Statistica Sinica **10**(2000), 731-749

EFFICIENT SCORE ESTIMATION AND ADAPTIVE M-ESTIMATORS IN CENSORED AND TRUNCATED REGRESSION MODELS

Chul-Ki Kim and Tze Leung Lai

Ewha Womans University and Stanford University

Abstract: An adaptive M-estimator of a regression parameter based on censored and truncated data is developed by using B-splines to estimate the efficient score function and a relatively simple cross validation method to determine the number of knots. An iterative algorithm to compute the estimator is also provided. The adaptive estimator is asymptotically efficient, and simulation studies of the finite-sample performance of the adaptive estimator shows that it is superior to other M-estimators for regression analysis of censored and truncated data in the literature. An asymptotic theory of cross validation in the presence of censoring and truncation is also developed in this connection.

Key words and phrases: Adaptation, B-splines, cross validation, censoring, truncation, efficient estimation.

1. Introduction

Consider the linear regression model

$$y_j = \beta^T x_j + \epsilon_j \quad (j = 1, 2, \ldots), \tag{1.1}$$

where the ϵ_j are i.i.d. random variables (representing unobservable disturbances) with a common distribution function F, β is a $d \times 1$ vector of unknown parameters, and the x_j are either nonrandom or independent $d \times 1$ random vectors independent of $\{\epsilon_n\}$. Suppose that the responses y_j in (1.1) are not completely observable due to left truncation and right censoring by random variables t_j and c_j such that $\infty > t_j \geq -\infty$ and $-\infty < c_j \leq \infty$. It will be assumed that (t_j, c_j) are i.i.d. and independent of (x_j, ϵ_j). Let $\tilde{y}_j = y_j \wedge c_j$ and $\delta_j = I_{\{y_j \leq c_j\}}$, where we use $\wedge$ and $\vee$ to denote minimum and maximum, respectively. In addition to right censorship of the responses y_j by c_j, we shall also assume left truncation in the sense that $(\tilde{y}_j, \delta_j, x_j)$ can be observed only when $\tilde{y}_j \geq t_j$. The data, therefore, consist of n observations $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$ with $\tilde{y}_i^o \geq t_i^o$, $i = 1, \ldots, n$. The special case $t_i \equiv -\infty$ corresponds to the "censored regression model" which is of basic importance in statistical modelling and analysis of failure time data (cf.

Kalbfleisch and Prentice (1980), Lawless (1982)). The special case $c_i \equiv \infty$ corresponds to the "truncated regression model" in econometrics (cf. Tobin (1958), Goldberger (1981), Amemiya (1985), Moon (1989)) and in astronomy (cf. Segal (1975), Nicoll and Segal (1980)), which assumes the presence of truncation variables τ_j so that (x_j, y_j) can be observed only when $y_j \le \tau_j$ (or equivalently, when $-y_j \ge -\tau_j = t_j$). Left truncated responses that are also right censored arise in prospective studies of a disease and other biomedical studies (cf. Andersen, Borgan, Gill and Keiding (1993), Keiding, Holst and Green (1989), Gross and Lai (1996)).

Lai and Ying (1991b, 1992) studied efficient estimation of β from the data $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$ by developing asymptotic minimax bounds for the semiparametric estimation problem and constructing rank estimators that attain these bounds. Assuming that F has a continuously differentiable density function f so that the hazard function $\lambda = f/(1-F)$ is also continuously differentiable, their construction of these rank estimators consists of (i) dividing the sample into two disjoint subsets and evaluating a preliminary consistent estimate $\hat{b}_j$ of β from the jth subsample ($j = 1, 2$), (ii) finding from the uncensored residuals in the jth subsample a smooth consistent estimate $\hat{\lambda}_j$ of the hazard function λ, (iii) smoothing $\hat{\lambda}_j'/\hat{\lambda}_j$ to obtain a smooth consistent estimate $\hat{\psi}_j$ of λ'/λ, and (iv) using $\hat{\psi}_1$ (respectively $\hat{\psi}_2$) as the weight function for the linear rank statistic of the second (respectively first) sample of residuals $\tilde{y}_i^o - b^T x_i^o$. The sum $S(b)$ of these two linear rank statistics is used to define the rank estimator as the minimizer of $\|S(b)\|$. There are, however, practical difficulties in carrying out this procedure.

First, rank estimators are difficult to compute when β is multidimensional. As noted by Lin and Geyer (1992), rank estimators of multidimensional β "require minimizing discrete objective functions with multiple local minima" and "conventional optimization algorithms cannot be used to solve such optimization problems". Computationally intensive search algorithms, such as the simulated annealing algorithm used by Lin and Geyer (1992), are needed to minimize $\|S(b)\|$. Another difficulty lies in estimation of λ'/λ to form the $\hat{\psi}_j$. Although there is an extensive literature on estimation of the hazard function λ and its derivative λ' for censored and truncated data, the problem of estimating λ'/λ from left truncated and right censored (l.t.r.c.) data is relatively unexplored. As will be shown in Section 2, simply plugging in $\hat{\lambda}_j'/\hat{\lambda}_j$ and smoothing the plugged-in estimate do not give good results unless the sample size is very large.

The present paper addresses these issues in constructing asymptotically efficient estimates of β from l.t.r.c. data. Instead of using rank estimators, we use M-estimators which have much lower computational complexity (cf. Kim and Lai (1999)). These M-estimators are defined for l.t.r.c. data by the estimating

equation

$$\sum_{i=1}^{n} x_i^o \{\delta_i^o \psi(\tilde{y}_i^o(b)) + (1-\delta_i^o) \int_{u>\tilde{y}_i^o(b)} \psi(u) d\hat{F}_b(u|\tilde{y}_i^o(b)) - \int_{u\geq t_i^o(b)} \psi(u) d\hat{F}_b(u|t_i^o(b)-)\} = 0, \tag{1.2}$$

where $\tilde{y}_i^o(b) = \tilde{y}_i^o - b^T x_i^o$, $t_i^o(b) = t_i^o - b^T x_i^o$, ψ is the score function associated with the M-estimator, and

$$\hat{F}_b(u|v) = 1 - \prod_{i:v<\tilde{y}_i^o(b)\leq u, \delta_i^o=1} \{1 - \Delta(b, \tilde{y}_i^o(b))/N(b, \tilde{y}_i^o(b))\}, \tag{1.3}$$

$$N(b,u) = \sum_{i=1}^{n} I(t_i^o(b) \leq u \leq \tilde{y}_i^o(b)), \quad \Delta(b,u) = \sum_{i=1}^{n} I(\tilde{y}_i^o(b) = u, \delta_i^o = 1), \tag{1.4}$$

cf. Lai and Ying (1994). The notation $\hat{F}_b(u|v-)$ in (1.2) is used to denote (1.3) in which "$v < \tilde{y}_i^o(b)$" is replaced by "$v \leq \tilde{y}_i^o(b)$". The function $\hat{F}_b(u|-\infty)$ is the product-limit estimate of the common distribution function $F(u)$ of the ϵ_j in (1.1). Note that $\hat{F}_b(u|v)$ is the product limit estimate of

$$F(u|v) = P\{\epsilon_j \leq u | \epsilon_j > v\}. \tag{1.5}$$

Lai and Ying (1994) have shown that an asymptotically optimal choice of ψ in (1.2) is

$$\psi^* = (\lambda'/\lambda) - \lambda, \tag{1.6}$$

for which the M-estimator of β is asymptotically normal with covariance matrix equal to that given by the information bound of the semiparametric estimation problem. Indeed this M-estimator of β has the same asymptotic properties as the asymptotically efficient rank estimator. Since the M-estimator has much lower computational complexity than the rank estimator, it will be used for adaptive estimation in which the asymptotically efficient score function (1.6) is not assumed to be known *a priori* but has to be estimated from the data.

