Harvard University
Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 21

Implementation Of Estimating-Function Based Inference Procedures With MCMC Sampler

Lu Tian* Jun S. Liu†
L. J. Wei‡

*Northwestern University, lutian@northwestern.edu
†Harvard University, jliu@stat.harvard.edu
‡Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper21

IMPLEMENTATION OF ESTIMATING-FUNCTION BASED INFERENCE PROCEDURES WITH MCMC SAMPLER

By Lu Tian

Department of Preventive Medicine, Northwestern University, 680 N. Lake Shore Drive, Chicago, Illinois 60611, U.S.A.

lutian@northwestern.edu

Jun S. Liu

Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, U.S.A.

jliu@stat.harvard.edu

L.J. Wei

Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

wei@hsph.harvard.edu

Summary

Under a semiparametric or nonparametric setting, inferences about the unknown parameter are often made based on a non-smooth 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 paper, we propose a simple, flexible procedure which provides such approximations via 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 the new proposal extensively with three examples.

Some key words: Bootstrap; Median regression; Metropolis algorithm; Normal approximation; Resampling.

1. INTRODUCTION

Under a nonparametric or semiparametric setting, inferences about a $p \times 1$ unknown vector θ_0 of parameters are often based on a p-dimensional estimating function $\tilde{S}_X(\theta)$, where X is the observable random quantity with sample size n. Let $\hat{\theta}_X$ be a consistent root to the equation

$$\tilde{S}_X(\theta) \approx 0. \tag{1.1}$$

If the estimating function is locally linear around θ_0, and for large n, the distribution of $\tilde{S}_X(\theta_0)$ can be approximated well by a zero-mean normal with covariance matrix $\Pi_X(\theta_0)$, then the random vector $W_X = n^{1/2}(\hat{\theta}_X - \theta_0)$ is asymptotically normal. Generally the matrix $\Pi_X(\theta)$ can be obtained easily, but the covariance matrix of W_X may be rather difficult to estimate well directly when $S_X(\theta)$ is not smooth in θ. Note that

$$S_X(\theta_0) = \{\Pi_X(\theta_0)\}^{-1/2}\tilde{S}_X(\theta_0) \tag{1.2}$$

is asymptotically pivotal and is approximately $\mathrm{MN}(0, I_p)$-distributed, where I_p is the $p \times p$ identity matrix.

The bootstrap method (Efron & 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 $\{X_1, \cdots, X_n\}$ and the estimating function is

$$n^{-1/2}\sum_{i=1}^{n}\tilde{S}_{X_i}(\theta), \tag{1.3}$$

where the random part of $\tilde{S}_{X_i}(\cdot)$ depends on X_i only, then for large n, it has been shown that the bootstrap distribution centered by $\hat{\theta}_X$ can approximate the distribution of $(\hat{\theta}_X - \theta_0)$ well (Arcone & Gine, 1992). Recently, Hu & Kalbfleisch (2000) proposed a novel estimating function bootstrap method based on (1.3).

The implementation of the bootstrapping can be problematic when the estimating equation $\tilde{S}_X(\theta) = 0$ and its bootstrap counterparts are difficult to solve numerically. For this case, one may utilize a "parametric bootstrap" method, which takes advantage of the

pivotal feature of the estimating function $S_X(\theta_0)$, to approximate the distribution of W_X. To be specific, let x be the observed value of X and let θ_x^* be a random vector such that

$$S_x(\theta_x^*) \approx Z, \tag{1.4}$$

where Z is $\mathrm{N}(0, I_p)$. If θ_X^* is a consistent estimator for θ_0, it follows from Parzen et al. (1994) that the distribution of W_X can be approximated well by the conditional distribution of $W_x^* = n^{1/2}(\theta_x^* - \hat{\theta}_x)$. This resampling method has been justified theoretically for a class of general estimating functions, which includes (1.3) as a special case. Moreover, realizations of θ_x^* in (1.4) can be generated without solving any estimating equations, for example, via an adaptive importance sampling technique (Tian et al., 2004).

In this article, we propose a procedure via the standard Metropolis algorithm to generate the distribution of θ_x^* without the need of solving (1.4). The procedure only involves computing $S_x(\theta)$ and is more flexible to implement in practice than the one 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 new method extensively with three examples.

Recently, He & Hu (2002) proposed a novel Markov chain marginal bootstrap method to estimate the covariance matrix of $\hat{\theta}_X$ based on a specific type of estimating functions (1.3). Also, Lee, Kosorok & Fine (2005) studied an intriguing stochastic numerical algorithm for the semiparametric profile likelihood estimation problem. More discussions about these two procedures are given in the Remarks Section.

