of nearly 1 out of 4 proposals. The proposed value of \(\gamma^{j+1}\) was obtained by drawing a random variate from a log-normal distribution with location parameter \(\log(\gamma^j)\), and a scale parameter chosen to maintain the acceptance rate of approximately 0.25. Thus, the Hastings adjustment to the Metropolis algorithm was needed for this asymmetric proposal distribution. The proposal for \(\alpha^{j+1}\) was generated by
adding independent Gaussian noise to each component of $\alpha^j$, and the prior measure $\pi(\cdot)$ was used to ensure that only proposals which satisfied the order constraint had a positive probability of acceptance. Components of $\theta$ were updated cluster-by-cluster by generating proposals from Gaussian distributions centered at the current values.

This process was carried out using a burn-in phase of 5,000 iterations and an estimation phase of 20,000 iterations. The length of the estimation phase was chosen to satisfy the property that the Monte Carlo variance would be less than 1 percent of the estimated variance of the posterior mean. Because our intention was to make certain that the Monte Carlo variance would be negligible, rather than to adjust for it, we used the conservative estimator $\hat{H}_1$ of Kosorok (1998). For this model, the latent variable density was simply the probability density function of a standard normal random variable, and a sample of $\nu = 1000$ random variates could easily be generated without using MCMC. Two separate covariance matrices described in section 3 were computed and will be denoted $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. Suppressing the superscripts and subscripts on $\hat{W}$ and $\hat{V}$, these two covariance matrices are the naive estimator $\hat{\Sigma}_1 = \frac{1}{n} \hat{W}$ and the robust estimator $\hat{\Sigma}_2 = \frac{1}{n} \hat{W} \hat{V} \hat{W}$.

5.3 Results

The parameter estimates and standard errors are summarized in Table 1. Here it is seen that patients with high Ritchie joint counts have much greater odds of having more severe arthritis. The same is true of patients with osteoarthritis and patients whom were diagnosed long ago. Note that the standard error estimates obtained from $\hat{\Sigma}_1$ are generally smaller but similar to those obtained from the robust estimator $\hat{\Sigma}_2$. One significant exception is the standard error estimate for $\omega_4$, the coefficient for the Ritchie swollen joint index. The robust standard error is about 50 percent larger in this case. The ratios of the estimated Monte Carlo variances to the variances of the posterior means for the parameters suggest...
that 20,000 iterations was long enough to consider the Monte Carlo error negligible. For this Markov chain we had \( m_1 = 17 \) so that 17 terms \((\overline{F}_0, \overline{F}_2, \ldots, \overline{F}_{16})\), were used to compute \( \overline{H}_1 \), utilizing sample autocovariance matrices of lags up to and including 33. In order to investigate the normal approximation to the normalized working likelihood function, kernel density estimates of the marginal normalized working likelihood functions were computed for each parameter and are displayed in Figure 1. These indicate that a normal approximation to the working likelihood surface might be close, although multivariate diagnostics might be needed for a more rigorous justification. Condition \( (M^2) \) is examined in Figure 2 by inspecting a quantile-quantile plot of \[ n \left\{ \beta^j - \overline{\beta} \right\}^T \left( W^{-1} \right) \left( \beta^j - \overline{\beta} \right) \] and the quantiles of a chi-squared random variable with 7 degrees of freedom and raised to the 3/2 power. The linearity of the plot suggests that these two distributions approximately agree. The computer time required for this analysis was approximately 29 minutes on a SUN Ultra 2 Systems Model 2200 multiprocessor.

6. DISCUSSION

The worked example based on a simple linear regression model with random effects demonstrates that the difference between the naive (non-robust) ML variance estimator and the robust MWL variance estimator can be quite large. Although the difference between these variance estimates is not large for most of the parameters considered in the arthritis data example, the difference is in fact substantial for the parameter \( \omega_4 \) (see Table 1). Given this, it may be prudent in large studies to use robust methods as a general rule, at least for diagnostic purposes. The computation time required for this method is primarily dictated by executing the MHA described in Section 3.1.1. However, the overall computational intensity will be slightly greater than that of a general application of Markov chain Monte Carlo
because of an additional Monte Carlo step involved in computing the marginal information and marginal score functions given in Section 3.1.2.

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.

