JIRSS (2003)

Vol. 2, No. 1, pp 21 - 42

On the Second Order Behaviour of the Bootstrap of L_1 Regression Estimators

Keith Knight

Department of Statistics, University of Toronto, Toronto, Ont. M5S 3G3. (keith@utstat.toronto.edu)

Abstract. We consider the second-order asymptotic properties of the bootstrap of L_1 regression estimators by looking at the difference between the L_1 estimator and its first-order approximation, where the latter is the minimizer of a quadratic approximation to the L_1 objective function. It is shown that the bootstrap distribution of the normed difference does not converge (either in probability or with probability 1) to the "correct" limiting distribution but rather converges in distribution to a random distribution. A characterization of this random distribution is given. Some applications and extensions are given.

Received: January 2003

Key words and phrases: Bahadur-Kiefer representations; bootstrap; L_1 estimation; linear regression.

1 Introduction

Consider the linear model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad \text{for } i = 1, \cdots, n \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \end{aligned} \tag{1}$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are independent, identically distributed (iid) random variables which with median 0.

We define the L_1 estimator of $\boldsymbol{\beta}$ in (1) to minimize the objective function

$$g_n(\boldsymbol{\phi}) = \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}|; \tag{2}$$

we denote the minimizer of g_n by $\widehat{\boldsymbol{\beta}}_n$. In the model (1), L_1 estimation is sometimes considered as a robust alternative to least squares estimation (although its robustness is not clear). L_1 estimation is a special case of regression quantile estimation, introduced by Koenker and Bassett (1978), in which the absolute value function in (2) is replaced by ρ_τ (where $\rho_\tau(x) = x\{\tau - I(x < 0)\}$) in order to estimate the τ quantile of the response given the covariates; L_1 estimation corresponds to the conditional median. Although we will focus on L_1 estimation here, all the results will hold *mutatis mutandis* for regression quantile estimators for $0 < \tau < 1$.

The asymptotics of $\widehat{\boldsymbol{\beta}}_n$ are complicated somewhat by the lack of smoothness in the objective function in (2). Throughout this paper, we will assume that

(A1) the iid random variables $\varepsilon_1, \cdots, \varepsilon_n$ have common distribution function F with $F(0) = 1/2$ and $F'(0) = \lambda > 0$;

(A2) as $n \to \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = C_0 \quad \text{and} \quad \lim_{n\to\infty} \max_{1\le i\le n} \frac{\|\boldsymbol{x}_i\|}{\sqrt{n}} = 0$$

where C_0 is positive definite.

Under assumptions (A1) and (A2), it can be shown that

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, C_0^{-1}/(4\lambda^2)).$$

This result was originally proved by Bassett and Koenker (1978) under the assumption that F was absolutely continuous (with respect to Lebesgue measure) with density f although it is only necessary to assume the differentiability of F at 0.

The asymptotic distribution of the L_1 estimator can be derived easily by exploiting the convexity of the objective function as shown by Pollard (1991), Hjørt and Pollard (1993) and Geyer (1996). First define

$$Z_n(\boldsymbol{u}) = \sum_{i=1}^{n} \left(|\varepsilon_i - n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{u}| - |\varepsilon_i| \right)$$

and note that Z_n is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. Under conditions (A1) and (A2), Z_n can be approximated by a quadratic function as follows:

$$Z_n(\boldsymbol{u}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i^T\boldsymbol{u}\{I(\varepsilon_i > 0) - I(\varepsilon_i < 0)\} + \lambda\boldsymbol{u}^T C_0 \boldsymbol{u} + R_n(\boldsymbol{u}) \tag{3}$$

where $R_n(\boldsymbol{u}) \xrightarrow{p} 0$ uniformly over $\boldsymbol{u}$ in compact sets. From (3), it can be deduced that

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) &= \frac{1}{2\lambda\sqrt{n}} C_0^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i\{I(\varepsilon_i > 0) - I(\varepsilon_i < 0)\} + \boldsymbol{\xi}_n \\ &= \boldsymbol{W}_n + \boldsymbol{\xi}_n \end{aligned} \tag{4}$$

where $\boldsymbol{\xi}_n \xrightarrow{p} \mathbf{0}$. Hence, it follows that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, C_0^{-1}/(4\lambda^2))$. This approach can be extended fairly trivially to derive more exotic limiting distributions when F is not differentiable at 0 (Knight, 1998, 1999).

