

Statistical inference based on non-smooth estimating functions

BY L. TIAN

*Department of Preventive Medicine, Northwestern University, 680 N. Lake Shore Drive,
Suite 1102, Chicago, Illinois 60611, U.S.A.*

lutian@northwestern.edu

J. LIU

*Department of Statistics, Harvard University, Science Center 6th floor, 1 Oxford Street,
Cambridge, Massachusetts 02138, U.S.A.*

jliu@stat.harvard.edu

Y. ZHAO AND L. J. WEI

*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A.*

mzhao@hsph.harvard.edu wei@sdac.harvard.edu

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 estimator by solving the corresponding estimating equation using standard numerical techniques. In this paper, we propose a simple inference procedure via the importance sampling technique, which provides a consistent root of the estimating equation and also an approximation to its distribution without solving any equations or involving nonparametric function estimates. The new proposal is illustrated and evaluated via two extensive examples with real and simulated datasets.

Some key words: Importance sampling; L_1 -norm; Linear regression for censored data; Resampling.

1. INTRODUCTION

Suppose that inferences about a vector θ_0 of p unknown parameters are to be based on a non-smooth estimating function $S_X(\theta)$, where X is the observable random quantity. Often it is difficult to solve the corresponding estimating equation $S_X(\theta) \simeq 0$ numerically, especially 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 θ_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, it is difficult to implement all existing inference procedures for θ_0 , including resampling methods, without additional information about θ_0 .

Suppose that there is a consistent, but not necessarily efficient, estimator $\hat{\theta}_X^\dagger$ readily available for θ_0 from a relatively simple estimating function. It is usually not difficult to obtain such an estimator. 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 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 uses $\hat{\theta}_X^\dagger$ to locate a consistent estimator $\hat{\theta}_X$ such that $S_X(\hat{\theta}_X) \doteq 0$, and draws inferences about θ_0 . Our procedure does not involve the need to solve any complicated equations. Moreover, it does not involve nonparametric function estimates or numerical derivatives (van der Vaart, 1998, § 5.7).

We illustrate the new proposal with two 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 procedure is evaluated via a real dataset 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, § 7). Our procedure is much simpler and also more general than that recently proposed by Jin et al. (2003) for analysing 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 THE 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 $n \rightarrow \infty$, the random vector $S_X(\theta_0)$ converges weakly to a multivariate normal distribution $N(0, I_p)$, where I_p is the $p \times p$ identity matrix. Furthermore, for large n , assume that, as a function of θ , $S_X(\theta)$ is approximately linear in a small neighbourhood of θ_0 . The formal definition of the local linearity property of $S_X(\theta)$ is given in (A.1) of the Appendix. It follows that, for a consistent estimator $\hat{\theta}_X$ such that $S_X(\hat{\theta}_X) \doteq 0$, the random vector $n^{\frac{1}{2}}(\hat{\theta}_X - \theta_0)$ is asymptotically normal. When the above limiting covariance matrix involves a completely unknown density-like function and direct estimation is difficult, various resampling methods may be used to make inference about θ_0 (Efron & Tibshirani, 1993; Hu & Kalbfleisch, 2000). Recently, Parzen et al. (1994) and Goldwasser et al. (2004) studied a resampling procedure which takes advantage of the pivotal feature of $S_X(\theta_0)$. To be specific, let x be the observed value of X and let the random vector θ_x^* be a solution to the stochastic equation $S_x(\theta_x^*) \doteq G$, where G is $N(0, I_p)$. If θ_x^* is consistent for θ_0 , then the distribution of $n^{\frac{1}{2}}(\hat{\theta}_X - \theta_0)$ can be approximated well by the conditional distribution of $n^{\frac{1}{2}}(\theta_x^* - \hat{\theta}_X)$. In practice, one can generate a large random sample $\{g_m, m = 1, \dots, N\}$ from G and obtain a large number of independent realisations of θ_x^* by solving the equations $S_x(\theta) \doteq g_m$, for $m = 1, \dots, N$. The sample covariance matrix based on those N realisations of θ_x^* can then be used to estimate the covariance matrix of $\hat{\theta}_X$.

