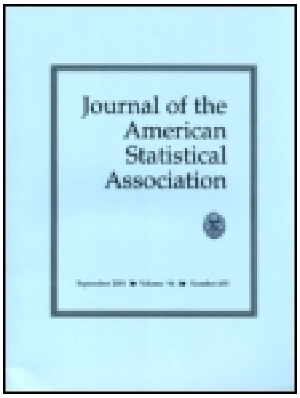

Journal of the American Statistical Association

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Implementation of Estimating Function-Based Inference Procedures With Markov Chain Monte Carlo Samplers

Lu Tian, Jun S. Liu and L. J. Wei

Lu Tian is Assistant Professor, Department of Preventive Medicine, Northwestern University, Chicago, IL 60611 . Jun S. Liu is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 . L. J. Wei is Professor, Department of Biostatistics, Harvard University, Boston, MA 02115 . The authors thank the referees, associate editor, and joint editors for constructive comments on the manuscript. This work was supported in part by grants from the National Science Foundation (DSM-02-44638, to Liu) and National Institutes of Health (R01-AI052817, to Wei). We are very grateful to Professor Chernozhukov and the joint editors for informing us that a similar approach was proposed by Chernozhukov and Han (2003) to handle our problem. The idea of using Monte Carlo methods for locating point estimates via estimating functions and making corresponding statistical inferences from the frequenstist's point of view can be traced back to the seminal work of Lin and Geyer (1992). Recently these methods have been successfylly utilized under various settings, for example, by He and Hu (2002), Lee, Kosorok, and Fine (2005), and Tian and Cai (2006). A rather unique feature of our proposal is the usage of the pivotal property of the estimating function to guide us for selecting realizations of the MCMC sampler as detailed in the present article.

Available online: 01 Jan 2012

Implementation of Estimating Function-Based Inference Procedures With Markov Chain Monte Carlo Samplers

Lu TIAN, Jun S. LIU, and L. J. WEI

Under a semiparametric or nonparametric setting, inferences about the unknown parameter are often made based on a nonsmooth estimating function. Resampling methods are quite handy for obtaining good approximations to the distribution of the consistent estimator when the estimating equation and its resampled counterparts are not difficult to solve numerically. In this article we propose a simple, flexible procedure that provides such approximations through the standard Markov chain Monte Carlo sampler without solving any equations. More generally, the procedure may locate all possible roots of the estimating equation and provides an approximation to the distribution of each root. We illustrate our proposed procedure extensively with three examples and evaluate its performance comprehensively through a simulation study.

KEY WORDS: Bootstrap; Median regression; Metropolis algorithm; Normal approximation; Resampling; Survival analysis.

1. INTRODUCTION

Under a nonparametric or semiparametric setting, inferences about a $p \times 1$ unknown vector $\boldsymbol{\theta}_0$ of parameters are often based on a p-dimensional estimating function $\tilde{\mathbf{S}}_X(\boldsymbol{\theta})$, where X is the observable random quantity with sample size n. Let $\hat{\boldsymbol{\theta}}_X$ be a consistent root to the equation

$$\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) \approx 0. \tag{1}$$

If the estimating function is locally linear around $\boldsymbol{\theta}_0$, and for large n, the distribution of $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}_0)$ can be well approximated by a mean-0 normal with covariance matrix $\boldsymbol{\Pi}_X(\boldsymbol{\theta}_0)$, then the random vector $W_X = n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$ is asymptotically normal. In general, the matrix $\boldsymbol{\Pi}_X(\boldsymbol{\theta})$ can be easily obtained, but the covariance matrix of W_X may be rather difficult to estimate well directly when $\tilde{\mathbf{S}}_X(\boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$ in the entire parameter space of interest. Note that

$$\mathbf{S}_X(\boldsymbol{\theta}_0) = \{\boldsymbol{\Pi}_X(\boldsymbol{\theta}_0)\}^{-1/2}\tilde{\mathbf{S}}_X(\boldsymbol{\theta}_0) \tag{2}$$

is asymptotically pivotal and is approximately $\mathrm{N}(0, \mathbf{I}_p)$-distributed, where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

The bootstrap method (Efron and Tibshirani 1993) is the standard resampling procedure, which provides a good approximation to the distribution of W_X. When the data X consist of n independent random quantities $\{X_1, \ldots, X_n\}$ and the estimating function is

$$n^{-1/2}\sum_{i=1}^{n}\tilde{\mathbf{S}}_{X_i}(\boldsymbol{\theta}), \tag{3}$$

where the random part of $\tilde{\mathbf{S}}_{X_i}(\cdot)$ depends on X_i only, then for large n, it has been shown that the bootstrap distribution centered by $\hat{\boldsymbol{\theta}}_X$ can closely approximate the distribution of $(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$ (Arcones and Gine 1992). Recently, Hu and Kalbfleisch (2000) proposed a novel estimating function bootstrap method based directly on (3).

Implementation of the bootstrapping can be problematic when the estimating equation $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) = 0$ and its bootstrap counterparts are difficult to solve numerically. For this case, one may use a "parametric bootstrap" method, which takes advantage of the pivotal feature of the estimating function $\mathbf{S}_X(\boldsymbol{\theta}_0)$, to approximate the distribution of W_X. To be specific, let x be the observed value of X and let $\boldsymbol{\theta}_x^*$ be a random vector such that

$$\mathbf{S}_x(\boldsymbol{\theta}_x^*) \approx \mathbf{Z}, \tag{4}$$

where $\mathbf{Z}$ is $\mathrm{N}(0, \mathbf{I}_p)$. If $\boldsymbol{\theta}_X^*$ is a consistent estimator for $\boldsymbol{\theta}_0$, then it follows from work of Parzen, Wei, and Ying (1994) that the distribution of W_X can be well approximated by the conditional distribution of $W_x^* = n^{1/2}(\boldsymbol{\theta}_x^* - \hat{\boldsymbol{\theta}}_x)$. This resampling method has been justified theoretically for a class of general estimating functions, which includes (3) as a special case. Moreover, realizations of $\boldsymbol{\theta}_x^*$ in (4) can be generated without solving any estimating equations, for example, through an adaptive importance sampling technique (Tian, Liu, Zhao, and Wei 2004).

In this article we propose a procedure through the standard Metropolis algorithm to generate the distribution of $\boldsymbol{\theta}_x^*$ without the need to solve (4). The procedure only involves computing $\mathbf{S}_x(\boldsymbol{\theta})$ and is more flexible to implement in practice than that proposed by Tian et al. (2004). Moreover, the new proposal may locate all possible roots of the estimating equation and provides an approximation to the distribution of each root. We illustrate the proposal extensively with three examples. We also conducted a comprehensive simulation study to examine the properties of the new procedure.

Recently, He and Hu (2002) proposed a novel Markov chain marginal bootstrap (MCMB) method to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_X$ based on a specific type of estimating function (3).

Lu Tian is Assistant Professor, Department of Preventive Medicine, Northwestern University, Chicago, IL 60611 (E-mail: *lutian@northwestern.edu*). Jun S. Liu is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *jliu@stat.harvard.edu*). L. J. Wei is Professor, Department of Biostatistics, Harvard University, Boston, MA 02115 (E-mail: *wei@hsph.harvard.edu*). The authors thank the referees, associate editor, and joint editors for constructive comments on the manuscript. This work was supported in part by grants from the National Science Foundation (DSM-02-44638, to Liu) and National Institutes of Health (R01-AI052817, to Wei). We are very grateful to Professor Chernozhukov and the joint editors for informing us that a similar approach was proposed by Chernozhukov and Han (2003) to handle our problem. The idea of using Monte Carlo methods for locating point estimates via estimating functions and making corresponding statistical inferences from the frequenstist's point of view can be traced back to the seminal work of Lin and Geyer (1992). Recently these methods have been successfylly utilized under various settings, for example, by He and Hu (2002), Lee, Kosorok, and Fine (2005), and Tian and Cai (2006). A rather unique feature of our proposal is the usage of the pivotal property of the estimating function to guide us for selecting realizations of the MCMC sampler as detailed in the present article.

© 2007 American Statistical Association
Journal of the American Statistical Association
September 2007, Vol. 102, No. 479, Theory and Methods
DOI 10.1198/016214506000000122

In addition, Lee, Kosorok, and Fine (2005) studied an intriguing stochastic numerical algorithm for the semiparametric profile likelihood estimation problem. These two procedures are discussed further in Section 5.