The asymptotic behaviour of $\boldsymbol{\xi}_n$ in (4) can be determined by examining the asymptotic behaviour of R_n in (3); this gives us a weak Bahadur-Kiefer (Bahadur, 1966; Kiefer, 1967) representation for $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. We can write

$$n^{1/4}\{R_n(\boldsymbol{u}+t\boldsymbol{w}) - R_n(\boldsymbol{u})\} = -\int_0^t \boldsymbol{w}^T \boldsymbol{D}_n(\boldsymbol{u}+s\boldsymbol{w})\,ds$$

and so if $\{\boldsymbol{D}_n\}$ converges in distribution to $\boldsymbol{D}$ then it follows from Lemma A of Knight (1997) that we have the weak Bahadur-Kiefer representation

$$n^{1/4}\boldsymbol{\xi}_n \xrightarrow{d} \frac{1}{\lambda}C_0^{-1}\boldsymbol{D}(\boldsymbol{W}) \tag{5}$$

where $\boldsymbol{W}$ has the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ and is independent of $\boldsymbol{D}$. Under slightly stronger conditions that (A1) and (A2), it can be shown that $\boldsymbol{D}_n \xrightarrow{d} \boldsymbol{D}$ where $\boldsymbol{D}$ is a zero mean Gaussian process with $\boldsymbol{D}(\boldsymbol{0}) = \boldsymbol{0}$, $E\{\boldsymbol{D}(\boldsymbol{u})\boldsymbol{D}(\boldsymbol{v})^T\} = E\{\boldsymbol{D}(\boldsymbol{v})\boldsymbol{D}(\boldsymbol{u})^T\}$ and

$$E[\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}^T] = \lim_{n\to\infty}\frac{\lambda}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T(\boldsymbol{u}-\boldsymbol{v})|\boldsymbol{x}_i\boldsymbol{x}_i^T \tag{6}$$

(where it is assumed that the limit in (6) exists and is finite for each $\boldsymbol{u}$ and $\boldsymbol{v}$). The convergence in (5) is more precisely written as

$$\left(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), n^{1/4}\boldsymbol{\xi}_n\right) \xrightarrow{d} \left(\boldsymbol{W}, \frac{1}{\lambda}C_0^{-1}\boldsymbol{D}(\boldsymbol{W})\right). \tag{7}$$

An example of $n^{1/4}R_n(\boldsymbol{u})$ is given in Figure 1; in this case, we have a simple linear regression model with $n = 200$. Weak and strong Bahadur-Kiefer representations for L_1 regression estimators are given in Arcones (1998) and Knight (1997).

The purpose of this paper is to investigate similar representations for the bootstrap. These results will complement those of De Angelis *et al* (1993) on the approximation error of the bootstrap for L_1 estimators. They consider approximating the distribution of the L_1 estimator $\widehat{\boldsymbol{\beta}}_n$ using analytical (for example, Edgeworth expansions) and bootstrap methods. They show that the error of the usual (unsmoothed) bootstrap distribution function of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is $O_p(n^{-1/4})$

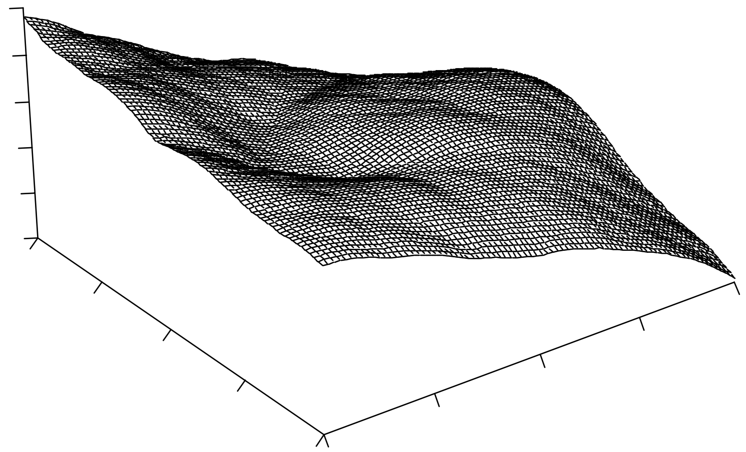

Figure 1: An example of $n^{1/4}R_n(\boldsymbol{u})$ in a simple linear regression model with $n = 200$.