When it is difficult to solve $S_x(\theta) \doteq g$ numerically, none of the resampling methods in the literature works 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 we generate M vectors $\theta_x^{(m)}$ ($m = 1, \dots, M$) in a small neighbourhood of θ_0 , where

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

$n^{\frac{1}{2}} \Sigma_x$ converges to a $p \times p$ deterministic matrix as $n \rightarrow \infty$, $\tilde{g}_m = g_m$, if $\|g_m\| \leq c_n$, and is 0, otherwise, $c_n \rightarrow \infty$, and $c_n = o(n^{\frac{1}{2}})$. Note that \tilde{g}_m in (2.1) is a slightly truncated g_m , which is a realisation from G . By the local linearity property of $S_x(\theta)$ around θ_0 , $\{S_x(\theta_x^{(m)}), m = 1, \dots, M\}$ is a set of independent realisations from a distribution which can be approximated by a multivariate normal with mean $\mu_x = S_x(\hat{\theta}_x^\dagger)$ and covariance matrix Λ_x , the sample covariance matrix constructed from M observations

$$\{S_x(\theta_x^{(m)}), m = 1, \dots, M\}.$$

Let θ_x be the random vector which is uniformly distributed on the discrete set $\{\theta_x^{(m)}, m = 1, \dots, M\}$. Then, approximately, $S_x(\theta_x) \sim N(\mu_x, \Lambda_x)$. For the resampling method of Parzen et al. (1994), one needs to construct a random vector θ_x^* such that the distribution of $S_x(\theta_x^*)$ is approximately $N(0, I_p)$. This can be done using the importance sampling idea discussed in Liu (2001, § 2) and Rubin (1987) in the context of Bayesian analysis and multiple imputation. For this, let θ_x^* be a random vector defined on the same support as θ_x , but let its mass at $t = \theta_x^{(m)}$ be proportional to

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

where $\phi(\cdot)$ is the density function of $N(0, I_p)$. Note that the numerator of (2.2) is the density function of the target distribution $N(0, I_p)$, and the denominator is the normal approximation to the density function of $S_x(\theta_x)$. In the Appendix, we show that $S_x(\theta_x^*) \sim N(0, I_p)$, approximately, for large M and n , and, with g_m in (2.1) being truncated by c_n , θ_x^* is consistent. Moreover, if we let $\hat{\theta}_x$ be the mean of θ_x^* , then $S_x(\hat{\theta}_x) \doteq 0$, and the unconditional distribution of $n^{\frac{1}{2}}(\hat{\theta}_x - \theta_0)$ can be approximated well by the conditional distribution of $n^{\frac{1}{2}}(\theta_x^* - \hat{\theta}_x)$.

The choice of Σ_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 & Berger (1992) in a different context. That is, one may start with an initial matrix Σ_x , such as $n^{-\frac{1}{2}} I_p$, to generate $\{\theta_x^{(l)}, l = 1, \dots, L\}$ via (2.1) for obtaining an intermediate θ_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 $N(0, I_p)$, we generate additional $\{\theta_x^{(m)}, m = 1, \dots, (M - L)\}$ under the same setting to obtain an accurate normal approximation to the distribution of $S_x(\theta_x)$ and a final θ_x^* . Otherwise, we generate a fresh set of $\{\theta_x^{(l)}, l = 1, \dots, L\}$ via (2.1) with an updated Σ_x , which, for example, is the covariance matrix of θ_x^* from the previous iteration, construct a new intermediate θ_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 $N(0, I_p)$ can be evaluated numerically or graphically. For each iteration, the standard coefficient of variation of the unnormalised weight (2.2) can also be used to monitor the adaptive procedure (Liu, 2001, § 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 Σ_x in (2.1). In § 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 § 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, $i = 1, \dots, n$. Here, $X = \{(T_i, z_i), i = 1, \dots, n\}$. Assume that

$$T_i = \theta'_0 z_i + \varepsilon_i, \quad (3.1)$$

where ε_i ($i = 1, \dots, n$) are mutually independent and have mean 0, but the distribution of ε_i may depend on z_i . Under this setting, the least squares estimator $\hat{\theta}_X^+$ is consistent for θ_0 .