2. INFERENCES FOR $\boldsymbol{\theta}_0$ VIA THE METROPOLIS ALGORITHM

Note that if $\mathbf{S}_x(\boldsymbol{\theta})$ is a one-to-one mapping and differentiable in $\boldsymbol{\theta}$ for large n, then the density function of $\boldsymbol{\theta}_x^*$ defined in (4) is approximately proportional to

$$g(\boldsymbol{\theta}) = \exp\left\{-\frac{1}{2}\mathbf{S}_x'(\boldsymbol{\theta})\mathbf{S}_x(\boldsymbol{\theta})\right\}. \tag{5}$$

Here we show how to obtain a good approximation to the distribution of $\boldsymbol{\theta}_x^*$ through (5) even when $\mathbf{S}_x(\boldsymbol{\theta})$ is neither smooth nor a one-to-one function. First, suppose that there exists a consistent estimator $\boldsymbol{\theta}_X^\dagger$ for $\boldsymbol{\theta}_0$, that may be obtained from a relatively simple estimating function of $\boldsymbol{\theta}$. In the Appendix we show that if we can construct a random vector $\tilde{\boldsymbol{\theta}}_x$ whose realizations are generated from the density function proportional to (5) in a c_n neighborhood Ω_x of $\boldsymbol{\theta}_x^\dagger$, where $c_n^{-1} = o(n^{1/2})$ and $c_n = o(1)$, then for large n, the distribution of $\tilde{\boldsymbol{\theta}}_x$ is a good approximation to that of $\boldsymbol{\theta}_x^*$.

To generate realizations from $\tilde{\boldsymbol{\theta}}_x$, we use the standard Metropolis algorithm (Liu 2001). Toward this end, we construct a sequence $\{\boldsymbol{\theta}_{(k)}, k \geq 1\}$ with an initial value $\boldsymbol{\theta}_{(1)}$ such that for $k > 1$,

$$\boldsymbol{\theta}_{(k)} = \begin{cases} \boldsymbol{\theta}_{(k-1)} & \text{with probability } 1 - \tau_k \\ \mathbf{v} & \text{with probability } \tau_k, \end{cases}$$

where $\mathbf{v}$ is generated from $\mathrm{N}(\boldsymbol{\theta}_{(k-1)}, \boldsymbol{\Sigma}_x)$, $\boldsymbol{\Sigma}_X = O_p(n^{-1/2})$, a prespecified nonsingular $p \times p$ matrix, and $\tau_k = \min\{1, g(\mathbf{v})/g(\boldsymbol{\theta}_{(k-1)})\}$. Note that if $\mathbf{v}$ is not in Ω_x, then we let $\boldsymbol{\theta}_{(k)} = \boldsymbol{\theta}_{(k-1)}$. In theory, for large K and M, we expect the empirical distribution constructed from $\mathcal{J} = \{\boldsymbol{\theta}_{(K)}, \ldots, \boldsymbol{\theta}_{(K+M)}\}$ to be a good approximation to the distribution of $\tilde{\boldsymbol{\theta}}_x$. To be specific, the distribution of $\boldsymbol{\theta}_x^*$ can be approximated by a p-dimensional normal with mean $\hat{\boldsymbol{\theta}}_x$ and covariance matrix Λ_x. Here we let $\hat{\boldsymbol{\theta}}_x$ be the sample mean of $\boldsymbol{\theta}_{(k)}$ or the $\boldsymbol{\theta}_{(k)}$ that gives the smallest value of $\{\|\mathbf{S}_x(\boldsymbol{\theta}_{(k)})\|^2, k = K+1, \ldots, K+M\}$, and let Λ_x be the sample covariance matrix based on those M dependent $\boldsymbol{\theta}_{(k)}$'s in $\mathcal{J}$. Note that to obtain robust $\hat{\boldsymbol{\theta}}_x$ and Λ_x, we may delete outliers of the realizations in $\mathcal{J}$, as illustrated with an example in the next section.

In practice, the choices of the matrix $\boldsymbol{\Sigma}_x$ in the proposal distribution for the foregoing Markov chain, the neighborhood Ω_x, and K and M in the sequence $\mathcal{J}$ affect the efficiency of the algorithm. Suppose that the covariance matrix of $\boldsymbol{\theta}_X^\dagger$ can be estimated by $\boldsymbol{\Gamma}_X = (\gamma_{lm})$. Generally, we would expect the target covariance matrix Λ_X of $\boldsymbol{\theta}_X^*$ to not be drastically different from $\boldsymbol{\Gamma}_X$. Let θ_l and $\theta_{xl}^\dagger$ be the lth components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_x^\dagger$, $l = 1, \ldots, p$. Then we may choose

$$\Omega_x = \left\{\boldsymbol{\theta} : |\theta_l - \theta_{xl}^\dagger| \leq \Phi^{-1}(\alpha_n)\gamma_{ll}^{1/2}, l = 1, \ldots, p\right\}, \tag{6}$$

where $\Phi(\cdot)$ is the distribution function of the univariate standard normal and $1 - \alpha_n = O(n^{-r})$, for a given $r > 0$. Furthermore, if the covariance matrix Λ_x of the normal target distribution were known, then we would let $\boldsymbol{\Sigma}_x$ in the proposal distribution be proportional to Λ_x (Gelman, Carlin, Stern, and Rubin 2003, p. 306). Therefore, for our procedure, we can let $\boldsymbol{\Sigma}_x = c\boldsymbol{\Gamma}_x$ and choose c adaptively in a batch fashion until the acceptance rate of $\mathbf{v}$ in the Metropolis algorithm is about 25–50% (Liu 2001, p. 115). We then use the final value of c to generate the foregoing sequence $\mathcal{J}$. Furthermore, after discarding the first K iterations from the generated sequence to complete sampler "burn-in," we may examine the autocorrelations for the sequence $\mathcal{J}$ to estimate M based on a prespecified effective sample size (Liu 2001, pp. 125–126). Finally, we may examine whether the empirical distribution of the realizations $\{\mathbf{S}_x(\boldsymbol{\theta}_{(k)}), k = K+1, \ldots, K+M\}$ is close to $\mathrm{N}(0, \mathbf{I}_p)$. Note that the choice of the initial starting point $\boldsymbol{\theta}_{(1)}$ does not seem critical for implementing our procedure.

Because the estimating function may not be smooth, we do not expect $\mathbf{S}_x(\hat{\boldsymbol{\theta}}_x) = 0$. In theory, any $\boldsymbol{\theta}$ such that $\mathbf{S}_X(\boldsymbol{\theta}) = o_p(1)$ is a root to the estimating equation. Empirically, an objective way to evaluate whether the resulting $\hat{\boldsymbol{\theta}}_x$ from the foregoing search is a possible root is to use the metric $\|\mathbf{S}_x(\boldsymbol{\theta})\|^2$ to compare the observed value of $\|\mathbf{S}_x(\hat{\boldsymbol{\theta}}_x)\|^2$ with the distribution of $\|\mathbf{S}_X(\boldsymbol{\theta}_0)\|^2$, which is χ_p^2.

Now consider the case where there is no initial consistent estimate $\boldsymbol{\theta}_X^\dagger$ available and the estimating equation may have multiple roots whose limits are interior points of the parameter space. Then, under the locally linear condition for $\mathbf{S}_X(\boldsymbol{\theta})$ around the limit of each root, the distribution of $\boldsymbol{\theta}_x^*$ in (4) is approximately a mixture of normals. Each normal is centered around one of the roots, and one of these normals would be a good approximation to the distribution of $(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$. Under a semiparametric setting, to implement the foregoing iterative procedure, we may fit the data with a parametric submodel to obtain a point estimator for $\boldsymbol{\theta}_0$ and its estimated covariance matrix as the surrogates of $\boldsymbol{\theta}_x^\dagger$ and $\boldsymbol{\Gamma}_x$ to generate realizations from (5). In the absence of an initial consistent estimate, we suggest considering a large parameter space Ω_x by, for example, choosing a fairly large α_n in (6), to obtain a relatively complete profile of the distribution of $\boldsymbol{\theta}_x^*$ for making inferences about $\boldsymbol{\theta}_0$.