2. INFERENCES FOR θ_0 VIA THE METROPOLIS ALGORITHM

Note that if $S_x(\theta)$ is a one-to-one mapping and differentiable in θ, for large n, the density function of θ_x^* defined in (1.4) is approximately proportional to

$$\exp\{-\frac{1}{2}S_x'(\theta)S_x(\theta)\}. \tag{2.1}$$

Here, we show how to obtain a good approximation to the distribution of θ_x^* via (2.1) even when $S_x(\theta)$ is neither smooth nor a one-to-one function. First, suppose that there

exists a consistent estimator $\theta_X^\dagger$ for θ_0, which may be obtained from a relatively simple estimating function of θ. In the Appendix, we show that if one can construct a random vector $\tilde{\theta}_x$ whose realizations are generated from the density function proportional to (2.1) in a c_n-neighborhood Ω_x of $\theta_x^\dagger$, where $c_n^{-1} = o(n^{1/2})$ and $c_n = o(1)$, then for large n, the distribution of $\tilde{\theta}_x$ is a good approximation to that of θ_x^*.

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

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

where v is generated from $\mathrm{N}(\theta_{(k-1)}, \Sigma_x)$, $\Sigma_X = O_p(n^{-1/2})$, a pre-specified non-singular $p \times p$ matrix, and $\tau_k = \min\{1, g(v)/g(\theta_{(k-1)})\}$. Note that if v is not in Ω_x, we let $\theta_{(k)} = \theta_{(k-1)}$. In theory, for large K and M, we expect that the empirical distribution constructed from $\mathcal{J} = \{\theta_{(K)}, \cdots, \theta_{(K+M)}\}$ is a good approximation to the distribution of $\tilde{\theta}_x$. To be specific, the distribution of θ_x^* can be approximated by a p-dimensional normal with mean $\hat{\theta}_x$ and covariance matrix Λ_x. Here, we let $\hat{\theta}_x$ be $\theta_{(k)}$ which gives the smallest value of $\{S_x'(\theta_{(k)})S_x(\theta_{(k)}), k = K+1, \cdots, K+M\}$ and Λ_x be the sample covariance matrix based on those M dependent θ's in $\mathcal{J}$. Note that to obtain robust $\hat{\theta}_x$ and Λ_x, one may delete outliers of the realizations in $\mathcal{J}$. This is illustrated with an example in the next section.

In practice, the choices of the matrix Σ_x in the proposal distribution for the above 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 $\theta_X^\dagger$ can be estimated by $\Gamma_X = (\gamma_{lm})$. Generally one expects that the target covariance matrix Λ_X of θ_X^* would not be drastically different from Γ_X. Let θ_l and $\theta_{xl}^\dagger$ be the lth components of θ and $\theta_x^\dagger$, $l = 1, \cdots, p$. Then, one may choose

$$\Omega_x = \{\theta : |\theta_l - \theta_{xl}^\dagger| \leq \Phi^{-1}(\alpha_n)\gamma_{ll}^{1/2}, l = 1, \cdots, p\}, \tag{2.2}$$

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 is known, one would let Σ_x in the proposal distribution be proportional to Λ_x (Gelman et al., 2003, p.306). Therefore, for our procedure, one may let $\Sigma_x = c\Gamma_x$ and choose c adaptively in a batch fashion until the acceptance rate of v in the Metropolis algorithm is about 25% to 50% (Liu, 2001, p.115). We then use the final value of c to generate the above sequence $\mathcal{J}$. Furthermore, one may choose K such that the empirical distribution function generated from the realizations $\{S_x(\theta_{(k)}), k = K+1, \cdots, K+M\}$ is as close as possible to $\mathrm{N}(0, I_p)$. Lastly, one may examine the auto-correlations for the sequence $\mathcal{J}$ to estimate M based on a pre-specified effective sample size (Liu, 2001, pp.125-26). Note that the choice of the initial starting point $\theta_{(1)}$ does not seem critical for implementing our procedure.

Since the estimating function may not be smooth, we do not expect $S_x(\hat{\theta}_x) = 0$. In theory, any θ such that $S_X(\theta) = o_p(1)$ is a root to the estimating equation. Empirically an objective way to evaluate if the resulting $\hat{\theta}_x$ from the above search is a possible root, one may use the metric

$$T(\theta) = S_x'(\theta) S_x(\theta) \tag{2.3}$$

to compare the observed value of $T(\hat{\theta}_x)$ with the distribution of $T(\theta_0)$, which is χ^2_p.

Now, consider the case that there is no initial consistent estimate $\theta_X^\dagger$ 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 $S_X(\theta)$ around the limit of each root, the distribution of θ_x^* in (1.4) 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 $(\hat{\theta}_X - \theta_0)$. Under a semiparametric setting, to implement the above iterative procedure, one may fit the data with a parametric submodel to obtain a point estimator for θ_0 and its estimated covariance matrix as the surrogates of $\theta_x^\dagger$ and Γ_x to generate realizations from (2.1). In the absence of an initial consistent estimate, we suggest to consider a large parameter space Ω_x, for example, by choosing a fairly large α_n in (2.2), to obtain a relatively complete profile of the distribution of θ_x^* for

making inferences about θ_0.

3. EXAMPLES

We use three examples to illustrate the new proposal. The first example is for a case that there exists a consistent estimator $\theta_X^{\dagger}$ for θ_0. The second example is to illustrate the case that there is no initial consistent estimator available, but for large n, the estimating equation, $S_x(\theta) = 0$, has a unique root. The third example is to show what our procedure would generate via (2.1) for a case that asymptotically the estimating equation may have multiple roots.