which can be improved to $O_p(n^{-2/5})$ by considering an appropriate smoothed version of the bootstrap, that is, by sampling from a smoothed version of the empirical distribution of the residuals (for example, using a kernel estimator). A possible disadvantage of using the smoothed bootstrap is that oversmoothing the empirical distribution function will lead to bias that will undermine the first order consistency of the bootstrap. In this paper, we develop an exact asymptotic representation for the error in the unsmoothed bootstrap case.

2 Second order asymptotics

As in the previous section, we define $\widehat{\boldsymbol{\beta}}_n$ to be the L_1 estimator of $\boldsymbol{\beta}$ minimizing (2). Given $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_n, Y_n)$, we define the residuals

$$e_i = Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n \quad \text{for } i = 1, \cdots, n$$

$$= \quad \varepsilon_i - \boldsymbol{x}_i^T(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$$

and their empirical distribution function

$$\begin{aligned} \widehat{F}_n(t) &= \frac{1}{n}\sum_{i=1}^{n} I(e_i \le t) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_i \le t + \boldsymbol{x}_i^T \boldsymbol{U}_n/\sqrt{n}) \end{aligned}$$

where $\boldsymbol{U}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. Now let $\varepsilon_1^*, \cdots, \varepsilon_n^*$ an iid sample from $\widehat{F}_n$ and define

$$Y_i^* = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n + \varepsilon_i^* \quad \text{for } i = 1, \cdots, n.$$

(Strictly speaking, we should draw ε_i^*'s from a distribution with median 0; however, if there is an intercept in the model then $\widehat{F}_n$ will have median 0 so that this is not a concern.) Now given $Y_1^*, \cdots, Y_n^*$, we define $\widehat{\boldsymbol{\beta}}_n^*$ to minimize

$$g_n^*(\boldsymbol{\phi}) = \sum_{i=1}^{n} |Y_i^* - \boldsymbol{x}_i^T \boldsymbol{\phi}|.$$

The idea of the bootstrap procedure is that the (bootstrap) distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)$ should approximate the sampling distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. Functionals of the distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ can then be estimated by corresponding functionals of the bootstrap distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)$.

As indicated in the previous section, we will examine the asymptotic behaviour of the bootstrap distribution by considering the asymptotic behaviour of the random function

$$Z_n^*(\boldsymbol{u}) = \sum_{i=1}^{n} \left(|\varepsilon_i^* - n^{-1/2}\boldsymbol{x}_i^T \boldsymbol{u}| - |\varepsilon_i^*| \right).$$

Note that Z_n^* is convex and is minimized at $\boldsymbol{u} = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)$. Using a fairly standard argument, it can be shown that

$$Z_n^*(\boldsymbol{u}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{u}\{I(\varepsilon_i^* > 0) - I(\varepsilon_i^* < 0)\} + \lambda \boldsymbol{u}^T C_0 \boldsymbol{u} + R_n^*(\boldsymbol{u})$$

where

$$P^*\left(\sup_{u\in K}|R_n^*(\boldsymbol{u})|>\epsilon\right)\xrightarrow{p}0$$

for any compact set K where P^* denotes the (random) probability measure induced by the bootstrap sampling; P^* could be viewed as a conditional probability measure given the observed data. With probability 1, we have for each $\boldsymbol{u}$,

$$P^*\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{u}\{I(\varepsilon_i^*>0)-I(\varepsilon_i^*<0)\}\leq x\right]\to\Phi\left(\frac{x}{\sqrt{\boldsymbol{u}^TC\boldsymbol{u}}}\right)$$

and so it follows that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^*-\widehat{\boldsymbol{\beta}}_n)$ has (with probability 1) the same limiting distribution as $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})$. In particular, we have the representation

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^*-\widehat{\boldsymbol{\beta}}_n) &= \frac{1}{2\lambda\sqrt{n}}C_0^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\{I(\varepsilon_i^*>0)-I(\varepsilon_i^*<0)\}+\boldsymbol{\xi}_n^*\\ &= \boldsymbol{W}_n^*+\boldsymbol{\xi}_n^* \qquad (8)\end{aligned}$$

Note that the bootstrap distribution of $\boldsymbol{W}_n^*$ is exactly the same as the distribution of the analogous term $\boldsymbol{W}_n$ in (4).

