*Harvard University*
Harvard University Biostatistics Working Paper Series

*Year* 2003 *Paper* 5

# STATISTICAL INFERENCES BASED ON NON-SMOOTH ESTIMATING FUNCTIONS

Lu Tian* Jun S. Liu†
Mary Zhao‡ L. J. Wei**

*Harvard University, ltian@hsph.harvard.edu
†Harvard University, jliu@stat.harvard.edu
‡Harvard University, ymzhao@hsph.harvard.edu
**Harvard University, wei@hsph.harvard.edu

# STATISTICAL INFERENCES BASED ON NON-SMOOTH ESTIMATING FUNCTIONS

Lu Tian
Department of Biostatistics
Harvard University

Jun Liu
Department of Statistics
Harvard University

Mary Zhao and L.J. Wei
Department of Biostatistics
Harvard University

## SUMMARY

When the estimating function for a vector of parameters is not smooth, it is often rather difficult, if not impossible, to obtain a consistent estimate by solving the corresponding estimating equation using the standard numerical techniques. In this article, we propose a simple inference procedure via the importance sampling technique, which provides a consistent root to the estimating equation and also an approximation to its distribution without solving any equations or involving non-parametric function estimates. The new proposal is illustrated and evaluated via two extensive examples with real and simulated data sets.

*Keywords*: Importance sampling; $L_1$-norm; Linear regression for censored data; Resampling method.

## 1. INTRODUCTION

Suppose that inferences to be made about a vector $\theta_0$ of $p$ unknown parameters are based on a *non-smooth* estimating function $S_X(\theta)$, where $X$ is the observable random quantity. Often it is rather difficult to solve the corresponding estimating equation $S_X(\theta) \approx 0$ numerically, especially for the case when $p$ is large. Moreover, the equation may have multiple solutions, and it is not clear how to identify a consistent root $\hat{\theta}_X$ for $\theta_0$. Furthermore, the covariance matrix of $\hat{\theta}_X$ may involve a completely unknown density-like function and may not be estimated well directly under a nonparametric setting. With such a non-smooth estimating function, all the existing inference procedures, including resampling methods, for $\theta_0$ are difficult to implement in practice without additional information on $\theta_0$.

Now, assume that there is a consistent estimator $\hat{\theta}_X^{\dagger}$ readily available for $\theta_0$ from a relatively simple estimating function. Such a simple consistent estimator, which may not be efficient, is usually not difficult to obtain. For example, in a recent paper, Bang & Tsiatis (2002) proposed a novel estimation method for the quantile regression model with censored medical cost data. Their estimating function $S_X(\theta)$ is neither smooth nor monotone. On the other hand, as indicated in Bang & Tsiatis (2002), a consistent estimator for the vector of the regression parameters can be obtained easily via the standard inverse probability weighted estimation procedure. Other similar examples can be found in Robins & Rotnitzky (1992) and Robins et al. (1994). In this paper we use the importance sampling idea to derive a general and simple inference procedure, which utilizes $\hat{\theta}_X^{\dagger}$ to locate a consistent estimator $\hat{\theta}_X$ such that $S_X(\hat{\theta}_X) \approx 0$, and draws inferences about $\theta_0$. Our procedure does not need to solve any complicated equations. Moreover, it does not involve nonparametric function estimates or numerical derivatives (van der Vaart,1998, Section 5.7).

We illustrate the new proposal with two extensive examples. The first example demonstrates how to obtain a robust estimator based on the $L_1$ norm for the regression coefficients of the heteroscedastic linear regression model. The performance of the new pro-

cedure is evaluated via a real data set and an extensive simulation study. The second example shows how to derive a general rank estimation procedure for the regression coefficients of the accelerated failure time model in survival analysis (Kalbfleisch & Prentice, 2002, Chapter 7). Our procedure is much simpler and also more general than that recently proposed by Jin et al. (2003) for analyzing this particular model. The new proposal is illustrated with the well-known Mayo primary cirrhosis data and is also evaluated via an extensive simulation study.

## 2. DERIVATION OF CONSISTENT ESTIMATOR $\hat{\theta}_X$ AND ITS DISTRIBUTION

Suppose that the random quantity $X$ in $S_X(\theta)$ is indexed implicitly by, for example, the sample size $n$. Assume that as $n \to \infty$, the random vector $S_X(\theta_0)$ converges weakly to a multivariate normal $\mathrm{MN}(0, I_p)$, where $I_p$ is the $p \times p$ identity matrix. Furthermore, for large $n$, assume that as a function of $\theta$, $S_X(\theta)$ is approximately linear in a small neighborhood of $\theta_0$. The formal definition of the local linearity property of $S_X(\theta)$ is given in (5.1) of the Appendix. It follows that for a consistent estimator $\hat{\theta}_X$ such that $S_X(\hat{\theta}_X) \approx 0$, the random vector $n^{1/2}(\hat{\theta}_X - \theta_0)$ is asymptotically normal. When the above limiting covariance matrix involves a completely unknown density-like function and is difficult to estimate well directly, various resampling methods may be utilized to make inferences about $\theta_0$ (Efron & Tibshirani, 1993; Hu & Kalbfleisch, 2000). Recently, Parzen et al. (1994) and Goldwasser et al. (2003) studied a resampling procedure which takes advantage of the pivotal feature of $S_X(\theta_0)$. Specifically, let $x$ be the observed value of $X$ and let the random vector $\theta_x^*$ be a solution to the stochastic equation: $S_x(\theta_x^*) \approx G$, where $G$ is $\mathrm{MN}(0, I_p)$. If $\theta_X^*$ is consistent for $\theta_0$, then the distribution of $n^{1/2}(\hat{\theta}_X - \theta_0)$ can be approximated well by the conditional distribution of $n^{1/2}(\theta_x^* - \hat{\theta}_x)$. In practice, one can generate a large random sample $\{g_m, m = 1, \cdots, M\}$ from $G$ and obtain a large number of independent realizations of $\theta_x^*$ by solving the equations $S_x(\theta) \approx g_m, m = 1, \cdots, M$. The sample covariance matrix based on those $M$ realizations of $\theta_x^*$ can then be used to estimate the covariance matrix

of $\hat{\theta}_X$.