We use a semiparametric, survival median regression model to generate these three cases. To this end, let T_i be the ith failure time or a transformation thereof, and V_i be the corresponding p-dimensional vector which consists of one for the intercept term and $(p-1)$ covariates, $i = 1, \cdots, n$. Assume that T_i and V_i are related via a median regression model. That is,

$$\mathrm{pr}(T_i \geq \theta_0' V_i \ \mid V_i) = 1/2. \tag{3.1}$$

Note that the distribution of the "error" term $T - \theta_0' V$ may depend on V. When T_i is subject to right censoring, one only observes (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, V). Here, the observable random quantity $X = \{(Y_i, \Delta_i, V_i), i = 1, \cdots, n\}$. Using the fact that

$$\mathrm{E}\left\{ \frac{I(Y_i \geq \theta_0' V_i)}{G(\theta_0' V_i)} - \frac{1}{2} \,\middle|\, V_i \right\} = 0,$$

Ying et al. (1995) proposed the following estimating function to make inferences about θ_0

$$\tilde{S}_X(\theta) = n^{-1/2} \sum_{i=1}^{n} V_i \left\{ \frac{I(Y_i \geq \theta' V_i)}{\hat{G}(\theta' V_i)} - \frac{1}{2} \right\}, \tag{3.2}$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimate for $G(\cdot)$. If there exists a t_0 such that $G(t_0) > 0$ and $\mathrm{pr}(\theta_0' V < t_0) = 1$, Ying et al. (1995) showed that for large n, the equation, $\tilde{S}_X(\theta) \approx 0$,

has a unique consistent root $\hat{\theta}_X$, and the distribution of $n^{1/2}(\hat{\theta}_X - \theta_0)$ can be approximated by a normal. However, since $S_x(\theta)$ is neither continuous nor monotone in θ, $\hat{\theta}_x$ is difficult to obtain via standard numerical methods. Moreover, the covariance matrix of $\hat{\theta}_X$, which involves unknown covariate-dependent density functions, cannot be estimated well directly with censored data.

Now, $\tilde{S}_X(\theta_0)$ can be approximated asymptotically by a mean-zero normal with covariance matrix $\Pi_X(\theta_0)$, where $\Pi_X(\theta) =$

$$n^{-1}\sum_{i=1}^{n}\left[V_i^{\otimes 2}\left\{ \frac{I(Y_i \geq \theta V_i)}{\hat{G}(\theta' V_i)} - \frac{1}{2} \right\}^2 - \frac{1-\Delta_i}{2}\left\{ \frac{\sum_{j=1}^{n} V_j I(\theta' V_j \geq Y_i)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)} \right\}^{\otimes 2}\right]. \qquad (3.3)$$

Then, $S_X(\theta_0) = \Pi_X^{-1/2}(\theta_0)\tilde{S}_X(\theta_0)$ is asymptotically $\mathrm{N}(0, I_p)$.

For the first example, we consider the case that 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 $\theta_X^{\dagger}$ by minimizing a convex function

$$\sum_{i=1}^{n} \frac{\Delta_i}{\hat{G}(Y_i)} |Y_i - \theta' V_i|. \qquad (3.4)$$

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

Let us use a lung cancer study data set analyzed by Ying et al. (1995) to illustrate the new procedure. For patients with small cell lung cancer, the standard therapy is to use 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 V has three components. The first component is one, corresponding to the intercept, the second is the patient's entry age, and the last one is the treatment indicator, which is one if the patient was assigned to A and zero otherwise. Since there are no loss-to-follow-ups 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 data set, the consistent estimate $\theta_X^\dagger$ from (3.4) is $(2.66, 0.0047, 0.073)'$ and its estimated covariance matrix Γ_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 $\hat{G}(3.27) \approx 0.1$, and $V_i'\theta_x^\dagger < t_0, i = 1, \cdots, n$. Furthermore, we let the covariance matrix $\Sigma_x = \Gamma_x$ as the initial proposal distribution. Also, we chose Ω_x using (2.2) with $\Phi(\alpha_n) = 6$. Based on the initial 1000 generated $\theta_{(k)}$, the acceptance rate is about 50%. We then used these Σ_x and Ω_x to generate 30000 $\theta_{(k)}$'s, but deleted the first 3000. The effective sample size based on these 27000 dependent $\theta_{(k)}$ is about 1400. Figure 1 provides a diagnostics quantile-quantile plot based on $\{T(\theta_{(k)}), k = 3001, \cdots, 30000\}$. 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 that the empirical distribution based on the above 27000 $\theta_{(k)}$'s is a good approximation to the distribution of θ_x^*. One may use the minimizer $\hat{\theta}_x$ described in Section 2 and the sample covariance matrix obtained from those 27000 $\theta_{(k)}$'s to estimate the mean and covariance matrix of θ_x^*. This results in $\hat{\theta}_x = (2.70, -0.039, 0.078)'$ with estimated standard errors of $0.039, 0.039$ and 0.040, respectively.