**APPENDIX: PROOFS**

We postpone the proof of Lemma 1 until after proving Theorem 2, and present first the following lemma which will be needed for the proof of Theorem 1:

**Lemma 2.** Suppose conditions (L1)–(L4) and (B1) and (B2) hold, then, for every compact \( T \subset B \),

\[
\lim_{n \to \infty} \sup_{\beta \in T} \left| n^{-1} L_n(\beta) - L(\beta) \right| = 0 \tag{A.1}
\]

almost surely and

\[
\lim_{n \to \infty} \sup_{\beta \in T} \left| n^{-1} \sum_{i=1}^{n} I_i(\beta) - E[ I(\beta) ] \right| = 0 \tag{A.2}
\]

almost surely.

**Proof.** This lemma follows from standard probability arguments, but an outline of the proof is included here for completeness. Notice that for any \( \beta_1, \beta_2 \in T \),

\[
\left| n^{-1} L_n(\beta_1) - n^{-1} L_n(\beta_2) \right| \leq \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} s_i (t\beta_1 + (1-t)\beta_2) \right| \cdot |\beta_1 - \beta_2|
\]

\[
\leq \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i) \right| \cdot |\beta_1 - \beta_2|
\]
by the mean value inequality followed by condition (L4). (A.1) now follows by the strong law of large numbers. (A.2) can be proved in a similar fashion. □

Proof of Theorem 1. ∀ε, η > 0, proceed as follows. The absolute continuity of L(·) and condition (L2) imply that ∃ε₀, ε₁, δ > 0 such that ε > ε₁ > ε₀, and, for any convex compact T ⊂ B such that β₀ ∈ T and

\[ \sup_{β ∈ T} |β - β₀| > ε₁, \]  \hspace{1cm} (A.3)

then

\[ \inf_{β ∈ T: |β - β₀| < ε₀} |L(β)| > \sup_{β ∈ T: |β - β₀| > ε₁} + δ. \]

Condition (L5) implies that ∃n₁ and a convex compact set T* such that β₀ ∈ T* ⊂ B and T* satisfies (A.3), and, ∀n ≥ n₁, \( P \left( \hat{β}_n ∈ T^* \right) > 1 - η/2. \) Lemma 3 implies that there also exists an \( n_2 ≥ n₁ \) such that ∀n ≥ n₂,

\[ P \left[ \sup_{β ∈ T^*} \left| n^{-1}L_n(β) - L(β) \right| > δ/3 \right] < η/2. \]

Therefore, with probability greater than 1 − η/2 and for every n ≥ n₂,

\[ \inf_{β ∈ T^*: |β - β₀| < ε₀} \left| n^{-1}L_n(β) \right| = \inf_{β ∈ T^*: |β - β₀| < ε₀} \left| L(β) + \left\{ n^{-1}L_n(β) - L(β) \right\} \right| \]

\[ ≥ \inf_{β ∈ T^*: |β - β₀| < ε₀} \left| L(β) \right| - \sup_{β ∈ T^*} \left| n^{-1}L_n(β) - L(β) \right| \]

\[ ≥ \sup_{β ∈ T^*: |β - β₀| > ε₁} \left| L(β) \right| + δ - \sup_{β ∈ T^*} \left| n^{-1}L_n(β) - L(β) \right| \]

\[ ≥ \sup_{β ∈ T^*: |β - β₀| > ε₁} \left| L(β) \right| + \sup_{β ∈ T^*} \left| n^{-1}L_n(β) - L(β) \right| \]

\[ ≥ \sup_{β ∈ T^*: |β - β₀| > ε₁} \left| n^{-1}L_n(β) \right|. \]

Notice that this inequality is a sufficient condition for \( \left| \hat{β}_n - β₀ \right| < ε \), when \( \hat{β}_n ∈ T^* \). Hence

\[ P \left[ \left| \hat{β}_n - β₀ \right| < ε \right] > 1 - η. \] □
Proof of Theorem 2. First observe that, for \( d > 0 \), (U') combined with (B2) implies that

\[
\lim_{c \to \infty} \limsup_{n \to \infty} \sqrt{n} \int_{\beta \in B} \left| \sqrt{n}(\beta - \hat{\beta}_n) \right|^d \exp \left\{ L_n(\beta) - L_n(\hat{\beta}_n) \right\} \frac{\pi(\beta)}{\pi(\beta_0)} d\beta = 0.
\]

(A.4)