When the equation $S_x(\theta) = g$ is difficult to solve numerically, all the resampling methods in the literature do not work well. Here we show how to take advantage of having an initial consistent estimator $\hat{\theta}_X^\dagger$ from a simple estimating function to identify $\hat{\theta}_X$ and approximate its distribution without solving any complicated equations. The theoretical justification of the new procedure is given in the Appendix.

First let us generate $M$ vectors $\theta_x^{(m)}, m = 1, \cdots, M$, in a small neighborhood of $\theta_0$, where

$$\theta_x^{(m)} = \hat{\theta}_x^\dagger + \Sigma_x \tilde{g}_m, \tag{2.1}$$

$n^{1/2}\Sigma_x$ converges to a $p \times p$ deterministic matrix as $n \to \infty$, $\tilde{g}_m = g_m$, if $||g_m|| \leq c_n$, and is 0, otherwise, $c_n \to \infty$, and $c_n = o(n^{1/2})$. Note that $\tilde{g}_m$ in (2.1) is a slightly truncated $g_m$, which is a realization from $G$. By the local linearity property of $S_X(\theta)$ around $\theta_0$, $\{S_x(\theta_x^{(m)}), m = 1, \cdots, M\}$ is a set of independent realizations from a distribution which can be approximated by a multivariate normal with mean $\mu_x = S_x(\hat{\theta}_x^\dagger)$ and covariance matrix $\Lambda_x$, the sample covariance matrix constructed from $M$ observations $\{S_x(\theta_x^{(m)}), m = 1, \cdots, M\}$.

Now, let $\theta_x$ be the random vector which is uniformly distributed on the discrete set $\{\theta_x^{(m)}, m = 1, \cdots, M\}$. Then, the distribution of $S_x(\theta_x)$ can be approximated by a normal with mean $\mu_x$ and covariance matrix $\Lambda_x$. For the resampling method by Parzen et al. (1994), one needs to construct a random vector $\theta_x^*$ such that the distribution of $S_x(\theta_x^*)$ is approximately $\mathrm{MN}(0, I_p)$. This can be done using the importance sampling idea discussed in Liu (2001, Chapter 2) and Rubin (1987) in the context of Bayesian analysis and multiple imputation. Specifically, let $\theta_x^*$ be a random vector defined on the same support of $\theta_x$, but let its mass at $t = \theta_x^{(m)}$ be proportional to

$$\frac{\phi(S_x(t))}{\phi(\Lambda_x^{-1/2}(S_x(t) - \mu_x))}, \tag{2.2}$$

where $\phi(\cdot)$ is the density function of $\mathrm{MN}(0, I_p)$. Note that the numerator of (2.2) is the density function of the target distribution $\mathrm{MN}(0, I_p)$, and the denominator is the normal

approximation to the density function of $S_x(\theta_x)$. In the Appendix, we show that the distribution of $S_x(\theta_x^*)$ is approximately MN$(0, I_p)$ for large $M$ and $n$, and with $g_m$ in (2.1) being truncated by $c_n$, $\theta_X^*$ is consistent. Moreover, if we let $\hat{\theta}_x$ be the mean of $\theta_x^*$, then $S_X(\hat{\theta}_X) \approx 0$, and the unconditional distribution of $n^{1/2}(\hat{\theta}_X - \theta_0)$ can be approximated well by the conditional distribution of $n^{1/2}(\theta_x^* - \hat{\theta}_x)$.

The choice of $\Sigma_x$ in (2.1) greatly affects the efficiency of the above procedure. Empirically we find that our proposal performs well in an iterative fashion similar to the adaptive importance sampling considered by Oh and Berger (1992) in a different context. That is, one may start with an initial matrix $\Sigma_x$, for example, $n^{-1/2}I_p$, to generate $\{\theta_x^{(l)}, l = 1, \cdots, L\}$ via (2.1) for obtaining an intermediate $\theta_x^*$ via (2.2), where $L$ is relatively smaller than $M$. If the distribution of $S_x(\theta_x^*)$ is "close enough" to that of MN$(0, I_p)$, we generate additional $\{\theta_x^{(m)}, m = 1, \cdots, (M - L)\}$ under the same setting to obtain an accurate normal approximation to the distribution of $S_x(\theta_x)$ and a final $\theta_x^*$. Otherwise, we generate a fresh set of $\{\theta_x^{(l)}, l = 1, \cdots, L\}$ via (2.1) with an updated $\Sigma_x$, which, for example, is the covariance matrix of $\theta_x^*$ from the previous iteration, construct a new intermediate $\theta_x^*$ via (2.2), and then decide if this adaptive process should be terminated at this stage or not. The "closeness" between the distributions of $S_x(\theta_x^*)$ and MN$(0, I_p)$ can be evaluated numerically or graphically. For each iteration, the standard coefficient of variation of the unnormalized weight (2.2) can also be used to monitor the adaptive procedure (Liu, 2001, Chapter 2). If the above sequential procedure does not stop within a reasonable number of iterations, we may repeat the entire process from the beginning with a new initial matrix $\Sigma_x$ in (2.1). In Section 3.1, we use an example to show how to modify this initial matrix for an entirely fresh run of the adaptive process.