For the second example, we consider 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 (3.4) is no longer consistent. To obtain an initial $\theta_{(1)}$ and the proposal distribution for the MCMC procedure, we fitted the data with a parametric model by assuming that the error term $T - \theta_0'V$ is a zero-mean normal with an unknown variance. The maximum likelihood

estimate for θ_0 in Model (3.1) is $(2.76, -0.063, 0.088)'$ with an estimated covariance matrix

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

Since we don't have an initial consistent estimate to locate a proper Ω_x, we chose a quite large Ω_x in (2.2) with $\Phi(\alpha_n) = 15$ and let the above matrix be Σ_x in the initial proposal distribution. Under this setting, the acceptance rate based on the first 1000 iterations is about 45%. We then generated 30000 $\theta_{(k)}$'s, but deleted the first 3000.

In Figure 2, we present marginal trace plots and histograms corresponding to three parameters, the intercept, age effect, and treatment difference, based on 27000 $\theta_{(k)}$'s. It appears that for the first and third components, there are a number of outliers. To obtain robust estimators for the mean and covariance matrix of θ_x^*, we deleted $\theta_{(k)}$ such that either its first component is larger than 2.86 or the third one is larger than 0.25 by visually examining the plots in Figure 2. This results in deleting 662 $\theta_{(k)}$'s. In Figure 3, we present two Q-Q plots, the quantiles of the observed $T(\theta_{(k)}) = S_x'(\theta_{(k)})S_x(\theta_{(k)})$ against the quantiles from χ_3^2. The dotted line is constructed with the original 27000 $\theta_{(k)}$'s, and the dashed line is based on those 26338 selected $\theta_{(k)}$'s. Figure 3 shows that the above ad hoc trimming works well. The effective sample size based on these 26338 dependent $\theta_{(k)}$ is about 1000. Now, with those selected $\theta_{(k)}$'s, $\hat{\theta}_x = (2.70, -0.038, 0.079)'$. In the original scale of the covariates, the regression coefficient estimates are $2.89, -0.004$, and 0.16 with the corresponding estimated standard errors of $0.044, 0.005$ and 0.084, respectively. These estimates are practically identical to those obtained by Ying et al. (1995) via a rather complex numerical procedure.

Lastly, the observed value of $T(\hat{\theta}_x) = 0.017$, which is the 0.06th percentile of χ_3^2. This provides a justification that $\hat{\theta}_x$ is a solution to the equation $S_x(\theta) \approx 0$. Moreover, the values of $T(\theta)$ for the above 662 deleted $\theta_{(k)}$ are substantially larger than 0.017. Also, since the parameter space Ω_x used for generating realizations from (2.1) is quite large, $\hat{\theta}_x$ appears to be the unique root to the estimating equation. This, coupled with the fact

that for large n, theoretically this estimating equation has a unique solution, implies that $\hat{\theta}_x$ is the consistent root to the estimating equation.

For the third example, we relax the assumption that there exists a t_0 such that $\text{pr}(\theta_0' V < t_0) = 1$ in the previous two cases. Here, we only require that

$$\text{pr}(\theta_0' V < t_0) \geq \xi, \tag{3.5}$$

where $\xi > 0$, a pre-specified constant. This weaker condition allows us to expand the parameter space substantially. Moreover, we modify the estimating function (3.2) 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. To this end, consider the following estimating function

$$\tilde{S}_X(\theta) = n^{-1/2} \sum_{\theta' V_i \leq t_0} V_i \left[\frac{I(Y_i \geq \theta' V_i)}{\hat{G}(\theta' V_i)} - \frac{1}{2} \right]. \tag{3.6}$$

It is not difficult to show that if $\text{pr}(\theta_0' V < t_0) > 0$, there exists a consistent root $\hat{\theta}_X$ to the equation $\tilde{S}_X(\theta) \approx 0$. Asymptotically the covariance matrix for $\tilde{S}_X(\theta_0)$ is $\Pi_X(\theta_0)$, where $\Pi_X(\theta) =$

$$n^{-1} \sum_{\theta' V_i \leq t_0} \left\{ \frac{I(Y_i \geq \theta V_i)}{\hat{G}(\theta' V_i)} - \frac{1}{2} \right\}^2 V_i^{\otimes 2} - n^{-1} \sum_{i=1}^{n} \frac{1 - \Delta_i}{2} \left\{ \frac{\sum_{j=1}^{n} V_j I(\theta' V_j \in [Y_i, t_0])}{\sum_{j=1}^{n} I(Y_j \geq Y_i)} \right\}^{\otimes 2}.$$

It is well-known that even under Type I censoring, asymptotically the equation, $\tilde{S}_X(\theta) = 0$, may have multiple roots (Khan and Powell, 2001). Using similar arguments given in 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 above lung cancer data to illustrate our procedure. In order to visualize the results better, we considered the case with a single covariate, the treatment indicator, in our analysis. Thus, θ is a 2×1 vector. Like the previous case, we fitted the data with a fully parametric normal model. The point estimate and its estimated

covariance matrix are $\theta_x^\dagger = (2.76, 0.95)'$ and

$$\Gamma_x = \begin{pmatrix} 9.5 \times 10^{-4} & 2.0 \times 10^{-5} \\ 2.0 \times 10^{-5} & 9.4 \times 10^{-4} \end{pmatrix}. \tag{3.7}$$

Note that this parametric point estimator may not be consistent. For our procedure, we let $t_0 = 3.27$ and $\xi = 0.4$ in (3.5), and let $\Phi^{-1}(\alpha_n) = 15$ for Ω_x in (2.2). We find that with $\Sigma_x = 2\Gamma_x$, given in (3.7), the acceptance rate is about 40%. Under this setting, we generated 30000 $\theta_{(k)}$ and deleted the first 3000 $\theta_{(k)}$.