Section 2 discusses how (1.6) can be estimated. We use a spline approximation to ψ and a cross validation method to choose the number of knots. This is shown to perform much better than the plug-in method in Lai and Ying (1991b) based on estimating λ and λ'. Simulation results reported in Section 3 show that the adaptive M-estimator of β, which uses this new approach to estimate the optimal score function, outperforms other M-estimators based on l.t.r.c. data in the literature. Because of the simplicity of the cross validation method that involves only cross validating the two subsamples with each other, using this

adaptive determination of the score function does not incur much increase in computational cost.

For complete data, Bickel (1982) showed how an adaptive estimate of β can be constructed so that it is asymptotically as efficient as the maximum likelihood estimate that requires specification of the density function f of the ϵ_j. The basic idea is to replace the unknown score function $(\log f)' = f'/f$ in the maximum likelihood estimate by $\hat{f}'_\eta/\hat{f}_\eta$, where $\hat{f}_\eta$ is a kernel estimate of f involving a bandwidth η that converges to 0 at a sufficiently slow rate as $n \to \infty$. Hsieh and Manski (1987) reported simulation studies showing that the behavior of an adaptive estimate can be changed dramatically in samples of moderate size by using different smoothing parameters η. They also proposed to choose the smoothing parameter that minimizes, over a preselected set of smoothing parameters, a bootstrap estimate of the mean squared error of $\hat{\beta}$. Faraway (1992) used B-splines to estimate $\log f$ so that the smoothing parameter is the number of knots (instead of the bandwidth in the kernel method) and estimated the mean squared error of $\hat{\beta}$ via an asymptotic formula instead of using the bootstrap. Jin (1992) used B-splines to estimate f'/f directly and proposed another cross validation method which we extend to l.t.r.c. data in Section 2, where alternative cross validation methods for l.t.r.c. data are also developed.

2. Estimation of the Score Function (1.6)

In this section we assume known $\beta = 0$, so that the $y_j (= \epsilon_j)$ are i.i.d. with a common distribution function F that has a continuously differentiable density function f and hazard function λ, and consider estimation of the score function (1.6) based on l.t.r.c. data. An obvious approach is to apply directly a method proposed by Uzunoğullari and Wang (1992) for estimating λ and λ' from l.t.r.c. data. The method estimates $\lambda^{(r)}(z)$ (the rth derivative for $r \geq 1$, with $\lambda^{(0)} = \lambda$) by $\hat{\lambda}^{(r)}(z) = \int K_{r,\eta}(z-u)d\hat{\Lambda}(u)$, where $\hat{\Lambda}$ is the estimated cumulative hazard function, $K_{r,\eta}(z) = \eta^{-(r+1)}K_r(z/\eta)$ and K_r is a kernel. The bandwidth $\eta (= \eta_{r,z})$ to estimate $\lambda^{(r)}(z)$ is chosen by a locally adaptive method that attempts to minimize the mean squared error of $\hat{\lambda}^{(r)}(z)$, replacing the unknown parameters in the mean squared error by their estimates. Once $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$ have been obtained by this procedure, one can estimate (1.6) by $\hat{\lambda}^{(1)}(z)/\hat{\lambda}(z) - \hat{\lambda}(z)$. However, this obvious estimate has difficulties when $\hat{\lambda}(z)$ is close to zero, as shown in Section 4.1. We next describe another method to estimate the score function (1.6). First note that if f is the common density function of the $y_j (= \epsilon_j)$ then

$$f'/f = \lambda'/\lambda - \lambda. \tag{2.1}$$

We shall approximate $\psi^* = f'/f$ by B-splines, with knots located at certain quantiles of the product-limit estimate of F and with the number of knots chosen by cross validation for l.t.r.c. data.

2.1. Spline approximations to the score function

In an interval (a,b), take k knots $a < \xi_{k,1} < \cdots < \xi_{k,k} < b$, and define the linear B-spline basis $\{B_{k,i}\colon\ i = 0, \ldots, k+1\}$ as follows:

$$B_{k,i}(y) = \begin{cases} (\xi_{k,1} - y)/(\xi_{k,1} - a) & \text{if } a \le y \le \xi_{k,1} \text{ and } i = 0, \\ (y - \xi_{k,i-1})/(\xi_{k,i} - \xi_{k,i-1}) & \text{if } \xi_{k,i-1} \le y \le \xi_{k,i} \text{ and } 1 \le i \le k, \\ (\xi_{k,i+1} - y)/(\xi_{k,i+1} - \xi_{k,i}) & \text{if } \xi_{k,i} < y \le \xi_{k,i+1} \text{ and } 1 \le i \le k, \\ (y - \xi_{k,k})/(b - \xi_{k,k}) & \text{if } \xi_{k,k} < y \le b \text{ and } i = k+1, \\ 0 & \text{otherwise,} \end{cases}$$

where we set $\xi_{k,0} = a$ and $\xi_{k,k+1} = b$. We can define $B_{k,i}$ on the whole real line by setting $B_{k,i}(y) = 0$ for $y \notin [a,b]$. Let $D_{k,i}(y)$ be the derivative of $B_{k,i}$ at $y \notin \{\xi_{k,i-1}, \xi_{k,i}, \xi_{k,i+1}\}$ (or $y \notin \{a, \xi_{k,1}\}$ if $i = 0$, $y \notin \{\xi_{k,k}, b\}$ if $i = k+1$). Denote $B_k(y) = (B_{k,0}(y), \ldots, B_{k,k+1}(y))^T$, $D_k(y) = (D_{k,0}(y), \ldots, D_{k,k+1}(y))^T$, $A_k(y) = B_k(y)B_k^T(y)$, and

$$A_k(F) = \int_a^b A_k(y)dF(y), B_k(F) = \int_a^b B_k(y)dF(y), D_k(F) = \int_a^b D_k(y)dF(y). \quad (2.2)$$

Given the knots $a < \xi_{k,1} < \cdots < \xi_{k,k} < b$, the best linear spline approximation to ψ^* is defined as $a_k^T(F)B_k(x)$, where $a_k(F)$ minimizes $\int_a^b (a_k^T B_k(y) - \psi^*(y)))^2 dF(y)$ over $a_k \in \mathbf{R}^{k+2}$. Since $\psi^* = f'/f$, integration by parts gives $\int_{-\infty}^{\infty} D_k(y)f(y)dy = \int_{-\infty}^{\infty} f dB_k = -\int_{-\infty}^{\infty} B_k \psi^* dF$, and therefore

$$\begin{aligned} &\int_a^b (a_k^T B_k(y) - \psi^*(y))^2 dF(y) \\ &= a_k^T \left(\int_a^b A_k(y)dF(y) \right) a_k - 2a_k^T \int_a^b B_k(y)\psi^*(y)dF(y) + \int_a^b (\psi^*(y))^2 dF(y) \\ &= a_k^T A_k(F) a_k + 2a_k^T D_k(F) + \int_a^b (\psi^*)^2 dF, \end{aligned} \quad (2.3)$$

noting that B_k vanishes outside $[a,b]$. Since minimizing (2.3) is equivalent to minimizing $a_k^T A_k(F)a_k + 2a_k^T D_k(F)$, it follows that $a_k(F) = -A_k^{-1}(F)D_k(F)$.

In the case $k = 0$, we shall also denote the best linear approximation to ψ^* on $[a,b]$ by $a_0^T(F)B_0(y)$ to unify the notation, where we set $B_0(y) = (1, y-a)^T$.