Now consider in more depth the asymptotic behaviour of $n^{1/4}R_n^*(\boldsymbol{u})$. To do this, we need the following regularity conditions.

(A3) For each compact set K,

$$n^{-3/4}\sup_{u\in K}\left\|\sum_{i=1}^{n}\boldsymbol{x}_i r_n(\boldsymbol{x}_i^T\boldsymbol{u})\right\|\to 0$$

where $r_n(t)=\sqrt{n}\{F(t/\sqrt{n})-F(0)\}-\lambda t$.

(A4) For C_0 defined in (A2),

$$n^{-3/4}\sum_{i=1}^{n}\left(\boldsymbol{x}_i\boldsymbol{x}_i^T-C_0\right)\to 0.$$

(A5) For some $\delta>0$,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^{3+\delta}<\infty.$$

(A6) For each $\boldsymbol{u}$,

$$\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{u}|\boldsymbol{x}_i\boldsymbol{x}_i^T \to B(\boldsymbol{u}) < \infty.$$

Conditions (A3), (A4), (A5) and (A6) are effectively moment conditions on the $\boldsymbol{x}_i$'s and are quite closely related. For example, suppose that the $\boldsymbol{x}_i$'s were i.i.d. vectors with $E(\|\boldsymbol{x}_i\|^{3+\delta}) < \infty$ for some $\delta > 0$; then (A5) is satisfied, $C_0 = E(\boldsymbol{x}_i\boldsymbol{x}_i^T)$ and conditions (A5) and (A6) follow from the finiteness of $E(\|\boldsymbol{x}_i\|^3)$. Moreover, if F has a Lipschitz continuous derivative (in a neighbourhood of 0) then $|r_n(t)| \le kt^2/\sqrt{n}$ and so (A3) is implied by (A6).

Theorem 2.1. *Define $\boldsymbol{\xi}_n^*$ as in (8) and assume conditions (A1) to (A6). Then for each $\boldsymbol{s} \in R^{p+1}$ and x, we have*

$$P^*\left(n^{1/4}\boldsymbol{s}^T\boldsymbol{\xi}_n^* \le x\right) \xrightarrow{d} P^*\left[\frac{1}{\lambda}\boldsymbol{s}^T C_0^{-1}\{\boldsymbol{D}^*(\boldsymbol{W}^*) + \boldsymbol{D}(\boldsymbol{W}^*)\} \le x\right] \tag{9}$$

where

(a) $\boldsymbol{D}^*$ and $\boldsymbol{D}$ are independent zero mean Gaussian processes with

$$\begin{aligned} & E^*\left[\{\boldsymbol{D}^*(\boldsymbol{u}) - \boldsymbol{D}^*(\boldsymbol{v})\}\{\boldsymbol{D}^*(\boldsymbol{u}) - \boldsymbol{D}^*(\boldsymbol{v})\}^T\right] \\ = \ & E\left[\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}^T\right] \\ = \ & \lambda B(\boldsymbol{u} - \boldsymbol{v}); \end{aligned}$$

(b) $\boldsymbol{W}^*$ has a Normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $C_0^{-1}/(4\lambda^2)$; moreover, $\boldsymbol{W}^*$ is independent of $\boldsymbol{D}^*$ and $\boldsymbol{D}$.

(Note that the limiting probability in (9) is a random variable; in this case, P^* and E^* are computed with respect to the conditional distribution of $(\boldsymbol{D}^*, \boldsymbol{W}^*)$ given $\boldsymbol{D}$ and hence depend on the stochastic process $\boldsymbol{D}$.)