Figure 4 gives the scatter diagram based on these 27000 $\theta_{(k)}$'s. It appears that there are two clusters of points, which can be separated well using any standard clustering method. The distribution of the points on the left hand side is approximately normal with mean $\hat{\theta}_x = (2.70, 0.098)'$ and estimated standard errors of 0.039 and 0.039. The effective sample size based on these points is about 1200. The corresponding value of $T(\hat{\theta}_x) = 0.017$, which is the 0.8th percentile of χ^2_2. The effective sample size based on these points is about 1200. Note that this normal distribution is very similar to its counterpart in Example 2.

For the cluster of points on the right hand side of the figure, $\hat{\theta}_x = (2.95, 0.34)'$, with $T(\hat{\theta}_x) = 0.55$, which is the 24th percentile of χ^2_2. The distribution of this set of points cannot be approximated by a complete normal. If $(2.95, 0.34)'$ is a root, 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 that the large sample approximation is not applicable. Note that due to the extremely discrete nature of the covariate for the present case, one cannot enlarge the parameter space further by choosing a smaller ξ in (3.5).

4. REMARKS

The novel Markov chain marginal bootstrap method proposed by He & Hu (2002) only works for a special class of estimating functions of (1.3). It is interesting to note that under some regularity conditions, his procedure can be viewed as a special Metropolis algorithm, the Gibbs sampler, to generate θ from the target density function which is proportional to (2.1).

Under the semiparametric setting, for large n, the profile likelihood function is approximately proportional to $\exp\{-\frac{1}{2}(\theta-\hat{\theta}_x)'B_x^{-1}(\theta-\hat{\theta}_x)\}$, for θ in a small neighborhood of θ_0, where B_x is a deterministic matrix. Therefore, one can generate observations from a density which is proportional to the profile likelihood function to obtain an approximation to the covariance matrix B_x of $(\hat{\theta}_X-\theta_0)$ (Lee, Kosorok & Fine, 2005). If one applies our proposal to the profile likelihood score function $\tilde{S}_X(\theta)$, the resulting covariance matrix of θ_x^* is a robust sandwich-type estimate, which can be quite different from the one obtained by Lee et al. (2005). Generalizing the results from 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.

5. APPENDIX
THEORETICAL JUSTIFICATION

Assume that $S_X(\theta)$ satisfies the local linearity condition around θ_0

$$\sup_{\|\theta^{(j)}-\theta_0\|\leq\epsilon_n;j=1,2}\frac{\|S_X(\theta^{(2)})-S_X(\theta^{(1)})-n^{1/2}A(\theta^{(2)}-\theta^{(1)})\|}{1+n^{1/2}\|\theta^{(2)}-\theta^{(1)}\|}=o_p(1), \tag{5.1}$$

where A is a deterministic matrix and $\epsilon_n=o_p(1)$. This implies that

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_n}\frac{\|S_X(\theta)-n^{1/2}A(\theta-\hat{\theta}_X)\|}{1+n^{1/2}\|\theta-\hat{\theta}_X\|}=o_p(1), \tag{5.2}$$

and

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_n}\frac{|S_X(\theta)'S_X(\theta)-n(\theta-\hat{\theta}_X)'A'A(\theta-\hat{\theta}_X)|}{1+n\|\theta-\hat{\theta}_X\|^2}=o_p(1). \tag{5.3}$$

First, recall that $\tilde{\theta}_x$ is a random vector generated from a density function which is proportional to (2.1). Since $\tilde{\theta}_X\in\Omega_X$, it is consistent with respect to θ_0. To claim that the distribution of $\tilde{\theta}_x$ is a good approximation to that of θ_x^*, one needs to show that $S_x(\tilde{\theta}_x)$ is asymptotically normal with mean 0 and covariance matrix I_p. That is,

$$|E[h\{S_X(\tilde{\theta}_X)\}|X]-E\{h(Z)\}|=o_p(1),$$

where $h(\cdot)$ is any uniformly bounded Lipschitz continuous function $R^p \to R^+$, and Z is $\mathrm{N}(0, I_p)$.