If the distribution of ε is symmetric about 0, an alternative way of estimating θ_0 is to use a minimiser $\hat{\theta}_X$ of the L_1 norm $\sum_{i=1}^n |T_i - \theta'_0 z_i|$. This estimator is asymptotically equivalent to a solution to the estimating equation $S_X(\theta) = 0$, where

$$S_X(\theta) = \Gamma^{-1} \sum_{i=1}^n z_i \{I(T_i - \theta'_0 z_i \leq 0) - \frac{1}{2}\}, \quad (3.2)$$

$I(\cdot)$ is the indicator function and $\Gamma = (\sum_{i=1}^n z_i z_i')^{\frac{1}{2}}/2$. It is easy to show that $S_X(\theta_0)$ is asymptotically $N(0, I_p)$. 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 θ_0 , and proposed a novel way of solving $S_X(\theta) = g$, for any given vector g , to generate realisations of θ_x^* . Our proposal is readily applicable to the present case and does not involve the need to solve the above equation repeatedly.

We use a small dataset on survival times in patients with a specific liver surgery (Neter et al., 1985, p. 419) to illustrate our proposal and we compare the results with those given by Parzen et al. (1994). This dataset has 54 files, and each file consists of the uncensored survival time of a patient with four covariates, namely blood clotting score, prognostic index, enzyme function test score and liver function test score. We let T be the base 10 logarithm of the survival time and let z be a 5×1 covariate vector with the first component being the intercept. We used the iterative procedure described at the end of § 2 with $L = 1000$ for each iteration, and $M = 3000$.

First, we let the initial matrix Σ_x in (2.1) be $n^{-\frac{1}{2}} I_5$. However, after 20 iterations, we found that the covariance matrix of $S_x(\theta_x^*)$ was markedly different from the matrix I_5 . 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 quite a

poor approximation to the target distribution $N(0, I_5)$. To search for a better choice of Σ_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 θ_0 is proportional to $(\sum_{i=1}^n z_i z_i')^{\frac{1}{2}}$. This suggests that, if one let

$$\Sigma_x = n^{\frac{1}{2}} \left(\sum_{i=1}^n z_i z_i' \right)^{-1} \tag{3.3}$$

in (2.1), the covariance matrix of the resulting $S_x(\theta_x)$ would be approximately diagonal and the corresponding distribution of $S_x(\theta_x)$ would be likely to be a better approximation to $N(0, I_5)$. With this initial Σ_x and $\hat{\theta}_x^\dagger$ being the least squares estimate for θ_0 , after three iterations, the maximum of the absolute values of the componentwise differences between the covariance matrix of $S_x(\theta_x^*)$ and I_5 was about 0.05. Based on 2000 additional $\theta_x^{(m)}$ generated from (2.1) under the setting at the beginning of the third 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). We also repeated the final stage of our iterative procedure with $M = 10\,000$, with results that are practically identical to those reported in Table 1. Moreover, 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 of Parzen et al. the corresponding value is 2.75. Note also 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.

Table 1: *Surgical unit data. L_1 estimates for heteroscedastic linear regression*

Parameter	New method		Parzen's method	
	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; Est, estimate; SE, standard error; Parzen's method, method of Parzen et al. (1994).

To examine further the performance of the new proposal for cases with small sample sizes, we fitted the above data with (3.1) assuming a zero-mean normal error based on the maximum likelihood estimation procedure. Here the logarithm of the variance for the error term is an unknown linear function of two covariates, blood clotting score and liver function. This results in a heteroscedastic linear model with four covariates in the deterministic component and the variance of the error term being a function of two covariates. With the set of the observed covariate vectors $\{z_i, i = 1, \dots, 54\}$ from the liver surgery example, we simulated 500 samples $\{T_i, i = 1, \dots, 54\}$ from this model. For each simulated

sample, we used the above iterative procedure to obtain $\hat{\theta}_x$ and its estimated covariance matrix. In Fig. 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 standardised estimates, each of which was centred by the corresponding component of θ_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 datasets, we computed the value $\|S_x(\hat{\theta}_x)\|$ and its counterpart from Parzen et al. In Fig. 2, we present the scatterplot 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.