Based on our extensive numerical studies for the two examples in Section 3, we find that the truncation of $g_m$ by $c_n$ in (2.1) is not essential in practice.

## 3. EXAMPLES

### 3.1 INFERENCES FOR HETEROSCEDASTIC LINEAR REGRESSION MODEL

Let $T_i$ be the $i$th response variable and $z_i$ be the corresponding covariate vector, $i = 1, \cdots, n$. Here, $X = \{(T_i, z_i), i = 1, \cdots, n\}$. Assume that

$$T_i = \theta_0' z_i + \epsilon_i, \tag{3.1}$$

where $\epsilon_i, i = 1, \cdots, n$, are mutually independent and have mean 0, but the distribution of $\epsilon_i$ may depend on $z_i$. Under this setting, the least squares estimate $\hat{\theta}_X^{\dagger}$ is consistent for $\theta_0$.

Now, if the distribution of $\epsilon$ is symmetric about 0, an alternative way to estimate $\theta_0$ is to use a minimizer $\hat{\theta}_X$ of the $L_1$ norm $\sum_{i=1}^n |T_i - \theta' z_i|$. This estimator is asymptotically equivalent to a solution to the estimating equation $S_X(\theta) = 0$, where

$$S_X(\theta) = \Gamma^{-1} \sum_{i=1}^n z_i \{I(T_i - \theta' z_i \leq 0) - 1/2\}, \tag{3.2}$$

$I(\cdot)$ is the indicator function and $\Gamma = \{\sum_{i=1}^n z_i z_i'\}^{1/2}/2$. It is easy to show that $S_X(\theta_0)$ is asymptotically $\mathrm{MN}(0, I_p)$. The point estimate $\hat{\theta}_X$ can be obtained via the linear programming technique (Barrodale & Roberts, 1977; Koenker & Bassett, 1978; Koenker & D'Orey, 1987). Furthermore, Parzen et al. (1994) demonstrated that $S_X(\theta)$ is locally linear around $\theta_0$, and proposed a novel way to solve $S_x(\theta) = g$, for any given vector $g$, to generate realizations of $\theta_x^*$. Our proposal is readily applicable to the present case and does not need to solve the above equation repeatedly.

Let us use a small data set on survival times in patients with a specific liver surgery (Neter et al., 1985, p.419) to illustrate our proposal and compare the results with those given by Parzen et al. (1994). This data set has 54 files, and each file consists of the uncensored survival time of a patient with four covariates: blood clotting score, prognostic index, enzyme function test score and liver function test score. Here, we let $T$ be the base

10 logarithm of the survival time and $z$ be a $5 \times 1$ covariate vector with the first component being the intercept. We used the iterative procedure described at the end of Section 2 with $L = 1000$ for each iteration, and $M = 3000$.

First, we let the initial matrix $\Sigma_x$ in (2.1) be $n^{-1/2} I_5$. However, after twenty iterations, we found that the covariance matrix of $S_x(\theta_x^*)$ was markedly different from the matrix $I_5$. We noticed that after the first iteration of the above process, the components of $\{S_x(\theta_x)\}$ were highly correlated and a large number of masses in (2.2) were almost zero, which gave a quite poor approximation to the target distribution $\mathrm{MN}(0, I_5)$. To search for a better choice of $\Sigma_x$, we observed that if the error term in (3.1) is free of $z_i$, for large $n$, the slope of $S_x(\theta)$ around $\theta_0$ is proportional to $\{\sum_{i=1}^n z_i z_i'\}^{1/2}$. This suggests that if one let

$$\Sigma_x = n^{1/2} \{\sum_{i=1}^{n} z_i z_i'\}^{-1} \tag{3.3}$$

in (2.1), the covariance matrix of the resulting $S_x(\theta_x)$ would be approximately diagonal and the corresponding distribution of $S_x(\theta_x)$ is expected to be a better approximation to $\mathrm{MN}(0, I_5)$. With this initial $\Sigma_x$ and $\hat{\theta}_x^{\dagger}$ being the least squares estimate for $\theta_0$, after three iterations, the maximum of the absolute values of the component-wise differences between the covariance matrix of $S_x(\theta_x^*)$ and $I_5$ was about 0.05. Based on additional 2000 $\theta_x^{(m)}$ generated from (2.1) under the setting at the beginning of the 3rd iteration, we obtained the point estimate $\hat{\theta}_x$ and the estimated standard error for each of its components. We report these estimates in Table 1 along with those from Parzen et al. (1994). It is interesting to note that for the present example, our procedure performs better than that by Parzen et al. (1994). With our point estimate $\hat{\theta}_x$, $||S_x(\hat{\theta}_x)|| = 1.71$, but with the method by Parzen et al., the corresponding value is 2.75. Also note that for our iterative procedure the coefficient of variation of the final weights is less than 0.5, indicating that it is appropriate to stop the process after the third iteration.