Now,

$$E[h\{S_X(\tilde{\theta}_X)\} \mid X = x] = \frac{\int_{\Omega_x} h\{S_x(\theta)\} \exp\{-\frac{1}{2} S_x(\theta)' S_x(\theta)\} d\theta}{\int_{\Omega_x} \exp\{-\frac{1}{2} S_x(\theta)' S_x(\theta)\} d\theta}.$$

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

$$\mathcal{C}_1 = \{x : \quad \|S_x(\theta) - n^{1/2} A(\theta - \hat{\theta}_x)\| \leq \epsilon(1 + n^{1/2}\|\theta - \hat{\theta}_x\|), \theta \in \Omega_x\}$$

and

$$\mathcal{C}_2 = \{x : \quad |S_x(\theta)' S_x(\theta) - n(\theta - \hat{\theta}_x)' A' A(\theta - \hat{\theta}_x)| \leq 2\epsilon(1 + n\|\theta - \hat{\theta}_x\|^2/2), \theta \in \Omega_x\}.$$

It follows from (5.2) and (5.3) that for large n, $\mathrm{pr}(\mathcal{C}_1 \cap \mathcal{C}_2) > 1 - \epsilon$.

For $x \in \mathcal{C}_2$, $I_1(x) \leq$

$$\int_{\Omega_x} h\{S_x(\theta)\} \exp\{\epsilon - \frac{n}{2}(\theta - \hat{\theta}_x)'(A'A - \epsilon I_p)(\theta - \hat{\theta}_x)\} d\theta. \tag{5.4}$$

Let $z = n^{1/2} A(\theta - \hat{\theta}_x)$. Then (5.4) =

$$n^{-1/2}\|A\|^{-1} \int_{\Omega^*} h\{S_x(\hat{\theta}_x + n^{-1/2} A^{-1} z)\} \exp\{\epsilon - \frac{1}{2} z' A'^{-1}(A'A - \epsilon I_p) A^{-1} z\} dz,$$

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

For $x \in \mathcal{C}_1$

$$\|S_x(A^{-1} n^{-1/2} z + \hat{\theta}_x) - z\| \leq \epsilon(1 + \|z\|),$$

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

$$n^{1/2} I_1(x) \leq a\epsilon + \|A\|^{-1} \int_{\Omega^*} h(z) \exp\{-\frac{1}{2} z'(I_p - \epsilon(AA')^{-1}) z\} dz$$

$$\leq a\epsilon + \|A\|^{-1} \int_{\Omega^*} h(z) \exp\{-\frac{1 - \epsilon\lambda_1}{2} z' z\} dz,$$

where λ_1 is the largest eigenvalue of $(AA')^{-1}$. Let $s = (1-\epsilon\lambda_1)^{1/2}z$, then

$$n^{1/2}I_1(x) \leq a\epsilon + \|A\|^{-1}(1-\epsilon\lambda_1)^{-1/2}\int_{\Omega^+} h\{(1-\epsilon\lambda_1)^{-1/2}s\}\exp\{-\frac{1}{2}s's\}ds,$$

where $\Omega^+ = \{s : \ ||A^{-1}(1-\epsilon\lambda_1)^{-1/2}s + n^{1/2}(\hat{\theta}_x - \theta_x^\dagger)|| \leq c_n\}$. Since for small ϵ, $(1-\epsilon\lambda_1)^{-1/2} \approx 1+\lambda_1\epsilon/2$, $|h\{(1-\epsilon\lambda_1)^{-1/2}s\} - h(s)| \leq \epsilon a$. Therefore, for large n,

$$n^{1/2}I_1(x) \leq a\epsilon + \|A\|^{-1}\int_{\Omega^+} h(s)\exp\{-\frac{1}{2}s's\}ds \leq a\epsilon + \|A\|^{-1}\int_{R^p} h(s)\exp\{-\frac{1}{2}s's\}ds.$$

Similarly, it can be shown that $n^{1/2}I_1(x) \geq \|A\|^{-1}\int_{R^p} h(s)\exp\{-\frac{1}{2}s's\}ds - a\epsilon$.

This implies that

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

Using the same argument, one can show that

$$\left| \ n^{1/2}I_2(x) - \|A\|^{-1}\int_{R^p} \exp\{-\frac{1}{2}s's\}ds \ \right| = |n^{1/2}I_2(x) - \|A\|^{-1}(2\pi)^{p/2}| \leq a\epsilon.$$

Therefore, for a large enough n,

$$|E[h\{S_x(\theta^*)\} \mid X] - E\{h(Z)\}| = \left| \ \frac{n^{1/2}I_1(X)}{n^{1/2}I_2(X)} - \int_{R^p}\frac{h(s)}{(2\pi)^{p/2}}\exp\{-\frac{1}{2}s's\}ds \ \right| \leq a\epsilon. \quad (5.5)$$

It follows that the left hand side of (5.5) converges to 0, in probability, as $n \to \infty$.

REFERENCES

Arcones, M. & Gine, E. (1992). On the bootstrap of M-estimators and other statistical functionals. *Exploring the Limit of Bootstrap*. 14-47. (Ed. R. LegPage & L. Billard) New York: Wiley.

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

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

He, X. & Hu, F. (2002). Markov Chain Marginal Bootstrap. *J. Am. Statist. Assoc.* **97**, 783-95.

Hu, F. & Kalbfleisch, J. D. (2000). The estimating function bootstrap (with Discussion). *Can. J. Statist.* **28**, 449-99.

Khan, S. & Powell, J. L. (2001). Two-step estimation of semiparametric censored regression models. *J. Econometrics* **103**, 73-110.