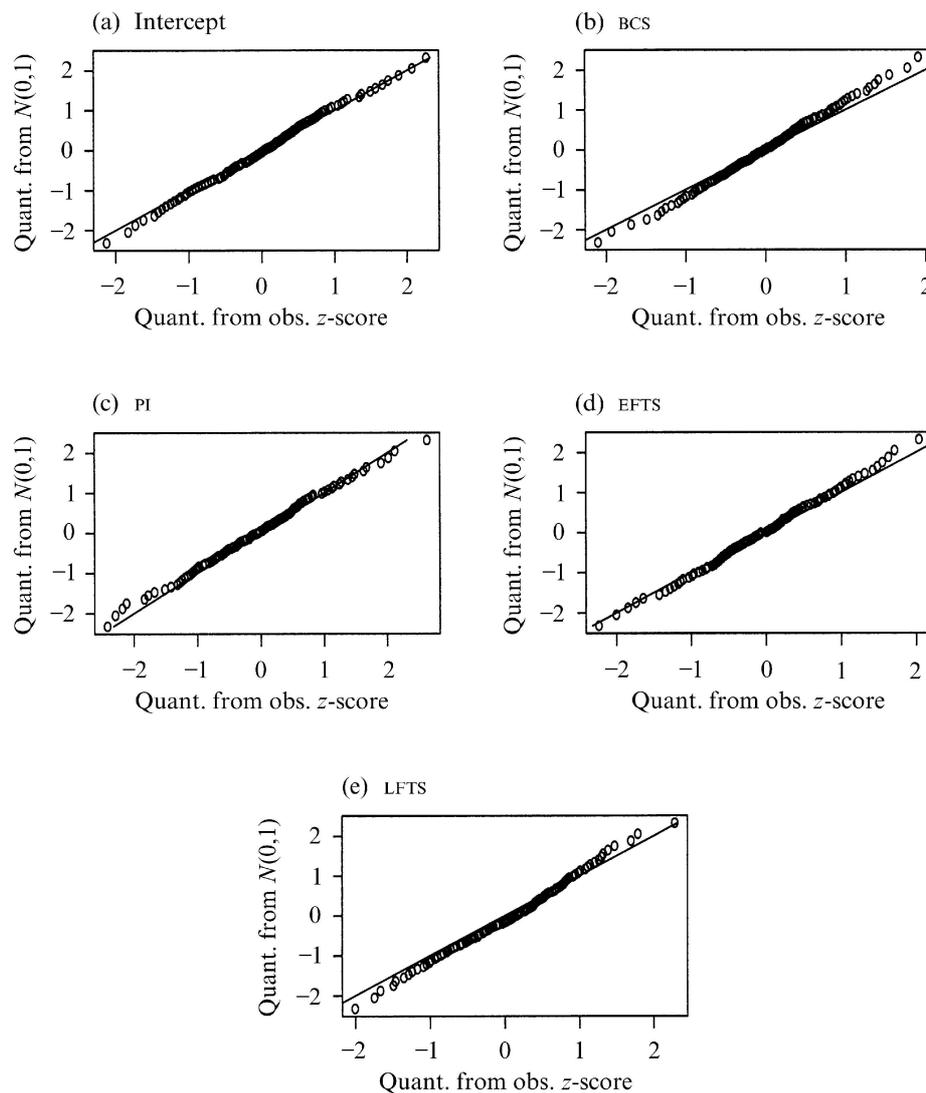


Fig. 1: The Q-Q plots, quantiles from the standard normal plotted against quantiles from the observed z-scores, based on 500 simulated surgical unit datasets, for estimates of parameters associated with (a) intercept, (b) blood clotting score, (c) prognostic index, (d) enzyme function test score and (e) liver function test score.

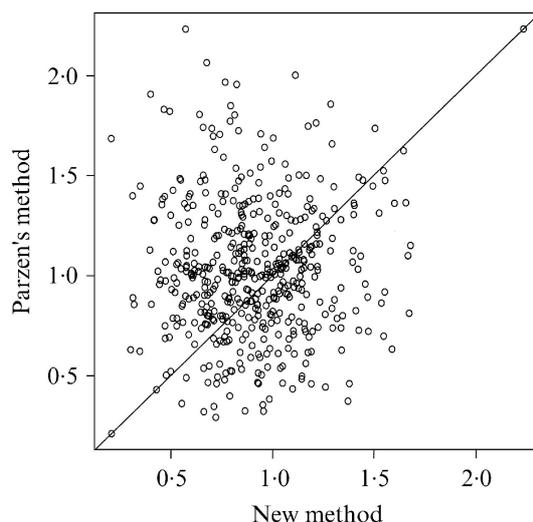


Fig. 2. The norms of the estimating functions evaluated at new estimates and estimates of Parzen et al. (1994) based on 500 simulated surgical unit datasets.