Note that the density associated with our Markov chain is

\[
\frac{\exp \left\{ L_n(\beta; \theta) \right\} \pi(\beta)}{\int_\beta \exp \left\{ L_n(b; t) \right\} \pi(b) \left\{ \prod_{i=1}^n \mu(dt_i) \right\} db},
\]

where \( t \equiv \{t_1, \ldots, t_n\}^T \). This density admits the marginal density

\[
\frac{\exp \left\{ L_n(\beta) \right\} \pi(\beta)}{\int_\beta \exp \left\{ L_n(b) \right\} \pi(b) db} = \frac{\sqrt{n} \exp \left\{ L_n(\beta) - L_n(\hat{\beta}_n) \right\} \pi(\beta)/\pi(\beta_0)}{\sqrt{n} \int_\beta \exp \left\{ L_n(b) - L_n(\hat{\beta}_n) \right\} \{\pi(b)/\pi(\beta_0)\} db}.
\]

By a Taylor's expansion,

\[
\sqrt{n} \exp \left\{ L_n(\beta) - L_n(\hat{\beta}_n) \right\} \pi(\beta)/\pi(\beta_0)
\]

\[= \sqrt{n} \exp \left\{ S_n^T(\hat{\beta}_n)(\beta - \hat{\beta}_n) - \frac{1}{2} \sqrt{n}(\beta - \hat{\beta}_n)^T \left[ \frac{L_n(\beta)}{n} \right] \sqrt{n}(\beta - \hat{\beta}_n) \right\} \frac{\pi(\beta)}{\pi(\beta_0)} \]

(A.5)

for some \( \beta^* \) on the line segment between \( \beta \) and \( \hat{\beta}_n \), and thus we have that \( \forall c < \infty \),

\[
\sqrt{n} \int_{\beta \in B, \left| \sqrt{n}(\beta - \hat{\beta}_n) \right| \leq c} \left| \sqrt{n}(\beta - \hat{\beta}_n) \right|^d \exp \left\{ L_n(\beta) - L_n(\hat{\beta}_n) \right\} \frac{\pi(\beta)}{\pi(\beta_0)} d\beta
\]

is asymptotically equivalent to

\[
\sqrt{n} \int_{\beta \in B, \left| \sqrt{n}(\beta - \hat{\beta}_n) \right| \leq c} \left| \sqrt{n}(\beta - \hat{\beta}_n) \right|^d \exp \left\{ -\frac{1}{2} \sqrt{n}(\beta - \hat{\beta}_n)^T I_0 \sqrt{n}(\beta - \hat{\beta}_n) \right\} \frac{\pi(\beta)}{\pi(\beta_0)} d\beta
\]

which, by (L5) and (B3) is asymptotically equivalent to

\[
\int_{u \leq c} |u|^d \exp \left\{ -\frac{1}{2} u^T I_0 u \right\} du,
\]
for finite $d > 0$. Since $c$ is arbitrary, we then have by (A.4) that

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

under condition (U1), and

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

under condition (U2), 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 Lemma 1. The assumptions for this lemma and the strong law give us that for some $n_0 < \infty$, that $\hat{\beta}_n \in T$ and $I_n(\beta) \geq C_0$ for all $\beta \in T$ almost surely, for all $n \geq n_0$. Combining this with the expansion given in (A.5), we obtain

$$\sqrt{n} \exp \left\{ L_n(\beta) - L_n(\hat{\beta}_n) \right\} \pi(\beta)/\pi(\beta_0) \leq M \sqrt{n} \exp \left\{ \frac{-1}{2} \sqrt{n}(\beta - \hat{\beta}_n)^T \left[ \frac{C_0}{2} \right] \sqrt{n}(\beta - \hat{\beta}_n) \right\},$$

for all $n \geq n_0$ almost surely, and thus condition (Md) holds for $0 \leq d < \infty$. □

Proof of Theorem 3. For each $n \geq 1$, and conditional on the data, we have by (L3) and (L4), and by the ergodic theorem, that $\forall \varepsilon > 0, \exists n, < \infty$ such that both

$$P \left[ n^{-1} \sum_{i=1}^{n} \bar{s}_i(\hat{\beta}) \bar{s}_i^T(\hat{\beta}) - n^{-1} \sum_{i=1}^{n} s_i(\hat{\beta}) s_i^T(\hat{\beta}) \right] > \varepsilon = 0$$

and

$$P \left[ n^{-1} \sum_{i=1}^{n} \bar{D}_i(\hat{\beta}) - n^{-1} \sum_{i=1}^{n} D_i(\hat{\beta}) \right] > \varepsilon = 0$$

for all $m \geq m_n$, where the probability is with respect to the Markov chain conditional on the data.