3. EXAMPLES

We use three examples to illustrate the new proposal. The first example is for a case where there exists a consistent estimator $\boldsymbol{\theta}_X^\dagger$ for $\boldsymbol{\theta}_0$. The second example illustrates the case where no initial consistent estimator is available, but for large n, the estimating equation, $\mathbf{S}_x(\boldsymbol{\theta}) = 0$, has a unique root. The third example shows what our procedure would generate through (5) for a case where asymptotically the estimating equation may have multiple roots.

We use a semiparametric survival median regression model to generate these three cases. Toward this end, let T_i be the ith failure time or a transformation thereof, and let $\mathbf{V}_i$ be the corresponding p-dimensional vector that consists of 1 for the intercept term and $(p-1)$ covariates, $i = 1, \ldots, n$. Assume that T_i and $\mathbf{V}_i$ are related through a median regression model, that is,

$$\Pr(T_i \geq \boldsymbol{\theta}_0'\mathbf{V}_i | \mathbf{V}_i) = 1/2. \tag{7}$$

Note that the distribution of the "error" term $T - \boldsymbol{\theta}_0'\mathbf{V}$ may depend on $\mathbf{V}$. When T_i is subject to right censoring, we observe only (Y_i, Δ_i), where $Y_i = \min\{T_i, C_i\}$, $\Delta_i = I(Y_i = T_i)$, $I(\cdot)$ is the indicator function and C_i is the censoring random variable with a common distribution survival function $G(\cdot)$. We assume that C is independent of $(T, \mathbf{V})$. Here the observable random quantity $X = \{(Y_i, \Delta_i, \mathbf{V}_i), i = 1, \ldots, n\}$. Using the fact that

$$E\left\{\frac{I(Y_i \geq \boldsymbol{\theta}_0'\mathbf{V}_i)}{G(\boldsymbol{\theta}_0'\mathbf{V}_i)} - \frac{1}{2}\bigg|\mathbf{V}_i\right\} = 0,$$

Ying, Jung, and Wei (1995) proposed the following estimating function to make inferences about $\boldsymbol{\theta}_0$:

$$\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\mathbf{V}_i\left\{\frac{I(Y_i \geq \boldsymbol{\theta}'\mathbf{V}_i)}{\hat{G}(\boldsymbol{\theta}'\mathbf{V}_i)} - \frac{1}{2}\right\}, \tag{8}$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimate for $G(\cdot)$. In cases where there exists a t_0 such that $G(t_0) > 0$ and $\Pr(\boldsymbol{\theta}_0'\mathbf{V} < t_0) = 1$, Ying et al. (1995) showed that for large n, the equation $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) \approx 0$ has a unique consistent root $\hat{\boldsymbol{\theta}}_X$ and that the distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$ can be approximated by a normal. But because $\mathbf{S}_x(\boldsymbol{\theta})$ is neither continuous nor monotone in $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_x$ is difficult to obtain through standard numerical methods. Moreover, the covariance matrix of $\hat{\boldsymbol{\theta}}_X$, which involves unknown covariate-dependent density functions, cannot be well estimated directly with censored data.

Now $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}_0)$ can be approximated asymptotically by a mean-0 normal with covariance matrix $\boldsymbol{\Pi}_X(\boldsymbol{\theta}_0)$, where

$$\boldsymbol{\Pi}_X(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\left[\mathbf{V}_i^{\otimes 2}\left\{\frac{I(Y_i \geq \boldsymbol{\theta}\mathbf{V}_i)}{\hat{G}(\boldsymbol{\theta}'\mathbf{V}_i)} - \frac{1}{2}\right\}^2 - \frac{1-\Delta_i}{4}\left\{\frac{\sum_{j=1}^{n}\mathbf{V}_j I(\boldsymbol{\theta}'\mathbf{V}_j \geq Y_i)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)}\right\}^{\otimes 2}\right]. \tag{9}$$

Then $\mathbf{S}_X(\boldsymbol{\theta}_0) = \boldsymbol{\Pi}_X^{-1/2}(\boldsymbol{\theta}_0)\tilde{\mathbf{S}}_X(\boldsymbol{\theta}_0)$ is asymptotically $\mathrm{N}(0, \mathbf{I}_p)$.

For the first example, we consider the case where the support of the censoring variable C is at least as large as that of the failure time T. Under this assumption, we can obtain a simple consistent estimator $\boldsymbol{\theta}_X^{\dagger}$ by minimizing a convex function,

$$\sum_{i=1}^{n}\frac{\Delta_i}{\hat{G}(Y_i)}|Y_i - \boldsymbol{\theta}'\mathbf{V}_i|. \tag{10}$$

In an unpublished thesis for Harvard School of Public Health, Tian showed that an estimate $\boldsymbol{\Gamma}_X$ for the covariance matrix of $\boldsymbol{\theta}_X^{\dagger}$ can be obtained easily through a resampling method. The proposal presented in Section 2 is readily applicable to the present case.

We use a lung cancer study dataset analyzed by Ying et al. (1995) to illustrate the new procedure. Standard therapy for patients with small-cell lung cancer is a combination of etoposide and cisplatin. This lung cancer study was designed to evaluate two regimens: arm A, cisplatin followed by etoposide, and arm B, etoposide followed by cisplatin. In the study, 121 lung cancer patients were randomly assigned to one of these two groups. Here the response variable is the base-10 logarithm of the time to death. The covariate vector $\mathbf{V}$ has three components; the first component is 1, corresponding to the intercept, the second is the patient's entry age, and the third is the treatment indicator, which is 1 if the patient was assigned to A and 0 otherwise. Because there is no loss to follow-up during the study, it is reasonable to assume that the censoring time C is independent of the failure time and the two covariates. Note that for numerical stability, in our analysis each observed covariate value is standardized; that is, it is centered by its sample mean and then divided by its sample standard deviation.

For this dataset, the consistent estimate $\boldsymbol{\theta}_X^{\dagger}$ from (10) is $(2.66, .0047, .073)'$, and its estimated covariance matrix $\boldsymbol{\Gamma}_x$ is

$$\begin{pmatrix} 7.9\times 10^{-4} & 8.2\times 10^{-6} & 1.1\times 10^{-4} \\ 8.2\times 10^{-6} & 6.4\times 10^{-4} & 2.5\times 10^{-4} \\ 1.1\times 10^{-4} & 2.5\times 10^{-4} & 8.0\times 10^{-4} \end{pmatrix}.$$

For illustration, we chose $t_0 = 3.27$. Note that $\hat{G}(3.27) \approx .1$ and $\mathbf{V}_i'\boldsymbol{\theta}_x^{\dagger} < t_0, i = 1, \ldots, n$. Furthermore, we used the covariance matrix $\boldsymbol{\Sigma}_x = \boldsymbol{\Gamma}_x$ as the initial proposal distribution and chose Ω_x using (6) with $\Phi^{-1}(\alpha_n) = 6$. Based on the initial 1,000 generated $\boldsymbol{\theta}_{(k)}$, the acceptance rate was about 50%. We then used these $\boldsymbol{\Sigma}_x$ and Ω_x to generate 30,000 $\boldsymbol{\theta}_{(k)}$'s, but deleted the first 3,000. The effective sample size based on these 27,000 dependent $\boldsymbol{\theta}_{(k)}$ is about 1,400. Figure 1 provides a diagnostics quantile–quantile plot based on $\{\|\mathbf{S}_x(\boldsymbol{\theta}_{(k)})\|^2, k = 3{,}001, \ldots, 30{,}000\}$. The y-axis is the quantile of χ_3^2, and the x-axis is the empirical quantile. In light of this plot, we expect the empirical distribution based on the foregoing 27,000 $\boldsymbol{\theta}_{(k)}$'s to be a good approximation to the distribution of $\boldsymbol{\theta}_x^*$. We could use the minimizer $\hat{\boldsymbol{\theta}}_x$ described in Section 2 and the sample covariance matrix obtained from those 27,000 $\boldsymbol{\theta}_{(k)}$'s to estimate the mean and covariance matrix of $\boldsymbol{\theta}_x^*$. This would result in $\hat{\boldsymbol{\theta}}_x = (2.70, -.039, .078)'$ with estimated standard errors of .039, .039, and .040.