To further examine the performance of the new proposal for cases with small sample sizes, we fitted the above data with (3.1) using the ordinary least squares estimation procedure. If we assume that the error terms are independent and identically distributed,

the variance estimate for the error is 0.002. We then considered a linear model with the true regression coefficients $\theta_0$ being the least squares estimates, but with the error term being a contaminated normal $\frac{2}{3}N(0, 0.002) + \frac{1}{3}N(0, 0.019)$. With the set of the observed covariate vectors $\{z_i, i = 1, \cdots, 54\}$ from the liver surgery example, we simulated 500 samples $\{T_i, i = 1, \cdots, 54\}$ from this model. For each simulated sample, we used the above iterative procedure to obtain $\hat{\theta}_x$ and its estimated covariance matrix. In Figure 1, we display five Q-Q plots. Each plot was constructed for a specific regression parameter to examine if the empirical distribution based on the above 500 standardized estimates, each of which was centered by the corresponding component of $\theta_0$ and divided by the estimated standard error, is approximately a univariate normal with mean 0 and variance one. Except for the extreme tails, the marginal normal approximation to the distribution of $\hat{\theta}_X$ seems quite satisfactory. To examine how well our point estimator performs, for each of the above simulated data sets, we computed the value $||S_x(\hat{\theta}_x)||$ and its counterpart from Parzen et al. In Figure 2, we present the scatter plot based on those 500 paired values. The new procedure tends to have a smaller norm of the estimating function evaluated at the observed point estimate than that of Parzen et al.

### 3.2. INFERENCES FOR LINEAR MODEL WITH CENSORED DATA

In this section, let $T_i$ be the logarithm of the time to a certain event for the $i$th subject in Model (3.1). Furthermore, we assume that the error terms of the model are independent and identically distributed with a completely unspecified distribution function. The vector $\theta_0$ of the regression parameters does not include the intercept term. Furthermore, $T$ may be censored by $C$, and conditional on $z$, $T$ and $C$ are independent. Here, the data $X = \{(Y_i, \Delta_i, z_i), i = 1, \cdots, n\}$, where $\Delta = I(T \leq C)$ and $Y = \min(T, C)$. In survival analysis, this log-linear model is called accelerated failure time model and has been extensively studied, for example, by Buckley & James (1979), Prentice (1978), Ritov (1990), Tsiatis (1990), Wei et al. (1990) and Ying (1993). An excellent review on this topic is given in Kalbfleisch & Prentice (2002).

A commonly used method for making inferences about this model is based on the rank estimation of $\theta_0$. Specifically, let $e_i(\theta) = Y_i - \theta' z_i$, $N_i(\theta; t) = \Delta_i I(e_i(\theta) \leq t)$ and $V_i(\theta; t) = I(e_i(\theta) \geq t)$. Also, let $S^{(0)}(\theta; t) = n^{-1} \sum_{i=1}^n V_i(\theta; t)$ and $S^{(1)}(\theta; t) = n^{-1} \sum_{i=1}^n V_i(\theta; t) z_i$. The rank estimating functions for $\theta_0$ is

$$\tilde{S}_X(\theta) = n^{-1/2} \sum_{i=1}^n \Delta_i w(\theta; e_i(\theta)) \{z_i - \bar{z}(\theta; e_i(\theta))\}, \tag{3.4}$$

where $\bar{z}(\theta; t) = S^{(1)}(\theta; t)/S^{(0)}(\theta; t)$, and $w$ is a possibly data-dependent weight function. Under the regularity conditions of Ying (1993, p.80), the distribution of $\tilde{S}_X(\theta_0)$ can be approximated by a normal with mean 0 and covariance matrix $\Gamma(\theta_0)$, where $\Gamma(\theta) = n^{-1} \sum_{i=1}^n \int_{-\infty}^{\infty} w^2(\theta; t) \{z_i - \bar{z}(\theta; t)\}^2 dN_i(\theta; t)$, and $\tilde{S}_X(\theta)$ is approximately linear in a small neighborhood of $\theta_0$. Note that under our setting, the estimating function is

$$S_X(\theta) = \Gamma(\theta)^{-1/2} \tilde{S}_X(\theta). \tag{3.5}$$

It follows that $S_X(\theta_0)$ is asymptotically $\mathrm{MN}(0, I_p)$.

When $w(\theta; t) = S^{(0)}(\theta; t)$, the Gehan-type weight function, the estimating function $\tilde{S}_X(\theta)$ is monotone, and the corresponding estimate $\hat{\theta}_X^{\dagger}$ can be obtained by the linear programming technique (Jin et al., 2003). When the weight function $w(\theta; t)$ is *monotone* in $t$, Jin et al. (2003) showed that one can use an iterative procedure with the Gehan-type estimate as the initial value to obtain a consistent root $\hat{\theta}_X$ to the equation: $\tilde{S}_X(\theta) = 0$.

With our new proposal, one can obtain a consistent estimator $\hat{\theta}_X$ for $\theta_0$ based on $S_X(\theta)$ in (3.5) and an approximation to its distribution *without* assuming that the weight function $w$ is monotone. A popular class of non-monotone weight functions is

$$w(\theta; t) = \{\hat{F}_\theta(t)\}^a \{1 - \hat{F}_\theta(t)\}^b, \tag{3.6}$$

where $a, b > 0$ and $1 - \hat{F}$ is the Kaplan-Meier estimate based on $\{(e_i(\theta), \Delta_i), i = 1, \cdots, n\}$ (Harrington & Fleming, 1991; Kosorok & Lin, 1999). Note that one may simplify the estimating function $S_X(\theta)$ by replacing $\theta$ in $\Gamma(\theta)$ of (3.5) with $\hat{\theta}_X^{\dagger}$.