3.2. Inferences for linear model with censored data

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 θ_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 are $X = \{(Y_i, \Delta_i, z_i), i = 1, \dots, n\}$, where $\Delta = I(T \leq C)$ and $Y = \min(T, C)$. In survival analysis, this log-linear model is called the 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 θ_0 . 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 function for θ_0 is

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

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

$$\Gamma(\theta) = n^{-1} \sum_{i=1}^n \int_{-\infty}^{\infty} w^2(\theta; t) \{z_i - \bar{z}(\theta; t)\} \{z_i - \bar{z}(\theta; t)\}' dN_i(\theta; t),$$

and $\tilde{S}_X(\theta)$ is approximately linear in a small neighbourhood of θ_0 . Note that, under our setting, the estimating function is

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

It follows that $S_X(\theta_0)$ is asymptotically $N(0, I_p)$.

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 $\hat{\theta}_X$ of the equation $\tilde{S}_X(\theta) = 0$.

With our new proposal, one can obtain a consistent estimator $\hat{\theta}_X$ for θ_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 $\{(e_i(\theta), \Delta_i), i = 1, \dots, n\}$ (Fleming & Harrington, 1991, p. 274; Kosorok & Lin, 1999). Note that one may simplify the estimating function $S_X(\theta)$ by replacing θ 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 dataset consists of 416 complete files, and each of them contains information about 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, namely oedema, age, log (albumin), log (bilirubin) and log (protime). The initial consistent estimate $\hat{\theta}_X^\dagger = (-0.878, -0.026, 1.59, -0.579, -2.768)'$, used in (2.1), was the one based on the Gehan weight function. First, we considered $S_X(\theta)$ with the log-rank weight function $w(\theta; t) = 1$. To generate $\{\theta_X^{(m)}\}$ via (2.1), we let $\Sigma_x = n^{-\frac{1}{2}}I_5$. Under the set-up of the iterative process discussed in § 3.1, the adaptive procedure was terminated at the second stage. The coefficient of variation of the weights (2.2) at this stage was 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 of 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 the non-monotone weight function w in (3.6) with $a = b = \frac{1}{2}$.

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

Parameter	Weight function					
	Log-rank		Jin's method		$\hat{F}_\theta(t)^{0.5} \{1 - \hat{F}_\theta(t)\}^{0.5}$	
	New method Est	SE	Est	SE	New method 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

Est, estimate; SE, standard error; Jin's method, method of Jin et al. (2003).

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

$$T = 13.73 - 0.898 \times \text{oedema} - 0.026 \times \text{age} + 1.533 \times \log(\text{albumin}) - 0.593 \times \log(\text{bilirubin}) - 2.428 \times \log(\text{protime}) + \varepsilon, \tag{3.7}$$

where ε is a normal random variable with mean 0 and variance 0.947. The regression coefficients and the variance of ε 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 $(0, \zeta)$, where ζ 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 dataset. With these fixed covariate vectors, we used (3.7) to simulate 500 sets of $\{T_i, i = 1, \dots, n\}$ 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 log-rank weight. For the case that $n = 200$ and the censoring rate is about 50%, the Q-Q plots with respect to five regression coefficients, not shown, indicate that 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 3, 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 log-rank weight function. The new procedure performs well for all the cases studied here.

Table 3. Empirical coverage probabilities of confidence intervals from the simulation study for the accelerated failure time model

n	Covariates	Censoring					
		0%		25%		50%	
		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.

4. REMARKS

In practice, the initial choice of Σ_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 in § 3.2, we find that, usually, our proposal with $n^{\frac{1}{2}}\Sigma_x$ 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 Σ_x may not work well.