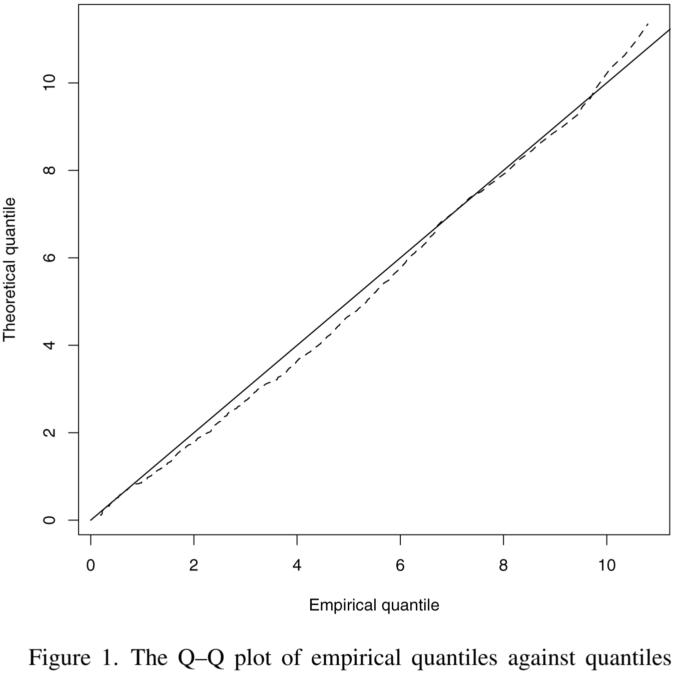

Figure 1. The Q–Q plot of empirical quantiles against quantiles from χ_3^2 based on 27,000 observed $\|\mathbf{S}_x(\boldsymbol{\theta})\|^2$ for Example 1.

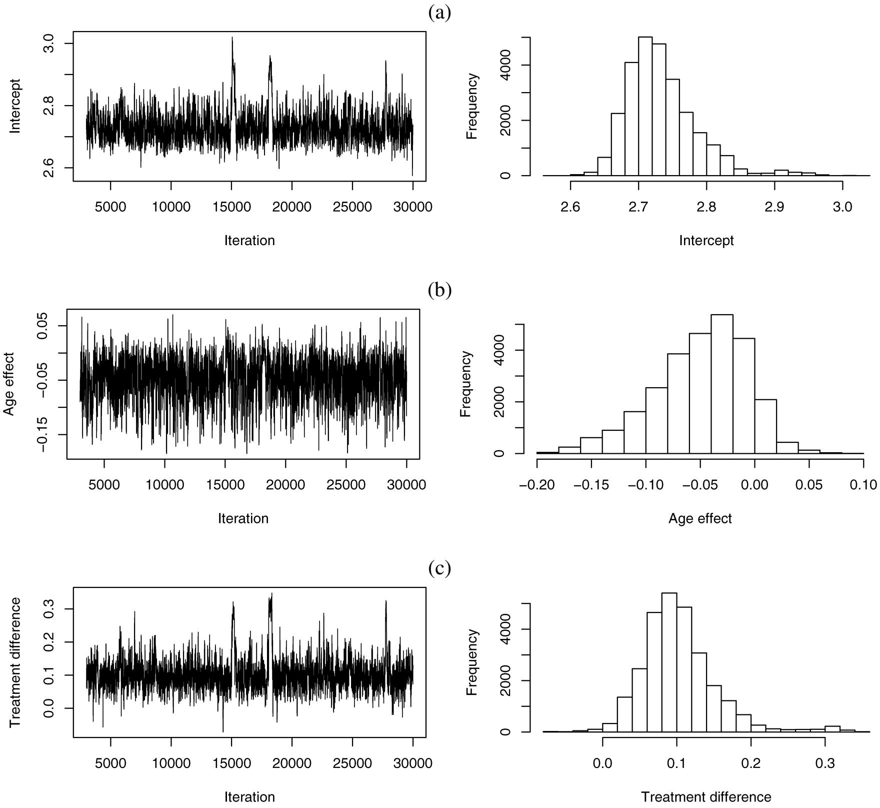

Figure 2. Marginal trace plots and histograms for (a) intercept, (b) age effect, and (c) treatment difference for Example 2.

For the second example, we considered a more realistic situation, that the support of the censoring is shorter than that of the failure time. For this case, the estimator derived from (10) is no longer consistent. To obtain an initial $\boldsymbol{\theta}_{(1)}$ and the proposal distribution for the Markov chain Monte Carlo (MCMC) procedure, we fitted the data with a parametric model by assuming that the error term $T - \boldsymbol{\theta}_0'\mathbf{V}$ is a mean-0 normal with an unknown variance. The maximum likelihood estimate for $\boldsymbol{\theta}_0$ in model (7) is $(2.76, -.063, .088)'$ with estimated covariance matrix

$$\begin{pmatrix} 9.2\times 10^{-4} & -2.2\times 10^{-5} & 1.7\times 10^{-5} \\ -2.2\times 10^{-5} & 9.1\times 10^{-4} & 1.0\times 10^{-4} \\ 1.7\times 10^{-5} & 1.0\times 10^{-4} & 9.2\times 10^{-4} \end{pmatrix}.$$

Because we do not have an initial consistent estimate to locate a proper Ω_x, we chose a quite large Ω_x in (6) with $\Phi^{-1}(\alpha_n) = 15$ and let the foregoing matrix be $\boldsymbol{\Sigma}_x$ in the initial proposal distribution. Under this setting, the acceptance rate based on the first 1,000 iterations is about 45%. We then generated 30,000 $\boldsymbol{\theta}_{(k)}$'s, but deleted the first 3,000.

Figure 2 presents marginal trace plots and histograms corresponding to three parameters (intercept, age effect, and treatment difference) based on 27,000 $\boldsymbol{\theta}_{(k)}$'s. It appears that for the first and third components, there are a number of outliers that likely are not in a $o_p(1)$ neighborhood of $\boldsymbol{\theta}_0$. To obtain robust estimators for the mean and covariance matrix of $\boldsymbol{\theta}_x^*$, we deleted $\boldsymbol{\theta}_{(k)}$ such that either its first component is larger than 2.86 or the third component is larger than .25 by visually examining the plots in Figure 2. This results in deleting 662 $\boldsymbol{\theta}_{(k)}$'s. Figure 3 presents two Q–Q plots, the quantiles of the observed $\|\mathbf{S}_x(\boldsymbol{\theta}_{(k)})\|^2$ against the quantiles from χ_3^2. The dotted line is constructed with the original 27,000 $\boldsymbol{\theta}_{(k)}$'s, and the dashed line is based on the 26,338 selected $\boldsymbol{\theta}_{(k)}$'s. Figure 3 shows that the foregoing ad hoc trimming works

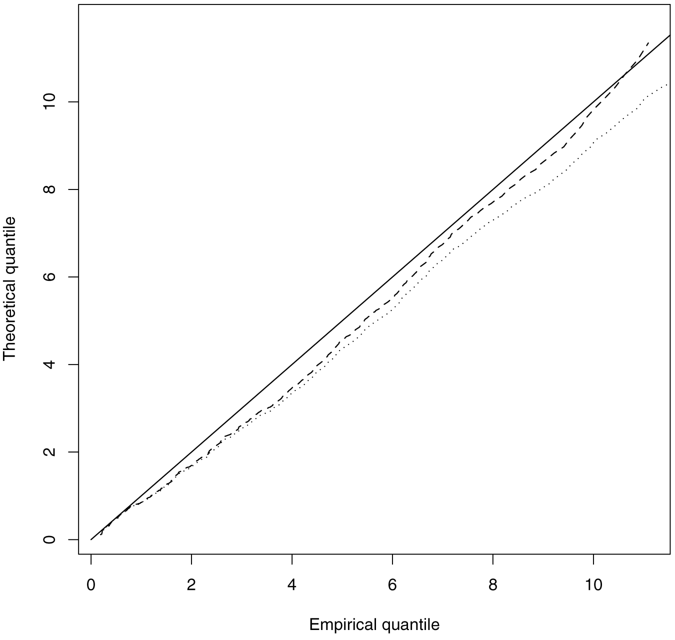

Figure 3. Q–Q plots of empirical quantiles against quantiles from χ_3^2 based on untrimmed (·········) and trimmed (— — ——) observed $\|\mathbf{S}_x(\boldsymbol{\theta})\|^2$ for Example 2.

well. The effective sample size based on these 26,338 dependent $\boldsymbol{\theta}_{(k)}$'s is about 1,000. Now, with those selected $\boldsymbol{\theta}_{(k)}$'s, $\hat{\boldsymbol{\theta}}_x = (2.70, -.038, .079)'$. In the original scale of the covariates, the regression coefficient estimates are 2.89, −.004, and .16, with corresponding estimated standard errors of .044, .005, and .084. These estimates are practically identical to those obtained by Ying et al. (1995) through a rather complex numerical procedure.