2.2. Knot placement and Jin's method for choosing the number of knots

Since F in (2.3) is unknown, we replace it by the product-limit estimate $\hat{F}$, which is defined in (1.3)–(1.4) with $\upsilon = -\infty$ and $(t_i^o(b),\ \tilde{y}_i^o(b))$ replaced by $(t_i^o,$

$\tilde{y}_i^o$). Take $0 < p < p^* < 1$ and set $a = \hat{F}^{-1}(p)$, $b = \hat{F}^{-1}(p^*)$. The knots $\xi_{k,i}$ $(i = 1, \ldots, k)$ are chosen to be the evenly spaced quantiles

$$\xi_{k,i} = \hat{F}^{-1}(p + (p^* - p)i/(k+1)). \tag{2.4}$$

Ideally we would like to choose the number of knots to minimize $\int_a^b (a_k^T(\hat{F})B_k(y) - \psi^*(y))^2 dF(y)$, or equivalently, to minimize $L(k, \hat{F}, F)$, where

$$L(k, G, F) = a_k^T(G)A_k(F)a_k(G) + 2a_k^T(G)D_k(F). \tag{2.5}$$

Since F in (2.5) is unknown, one approach to implement the minimization of (2.5) with $G = \hat{F}$ is to extend Jin's (1992) method for complete data to the l.t.r.c. situation as follows.

1. Split the data into two subsamples $\{(\tilde{y}_1^o, \delta_1^o, t_1^o), \ldots, (\tilde{y}_{n_1}^o, \delta_{n_1}^o, t_{n_1}^o)\}$, $\{(\tilde{y}_{n_1+1}^o, \delta_{n_1+1}^o, t_{n_1+1}^o), \ldots, (\tilde{y}_{n_1+n_2}^o, \delta_{n_1+n_2}^o, t_{n_1+n_2}^o)\}$, where $n_1 = [n/2]$ and $n_2 = n - n_1$. Let $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ be the product-limit estimates based on these two subsamples separately.
2. Compute $L(k, \hat{F}^{(1)}, \hat{F}^{(2)}) = a_k^T(\hat{F}^{(1)})A_k(\hat{F}^{(2)})a_k(\hat{F}^{(1)}) + 2a_k^T(\hat{F}^{(1)})D_k(\hat{F}^{(2)})$ for $k = 1, 2, \ldots$, and find the first local minimizer $\hat{k}_{cv}$ of $L(k, \hat{F}^{(1)}, \hat{F}^{(2)})$, i.e.
$$L(0, \hat{F}^{(1)}, \hat{F}^{(2)}) \geq \cdots \geq L(\hat{k}_{cv}, \hat{F}^{(1)}, \hat{F}^{(2)}) < L(\hat{k}_{cv} + 1, \hat{F}^{(1)}, \hat{F}^{(2)}).$$
3. Interchange $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ in Step 2, yielding the first local minimizer $\hat{k}'_{cv}$ of $L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$.
4. Suppose $\hat{k}'_{cv} \leq \hat{k}_{cv}$ for definiteness. Compute $ST(k, \hat{F}) = k^{-1} \sum_{j=0}^{k-1} \int_a^b (a_j^T(\hat{F})B_j(y) - a_k^T(\hat{F})B_k(y))^2 \, d\hat{F}$ and find the first local minimizer $\hat{k}$ of $ST(k, \hat{F})$ over $k \in I(n)$, where $I(n) = \{k : \hat{k}'_{cv} \leq k \leq \hat{k}^2_{cv}\}$. Thus $ST(\hat{k}'_{cv}, \hat{F}) \geq \cdots \geq ST(\hat{k}, \hat{F}) < ST(\hat{k} + 1, \hat{F})$. If there is no such $\hat{k}$ within $I(n)$, choose $\hat{k} = \hat{k}^2_{cv}$. This step is called "stationary correction" by Jin (1992), who explains its motivation as an attempt to ensure that $a_{k+1}^T(\hat{F})B_{k+1}$ does not differ too much from $a_k^T(\hat{F})B_k$ for the chosen k and thereby to reduce the variance of $\hat{k}$.

2.3. Alternative cross validation methods for truncated/censored data

To begin with, note that a simpler way to combine $L(k, \hat{F}^{(1)}, \hat{F}^{(2)})$ and $L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$ in Steps 2 and 3 above is to add them so that $\hat{k}$ is defined as the minimizer of $L(k, \hat{F}^{(1)}, \hat{F}^{(2)}) + L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$ over $0 \leq k \leq K_n$, some prescribed upper bound, instead of using Jin's stationary correction to combine the two subsample results. This is in fact tantamount to two-fold cross validation, as will be discussed below.

More generally, for m-fold cross validation, the dataset $\mathcal{S} = \{(\tilde{y}_1^o, \delta_1^o, t_1^o), \ldots, (\tilde{y}_n^o, \delta_n^o, t_n^o)\}$ is randomly divided into m disjoint subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$ with size

$[n/m]$ for the first $m-1$ subsets and $n-(m-1)[n/m]$ for $\mathcal{S}_m$. Let $\hat{F}^{(\nu)}$ be the product-limit estimate of F based on $\mathcal{S}_\nu$ and let G_ν denote the product-limit estimate of F based on $\mathcal{S}-\mathcal{S}_\nu$. We use $\mathcal{S}-\mathcal{S}_\nu$ as the "training sample", from which estimates of the coefficients of the linear spline approximation are computed, and use $\mathcal{S}_\nu$ as the "test sample", leading to the measure $L(k, G_\nu, \hat{F}^{(\nu)})$ of the mean squared error (of using the training sample estimates to predict the efficient scores of the test sample values) minus $\int_a^b (\psi^*)^2 dF$, in view of (2.3). The m-fold cross validation approach chooses $\hat{k}$ to be the minimizer of $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ over $k \le K_n$. This way of defining m-fold cross validation requires n/m to be large enough so that $\hat{F}^{(\nu)}$ estimates F reasonably well. In the case of complete data, such requirement is actually not needed and one can in fact carry out full ("leave one out") cross validation with $m=n$, since $h(y_i)$ is an unbiased estimate of $\int_{-\infty}^{\infty} h dF$. Suppose h vanishes outside an interval (a,b). When y_i is not completely observable due to censoring and truncation, we can replace the unobservable $h(y_i)$ by

$$h_F(\tilde{y}_i^o, \delta_i^o, t_i^o) = \delta_i^o h(\tilde{y}_i^o) + (1-\delta_i^o) \int_{\tilde{y}_i^o < y \le b} h(y) dF(y|\tilde{y}_i^o) + \int_{a \le y < t_i^o} h(y) dF(y)/(1-F(t_i^o-)), \tag{2.6}$$

where $F(u|v)$ is defined in (1.5); see Eq. (2.25) of Lai and Ying (1994). Although $F(y-) = F(y)$ since F is continuous, we still write $F(t_i^o-)$ in (2.6), where F will be replaced later by the product-limit estimate which is discrete. In (2.6), it is assumed that F is known; in fact, $E\{h_F(\tilde{y}_i^o, \delta_i^o, t_i^o)\} = (\int_a^b h dF)/P\{y_1 \wedge c_1 \ge t_1\}$, cf. Lemma 1 of Gross and Lai (1996). When F is unknown, we replace it in h_F by the product-limit estimate $\hat{F}^{(\nu)}$ based on the test sample $\mathcal{S}_\nu$ when n/m is not too small, or by the product-limit estimate $\hat{F}$ based on entire sample otherwise.