Proof. First of all, we have

$$n^{1/4}\{R_n^*(\boldsymbol{u}+t\boldsymbol{w})-R_n^*(\boldsymbol{u})\}=-2\int_0^t \boldsymbol{w}^T\boldsymbol{D}_n^*(\boldsymbol{u}+s\boldsymbol{w})\,ds$$

where

$$\boldsymbol{D}_n^*(\boldsymbol{u})=-n^{-3/4}\sum_{i=1}^{n}\left[\boldsymbol{x}_i\sqrt{n}\{I(\varepsilon_i^*\le \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})-I(\varepsilon_i^*\le 0)\}-\lambda C_0\boldsymbol{u}\right].$$

For convenience, define $\boldsymbol{U}_n=\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})$, $H(t)=F(t)-F(0)$ and

$$\begin{aligned}\widehat{H}_n(t) &= \widehat{F}_n(t)-\widehat{F}_n(0)\\ H_n(t) &= \frac{1}{n}\sum_{i=1}^{n}\{F(t+\boldsymbol{x}_i^T\boldsymbol{U}_n/\sqrt{n})-F(\boldsymbol{x}_i^T\boldsymbol{U}_n/\sqrt{n})\}.\end{aligned}$$

For t in a neighbourhood of 0, $H_n(t)$ and $\widehat{H}_n(t)$ are close to $H(t)$ for large n. Then we can write

$$\begin{aligned}\boldsymbol{D}_n^*(\boldsymbol{u}) = -\Bigg[& n^{-1/4}\sum_{i=1}^{n}\boldsymbol{x}_i\left\{I(\varepsilon_i^*\le \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})-I(\varepsilon_i^*\le 0))-\widehat{H}_n(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})\right\}\\ & +n^{-1/4}\sum_{i=1}^{n}\boldsymbol{x}_i\left\{\widehat{H}_n(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})-H_n(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})\right\}\\ & +n^{-1/4}\sum_{i=1}^{n}\boldsymbol{x}_i\left\{H_n(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})-H(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})\right\}\\ & +n^{-3/4}\sum_{i=1}^{n}\left\{\sqrt{n}H(\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})\boldsymbol{x}_i-\lambda C_0\boldsymbol{u}\right\}\Bigg]\\ = {} & \boldsymbol{D}_{n1}^*(\boldsymbol{u})+\boldsymbol{D}_{n2}(\boldsymbol{u})+\boldsymbol{D}_{n3}(\boldsymbol{u})+\boldsymbol{D}_{n4}(\boldsymbol{u}).\end{aligned}$$

Note that only $\boldsymbol{D}_{n1}^*(\boldsymbol{u})$ is random with respect to bootstrap sampling (hence the superscript $*$) while $\boldsymbol{D}_{n2}(\boldsymbol{u})$, $\boldsymbol{D}_{n3}(\boldsymbol{u})$ and $\boldsymbol{D}_{n4}(\boldsymbol{u})$ are constant (as they do not depend on the ε_i^*'s). By assumptions (A3) and (A4), $\boldsymbol{D}_{n3}(\boldsymbol{u}),\boldsymbol{D}_{n4}(\boldsymbol{u})\to\boldsymbol{0}$ uniformly over $\boldsymbol{u}$ in compact sets. $\boldsymbol{D}_{n1}^*$ converges in distribution (with probability 1) to a zero mean Gaussian process $\boldsymbol{D}^*$ with $\boldsymbol{D}^*(\boldsymbol{0})=\boldsymbol{0}$ whose covariance structure is given by

$$E\left[\{\boldsymbol{D}^*(\boldsymbol{u})-\boldsymbol{D}^*(\boldsymbol{v})\}\{\boldsymbol{D}^*(\boldsymbol{u})-\boldsymbol{D}^*(\boldsymbol{v})\}^T\right]=\lambda B(\boldsymbol{u}-\boldsymbol{v}).$$

Likewise, $\boldsymbol{D}_{n2}$ converges in distribution to a zero mean Gaussian process $\boldsymbol{D}$ (independent of $\boldsymbol{D}^*$) having the same covariance structure as $\boldsymbol{D}^*$. The result follows by applying Lemma A of Knight (1997). $\square$