Finally, the observed value of $\|\mathbf{S}_x(\hat{\boldsymbol{\theta}}_x)\|^2 = .017$, which is the .06th percentile of χ^2_3. This provides a justification that $\hat{\boldsymbol{\theta}}_x$ is a solution to the equation $\mathbf{S}_x(\boldsymbol{\theta}) \approx 0$. Moreover, the values of $\|\mathbf{S}_x(\boldsymbol{\theta})\|^2$ for the 662 deleted $\boldsymbol{\theta}_{(k)}$ are substantially larger than .017. Also, because the parameter space Ω_x used for generating realizations from (5) is quite large, $\hat{\boldsymbol{\theta}}_x$ appears to be the unique root to the estimating equation. This, coupled with the fact that for large n, this estimating equation theoretically has a unique solution implies that $\hat{\boldsymbol{\theta}}_x$ is the consistent root to the estimating equation.

The question of how to choose the outlier removing process is often raised for all resampling methods. For the current problem, when the root to the estimating equation is unique, we do not expect to have a large number of outliers $\boldsymbol{\theta}_{(k)}$. Moreover, contrary to the lack of objective criteria for choosing the trimming strategy for the general robust variance estimation problem (Wang and Raftery 2002), our outlier deletion process can be guided by the knowledge that the empirical distribution based on $\mathbf{S}_x(\boldsymbol{\theta}_{(k)})$'s from selected $\boldsymbol{\theta}_{(k)}$'s is $\mathrm{N}(0, \mathbf{I}_p)$ and that the distribution based on those $\boldsymbol{\theta}_{(k)}$'s is approximately normal. When interest lies in making inferences about individual components of $\boldsymbol{\theta}_0$, standard numerical methods may be used to obtain robust variance estimates of the parameter estimates. For example, letting $\theta^l_{(k)}$ be the lth component of $\boldsymbol{\theta}_{(k)}$, a popular robust estimate for the standard error based on the median absolute deviation is

$$\hat{\sigma}_l = 1.483 \times \operatorname{median}\Big\{\big|\theta^l_{(j)} - \operatorname{median}\big\{\theta^l_{(k)} : k = K+1, \ldots, K+M\big\}\big| : K+1 \le j \le K+M\Big\}.$$

For the foregoing example, the resulting variance estimates are almost identical to ours using the ad hoc visual deletion method. It is interesting to note that for the first and third components of the parameter vector, $3 \times \hat{\sigma}_l, l = 1, 3$, are 2.86 and .22, which are amazingly close to our cutoff points 2.86 and .25. Moreover, we find that the trimming process in general has little impact on the variance estimates. For example, even without any trimming, the resulting variance estimates for three components are .052, .005, and .10, which are only slightly larger than ours. In general, we suggest performing various sensitivity analyses through graphical and numerical methods of outlier deletion.

For the third example, we relax the assumption that there exists a t_0 such that $\Pr(\boldsymbol{\theta}_0'\mathbf{V} < t_0) = 1$ in the previous two cases. Here, we require only that

$$\Pr(\boldsymbol{\theta}_0'\mathbf{V} < t_0) \ge \xi, \tag{11}$$

where $\xi > 0$, a prespecified constant. This weaker condition allows us to expand the parameter space substantially. Moreover, we modify the estimating function (8) to accommodate the case with type I censoring, that is, the censoring variable C is a fixed time point. This type of censoring is quite common in the econometrics literature. Toward this end, consider the following estimating function:

$$\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^n I(\boldsymbol{\theta}'\mathbf{V}_i < t_0)\mathbf{V}_i \left[\frac{I(Y_i \ge \boldsymbol{\theta}'\mathbf{V}_i)}{\hat{G}(\boldsymbol{\theta}'\mathbf{V}_i)} - \frac{1}{2}\right]. \tag{12}$$

It is not difficult to show that if $\Pr(\boldsymbol{\theta}_0'\mathbf{V} < t_0) > 0$, then there exists a consistent root $\hat{\boldsymbol{\theta}}_X$ to the equation $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) \approx 0$. Asymptotically, the covariance matrix for $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}_0)$ is $\boldsymbol{\Pi}_X(\boldsymbol{\theta}_0)$, where

$$\boldsymbol{\Pi}_X(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \Bigg[I(\boldsymbol{\theta}'\mathbf{V}_i < t_0)\mathbf{V}_i^{\otimes 2} \left\{\frac{I(Y_i \ge \boldsymbol{\theta}\mathbf{V}_i)}{\hat{G}(\boldsymbol{\theta}'\mathbf{V}_i)} - \frac{1}{2}\right\}^2 - \frac{1-\Delta_i}{4} \left\{\frac{\sum_{j=1}^n \mathbf{V}_j I(\boldsymbol{\theta}'\mathbf{V}_j \in [Y_i, t_0])}{\sum_{j=1}^n I(Y_j \ge Y_i)}\right\}^{\otimes 2} \Bigg].$$

It is well known that even under type I censoring, asymptotically the equation $\tilde{\mathbf{S}}_X(\boldsymbol{\theta}) = 0$ may have multiple roots (Khan and Powell 2001). Using arguments similar to those given by Ying et al. (1995), this particular estimating function is locally linear around the limit of each root provided that the limit is an interior point of the parameter space. It follows that the distribution of each root can be approximated by a normal.

Now, we use the aforementioned lung cancer data to illustrate our procedure. To better visualize the results, we considered the case with a single covariate (the treatment indicator) in our analysis. Thus $\boldsymbol{\theta}$ is a 2×1 vector. As in the previous case, we fitted the data with a fully parametric normal model. The point estimate and its estimated covariance matrix are $\boldsymbol{\theta}_x^\dagger = (2.76, .95)'$ and

$$\boldsymbol{\Gamma}_x = \begin{pmatrix} 9.5 \times 10^{-4} & 2.0 \times 10^{-5} \\ 2.0 \times 10^{-5} & 9.4 \times 10^{-4} \end{pmatrix}. \tag{13}$$

Note that this parametric point estimator may not be consistent. For our procedure, we let $t_0 = 3.27$ and $\xi = .4$ in (11), and let $\Phi^{-1}(\alpha_n) = 15$ for Ω_x in (6). We find that with $\boldsymbol{\Sigma}_x = 2\boldsymbol{\Gamma}_x$, given in (13), the acceptance rate is about 40%. Under this setting, we generated 30,000 $\boldsymbol{\theta}_{(k)}$ and deleted the first 3,000 $\boldsymbol{\theta}_{(k)}$.