Let $\hat{h}^{(\nu)} = h_{\hat{F}^{(\nu)}}$, $\hat{h} = h_{\hat{F}}$, and denote the size of the subsample $\mathcal{S}_\nu$ by $\#(\mathcal{S}_\nu)$. Setting first $h = A_k$ and then $h = D_k$, an alternative to $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ as the criterion for m-fold cross validation is

$$C_m(k) = \sum_{\nu=1}^m \{ \sum_{(\tilde{y}_i^o, \delta_i^o, t_i^o) \in \mathcal{S}_\nu} a_k^T(G_\nu) \hat{A}_k(\tilde{y}_i^o, \delta_i^o, t_i^o) a_k(G_\nu) + 2 a_k^T(G_\nu) \hat{D}_k(\tilde{y}_i^o, \delta_i^o, t_i^o) \}/\#(\mathcal{S}_\nu). \tag{2.7}$$

In the censored case without truncation variables, if we replace $\hat{A}_k$ and $\hat{D}_k$ in (2.7) by $\hat{A}_k^{(\nu)}$ and $\hat{D}_k^{(\nu)}$, then (2.7) reduces to $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ as a consequence of the following identity due to Susarla, Tsai and Van Ryzin (1984):

$$\sum_{(y_j, \delta_j) \in \mathcal{S}_\nu} \hat{h}^{(\nu)}(y_j, \delta_j)/\#(\mathcal{S}_\nu) = \int h(y) d\hat{F}^{(\nu)}(y). \tag{2.8}$$

An advantage of (2.7) is that it does not involve $\hat{F}^{(\nu)}$, and therefore it is applicable also to the case of full cross validation (with $m = n$ and $\#(\mathcal{S}) = 1$).

We next establish some asymptotic properties of m-fold or full cross validation with $2 \le m \le n$. Let $H(y) = P\{t_j \le y \le c_j\}$. As pointed out by Lai and Ying (1991a), $F(y)$ may not be estimable over its entire support. In fact, $t_i^o \ge \tau$ and $\tilde{y}_i^o \le \tau^*$ with probability 1, where

$$\tau = \inf\{y : H(y) > 0\}, \quad \tau^* = \{y > \tau : H(y)(1 - F(y)) = 0\}, \tag{2.9}$$

and only the conditional distribution $F_\tau(y) = P\{Y \le y | Y \ge \tau\}$ can be nonparametrically estimated from the data, and then only for $y \le \tau^*$. With $0 < p < p^*$ chosen in (2.4) such that $F_\tau(\tau^*) > p^*$, $\hat{F}$ converges uniformly to F_τ in the interval $(F_\tau^{-1}(p), F_\tau^{-1}(p^*))$ with probability 1, cf. Lai and Ying (1991a). Making use of this, it is shown in the Appendix that

$$\lim_{n\to\infty} C_n(k)/n = L(k, F_\tau, F_\tau)/[(1 - F(\tau))P\{y_1 \wedge c_1 \ge t\}] \text{ a.s.}, \tag{2.10}$$

and that for any fixed $m \ge 2$ or for $m = m(n) \to \infty$ such that $m(n)/n \to 0$,

$$\lim_{n\to\infty} C_m(k)/n = L(k, F_\tau, F_\tau)/[(1 - F(\tau))P\{y_1 \wedge c_1 \ge t_1\}] \text{ a.s.}, \tag{2.11}$$

in which the k knots associated with $L(k, F_\tau, F_\tau)$ via (2.5) are placed at the population values $F_\tau^{-1}(p+(p^*-p)i/(k+1))$, instead of at the sample quantiles (2.4). Moreover, for any fixed $m \ge 2$ or for $m = m(n) \to \infty$ such that $m(n)/n \to 0$, $\lim_{n\to\infty} \sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})/m = L(k, F_\tau, F_\tau)$ a.s. , since $G^{(\nu)}$ and $\hat{F}^{(\nu)}$ converge uniformly to F_τ in the interval $(F_\tau^{-1}(p), F_\tau^{-1}(p^*))$ with probability 1. From (2.3) and (2.5), it follows that

$$L(k, F_\tau, F_\tau) + \int_{F_\tau^{-1}(p)}^{F_\tau^{-1}(p^*)} (\psi^*)^2 dF_\tau = \int_{F_\tau^{-1}(p)}^{F_\tau^{-1}(p^*)} \{a_k^T(F_\tau)B_k(y) - \psi^*(y)\}^2 dF_\tau(y), \tag{2.12}$$

where the k knots associated with B_k are placed at $F_\tau^{-1}(p + (p^* - p)i/(k + 1))$. Combining (2.10) or (2.11) with (2.12), we obtain the following.

Theorem 1. *Suppose $0 < p < p^*$ in (2.4) are so chosen that $F_\tau(\tau^*) > p^*$, where τ and τ^* are defined in (2.9). Let $\hat{k}$ be the minimizer of $C_n(k)$ (or of $C_m(k)$) over $k \le K_n$, where $K_n \to \infty$ and $m/n \to 0$. In the cross validation criterion (2.7) defining $C_m(k)$, we can replace $\hat{A}_k$ and $\hat{D}_k$ by $\hat{A}_k^{(\nu)}$ and $\hat{D}_k^{(\nu)}$ if $m = o(n)$. Alternatively we can also replace (2.7) by $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$.*

(i) *If $\psi^*(y) = a_{k^*}^T(F_\tau)B_{k^*}(y)$ for some $k^* \ge 0$ and all $F_\tau^{-1}(p) < y < F_\tau^{-1}(p^*)$, then $\hat{k} \to k^*$ a.s.*

(ii) *If $\int_{F_\tau^{-1}(p)}^{F_\tau^{-1}(p^*)} \{a_k^T(F_\tau)B_k(y) - \psi^*(y)\}^2 dF_\tau(y) > 0$ for every k, then $\hat{k} \to \infty$ a.s.*

In Theorem 1, p and p^* are assumed to be fixed positive constants with $p < p^*$ and $F_\tau(\tau^*) > p^*$, and the objective is to estimate ψ^* in the interval $(F_\tau^{-1}(p), F_\tau^{-1}(p^*))$. We can also let p and p^* be data-dependent and vary with n so that $\hat{F}^{-1}(p_n^*) \to \tau^*$, $p_n \to 0$ (so $\hat{F}^{-1}(p_n) \to \tau$) and

$$\int_{\hat{F}^{-1}(p_n)}^{\hat{F}^{-1}(p_n^*)} \{a_{\hat{k}}^T(\hat{F}) B_{\hat{k}}(y) - \psi^*(y)\}^2 dF(y) \to 0 \text{ a.s.};$$

see Section 3 of Lai and Ying (1991a,b) for the basic ideas underlying such choice of p_n and p_n^*. The details of the argument are quite technical and are omitted, as they are mostly similar to those in Lai and Ying (1991a,b). In practice, taking $p = 0.05$ and $p^* = 0.95$, which amounts to omitting 5% of either tail of $\hat{F}$, suffices to provide an adequate range of y's at which the score function (1.6) can be approximated by splines for use in adaptive estimation problems, while maintaining stability of the approximation. Some numerical results are presented in Section 4.1.

We have followed Jin (1992) to use linear splines because of their simplicity. Alternatively we can use smoother spline approximations to (1.6) instead. The preceding cross validation methods can also be applied to determine the number of knots for cubic or other B-splines used to approximate (1.6).

3. Adaptive M-estimators of Regression Parameters

Using the same notation and assumptions as those in the first paragraph of Section 1, we develop in this section adaptive M-estimators of the regression parameter β in model (1.1) based on l.t.r.c. data $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$, $i = 1, \ldots, n$. Following Lai and Ying (1991b) in their construction of asymptotically efficient adaptive rank estimators, we divide the observed sample randomly into two subsamples, of sizes $n_1 = [n/2]$ and $n_2 = n - n_1$ respectively. For notational simplicity, we shall relabel the original sample so that the two random subsamples can be written as $\{(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o) : 1 \le i \le n_1\}$ and $\{(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o) : n_1 < i \le n\}$, which will be referred to as the first and second subsample, respectively. The construction of adaptive M-estimators of β consists of several steps which are described in the next three subsections. Simulation results are also presented in the last subsection.