For illustration, we applied the new method to the Mayo primary biliary cirrhosis data (Fleming & Harrington, 1991, Appendix D.1). This data set consists of 416 complete files,

and each of them contains the information on the survival time and various prognostic factors. In order to compare our results with those presented in Jin et al. (2003), we only used five covariates in our analysis: oedema, age, log(albumin), log(bilirubin) and log(protime). The initial consistent estimate $\hat{\theta}_x^{\dagger} = (-0.878, -0.026, 1.59, -0.579, -2.768)'$, utilized in (2.1) was the one based on the Gehan weight function. First, we considered $S_X(\theta)$ with the logrank weight function $w(\theta; t) = 1$. To generate $\{\theta_x^{(m)}\}$ via (2.1), we let $\Sigma_x = n^{-1/2} I_5$. Under the set-up of the iterative process discussed in Section 3.1, the adaptive procedure was terminated at the second stage. The coefficient of variation of the weights (2.2) at this stage is less than 0.5. We report our point estimates and their estimated standard errors in Table 2 along with those obtained by Jin et al. (2003). For the present example, our procedure outperforms the iterative method by Jin et al. The norm of $S_x(\hat{\theta}_x)$ with our point estimate is 0.13, in contrast to 0.29 with the method by Jin et al. In Table 2, we also report the results based on a *non-monotone* weight function $w$ in (3.6) with $a = b = 1/2$.

To examine the performance of the new proposal under the accelerated failure time with various settings, we conducted an extensive simulation study. Specifically, we generated the logarithm of the failure times via the model

$$T = 13.73 - 0.898 \times \text{oedema} - 0.026 \times \text{age} + 1.533 \times \log(\text{albumin}) -$$

$$0.593 \times \log(\text{bilirubin}) - 2.428 \times \log(\text{protime}) + \epsilon, \tag{3.7}$$

where $\epsilon$ is a normal random variable with mean 0 and variance 0.947. The regression coefficients and the variance of $\epsilon$ in (3.7) were estimated from the parametric normal regression model with the Mayo liver data. For our simulation study, the censoring variable is the logarithm of the uniform distribution on $(0, \xi)$, where $\xi$ was chosen to yield a pre-specified censoring proportion. For each sample size $n$, we chose the first $n$ observed covariate vectors in the Mayo data set. With these fixed covariate vectors, we used (3.7) to simulate 500 sets of $\{T_i, i = 1, \cdots, n\}$ and created 500 corresponding sets of possibly censored failure time data with a desired censoring rate. We then applied the

above iterative method based on $S_x(\theta)$ in (3.5) with the logrank weight. For the case that $n = 200$ and the censoring rate is about 50%, we present the Q-Q plots with respect to five regression coefficients in Figure 3. Each plot was constructed based on 500 standardized estimates for a specific regression parameter derived from the vectors $\hat{\theta}_x$. The marginal normal approximation to the distribution of our estimator appears quite accurate for the case with a moderate sample size and heavy censoring. In Table 2, for various sample sizes $n$ and censoring rates, we report the empirical coverage probabilities of 0.95 and 0.90 confidence intervals for each of the five regression coefficients based on our iterative procedure with the logrank weight function. The new procedure performs well for all the cases studied here.

## 4. REMARKS

For an estimating function $S_X(\theta)$, which is neither smooth nor monotone in $\theta$, generally it is difficult, if not impossible, to identify the consistent roots to the equation $S_X(\theta) \approx 0$, especially when the dimension of $\theta$ is large. With an initial consistent estimator $\hat{\theta}_x^{\dagger}$ for $\theta_0$ based on a simple estimation procedure, one may identify a consistent root to $S_X(\theta) \approx 0$ and obtain an approximation to the distribution of such an estimator via the simple importance sampling scheme proposed in the paper.

In practice, the initial choice of $\Sigma_x$ in (2.1) for obtaining $\{\theta_x^{(m)}\}$ may have a significant impact on the efficiency of the adaptive procedure. When the sample size is moderate or large as for the Mayo primary bililary cirrhosis and simulated examples presented in Section 3.2, we find that generally, our proposal with $n^{1/2}\Sigma_x$ in (2.1) being the simple identity matrix performs well even with only a very few iterations. On the other hand, for a small sample case with a rather discrete estimating function $S_X(\theta)$, a naive choice of $\Sigma_x$ may not work well.

## 5. APPENDIX

Assume that for a sequence of constants, $\{\epsilon_n\} \to 0$, there exists a deterministic matrix

$D$ such that for $l = 1, 2$,

$$\sup_{\|\theta_l - \theta_0\| \le \epsilon_n} \frac{\|S_X(\theta_2) - S_X(\theta_1) - Dn^{1/2}(\theta_2 - \theta_1)\|}{1 + n^{1/2}\|\theta_2 - \theta_1\|} = o_{P_X}(1), \tag{5.1}$$

where $\mathrm{P}_X$ is the probability measure generated by $X$. Now, for the observed $x$ of $X$, let $\tilde{\theta}_x = \hat{\theta}_x^\dagger + \Sigma_x G I(\|G\| \le c_n)$, where $G$ is $\mathrm{MN}(0, I_p)$. Note that as $M \to \infty$, the distribution of $\theta_x$, which is the uniform random vector on the discrete set $\{\theta_x^{(m)}, m = 1, \cdots, M\}$ discussed in Section 2, is the same as that of $\tilde{\theta}_x$. Let $\mathrm{P}_G$ be the probability measure generated by $G$ and P be the product measure $\mathrm{P}_X \times \mathrm{P}_G$. Then, since $\tilde{\theta}_X$ and $\hat{\theta}_X^\dagger$ are in a $o_P(1)$-neighborhood of $\theta_0$, it follows from (5.1) that

$$S_X(\tilde{\theta}_X) - S_X(\hat{\theta}_X^\dagger) = n^{1/2} D \Sigma_X G + o_P(1).$$