Lee, B. L., Kosorok, M. & Fine, J. P. (2005). The profile sampler. *J. Am. Statist. Assoc.*, to appear.

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

Parzen, M. I., Wei, L. J. & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341-50.

Tian, L., Liu, J. S., Zhao, Y. & Wei, L. J. (2004). Statistical inferences based on non-smooth estimating functions. *Biometrika* **91**, 943-54.

Ying, Z., Jung, S. H., & Wei, L. J. (1995). Survival Analysis with Median Regression Model. *Biometrika* **90**, 178-84.

Figure 1: The Q-Q plot of empirical quantiles against quantiles from χ^2_3 based on 27000 observed $T(\theta) = S_x(\theta)' S_x(\theta)$ for Example 1

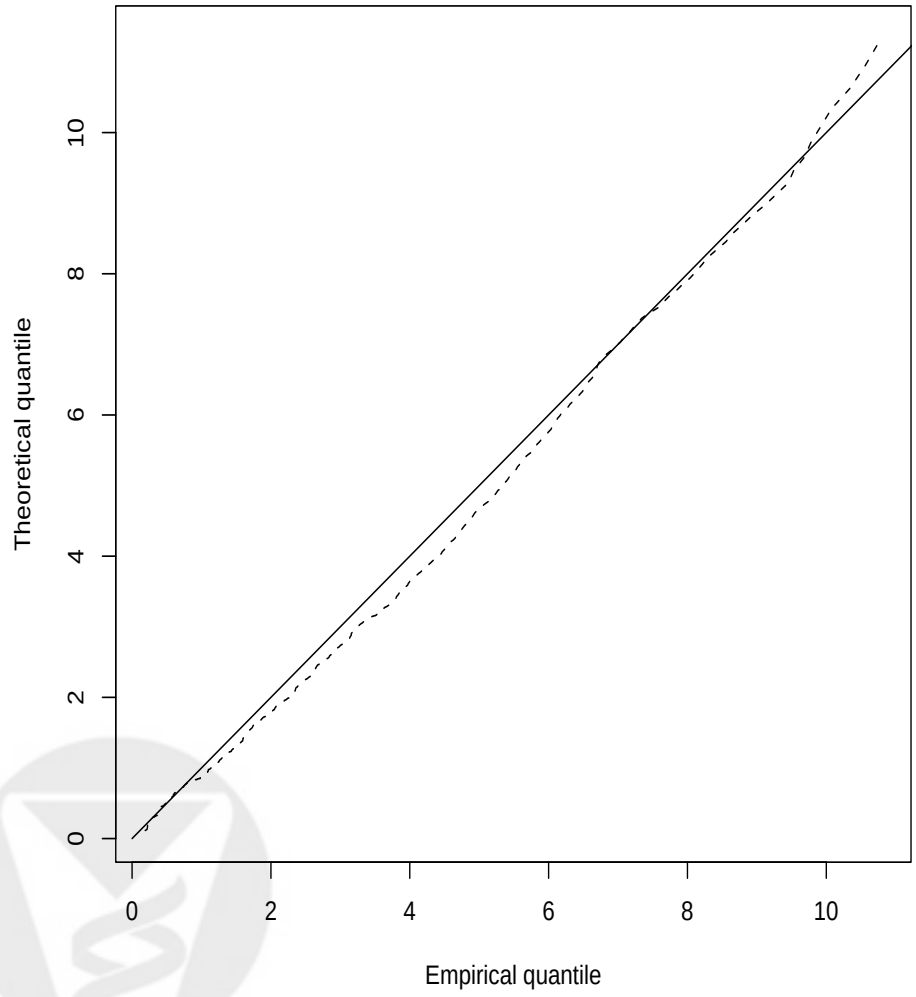

Figure 2: Marginal trace plots and histograms for intercept, age effect and treatment difference for Example 2