3.1. Preliminary regression estimates and score estimation based on the residuals from each subsample

To estimate the efficient score function (1.6), we use a modification of the method in Section 2 which assumes known $\beta = 0$ so that $\epsilon_j = y_j$. Since β is now unknown and need not be zero, we need preliminary estimates of β, one from each

subsample, before we can apply the method of Section 2 to the residuals that approximate the $\tilde{y}_i^o(\beta)$. The preliminary estimate $\hat{b}_1$ from the first subsample is defined by (1.2)-(1.4) with n replaced by n_1 and with the simple choice $\psi(u) = u$. It can be computed by an iterative algorithm described in Section 3.3. Likewise we can compute the preliminary estimate $\hat{b}_2$ from the second subsample. Having computed the preliminary estimates $\hat{b}_1$ and $\hat{b}_2$, we can evaluate the residuals $\{\tilde{y}_i^o(\hat{b}_1) : 1 \le i \le n_1\}$ and $\{\tilde{y}_i^o(\hat{b}_2) : n_1 < i \le n\}$ associated with the two subsamples.

With the two sets of residuals $\{\tilde{y}_i^o(\hat{b}_1) : 1 \le i \le n_1\}$ and $\{\tilde{y}_i^o(\hat{b}_2) : n_1 < i \le n\}$, compute the product-limit estimates $\hat{F}_1$ and $\hat{F}_2$ defined by (1.3)-(1.4) with $\upsilon = -\infty$ and $b = \hat{b}_j (j = 1, 2)$. Evaluate from the first subsample the score estimate $\hat{\psi}_1(x) = a_k^T(\hat{F}_1)B_k(x)$, with k chosen as the minimizer of $L(k, \hat{F}_1, \hat{F}_2)$ over $k \le K_n$, where a_k, B_k and L are defined in Sections 2.1 and 2.2. Likewise evaluate from the second subsample the score estimate $\hat{\psi}_2(x) = a_k^T(\hat{F}_2)B_k(x)$, with k chosen as the minimizer of $L(k, \hat{F}_1, \hat{F}_2)$ over $k \le K_n$. As pointed out in Section 2.3, we can use smoother B-splines such as cubic B-splines instead of linear B-splines.

3.2. Combining the two subsamples to form the adaptive M-estimator

As shown by Lai and Ying (1994), an asymptotically efficient estimator of β is the M-estimator defined by the estimating equation (1.2) with ψ given by (1.6). Since ψ is unknown, we replace $\psi(\tilde{y}_i^o(b))$ in (1.2) by $\hat{\psi}_2(\tilde{y}_i^o(b))$ for $i \le n_1$ and by $\hat{\psi}_1(\tilde{y}_i^o(b))$ for $i > n_1$, leading to the following estimating equation that defines the adaptive M-estimator:

$$\begin{aligned}
&\sum_{i=1}^{n_1} x_i^o\{\delta_i^o\hat{\psi}_2(\tilde{y}_i^o(b)) + (1-\delta_i^o)\int_{u>\tilde{y}_i^o(b)} \hat{\psi}_2(u)d\hat{F}_{b,1}(u|\tilde{y}_i^o(b))\\
&-\int_{u\ge t_i^o(b)} \hat{\psi}_2(u)d\hat{F}_{b,1}(u|t_i^o(b)-)\} + \sum_{i=n_1+1}^{n} x_i^o\{\delta_i^o\hat{\psi}_1(\tilde{y}_i^o(b))\\
&+(1-\delta_i^o)\int_{u>\tilde{y}_i^o(b)} \hat{\psi}_1(u)d\hat{F}_{b,2}(u|\tilde{y}_i^o(b)) - \int_{u\ge t_i^o(b)} \hat{\psi}_1(u)d\hat{F}_{b,2}(u|t_i^o(b)-)\} = 0, \quad (3.1)
\end{aligned}$$

where $\hat{F}_{b,j}$ is the product-limit estimate based on the $(\delta_i^o, \tilde{y}_i^o(b), \tilde{t}_i^o(b))$ from the jth subsample $(y = 1, 2)$. Thus, we associate with the $\tilde{y}_i^o(b)$ in the first subsample the score estimate $\hat{\psi}_2$ based on the second subsample, and the $\tilde{y}_i^o(b)$ in the second subsample with the score estimate $\hat{\psi}_1$ based on the first subsample. If p and p^* in Section 2.2 are chosen to vary with n such that $p_n \to 0$ and $\hat{F}^{-1}(p_n^*) \to \tau^*$ a.s. at some rate, then the arguments in Section 5 of Lai and Ying (1994) and Sections 3 and 4 of Lai and Ying (1991b) can be used to show that the adaptive M-estimator thus constructed using this sample splitting device is asymptotically

efficient. However, as pointed out near the end of Section 2.3, such elaborate choice of p and p^* is typically not needed and choosing $p = 0.05$, $p^* = 0.95$ already gives good results.

3.3. Implementation

To implement the adaptive estimation procedure described above, we need to compute preliminary estimates $\hat{b}_1$, $\hat{b}_2$ from the two subsamples and also the adaptive M-estimators defined by (3.1). Since they are all M-estimators defined by estimating equations of the type (1.2), we can compute them using the algorithm in Kim and Lai (1999), which is an extension to l.t.r.c. data of the standard method to compute M-estimators from complete data. Specifically, let X denote the $N \times p$ matrix with row vectors x_i^{oT}, in which $N = n_1$ (for the first subsample estimate $\hat{b}_1$), or n_2 (for the second subsample estimate $\hat{b}_2$), or n (for the adaptive estimate defined by (3.1)). Letting $\beta^{(k)}$ denote the result after the kth iteration, the algorithm consists of the following steps (in which $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$ may have been relabeled according to the subsample chosen and $\hat{F}_b$ may refer to $\hat{F}_{b,1}$ or $\hat{F}_{b,2}$ or the $\hat{F}_b$ for the whole sample).

1. Compute $\tilde{y}_i^o(\beta^{(k)})$, $i = 1, \ldots, N$.
2. Evaluate $\hat{F}_{\beta^{(k)}}(u|v)$ or $\hat{F}_{\beta^{(k)}}(u|v-)$ at $u \in \{\tilde{y}_i^o(\beta^{(k)}) : \delta_i^o = 1, i \le N\}$ and $v \in \{\tilde{y}_i^o(\beta^{(k)}) : i \le N\}$, $u \ge v$.
3. Compute the $N \times 1$ vector $\Psi^{(k)}$, whose ith component is $\psi(\tilde{y}_i^o(\beta^{(k)}))$, where $\psi(u) = u$ for $\hat{b}_1$ or $\hat{b}_2$, while $\psi(\tilde{y}_i^o(b)) = \hat{\psi}_2(\tilde{y}_i^o(b))$ if $i \le n_1$ and $\psi(\tilde{y}_i^o(b)) = \hat{\psi}_1(\tilde{y}_i^o(b))$ if $i > n_1$ for the adaptive estimate (3.1).
4. Solve the linear equation $X^T X z = X^T \Psi^{(k)}$ to find $z = z^{(k)}$.
5. Put $\beta^{(k+1)} = \beta^{(k)} + z^{(k)}$.
6. Increase counter from k to $k + 1$ and go to Step 1.

The rationale and the termination criteria for this iterative procedure are described in Kim and Lai (1999). Concerning the choice of $\beta^{(0)}$, Kim and Lai (1999) propose to use the weighted least squares estimate of Gross and Lai (1996), which we can apply here to initialize the algorithm for $\hat{b}_1$ and $\hat{b}_2$. However, since $\hat{b}_1$ and $\hat{b}_2$ have already been obtained before we can start the adaptive phase (3.1) of the estimation procedure, we can initialize the solution of (3.1) with the better estimate $\beta^{(0)} = (\hat{b}_1 + \hat{b}_2)/2$. This procedure is applied in the simulation study in Section 4.2.