This implies that

$$|\mathrm{E}\{h(S_X(\tilde{\theta}_X))|X\} - \int_{R^p} \frac{h(t)}{\sqrt{|\tilde{\Lambda}_X|}} \phi(\tilde{\Lambda}_X^{-1/2}(t - S_X(\hat{\theta}_X^\dagger)))dt| = o_{P_X}(1), \tag{5.2}$$

where $\tilde{\Lambda}_x$ is the limit of $\Lambda_x$, as $M \to \infty$, $h(\cdot)$ is any uniformly bounded, Lipschitz function, and the expectation E in (5.2) is taken under $\mathrm{P}_G$. Note that loosely speaking (5.2) indicates that for $X = x$, the distribution of $S_x(\tilde{\theta}_x)$ is approximately $\mathrm{MN}(\mu_x, \tilde{\Lambda}_x)$. Now, for given $x$, as $M \to \infty$, the distribution function of $\theta_x^*$ at $t$ converges to

$$c_x \mathrm{E}\{I(\tilde{\theta}_x \le t) \frac{\phi(S_x(\tilde{\theta}_x))}{\phi(\tilde{\Lambda}_x^{-1/2}(S_x(\tilde{\theta}_x) - S(\hat{\theta}_x^\dagger)))}\},$$

where $c_x$ is the normalized constant which is free of $t$. This implies that for large $M$,

$$\mathrm{E}(h(S(\theta_X^*))|X) \approx c_X \mathrm{E}\{h(S_X(\tilde{\theta}_X)) \frac{\phi(S_X(\tilde{\theta}_X))}{\phi(\tilde{\Lambda}_X^{-1/2}(S_X(\tilde{\theta}_X) - S_X(\hat{\theta}_X^\dagger)))}|X\}, \tag{5.3}$$

where $h$ is any uniformly bounded, Lipschitz continuous function. With (5.2), it is straightforward to show that the absolute value of the difference between the right hand side of (5.3) and $\int_{R^p} h(t)\phi(t)dt$ is $o_{P_X}(1)$. It follows that the conditional distribution of $S_X(\theta_X^*)$ is approximately $\mathrm{MN}(0, I_p)$ in a certain probability sense. That is, for any $p$-dimensional vector $t$,

$$|\mathrm{pr}(S_X(\theta_X^*) < t|X) - \Phi(t)| = o_{P_X}(1),$$

where $\Phi(\cdot)$ is the distribution function of $\mathrm{MN}(0, I_p)$.

The consistency for $\theta_X^*$ follows from the fact that $\theta_X^{(m)}, m = 1, \cdots, M$, are truncated by $c_n = o(n^{1/2})$.

## REFERENCES


Bang, H. and Tsiatis, A. A. (2002), Median regression with censored cost data, *Biometrics*, **3**, 643-649.

Barrodale, I. and Roberts, F. D. K. (1977), Algorithms for restricted least absolute value estimation, *Communications in Statistics, Part B-Simulation and Computation*, **6**, 353-364.

Buckley, J. and James I. (1979), Linear regression with censored data, *Biometrika*, **66**, 429-436.

Efron, B. and Tibshirani, R. (1993), An introduction to bootstrap, Chapman & Hall Ltd, London.

Fleming, T. R. and Harrington, D. (1991), Counting processes and survival analysis, John Weiley & Sons, New York.

Goldwasser, M. I., Tian, L., and Wei, L. J. (2003), Statistical inferences for infinite dimensional parameters via asymptotically pivotal estimating functions, *Biometrika*, to appear.

Hu, F. and Kalbfleisch, J. D. (2000), The estimating function bootstrap, *The Canadian Journal of Statistics*, **28**, 449-481.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying Z. (2003), Rank-based inference for the accelerated failure time model, *Biometrika*, **90**, 341-353.

Kalbfleisch, J. D. and Prentice, R. L. (2002), The statistical analysis of failure time

data, John Wiley & Sons, New York.

KOENKER, R. and BASSETT G. J. (1978), Regression quantiles, *Econometrica*, **46**, 33-50.

KOENKER, R. and D'OREY V. (1987), Computing regression quantiles, *Applied Statistics*, **36**, 383-393.

KOSOROK, M. R. and LIN, C. (1999), The versatility of function-indexed weighted log-rank statistics, *J. Am. Statist. Assoc.*, **94**, 320-332.

LIU, J. (2001), Monte Carlo strategies in scientific computing, Springer, New York.

NETER, J., WASSERMAN, W., and KUTNER, M. H. (1985), Applied linear statistical models: regression, analysis of variance, and experimental designs, Richard D. Irwin Inc, Homewood, IL.

OH, M.S. and BERGER, J.O. (1992), Adaptive importance sampling in Monte Carlo integration, *J. Am. Statist. Assoc.*, **41**, 143-168.

PARZEN, M. I., WEI, L. J., and YING Z. (1994), A resampling method based on pivotal estimating functions, *Biometrika*, **81**, 341-350.

PRENTICE, R. L. (1978), Linear rank tests with censored data, *Biometrika*, **65**, 167-180.

RITOV, Y. (1990), Estimation in a linear regression model with censored data, *The Annals of Statistics*, **18**, 303-328.

ROBINS, J. M. and ROTNITZKY, A. (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, *Aids Epidemiology, Methodological issues*, Birkhauser, 297-331.

ROBINS, J. M., RONTNITZKY, A., and ZHAO, L. P. (1994), Estimation of Regression Coefficients when some regressors are not always observed, *J. Am Statist. Assoc.*, **89**, 846-866.

RUBIN, D. (1987), A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm, *J. Am. Statist. Assoc.*, **82**, 543-546.

TSIATIS, A. A. (1990), Estimating regression parameters using linear rank tests for censored data, *The Annals of Statistics*, **18**, 354-372.

VAN DER VART, A. W. (1998), Asymptotic statistics, Cambridge University Press, New York.