Standard arguments can now be used to show that conditions (L1)–(L4) and the consistency of $\hat{\beta}$ yield the desired results. □
Proof of corollary 1. Taking a Taylor's expansion about \( \tilde{\beta}_n \), we obtain for every direction vector \( u \in \mathbb{R}^p \), that \( S_n(\tilde{\beta}_n) + S_\pi(\tilde{\beta}_n) = 0 \) implies

\[
\frac{u^T \{ S_n(\beta_0) + S_\pi(\beta_0) \}}{\sqrt{n}} = u^T \left\{ \frac{I_n(\beta^*) + I_\pi(\beta^*)}{n} \right\} \sqrt{n}(\tilde{\beta}_n - \beta_0),
\]

where \( \beta^* \) is on the line segment between \( \beta_0 \) and \( \tilde{\beta}_n \); and the results follow provided that \( \tilde{\beta}_n \) is consistent for \( \beta_0 \). After noting that for every convex compact \( T \subset B \),

\[
\sup_{\beta \in T} \left| n^{-1} \{ L_n(\beta) + \log \pi(\beta) \} - L(\beta) \right|
\]

\[
\leq \sup_{\beta \in T} \left| n^{-1} L_n(\beta) - L(\beta) \right| + \left| \log \frac{\pi(\beta_0)}{n} \right| + \sup_{\beta \in T} \left| \log \frac{\pi(\beta)}{\pi(\beta_0)} \right|
\]

\[
\to 0,
\]

as \( n \to \infty \), by lemma 3 and (B2), we can then use arguments similar to those used in Theorem 1 to obtain the desired results. \( \square \)
REFERENCES


Table 1: Modes are based on kernel density estimates of the marginal normalized working likelihood functions for each parameter using the 20,000 elements of the Markov chain. The standard error estimates, $\hat{\sigma}_i$, are defined as the square root of the diagonal elements of $\tilde{\Sigma}_i$. The ratio $\sigma^2_{\tilde{\sigma}}/\sigma^2_i$, is the ratio of the estimated Monte Carlo variance and posterior mean variance.

<table>
<thead>
<tr>
<th>parameter</th>
<th>mean ($\hat{\theta}_{576}$)</th>
<th>mode</th>
<th>$\hat{\sigma}_1$</th>
<th>$\hat{\sigma}_2$</th>
<th>$\sigma^2_{\tilde{\sigma}}/\sigma^2_i$</th>
<th>robust t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>-2.305</td>
<td>-2.315</td>
<td>0.252</td>
<td>0.264</td>
<td>0.004</td>
<td>-8.731</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.379</td>
<td>0.355</td>
<td>0.234</td>
<td>0.237</td>
<td>0.004</td>
<td>1.599</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>-0.353</td>
<td>-0.303</td>
<td>0.208</td>
<td>0.203</td>
<td>0.004</td>
<td>-1.738</td>
</tr>
<tr>
<td>$\omega_2$</td>
<td>1.341</td>
<td>1.334</td>
<td>0.226</td>
<td>0.227</td>
<td>0.004</td>
<td>5.022</td>
</tr>
<tr>
<td>$\omega_3$</td>
<td>0.312</td>
<td>0.314</td>
<td>0.092</td>
<td>0.092</td>
<td>0.003</td>
<td>3.391</td>
</tr>
<tr>
<td>$\omega_4$</td>
<td>5.740</td>
<td>5.460</td>
<td>0.588</td>
<td>0.580</td>
<td>0.004</td>
<td>6.074</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>1.936</td>
<td>1.919</td>
<td>0.123</td>
<td>0.137</td>
<td>0.002</td>
<td>14.13</td>
</tr>
</tbody>
</table>
Figure 1: Kernel density estimates of the marginal normalized working likelihood functions of a) $\alpha_1$, b) $\alpha_2$, c) $\omega_1$, d) $\omega_2$, e) $\omega_3$, f) $\omega_4$, and g) $\gamma$. Estimates were computed from the 20,000 elements of the Markov chain, using a Gaussian kernel with a bandwidth equal to one-half the sample standard deviation of the elements of the chain corresponding to the parameter.
Figure 2: QQ plot of the sample values of \( n \{ \beta^j - \bar{\beta} \}^T (W)^{-1} \{ \beta^j - \bar{\beta} \} \)^{3/2} and the quantiles of a chi-squared random variable with 7 degrees of freedom and raised to the 3/2 power.