3.4. Discussion

If the efficient score function ψ^* were known, then taking $\psi = \psi^*$ in (1.2) would lead to an asymptotically normal estimate of β with covariance matrix $n^{-1}V^*$ such that $V_T - V^*$ is nonnegative definite, for the limiting covariance matrix V_T of $\sqrt{n}(T_n - \beta)$ of any regular estimate T_n of β, cf. Lai and Ying

(1994). In ignorance of ψ^*, if we can choose ψ in (1.2) such that ψ is close to ψ^* in the L_1 sense that $\int_\tau^{\tau^*} |\psi - \psi^*| dF$ is small, then the asymptotic variance of the M-estimator defined by (1.2) with score function ψ is close to $n^{-1}V^*$. This follows from Theorem 1 of Lai and Ying (1994), noting that integration by parts yields

$$\int_u^{\tau^*} (1 - F_u(s|u))\psi'(s)ds = -F_u(\tau^*|u) + \int_u^{\tau^*} \psi(s)dF_u(s|u) \ .$$

Hence we need only estimate ψ well in an average (instead of pointwise) sense.

The sample splitting method used in (3.1) results in halving the sample size in estimating ψ^* for each subsample. To achieve reasonable (L_1-) accuracy in estimating ψ^*, $n/2$ cannot be too small. Our experience is that for $n/2 \geq 50$ the adaptive M-estimator defined by (3.1) is quite insensitive to the random selection of the two subsamples. A simulation study involving a sample size of $n = 100$ is given in Section 4.2 as an illustration.

Lai and Ying (1991b, 1994) proposed to use kernel estimates of λ and of its derivative and to smooth the resultant $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$ via a kernel method. Although their asymptotic theory allows a wide choice of bandwidths and although Uzunoğullari and Wang (1992) have provided locally adaptive methods for choosing the bandwidths for $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$, it is not clear how the bandwidth should be chosen to smooth $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$ in practice. Moreover, the locally adaptive determination of bandwidths for $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$ involves considerable computational effort. The use of B-spline approximations to ψ^* and the associated cross validation method to choose the number of knots provides a practical alternative to kernel methods (that involve bandwidth choices for $\hat{\lambda}, \hat{\lambda}^{(1)}$ and $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$, the last of which is the most difficult) in constructing the $\hat{\psi}_1$ and $\hat{\psi}_2$ in (3.1).

4. Numerical Examples

4.1. Estimation of the score function (1.6)

To estimate the score function (1.6) in the case of i.i.d. y_j, Section 2 first uses kernel estimates $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$ to estimate the hazard function λ and its derivative λ', thereby estimating (1.6) by $\hat{\psi} = \hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$. As pointed out by Lai and Ying (1991b), some smoothing of $\hat{\psi}$ is needed, particularly when $\hat{\lambda}(z)$ is near 0. Figure 1 shows a steep peak in $\hat{\psi}$ due to dividing by a small number. Such peaks can be dampened by applying some smoother to the estimate. The l.t.r.c. data in the figure consist of 200 observations $(\tilde{y}_i^o, \delta_i^o, t_i^o)$ generated from the model of independent $\log y_j \sim N(0,1)$, $\log t_j \sim N(-1,1)$ and $\log c_j \sim N(1,1)$, with about 30% of the original sample being censored and truncated. The uncensored

observations are represented by vertical bars along the horizontal axis. The kernels used in $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$ are

$$K_0(y)=\begin{cases}(15/16)(1-y^2)^2 \text{ if } |y|\le 1\\ 0 \qquad\qquad\qquad \text{otherwise}\end{cases},\quad K_1(y)=\begin{cases}(-15/4)y(1-y^2) \text{ if } |y|\le 1\\ 0 \qquad\qquad\qquad\quad \text{otherwise}\end{cases}.$$

Friedman's supersmoother (cf. Härdle (1990)) is used to smooth $\hat{\lambda}^{(1)}/\hat{\lambda}-\hat{\lambda}$.

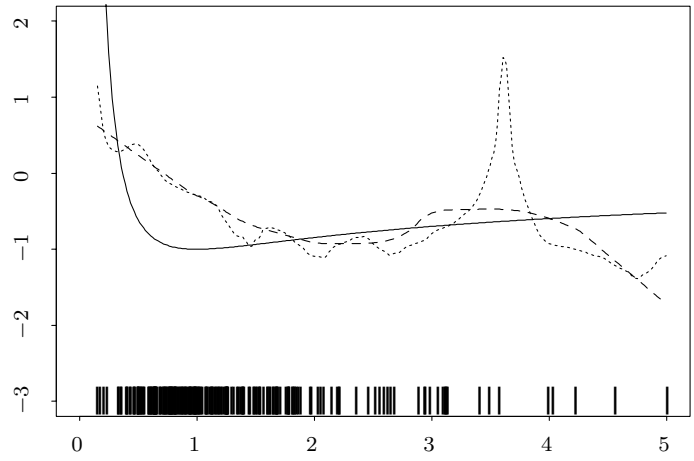

Figure 1. True score function (solid curve) and its estimate $\hat{\lambda}^{(1)}/\hat{\lambda}-\hat{\lambda}$ (dotted curve). The broken curve is obtained by smoothing the dotted curve using Friedman's supersmoother.

Section 2 considers spline approximations to the score function (1.6) and develops several versions of cross validation for l.t.r.c. data to choose the number of knots. Figure 2(a)–(d) represent the true and the estimated score functions based on a simulated dataset of 200 observations $(\tilde{y}_i^o, \delta_i^o, t_i^o)$ generated from each of the following models. The vertical line segments along the horizontal axis represent the uncensored observations.

(a) Normal: $y_j \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $t_j \overset{\text{i.i.d.}}{\sim} N(-1,1)$,
$c_j = t_j + u_j \max\{0.5, e^{-t_j}\}$ with $u_j \overset{\text{i.i.d.}}{\sim} U[0,0.1]$.

(b) Contaminated normal: $y_j \overset{\text{i.i.d.}}{\sim} 0.9N(0,1/9)+0.1N(0,9)$ and $t_j \overset{\text{i.i.d.}}{\sim} N(-1,1)$,
$c_j \overset{\text{i.i.d.}}{\sim} N(0.8,1)$.

(c) Lognormal: $\log y_j \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\log t_j \overset{\text{i.i.d.}}{\sim} N(-1,1)$,
$\log c_j \overset{\text{i.i.d.}}{\sim} N[1,1]$.

(d) Beta: $y_j \overset{\text{i.i.d.}}{\sim} Beta(2,2)$ and $t_j \overset{\text{i.i.d.}}{\sim} U[-1,1]$,
$c_j \overset{\text{i.i.d.}}{\sim} U[0.5,1]$.

There are four estimated score functions in each plot. They are labeled Kernel, Spline(J), Spline(4) and Spline(CV) respectively. "Kernel" refers to the estimate obtained by smoothing $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$ using Friedman's supersmoother. The "Spline(·)" estimates refer to the estimated linear spline approximations using different methods to choose the number of knots: Spline(J) uses the extension of Jin's method in Section 2.2; Spline(4) uses the 4-fold cross validation criterion $\sum_{\nu=1}^{4} L(k, G_\nu, F^{(\nu)})$; Spline(CV) uses the full cross validation criterion $C_n(k)$.