WEI, L. J., YING, Z., and LIN, D. Y. (1990), Linear regression analysis of censored survival data based on rank tests, *Biometrika*, **77**, 845-851.

YING, Z. (1993), A large sample study of rank estimation for censored regression data, *The Annals of Statistics*, **21**, 76-99.

Table 1: $L_1$ estimates for heteroscedastic linear regression with the surgical unit data

| | New method | | Parzen's method | |
|---|---|---|---|---|
| Parameter | Est | SE | Est | SE |
| Intercept | 0.4146 | 0.0535 | 0.4151 | 0.0649 |
| BCS* | 0.0735 | 0.0058 | 0.0710 | 0.0075 |
| PI | 0.0096 | 0.0004 | 0.0098 | 0.0005 |
| EFTS | 0.0098 | 0.0003 | 0.0097 | 0.0003 |
| LFTS | 0.0029 | 0.0071 | 0.0029 | 0.0092 |

*BCS: blood clotting score; PI: prognostic index;

EFTS: enzyme function test score; LFTS: liver function test score

Table 2: Accelerated failure time regression for the Mayo primary biliary cirrhosis data

| | Weight function | | | | | |
|---|---|---|---|---|---|---|
| | Logrank | | | | $\hat{F}_\theta(t)^{0.5}(1-\hat{F}_\theta(t))^{0.5}$ | |
| | New method | | Jin's method | | New method | |
| Parameter | Est | SE | Est | SE | Est | SE |
| Oedema | -0.7173 | 0.2385 | -0.7338 | 0.1781 | -0.5903 | 0.2798 |
| Age | -0.0266 | 0.0054 | -0.0265 | 0.0042 | -0.0268 | 0.0063 |
| Log(albumin) | 1.6157 | 0.4939 | 1.6558 | 0.3683 | 1.5576 | 0.4978 |
| Log(bilirubin) | -0.5773 | 0.0559 | -0.5849 | 0.0455 | -0.5732 | 0.0589 |
| Log(protime) | -1.8800 | 0.5620 | -1.9439 | 0.4622 | -1.4995 | 0.6280 |

Table 3: Empirical coverage probabilities of confidence intervals from the simulation study for the AFT model

| | | Censoring | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0% | | 25% | | 50% | |
| n | Covariates | 0.90 CL* | 0.95 CL | 0.90 CL | 0.95 CL | 0.90 CL | 0.95 CL |
| 200 | Oedema | 0.90 | 0.94 | 0.90 | 0.94 | 0.90 | 0.95 |
| | Age | 0.89 | 0.95 | 0.89 | 0.94 | 0.90 | 0.95 |
| | Log(albumin) | 0.89 | 0.93 | 0.89 | 0.93 | 0.92 | 0.96 |
| | Log(bilirubin) | 0.90 | 0.96 | 0.90 | 0.96 | 0.88 | 0.95 |
| | Log(protime) | 0.89 | 0.94 | 0.89 | 0.93 | 0.88 | 0.93 |
| 400 | Oedema | 0.90 | 0.94 | 0.90 | 0.95 | 0.89 | 0.94 |
| | Age | 0.91 | 0.96 | 0.90 | 0.96 | 0.90 | 0.95 |
| | Log(albumin) | 0.90 | 0.95 | 0.92 | 0.97 | 0.92 | 0.95 |
| | Log(bilirubin) | 0.91 | 0.96 | 0.92 | 0.96 | 0.93 | 0.96 |
| | Log(protime) | 0.90 | 0.95 | 0.90 | 0.94 | 0.88 | 0.95 |