Theorem 1 also holds jointly for a finite number of $\boldsymbol{s}$'s and x's; for example,

$$\left\{P^*\left(n^{1/4}\boldsymbol{s}_\ell^T\boldsymbol{\xi}_n^* \le x_\ell\right), \ell = 1,\cdots,k\right\} \xrightarrow{d} \left(P^*\left[\frac{1}{\lambda}\boldsymbol{s}_\ell^T C_0^{-1}\{\boldsymbol{D}^*(\boldsymbol{W}^*) + \boldsymbol{D}(\boldsymbol{W}^*)\} \le x_\ell\right], \ell = 1,\cdots,k\right)$$

for any $\boldsymbol{s}_1,\cdots,\boldsymbol{s}_k$ and $x_1,\cdots,x_k$. An expression for the limiting distribution function is fairly simple to derive. Given $\boldsymbol{W}^* = \boldsymbol{w}$, $\lambda^{-1}\boldsymbol{s}^T C_0^{-1}\{\boldsymbol{D}^*(\boldsymbol{W}^*)+\boldsymbol{D}(\boldsymbol{W}^*)\}$ has a Normal distribution with mean $\lambda^{-1}\boldsymbol{s}^T C_0^{-1}\boldsymbol{D}(\boldsymbol{w})$ and variance $\lambda^{-1}\boldsymbol{s}^T C_0^{-1}B(\boldsymbol{w})C_0^{-1}\boldsymbol{s}$. Thus if Φ is the standard Normal distribution function, we have

$$P^*\left[\frac{1}{\lambda}\boldsymbol{s}^T C_0^{-1}\{\boldsymbol{D}^*(\boldsymbol{W}^*) + \boldsymbol{D}(\boldsymbol{W}^*)\} \le x\right] = \int\cdots\int \Phi\left(\frac{\lambda^{1/2}\left\{x - \lambda^{-1}\boldsymbol{s}^T C_0^{-1}\boldsymbol{D}(\boldsymbol{w})\right\}}{\left\{\boldsymbol{s}^T C_0^{-1}B(\boldsymbol{w})C_0^{-1}\boldsymbol{s}\right\}^{1/2}}\right) f(\boldsymbol{w})\,d\boldsymbol{w}$$

where $f(\boldsymbol{w})$ is a $\mathcal{N}(\boldsymbol{0}, C_0^{-1}/(4\lambda^2))$ density. On the other hand, we have

$$P\left(n^{1/4}\boldsymbol{s}^T\boldsymbol{\xi}_n \le x\right) \to \int\cdots\int \Phi\left(\frac{\lambda^{1/2}x}{\{\boldsymbol{s}^T C_0^{-1}B(\boldsymbol{w})C_0^{-1}\boldsymbol{s}\}^{1/2}}\right) f(\boldsymbol{w})\,d\boldsymbol{w}.$$

The difference between these two limits essentially depends on the function $\boldsymbol{D}$, which is related to how well (in the limit) the objective function Z_n in (3) is approximated by a quadratic function.

Example 2.1. Consider a two sample estimation problem, which can be written as the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1,\cdots,n)$$

where $x_i = 0$ or 1 each for approximately $n/2$ observations and the ε_i's are i.i.d. with $F'(0) = \lambda > 0$. Then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, C_0^{-1}/(4\lambda^2))$$

where

$$C_0 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

Then $B(\boldsymbol{u}) = B(u_0, u_1)$ defined in (A6) becomes

$$B(\boldsymbol{u}) = \frac{1}{2}\begin{pmatrix} |u_0| + |u_0 + u_1| & |u_0 + u_1| \\ |u_0 + u_1| & |u_0 + u_1| \end{pmatrix}.$$

In this case, the vector process $\boldsymbol{D}$ can be expressed in terms of two real-valued two-sided Brownian motion processes so that it is fairly straightforward to draw random samples (of distribution functions) from the limiting distribution for a given vector $\boldsymbol{s}$. To illustrate, we will take $\boldsymbol{s} = (0,1)^T$. A random sample of 10 such distribution functions is given in Figure 2 (taking $\lambda = 1$); the corresponding density functions are shown in Figure 3.