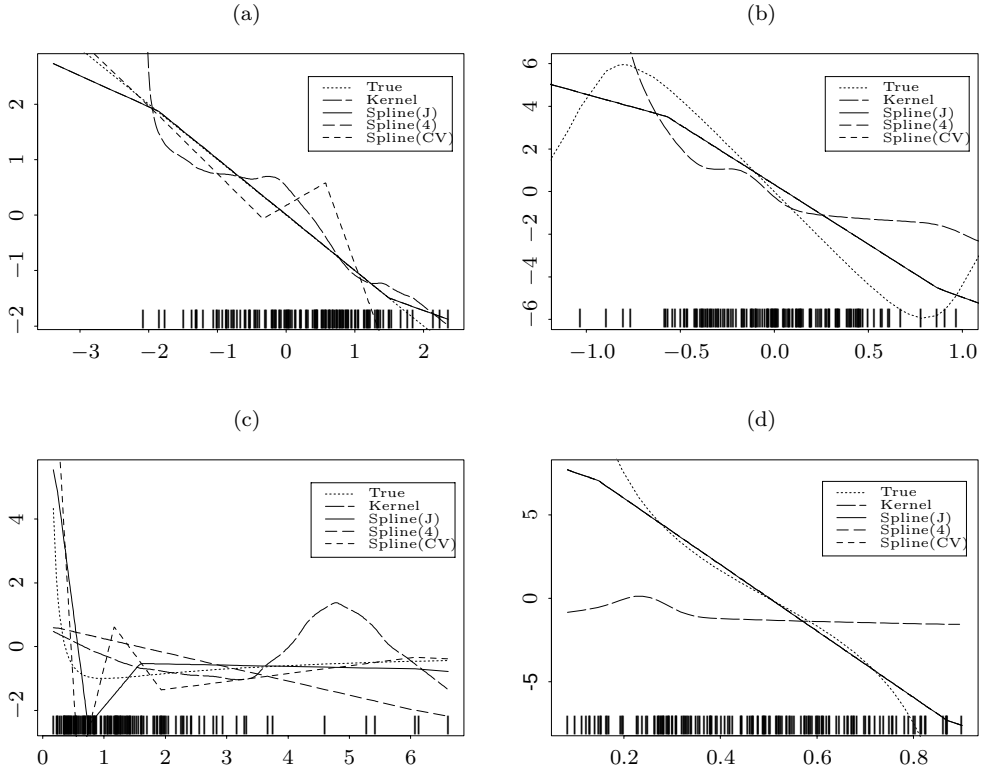

Figure 2. True and estimated score functions based on censored and truncated data from (a) normal, (b) contaminated normal, (c) lognormal, and (d) Beta populations.

In Figure 2(b) and (d), all spline estimates choose the same number of knots and therefore coincide with each other because of the way, (2.4), in which knots are placed. In Figure 2(a), Jin's method and 4-fold cross validation pick no knot between a and b while full cross validation chooses 2 knots. In Figure 2(a)–(c), all estimates are quite close to the true score function. However, in Figure 2(d), the kernel estimate is relatively flat and differs substantially from the true score function, which is well approximated by the spline estimates that coincide with each other.

Table 1 compares the mean squared errors MSE $= E(\hat{\psi}(Y) - \psi^*(Y))^2$ of the estimates $\hat{\psi}$ obtained from $(\tilde{y}_i^o, \delta_i^o, t_i^o)$, $i = 1, \ldots, 200$, by these four methods, where Y is generated from F and is independent of the (y_j, c_j, t_j) underlying the observed $(\tilde{y}_i^o, \delta_i^o, t_i^o)$. Each MSE in Table 1 is based on 100 simulations, and its associated standard deviation (SD) is also included in the table. The results in Table 1 show that the kernel estimate has considerably larger MSE than the spline estimates, and that the spline estimate with full cross validation has smaller MSE than the other spline estimates.

Table 1. Comparison of the mean squared errors (MSE) of different estimates of the efficient score function in four models, whose censoring proportion p_c and truncation proportion p_t are also indicated.

Model	**Estimate**	**MSE**	**SD**
Normal	Kernel	0.140	0.097
$p_c = 0.25$	Spline(J)	0.071	0.024
$p_t = 0.24$	Spline(4)	0.073	0.023
	Spline(CV)	0.045	0.025
Contaminated	Kernel	2.953	1.042
normal	Spline(J)	1.597	0.739
$p_c = 0.23$	Spline(4)	1.260	1.010
$p_t = 0.25$	Spline(CV)	0.450	0.120
Lognormal	Kernel	4.373	2.691
$p_c = 0.26$	Spline(J)	1.429	0.538
$p_t = 0.28$	Spline(4)	1.869	0.704
	Spline(CV)	1.516	0.801
Beta	Kernel	19.878	10.149
$p_c = 0.28$	Spline(J)	7.482	7.288
$p_t = 0.27$	Spline(4)	8.802	6.797
	Spline(CV)	4.506	2.267

4.2. Estimation of regression parameters

Consider the simple linear regression model $y_j = \beta x_j + \epsilon_j$, $j = 1, \ldots, 100$, where $\beta = 1$, the x_j are i.i.d. $U[-2, 1]$, and the ϵ_j are i.i.d. $pN(-3, 1) + (1 - p)N(3, 1)$. Suppose there are left truncation variables t_j which are i.i.d. $N(\mu_1 + x_j, 1)$ and right censoring variables c_j are i.i.d. $N(\mu_2 + x_j, 1)$. Table 2 considers five choices of (p, μ_1, μ_2) and compares the adaptive M-estimator (AdM) defined by (3.1) with the weighted least squares estimate (WLS) of Gross and Lai (1996), the Buckley-James-type M-estimator (BJ) corresponding to $\psi(u) = u$, and the Huber-type M-estimator (H) corresponding to $\psi(u) = (u \wedge 1) \vee (-1)$. It gives the mean squared error $E(\hat{\beta} - \beta)^2$, $E\hat{\beta}$ and the first, second and third quartiles of the sampling distribution of $\hat{\beta}$, for the estimate $\hat{\beta}$ of $\beta(= 1)$ based on each of the

four methods. The case $p = 1$ corresponds to normal ϵ_j, in which BJ and AdM are both asymptotically efficient, and the results in the top panel of the table are consistent with this. As p decreases, the amount of contamination increases and BJ has an MSE which is over 3 times that of AdM when $p = 0.5$ (bottom panel of the table). It is surprising that H, which is known to be effective for the contaminated normal model in the case of complete data, has even worse performance here than BJ in the bottom panel of the table. The reason is that we have not used concomitant estimation of scale in H. It should be noted that AdM automatically adapts to the scale and other features of the underlying distribution F of the ϵ_j, and is therefore a significant improvement over H with the wrong scale. Although WLS is computationally simplest among the four methods as it has a closed-form solution that does not require iterations, its computational simplicity comes at the expense of inferior performance. However, its computational simplicity can be exploited to initialize the iterative scheme to compute the BJ or H estimates, as is done in the algorithm of Kim and Lai (1999).

Table 2. Comparison of the mean squared error (MSE) and other summaries in the sampling distribution of the adaptive M-estimate (AdM) with those of three other M-estimates (BJ, H, WLS).

(p, μ_1, μ_2)	ψ	**MSE**	**Mean**	**First quartile**	**Median**	**Third quartile**
(1, -4, -2)	BJ	0.006	1.003	0.958	1.005	1.046
	H	0.006	1.001	0.959	1.006	1.042
	WLS	0.034	1.141	1.060	1.127	1.213
	AdM	0.006	1.003	0.963	0.998	1.047
(0.8, -4, 1)	BJ	0.064	1.040	0.888	1.015	1.190
	H	0.013	1.023	0.926	1.026	1.085
	WLS	0.126	1.278	1.109	1.203	1.427
	AdM	0.040	1.037	0.943	1.032	1.148
(0.6,-3.2,3.2)	BJ	5.003	2.930	1.925	2.868	3.603
	H	5.115	3.009	1.955	3.178	3.954
	WLS	0.968	1.953	1.805	1.964	2.135
	AdM	0.857	1.773	1.425	1.860	2.098
(0.6,-3.5,3.5)	BJ	2.817	2.358	1.812	1.954	2.745
	H	2.755	2.413	1.831	1.948	3.283
	WLS	0.788	1.850	1.686	1.868	2.005
	AdM	0.582	1.614	1.182	1.742	1.911
(0.5,-3.5,3.5)	BJ	1.557	1.895	1.524	1.909	2.399
	H	3.948	2.797	1.941	3.016	3.438
	WLS	0.930	1.914	1.696	1.919	2.152
	AdM	0.457	1.440	1.028	1.279	1.954