*CL: Nominal confidence level

Figure 1: The Q-Q plots based on 500 simulated surgical unit data sets

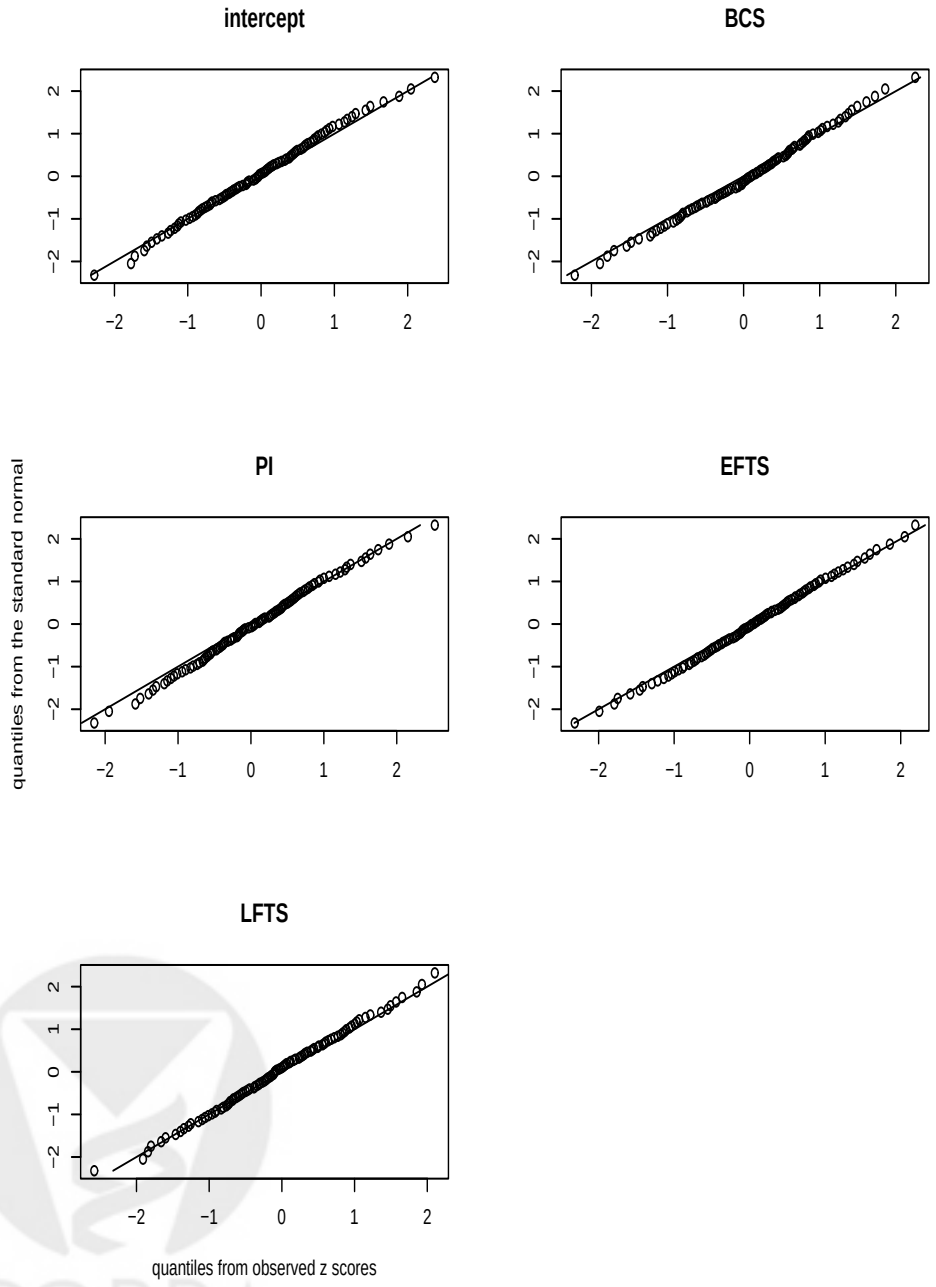

Figure 2: The norms of the estimating functions evaluated at new and Parzen's estimates based on 500 simulated surgical unit data sets

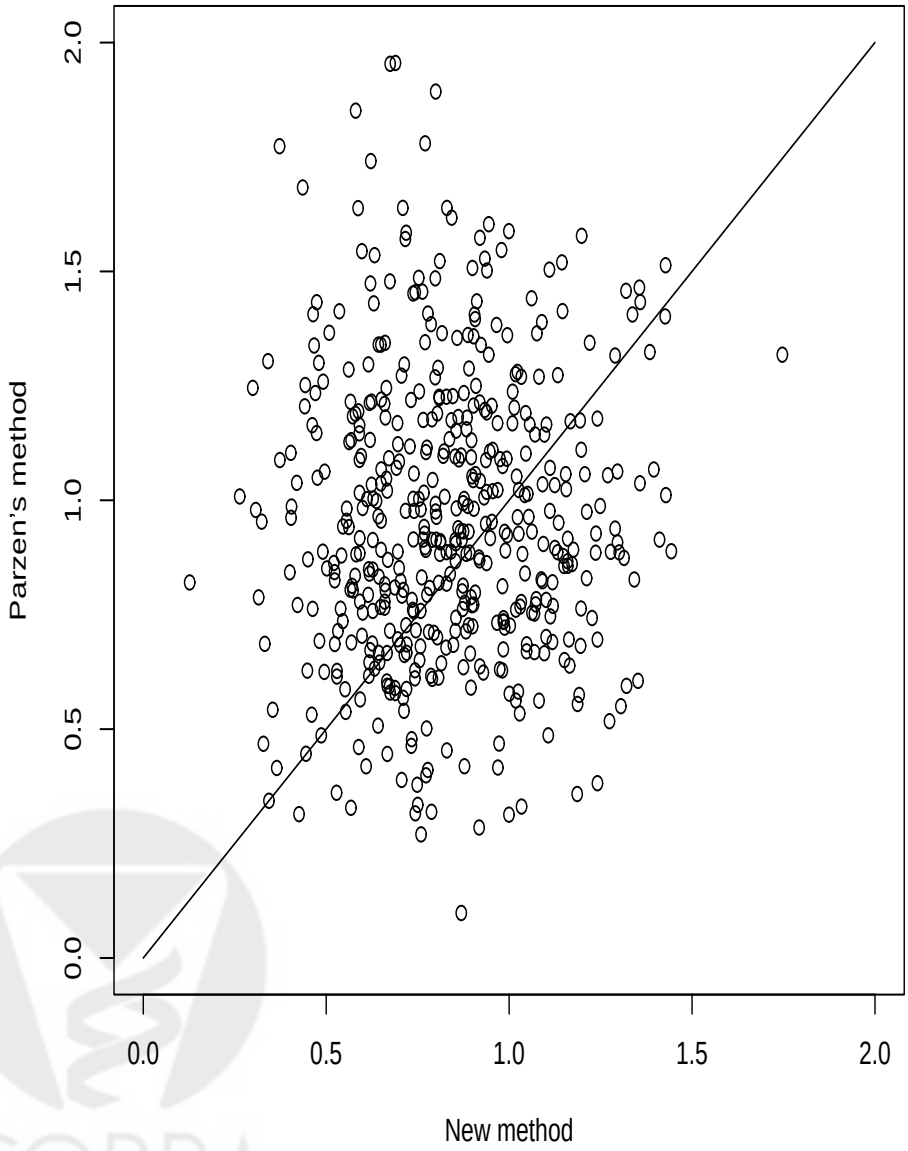

Figure 3: The Q-Q plots based on simulated Mayo primary biliary cirrhosis data