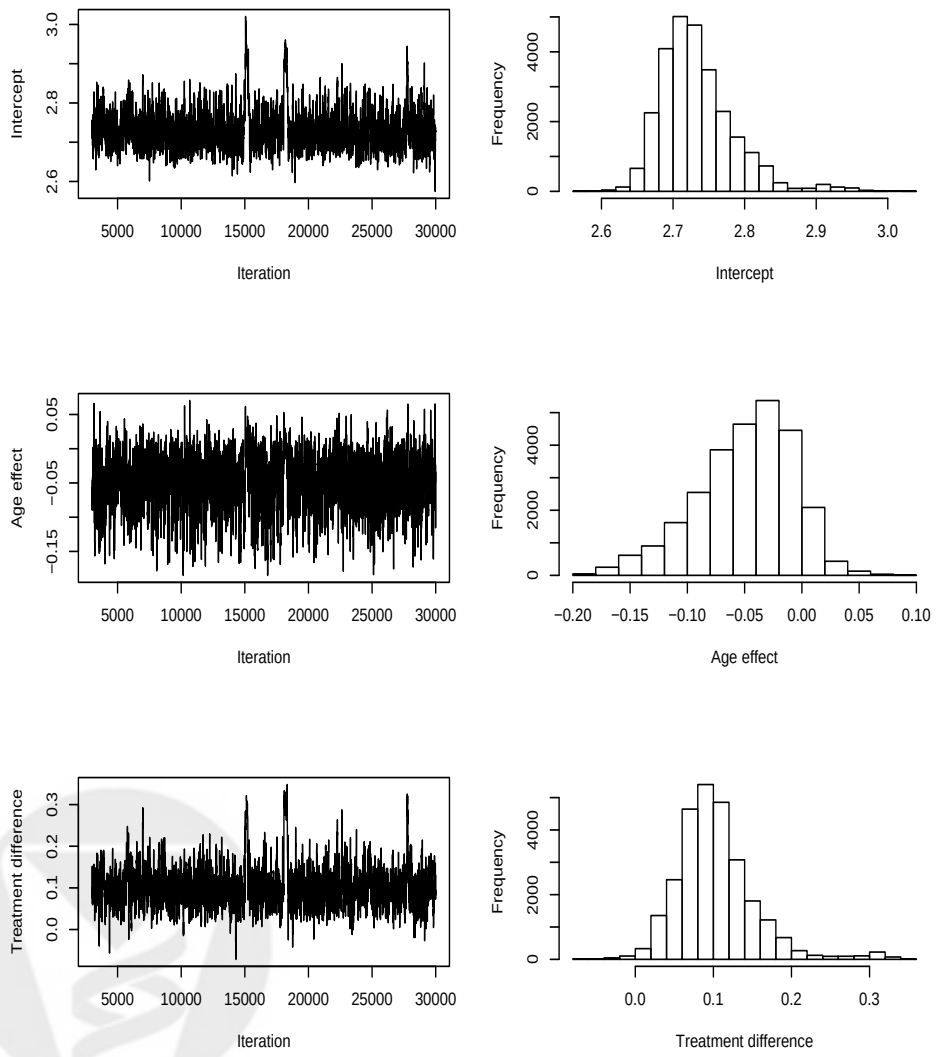

Figure 3: The Q-Q plots of empirical quantiles against quantiles from χ^2_3 based on untrimmed (dotted) and trimmed (dashed) observed $T(\theta)$ for Example 2

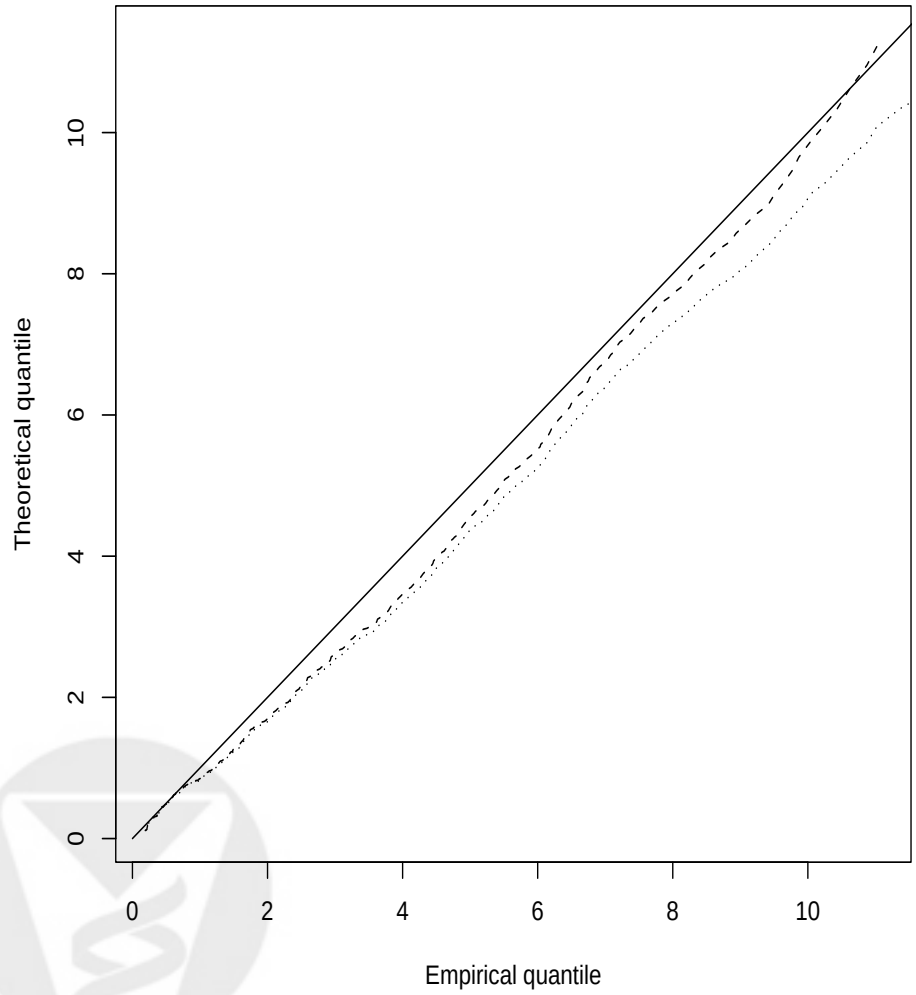

Figure 4: Scatter diagram for the intercept against the treatment difference