To implement our procedure, at each step of the iterative stage, we closely monitor numerically and/or graphically whether or not the distribution of the current $S_x(\theta_x^*)$ is closer than that from the previous step to the target $N(0, I_p)$. We recommend terminating the iteration when the maximum of the absolute values of the componentwise differences between the covariance matrix of $S_x(\theta_x^*)$ and I_p is less than 0.1 and the coefficient of variation of the weights is less than 0.5. On the other hand, if the distributions of $S_x(\theta_x^*)$

seem to be drifting away from the target distribution during, say, twenty iterations, one may repeat the entire iterative process from the beginning with the original or a new Σ_x . For this exploratory stage, we find that the choice of L , the number of θ_x^* generated at each step, can be quite flexible. For the two examples presented in this paper, we have repeated the same type of analysis with different values of L ranging from 500 to 1000, and all the results are practically identical. For the final stage of our process, one may use the following heuristic argument to choose M , the number of θ_x^* , to obtain a reasonably good approximation to the distribution of $(\hat{\theta}_x - \theta_0)$. First, suppose that θ_0 is a scalar parameter and assume that we can draw a random sample with size B from a normal with mean $\hat{\theta}_x$ and variance s_x^2 , a consistent estimate for the variance of $\hat{\theta}_x$. Let the resulting sample mean and variance be denoted by $\tilde{\theta}$ and \tilde{s}^2 . Then, if $B \geq \{\Phi^{-1}(1 - \alpha/4)/d\}^2$, the probability of the joint event that $|\tilde{\theta} - \hat{\theta}_x|/s_x < d$ and $|\tilde{s} - s_x|/s_x < d$ is greater than $1 - \alpha$, where $\Phi(\cdot)$ is the distribution function of the standard normal, d is a small number and $0 < \alpha < 1$; that is, with the sample size B , we expect that the sample mean and variance are very close to $\hat{\theta}_x$ and s_x^2 . For the case of a p -dimensional θ_0 , one may choose $B = \lceil \Phi^{-1}\{1 - \alpha/(4p)\}/d \rceil^2$ using the Bonferroni adjustment argument. At the end of our iterative stage, however, we do not know the exact values of $\hat{\theta}_x$ and s_x^2 and cannot take a random sample from $N(\hat{\theta}_x, s_x^2)$ directly. One may follow a general 'rule of thumb' for importance sampling by inflating B by a factor of $\{1 + cv\}$, where cv is the coefficient of variation at the last step of the iteration, to set a possible value for M (Liu, 2001, § 2.5.3); that is, $M = (1 + cv)B$. For the two examples discussed in § 3, $p = 5$. If we let $\alpha = 0.2$ and $d = 0.05$, with the coefficient of variation being 0.5 when we terminate the iterative stage, M is roughly 3000. Before stating the final conclusions about θ_0 , we recommend generating M additional θ_x^* and using a total of $2M$ such realisations to construct the desired interval estimates for the parameters of interest. If the resulting intervals are practically identical to those obtained from the initial M realisations, further sampling of θ_x^* may not be needed.

ACKNOWLEDGEMENT

The authors are grateful to a referee, an associate editor and the editor for helpful comments on the paper. This work is partially supported by grants from the U.S. National Institute of Health.

APPENDIX

Large-sample theory

Assume that, for any sequence of constants, $\{\varepsilon_n\} \rightarrow 0$, there exists a deterministic matrix D such that, for $l = 1, 2$,

$$\sup_{\|\theta_l - \theta_0\| \leq \varepsilon_n} \frac{\|S_X(\theta_2) - S_X(\theta_1) - Dn^{1/2}(\theta_2 - \theta_1)\|}{1 + n^{1/2}\|\theta_2 - \theta_1\|} = o_{P_X}(1), \quad (\text{A}\cdot 1)$$

where P_X is the probability measure generated by X . For the observed x of X , let $\tilde{\theta}_x = \hat{\theta}_x + \Sigma_x GI(\|G\| \leq c_n)$, where $G \sim N(0, I_p)$. Note that, as $M \rightarrow \infty$, the distribution of θ_x , which is the uniform random vector on the discrete set $\{\theta_x^{(m)}, m = 1, \dots, M\}$ discussed in § 2, is the same as that of $\tilde{\theta}_x$. Let P_G be the probability measure generated by G and let P be the product measure $P_X \times P_G$. Then, since $\tilde{\theta}_x$ and $\hat{\theta}_x^*$ are in a $o_P(1)$ -neighbourhood of θ_0 , it follows from (A.1) that