We could use the standard cluster analysis technique to identify the potential well-separated "mixture normals" corresponding to multiple roots, especially for large sample sizes. For example, we could use the R function "Mclust" to implement such analysis (Fraley and Raftery 2002). With the foregoing 27,000 $\boldsymbol{\theta}_{(k)}$'s, Mclust locates two clusters (Fig. 4). The points in one cluster are denoted by open circles, and those in the other cluster are denoted by crosses. Then, for each cluster, we can apply the robust variance estimation procedure discussed in Example 2 with those "data points" in the cluster. Here we used all of the data points to estimate the standard error of each parameter estimator. The resulting distribution of the points on the left side is approximately normal with mean $\hat{\boldsymbol{\theta}}_x = (2.70, .098)'$ and estimated standard errors of .039 and .039. The effective sam-

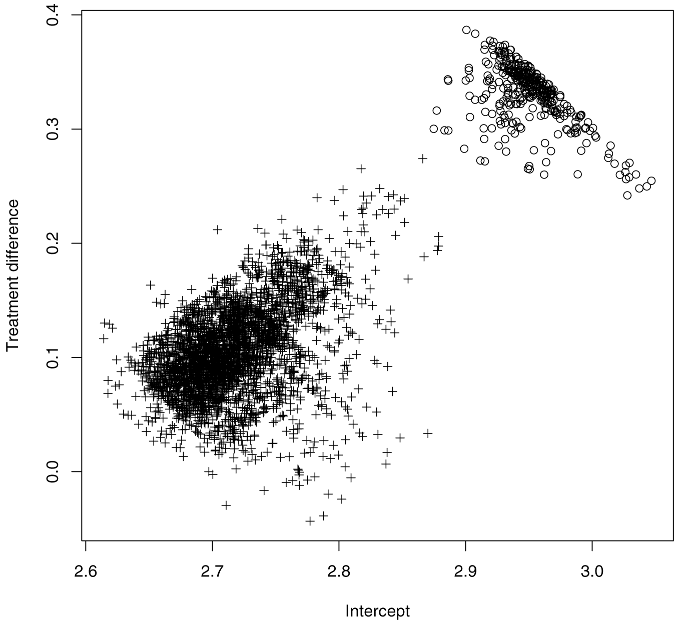

Figure 4. Scatter diagram for the intercept against the treatment difference. “+” and “○” represent points in the two clusters identified by the “Mclust” function.

ple size based on these points is about 1,200. The corresponding value of $\|\mathbf{S}_x(\hat{\boldsymbol{\theta}}_x)\|^2 = .017$, which is the .8th percentile of χ_2^2. Note that this normal distribution is very similar to its counterpart in Example 2.

For the cluster of points on the right side of the figure, $\hat{\boldsymbol{\theta}}_x = (2.95, .34)'$, with $\|\mathbf{S}_x(\hat{\boldsymbol{\theta}}_x)\|^2 = .55$, which is the 24th percentile of χ_2^2. The distribution of this set of points cannot be approximated by a complete normal. If $(2.95, .34)'$ is a root, then this suggests that its limit may be very close to the boundary of the parameter space or the sample size of the study may be too small so as to make the large-sample approximation not applicable.

It is interesting to note that the foregoing two estimates for the intercept term are quite similar, but their counterparts for the treatment difference appear to be markedly different. Although we cannot determine which estimate is consistent with $\boldsymbol{\theta}_0$ without additional information, we may choose .098 as a conservative estimate for the contrast between treatments A and B, indicating that patients treated by A tended to live longer than those treated by B. In any event, it is highly desirable for the present case to explore an alternative estimating equation that provides a globally consistent estimate. In addition, note that due to the extremely discrete nature of the covariate for the present case, the parameter space cannot be further enlarged by choosing a smaller ξ in (11).

4. EVALUATING THE PERFORMANCE OF THE NEW PROCEDURE THROUGH A SIMULATION STUDY

We conducted an extensive simulation study to examine the performance of the proposed inference procedure based on the MCMC sampler. Specifically, we compared the standard bootstrap method, the pivotal resampling method of Parzen et al. (1994), the MCMB method of He and Hu (2002), and our method for analyzing the heterogeneous median regression model (7) with a noncensored response variable T and its covariate vector $\mathbf{V}$. Note that for the present case, the bootstrap and the method of Parzen et al. can be implemented efficiently through the standard linear programming technique, but such an efficient algorithm may not exist for more complicated estimating equations.

The observations $\{(T_i, \mathbf{V}_i), i = 1, \ldots, n\}$ were generated from each of the following four models, which were used by Kocherginsky, He, and Mu (2005) for evaluating the MCMB method for general quantile regression. The four models considered here are in the form of

$$T = \boldsymbol{\theta}_0'\mathbf{V} + c(\mathbf{V})\epsilon,$$

where $\mathbf{V} = (v_1, v_2, \ldots, v_p)'$, $v_1 = 1$, $c(\cdot)$ is a known function, and ϵ is the covariate-free error term. Here $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$.

For model 1, $p = 3$, ϵ, v_2 and v_3 are $N(0, 1)$, $c(\mathbf{V}) = 1$, and $n = 400$. For model 2, $p = 4$, v_2 and ϵ are $N(0, 1)$, v_4 is uniformly distributed at $[0, 1]$, $v_3 = v_2 + v_4 + e$, e is $N(0, 1)$, $c(\mathbf{V}) = 1 + v_4$, and $n = 400$. For model 3, $p = 8$; v_2 and v_3 are Bernoulli(.4); v_4 and v_5 are the standard lognormal; $(v_6, v_7)'$ is bivariate normal with mean $(2, 2)'$, variance $(1, 1)'$, and a correlation coefficient of .8; v_8 is χ_1^2, ϵ is t_2, $c(\mathbf{V}) = 1$, and $n = 200, 500$. Model 4 is similar to model 3, except that $c(\mathbf{V}) = 1 + v_4 + v_6 + v_8$. All regression coefficients in models 1, 2, 3, and 4 are 1.

For each simulated dataset $\{(T, \mathbf{V})\}$ from each model, the regression parameters were estimated based on the estimating function (8) with $\hat{G}(\cdot) = 1$. The bootstrap and Parzen's resampling methods were implemented using the R function “boot.rq,” and the MCMB was implemented through “rqmcmb.” With default values of R functions, the number of independent resampling replications is 200 for the bootstrap method and the pivotal resampling method, and the length of Markov chain is 100 for the MCMB method. To implement the new method, we used the least squares estimate as the initial $\boldsymbol{\theta}_{(1)}$ and let the proposal distribution be the multivariate normal whose covariance matrix is proportional to the estimated covariance matrix of the foregoing least squares estimate. Similar to the setup in Section 3, for each generated dataset, the proportion parameter c was chosen adaptively with an acceptance rate of 20–40%. After selecting a proper c and a burn-in period, we then generated 3,000 samplers for models 1 and 2 and 5,000 samplers for models 3 and 4, to approximate the distribution of $\boldsymbol{\theta}_x^*$. For each of the four methods, we then constructed a $(1 - \alpha)$ Wald-type confidence interval for each regression parameter and checked whether the interval contained the true parameter value and also recorded its interval length. We repeated this process 500 times and computed the empirical coverage probability and the average length for each interval estimation procedure. For the proposed MCMC method, we find that the empirical coverage levels are practically identical to their nominal counterparts and their average lengths are similar to those based on the existing methods. The results for various components of $\boldsymbol{\theta}_0$ are summarized in Tables 1–3 with $(1 - \alpha) = .95$. Note that for model 4 with eight parameters involved, the new procedure performs quite well even for cases with relatively moderate sample sizes.

Table 1. Simulation results for models 1 and 2 ($n = 400$)

	Model 1, θ_2		Model 2, θ_3	
Method	AL	CP	AL	CP
MCMC	.26	.954	.37	.950
MCMB	.25	.932	.36	.944
PWY	.26	.929	.38	.956
BT	.26	.929	.36	.962

NOTE: AL, average length of the .95 confidence interval; CP, empirical coverage probability; MCMC, new procedure based on the MCMC sampler; PWY, Parzen et al.; BT, unconditional bootstrap method.

Note that in the simulation study, for 1% ∼ 2% of the 500 generated datasets from model 3 (also model 4), our method needed to run longer chains to obtain stabilized realizations. For each of these cases, we generated a Markov chain of length 15,000 and used the last 10,000 for constructing the confidence interval.

5. REMARKS

The novel MCMB method proposed by He and Hu (2002) only works for a special class of estimating functions of (3). It is interesting to note that under some regularity conditions, for each fixed iteration, the MCMB is asymptotically equivalent to the Gibbs sampler, a special MCMC algorithm. That is, approximately, at each iteration, the updated $\boldsymbol{\theta}$ value is generated recursively from the conditional densities resulting from the normalized joint density (5). It is not clear, however, whether globally the MCMB procedure would produce similar results as a Gibbs sampler with a large number of iterations. This is due in part to the fact that for each step, components of the generated $\boldsymbol{\theta}$ in the MCMB procedure only approximately follow the corresponding conditional distributions. Therefore, as indicated by He and Hu, it is advisable to implement the MCMB with a moderate number of iterations.