5. Conclusion

For regression analysis with complete data, the least squares estimate that is widely used because of its simplicity may have inferior performance if the errors are non-normal. Using a nonlinear score function that differs from $\psi(x) = x$ used in the least squares estimate leads to an M-estimator with greater computational complexity, but with better robustness properties. For l.t.r.c. data, there are no computational advantages in choosing $\psi(x) = x$ for the estimating equation (1.2) defining M-estimators, and the simulation results in Lai and Ying (1994) show that it can perform much worse than Huber's score function. However, for Huber's score function to perform well, one needs a concomitant estimation of scale, and in general proper choice of ψ depends on the underlying distribution F of the ϵ_j. Our numerical study shows that for samples of size 100 and larger, one can estimate ψ reasonably well and achieve good performance of the adaptive M-estimator by using regression splines and cross validation, despite substantial truncation and censoring of the response variable. In an extensive simulation study, Moon (1989) has compared the Buckley-James estimator with several other semiparamertic estimators in truncated regression models (which are called "Tobit models" in econometrics) and has found that it "stands out, in terms of short computation time and a smaller root mean square", except in highly nonnormal error distributions. The adaptive M-estimator developed here incurs at most a few times the computational cost of the Buckley-James estimator but can adapt to *any* underlying distribution of the ϵ_j. In this connection, a relatively simple cross validation method is also developed for determining the number of knots in regression splines based on l.t.r.c. data.

Acknowledgement

The research of Chul-Ki Kim is supported by a grant for the promotion of scientific research in women's universities of the Republic of Korea. The research of Tze Leung Lai is supported by the National Science Foundation, the National Security Agency and the Center for Advanced Study in the Behavioral Sciences.

Appendix

Proof of (2.10) and (2.11)

Since $\hat{F}$ converges uniformly to F_τ in the interval $(F_\tau^{-1}(p), F_\tau^{-1}(p^*))$ with probability 1, it follows from (2.4) that $\xi_{k,i}$ converges a.s. to $F_\tau^{-1}(p + (p^* - p)i/(k+1))$ for $i = 1, \ldots, k$, and for every $k \geq 1$. Moreover, letting $h = A_k$ (which is bounded and continuous), $h_{\hat{F}}$ converges a.s. to h_F defined in (2.6) but with a replaced by $F_\tau^{-1}(p)$ and b replaced by $F_\tau^{-1}(p^*)$, uniformly in the

arguments $(\tilde{y}_i^o, \delta_i^o, t_i^o)$, noting that the last term in (2.6) vanishes if $t_i^o \le a$. For any $\nu = 1, \ldots, n$, let G_ν be the product-limit estimate based on all the observations with the exception of $(\tilde{y}_\nu^o, \delta_\nu^o, t_\nu^o)$. Then as $n \to \infty$, $a_k(G_\nu)$ converges to $a_k(F_\tau)$ uniformly in $1 \le \nu \le n$ with probability 1, as can be shown by using classical exponential bounds (cf. proof of Lemma 1 of Lai and Ying (1991a)). Hence, as $n \to \infty$,

$$n^{-1}\sum_{\nu=1}^{n}\{\sum_{(\tilde{y}_i^o,\delta_i^o,t_i^o)\in\mathcal{S}_\nu} a_k^T(G_\nu)\hat{A}_k(\tilde{y}_i^o,\delta_i^o,t_i^o)a_k(G_\nu)\}$$

$$= a_k^T(F_\tau)\{\frac{1}{n}\sum_{\nu=1}^{n} h_F(\tilde{y}_\nu^o,\delta_\nu^o,t_\nu^o)\}a_k^T(F_\tau) + o(1) = \frac{a_k^T(F_\tau)A_k(F_\tau)a_k(F_\tau)}{(1-F(\tau))P\{y_1\wedge c_1 \ge t_1\}} + o(1)$$

with probability 1, recalling that $h = A_k$ and noting that the $(\tilde{y}_i^o, \delta_i^o, t_i^o)$ are i.i.d. with $E\{h_F(\tilde{y}_i^o, \delta_i^o, t_i^o)\}$
$= \int_a^b A_k dF / P\{y_1 \wedge c_1 \ge t_1\}$ by Lemma 1 of Gross and Lai (1996). Although $D_{k,i}$ has three (or two in the case $i = 0$ or $k+1$) jump discontinuities, we can use the continuity of F_τ and a slight modification of the preceding argument to show that as $n \to \infty$,

$$\frac{1}{n}\sum_{\nu=1}^{n}\{\sum_{(\tilde{y}_i^o,\delta_i^o,t_i^o)\in\mathcal{S}_\nu} a_k^T(G_\nu)\hat{D}_k(\tilde{y}_i^o,\delta_i^o,t_i^o)\} \to \frac{a_k^T(F_\tau)D_k(F_\tau)}{(1-F(\tau))P\{y_1\wedge c_1 \ge t_1\}} \text{ a.s.}$$

Since $\#(\mathcal{S}_\nu) = 1$ in the case of full cross validation (with $m = n$), (2.10) then follows from (2.7) and (2.5). The proof of (2.11) is similar.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Faraway, J. J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20**, 414-427.

Goldberger, A. S. (1981). Linear regression after selection. *J. Econometrics* **15**, 357-366.

Gross, S. and Lai, T. L. (1996). Nonparametric estimation and regression analysis with left truncation and right censored data. *J. Amer. Statist. Assoc.* **91**, 1166-1180.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hsieh, D. A. and Manski, C. F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Statist.* **15**, 541-551.

Jin, K. (1992) Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.* **20**, 1844-1874.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Keiding, N., Holst, C. and Green, A. (1989). Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality. *Amer. J. Epidemiol.* **130**, 588-600.

Kim, C. K. and Lai, T. L. (1999). Robust regression with censored and truncated data. In *Multivariate Analysis, Design of Experiments and Survey Sampling* (Edited by S. Ghosh), 231-264. Marcel Dekker, New York.

Lai, T. L. and Ying, Z. (1991a). Estimating a distribution function with truncated and censored data. *Ann. Statist.* **19**, 417-442.

Lai, T. L. and Ying, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* **19**, 531-556.

Lai, T. L. and Ying, Z. (1992). Asymptotically efficient estimation in censored and truncated regression models. *Statist. Sinica* **2**, 17-46.

Lai, T. L. and Ying, Z. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data. *Ann. Statist.* **22**, 1222-1255.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lin, D. Y. and Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *J. Comput. Graph. Statist.* **1**, 77-90.

Moon, C.-G. (1989). A Monte Carlo comparison of semiparametric Tobit estimators. *J. Appl. Econometrics* **4**, 361-382.

Nicoll, J. F. and Segal, I. E. (1980). Nonparametric estimation of the observational cutoff bias. *Astron. Astrophy.* **82**, L3-L6.

Segal, I. E. (1975). Observational validation of the chronometric cosmology. I. Preliminaries and the redshift magnitude relation. *Proc. Natl. Acad. Sci. USA*. **72**, 2437-2477.

Susarla, V., Tsai, W. Y. and Van Ryzin, J. (1984). A Buckley-James-type estimator for the mean with censored data. *Biometrika* **71**, 624-625.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24-36.

Uzunoğullari, Ü. and Wang, J. L. (1992). A comparison of hazard rate estimators for left-truncated and right-censored data. *Biometrika* **79**, 297-310.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, U.S.A.

Email: lait@stat.stanford.edu

Department of Statistics, Ewha Womans University, Seoul, 120-750, Korea.

Email: iron@mm.ewha.ac.kr

(Received February 1999; accepted December 1999)