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

This implies that

$$\left| E[h\{S_X(\tilde{\theta}_X)\} | X] - \int_{R^p} \frac{h(t)}{\sqrt{|\tilde{\Lambda}_X|}} \phi[\tilde{\Lambda}_X^{-1/2}\{t - S_X(\hat{\theta}_X^*)\}] dt \right| = o_{P_X}(1), \quad (\text{A} \cdot 2)$$

where $\tilde{\Lambda}_X$ is the limit of Λ_X , as $M \rightarrow \infty$, $h(\cdot)$ is any uniformly bounded, Lipschitz function, and the expectation E in (A·2) is taken under P_G . Note that, loosely speaking, (A·2) indicates that, for $X = x$, the distribution of $S_X(\tilde{\theta}_X)$ is approximately $N(\mu_x, \tilde{\Lambda}_X)$. Now, for given x , as $M \rightarrow \infty$, the distribution function of θ_x^* at t converges to

$$c_x E \left(I(\tilde{\theta}_X \leq t) \frac{\phi\{S_X(\tilde{\theta}_X)\}}{\phi[\tilde{\Lambda}_X^{-1/2}\{S_X(\tilde{\theta}_X) - S_X(\hat{\theta}_X^*)\}]} \right),$$

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

$$E[h\{S(\theta_x^*)\} | X] \simeq c_x E \left(h\{S_X(\tilde{\theta}_X)\} \frac{\phi\{S_X(\tilde{\theta}_X)\}}{\phi[\tilde{\Lambda}_X^{-1/2}\{S_X(\tilde{\theta}_X) - S_X(\hat{\theta}_X^*)\}]} \Big| X \right), \quad (\text{A} \cdot 3)$$

where h is any uniformly bounded, Lipschitz continuous function. With (A·2), it is straightforward to show that the absolute value of the difference between the right-hand side of (A·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 $N(0, I_p)$ in a certain probability sense; that is, for any p -dimensional vector t ,

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

where $\Phi(\cdot)$ is the distribution function of $N(0, I_p)$.

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

REFERENCES

- BANG, H. & TSIATIS, A. A. (2002). Median regression with censored cost data. *Biometrics* **58**, 643–9.
- BARRODALE, I. & ROBERTS, F. D. K. (1977). Algorithms for restricted least absolute value estimation. *Commun. Statist. B* **6**, 353–64.
- BUCKLEY, J. & JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–36.
- EFRON, B. & TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
- FLEMING, T. R. & HARRINGTON, D. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons.
- GOLDWASSER, M. I., TIAN, L. & WEI, L. J. (2004). Statistical inference for infinite-dimensional parameters via asymptotically pivotal estimating functions. *Biometrika* **91**, 81–94.
- HU, F. & KALBFLEISCH, J. D. (2000). The estimating function bootstrap. *Can. J. Statist.* **28**, 449–81.
- JIN, Z., LIN, D. Y., WEI, L. J. & YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–53.
- KALBFLEISCH, J. D. & PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.
- KOENKER, R. & BASSETT, G. J. (1978). Regression quantiles. *Econometrica* **46**, 33–50.
- KOENKER, R. & D'OREY, V. (1987). Computing regression quantiles. *Appl. Statist.* **36**, 383–93.
- KOSOROK, M. R. & LIN, C. (1999). The versatility of function-indexed weighted log-rank statistics. *J. Am. Statist. Assoc.* **94**, 320–32.
- LIU, J. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
- NETER, J., WASSERMAN, W. & KUTNER, M. H. (1985). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL.: Richard D. Irwin, Inc.
- OH, M. S. & BERGER, J. O. (1992). Adaptive importance sampling in Monte Carlo integration. *J. Am. Statist. Assoc.* **41**, 143–68.
- PARZEN, M. I., WEI, L. J. & YING, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–50.
- PRENTICE, R. L. (1978). Linear rank tests with censored data. *Biometrika* **65**, 167–80.
- RITOV, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18**, 303–28.
- ROBINS, J. M. & ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *Aids Epidemiology*, Ed. N. P. Jewell, K. Dietz and V. T. Farewell, pp. 297–331. Boston: Birkhauser.

- ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
- 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**, 543–6.
- TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354–72.
- VAN DER VAART, A. W. (1988). *Asymptotic Statistics*. New York: Cambridge University Press.
- WEI, L. J., YING, Z. & LIN, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–51.
- YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76–99.

[Received September 2003. Revised April 2004]