Under the semiparametric setting, for large n, the profile likelihood function is approximately proportional to $\exp\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)'\mathbf{B}_0^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)\}$ for $\boldsymbol{\theta}$ in a small neighborhood of $\boldsymbol{\theta}_0$, where $\mathbf{B}_0$ is the inverse of the information matrix and $\hat{\boldsymbol{\theta}}_x$ is the maximum profile likelihood estimator of $\boldsymbol{\theta}_0$. Therefore, we can generate observations from a density that is proportional to the profile likelihood function to obtain an approximation to the covariance matrix $\mathbf{B}_0$ of $(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$ (Lee et al. 2005). If we apply our proposal to the profile likelihood score function $\tilde{\mathbf{S}}_X(\boldsymbol{\theta})$, then the resulting covariance matrix of $\hat{\boldsymbol{\theta}}_X$ is a robust sandwich-type estimate of $\mathbf{B}_0$, which can be quite different from the one obtained by Lee et al. (2005) for finite-sample cases. Generalizing the results of Lee et al. (2005) and our procedure to the case with a maximand, which may not be a likelihood function and whose "score function" is difficult to obtain, warrants further investigation.

Table 2. Simulation results for model 3

	$\theta_4, n = 200$		$\theta_6, n = 200$		$\theta_4, n = 500$		$\theta_6, n = 500$	
Method	AL	CP	AL	CP	AL	CP	AL	CP
MCMC	.33	.970	.87	.970	.17	.964	.44	.942
MCMB	.31	.970	.76	.960	.15	.950	.42	.952
PWY	.27	.978	.80	.980	.15	.972	.45	.960
BT	.26	.980	.78	.974	.14	.958	.45	.958

NOTE: AL, average length of the .95 confidence interval; CP, empirical coverage probability; MCMC, new procedure based on the MCMC sampler; PWY, Parzen et al.; BT, unconditional bootstrap method.

Table 3. Simulation results for model 4

	$\theta_4, n = 200$		$\theta_6, n = 200$		$\theta_4, n = 500$		$\theta_6, n = 500$	
Method	AL	CP	AL	CP	AL	CP	AL	CP
MCMC	2.96	.915	5.31	.984	1.54	.946	2.82	.958
MCMB	2.76	.948	4.91	.970	1.50	.920	2.49	.946
PWY	2.70	.960	4.63	.970	1.59	.934	2.61	.958
BT	2.60	.943	4.52	.970	1.57	.922	2.58	.956

NOTE: AL, average length of the .95 confidence interval; CP, empirical coverage probability; MCMC, new procedure based on the MCMC sampler; PWY, Parzen et al.; BT, unconditional bootstrap method.

APPENDIX: THEORETICAL JUSTIFICATION

Theorem A.1. Assume that the estimating function $\mathbf{S}_X(\boldsymbol{\theta})$ satisfies the local linearity condition around $\boldsymbol{\theta}_0$; that is,

$$\sup_{\|\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}_0\|\le\epsilon_n; j=1,2} \frac{\|\mathbf{S}_X(\boldsymbol{\theta}^{(2)}) - \mathbf{S}_X(\boldsymbol{\theta}^{(1)}) - n^{1/2}\mathbf{A}(\boldsymbol{\theta}^{(2)} - \boldsymbol{\theta}^{(1)})\|}{1 + n^{1/2}\|\boldsymbol{\theta}^{(2)} - \boldsymbol{\theta}^{(1)}\|} = o_p(1), \quad \text{(A.1)}$$

where $\mathbf{A}$ is a nonsingular deterministic matrix and $\epsilon_n = o_p(1)$. Also assume that $\mathbf{S}_X(\boldsymbol{\theta}_0)$ converges weakly to the standard normal distribution. For $X = x$, $\tilde{\boldsymbol{\theta}}_x$ defined in Section 2 is a random vector with density function proportional to $\exp\{-\frac{1}{2}\mathbf{S}'_x(\boldsymbol{\theta})\mathbf{S}_x(\boldsymbol{\theta})\}I(\boldsymbol{\theta} \in \Omega_x)$. Then

$$\big|E\big[h\{\mathbf{S}_X(\tilde{\boldsymbol{\theta}}_X)\}|X\big] - E\{h(\mathbf{Z})\}\big| = o_p(1), \quad \text{(A.2)}$$

where $h(\cdot)$ is any uniformly bounded Lipschitz continuous function $\mathbb{R}^p \to \mathbb{R}^+$ and $\mathbf{Z}$ is $\mathrm{N}(0, \mathbf{I}_p)$. Loosely speaking, the distribution of $\mathbf{S}_x(\tilde{\boldsymbol{\theta}}_x)$ converges to $\mathrm{N}(0, \mathbf{I}_p)$. Moreover,

$$\big|E\big[h\{n^{1/2}(\tilde{\boldsymbol{\theta}}_X - \hat{\boldsymbol{\theta}}_X)\}|X\big] - E\big[h\{n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)\}\big]\big| = o_p(1).$$

Loosely speaking, the conditional distribution of $n^{1/2}(\tilde{\boldsymbol{\theta}}_x - \hat{\boldsymbol{\theta}}_x)$ is a good approximation to the distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)$.

Proof. The local linearity condition (A.1) implies that

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\epsilon_n} \frac{\|\mathbf{S}_X(\boldsymbol{\theta}) - n^{1/2}\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X)\|}{1 + n^{1/2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X\|} = o_p(1) \quad \text{(A.3)}$$

and

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\epsilon_n} \frac{|\mathbf{S}_X(\boldsymbol{\theta})'\mathbf{S}_X(\boldsymbol{\theta}) - n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X)'\mathbf{A}'\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X)|}{1 + n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X\|^2} = o_p(1) \quad \text{(A.4)}$$

for any $\epsilon_n = o_p(1)$. Note that because $\tilde{\boldsymbol{\theta}}_x$ is restricted at Ω_x,

$$E\big[h\{\mathbf{S}_X(\tilde{\boldsymbol{\theta}}_X)\}|X = x\big] = \frac{\int_{\Omega_x} h\{\mathbf{S}_x(\boldsymbol{\theta})\}\exp\{-\frac{1}{2}\mathbf{S}_x(\boldsymbol{\theta})'\mathbf{S}_x(\boldsymbol{\theta})\}\,d\boldsymbol{\theta}}{\int_{\Omega_x} \exp\{-\frac{1}{2}\mathbf{S}_x(\boldsymbol{\theta})'\mathbf{S}_x(\boldsymbol{\theta})\}\,d\boldsymbol{\theta}}.$$

Let the nominator and denominator of the foregoing ratio be denoted by $I_1(x)$ and $I_2(x)$. For any arbitrarily small $\epsilon > 0$, define the two regions $\mathcal{C}_1$ and $\mathcal{C}_2$ for the sample space of X, where

$$\mathcal{C}_1 = \bigl\{x : \bigl\|\mathbf{S}_x(\boldsymbol{\theta}) - n^{1/2}\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)\bigr\| \le \epsilon\bigl(1 + n^{1/2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x\|\bigr), \boldsymbol{\theta} \in \Omega_x\bigr\}$$

and

$$\mathcal{C}_2 = \bigl\{x : |\mathbf{S}_x(\boldsymbol{\theta})'\mathbf{S}_x(\boldsymbol{\theta}) - n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)'\mathbf{A}'\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)| \\ \le 2\epsilon(1 + n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x\|^2/2), \boldsymbol{\theta} \in \Omega_x\bigr\}.$$

It follows from (A.3) and (A.4) that for large n, $\Pr(\mathcal{C}_1 \cap \mathcal{C}_2) > 1 - \epsilon$. Now, for $x \in \mathcal{C}_2$,

$$I_1(x) \le \int_{\Omega_x} h\{\mathbf{S}_x(\boldsymbol{\theta})\} \exp\left\{\epsilon - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)'(\mathbf{A}'\mathbf{A} - \epsilon\mathbf{I}_p)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)\right\} d\boldsymbol{\theta}. \tag{A.5}$$

Let $\mathbf{z} = n^{1/2}\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_x)$. Then

$$\text{(A.5)} = n^{-1/2}\|\mathbf{A}\|^{-1} \int_{\Omega^*} h\bigl\{\mathbf{S}_x\bigl(\hat{\boldsymbol{\theta}}_x + n^{-1/2}\mathbf{A}^{-1}\mathbf{z}\bigr)\bigr\} \\ \times \exp\left\{\epsilon - \frac{1}{2}\mathbf{z}'\mathbf{A}'^{-1}(\mathbf{A}'\mathbf{A} - \epsilon\mathbf{I}_p)\mathbf{A}^{-1}\mathbf{z}\right\} d\mathbf{z},$$

where $\Omega^* = \{\mathbf{z} | \|\mathbf{A}^{-1}\mathbf{z} + n^{1/2}(\hat{\boldsymbol{\theta}}_x - \boldsymbol{\theta}_x^{\dagger})\| \le c_n\}$.

Moreover, for $x \in \mathcal{C}_1$

$$\bigl\|\mathbf{S}_x\bigl(\mathbf{A}^{-1}n^{-1/2}\mathbf{z} + \hat{\boldsymbol{\theta}}_x\bigr) - \mathbf{z}\bigr\| \le \epsilon(1 + \|\mathbf{z}\|),$$

which implies that $|h\{\mathbf{S}_x(\mathbf{A}^{-1}n^{-1/2}\mathbf{z} + \hat{\boldsymbol{\theta}}_x)\} - h(\mathbf{z})| \le a\epsilon(1 + \|\mathbf{z}\|)$, where "$a$" is a generic notation for a positive constant. It follows that

$$n^{1/2}I_1(x) \le a\epsilon + \|\mathbf{A}\|^{-1} \int_{\Omega^*} h(\mathbf{z}) \exp\left\{-\frac{1}{2}\mathbf{z}'\bigl(\mathbf{I}_p - \epsilon(\mathbf{A}\mathbf{A}')^{-1}\bigr)\mathbf{z}\right\} d\mathbf{z} \\ \le a\epsilon + \|\mathbf{A}\|^{-1} \int_{\Omega^*} h(\mathbf{z}) \exp\left\{-\frac{1 - \epsilon\lambda}{2}\mathbf{z}'\mathbf{z}\right\} d\mathbf{z},$$

where λ is the largest eigenvalue of $(\mathbf{A}\mathbf{A}')^{-1}$. Let $\mathbf{s} = (1 - \epsilon\lambda)^{1/2}\mathbf{z}$; then

$$n^{1/2}I_1(x) \le a\epsilon + \|\mathbf{A}\|^{-1}(1 - \epsilon\lambda)^{-1/2} \\ \times \int_{\Omega^+} h\bigl\{(1 - \epsilon\lambda)^{-1/2}\mathbf{s}\bigr\} \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s},$$

where $\Omega^+ = \{\mathbf{s} : \|\mathbf{A}^{-1}(1 - \epsilon\lambda)^{-1/2}\mathbf{s} + n^{1/2}(\hat{\boldsymbol{\theta}}_x - \boldsymbol{\theta}_x^{\dagger})\| \le c_n\}$. Because for a small ϵ, $(1 - \epsilon\lambda)^{-1/2} \approx 1 + \lambda\epsilon/2$, $|h\{(1 - \epsilon\lambda)^{-1/2}\mathbf{s}\} - h(\mathbf{s})| \le a\epsilon$. Therefore, for large n,

$$n^{1/2}I_1(x) \le a\epsilon + \|\mathbf{A}\|^{-1} \int_{\Omega^+} h(\mathbf{s}) \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s} \\ \le a\epsilon + \|\mathbf{A}\|^{-1} \int_{\mathbb{R}^p} h(\mathbf{s}) \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s}.$$

Similarly, it can be shown that $n^{1/2}I_1(x) \ge -a\epsilon + \|\mathbf{A}\|^{-1} \int_{\mathbb{R}^p} h(\mathbf{s}) \times \exp\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\} d\mathbf{s}$. This implies that

$$\left|n^{1/2}I_1(x) - \|\mathbf{A}\|^{-1} \int_{\mathbb{R}^p} h(\mathbf{s}) \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s}\right| \le a\epsilon.$$

Using the same argument, we can show that

$$\left|n^{1/2}I_2(x) - \|\mathbf{A}\|^{-1} \int_{\mathbb{R}^p} \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s}\right| \\ = \bigl|n^{1/2}I_2(x) - \|\mathbf{A}\|^{-1}(2\pi)^{p/2}\bigr| \le a\epsilon.$$

Therefore, for a large n,

$$\bigl|E\bigl[h\{\mathbf{S}_X(\tilde{\boldsymbol{\theta}}_X)\}\bigr] - E\{h(\mathbf{Z})\}\bigr| \\ = \left|\frac{n^{1/2}I_1(X)}{n^{1/2}I_2(X)} - \int_{\mathbb{R}^p} \frac{h(\mathbf{s})}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\mathbf{s}'\mathbf{s}\right\} d\mathbf{s}\right| \le a\epsilon. \tag{A.6}$$

It follows that the left side of (A.6) converges to 0 in probability, as $n \to \infty$, and (A.2) holds true.

To show the second part of theorem, first it follows from the local linearity condition that

$$n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0) = \mathbf{A}^{-1}\mathbf{S}_X(\boldsymbol{\theta}_0) + o_p\bigl(1 + n^{1/2}\|\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0\|\bigr).$$

This implies that $\mathbf{A}n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0) = \mathbf{S}_X(\boldsymbol{\theta}_0) + o_p(1)$. Furthermore, because $\|\tilde{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0\| = o_p(1)$, it is straightforward to show that $\mathbf{A}n^{1/2}(\tilde{\boldsymbol{\theta}}_X - \hat{\boldsymbol{\theta}}_X) = \mathbf{S}_X(\tilde{\boldsymbol{\theta}}_X) + o_{p^*}(1)$, where p^* is the product probability measure generated by that for X and $\tilde{\boldsymbol{\theta}}_x$. Therefore, it follows from (A.2) that

$$\bigl|E\bigl[h\bigl\{n^{1/2}(\tilde{\boldsymbol{\theta}}_X - \hat{\boldsymbol{\theta}}_X)\bigr\}|X\bigr] - E\bigl[h\bigl\{n^{1/2}(\hat{\boldsymbol{\theta}}_X - \boldsymbol{\theta}_0)\bigr\}\bigr]\bigr| = o_p(1).$$

[Received March 2005. Revised November 2005.]

REFERENCES

Arcones, M., and Gine, E. (1992), "On the Bootstrap of M-Estimators and Other Statistical Functionals," in *Exploring the Limit of Bootstrap*, eds. R. LePage and L. Billard, New York: Wiley, pp. 14–47.

Chernozhukov, V., and Hong, H. (2003), "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall.

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Gelman, A., Carlin, H., Stern, S., and Rubin, D. B. (2003), *Bayesian Data Analysis*, London: Chapman & Hall.

He, X., and Hu, F. (2002), "Markov Chain Marginal Bootstrap," *Journal of the American Statistical Association*, 97, 783–795.

Hu, F., and Kalbfleisch, J. D. (2000), "The Estimating Function Bootstrap" (with discussion), *Canadian Journal of Statistics*, 28, 449–499.

Khan, S., and Powell, J. L. (2001), "Two-Step Estimation of Semiparametric Censored Regression Models," *Journal of Econometrics*, 103, 73–110.

Kocherginsky, M., He, X., and Mu, Y. (2005), "Practical Confidence Intervals for Regression Quantiles," *Journal of Computational and Graphical Statistics*, 14, 41–55.

Lee, B. L., Kosorok, M., and Fine, J. P. (2005), "The Profile Sampler," *Journal of the American Statistical Association*, 100, 960–969.

Lin, D. Y., and Geyer, C. J. (1992), "Computational Methods for Semiparametric Linear Regression With Censored Data," *Journal of Computational and Graphical Statistics*, 1, 77–90.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994), "A Resampling Method Based on Pivotal Estimating Functions," *Biometrika*, 81, 341–350.

Tian, L., and Cai, T. (2006), "On the Accelerated Failure Time Model for Current Status and Interval Censored Data," *Biometrika*, 93, 329–342.

Tian, L., Liu, J. S., Zhao, Y., and Wei, L. J. (2004), "Statistical Inferences Based on Non-Smooth Estimating Functions," *Biometrika*, 91, 943–954.

Wang, N., and Raftery, A. E. (2002), "Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest Neighbor-Cleaning" (with discussion), *Journal of the American Statistical Association*, 97, 994–1019.

Ying, Z., Jung, S. H., and Wei, L. J. (1995), "Survival Analysis With a Median Regression Model," *Biometrika*, 90, 178–184.