The effect of taking bootstrap samples from a smoothed version of $\widehat{F}_n$ (call it $\widetilde{F}_n$) is easily seen from the proof above. For an appropriate choice of $\widetilde{F}_n$, we will have $\boldsymbol{D}_{n2}(\boldsymbol{u}) \xrightarrow{p} \mathbf{0}$, in which case,

$$P^*\left(n^{1/4}\boldsymbol{s}^T\boldsymbol{\xi}_n^* \le x\right) \xrightarrow{p} P^*\left\{\frac{1}{\lambda}\boldsymbol{s}^T C_0^{-1}\boldsymbol{D}^*(\boldsymbol{W}^*) \le x\right\}$$

where the limit above is fixed and agrees with the limit of $P(n^{1/4}\boldsymbol{s}^T\boldsymbol{\xi}_n \le x)$. As mentioned above, the downside of smoothing $\widehat{F}_n$ is the possibility that the smoothed version may not be a good estimator of the underlying F (for example, if $\widehat{F}_n$ is oversmoothed) in which case the bootstrap may not be correct to first order.

It is easy to extend Theorem 1 to the case where bootstrap samples are generated by sampling $(\boldsymbol{x}_1^*, Y_1^*), \cdots, (\boldsymbol{x}_n^*, Y_n^*)$ with replacement from $\{\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_n, Y_n)\}$. The approach to proving this modification of Theorem 1 is essentially the same except that Z_n^* must

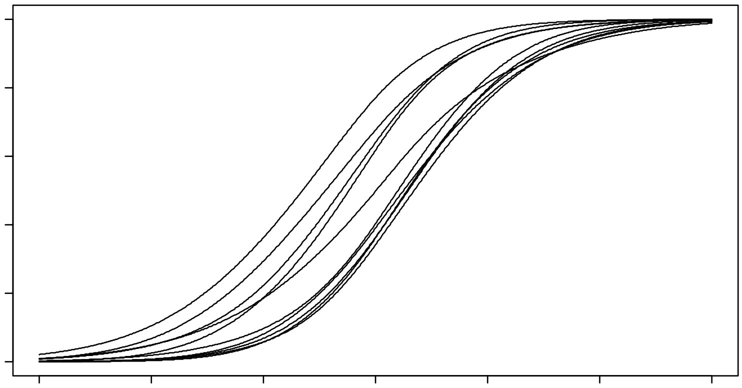

Figure 2: Sample of 10 distribution functions

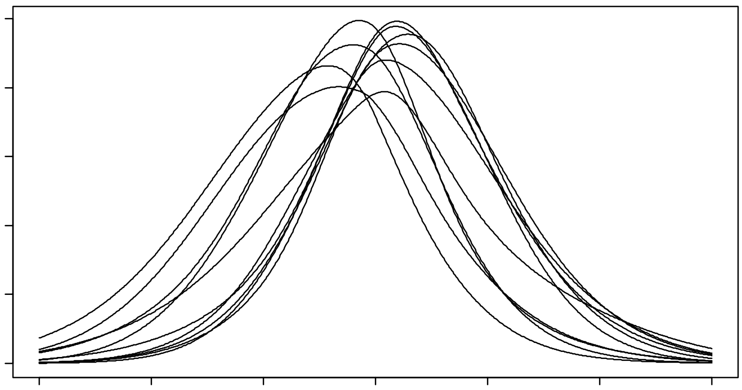

Figure 3: The 10 density functions corresponding to the distribution functions in Figure 2

now be defined as

$$Z_n^*(\boldsymbol{u}) = \sum_{i=1}^n M_{ni}^* \left(|\varepsilon_i - n^{-1/2}\boldsymbol{x}_i^T \boldsymbol{u}| - |\varepsilon_i| \right)$$

where $(M_{n1}^*, \cdots, M_{nn}^*)$ is a Multinomial random vector with $M_{n1}^* + \cdots + M_{nn}^* = n$; Z_n^* is now minimized at

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}) = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n) + \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}).$$

The difference between Z_n^* and the appropriate quadratic function has a similar asymptotic representation and the result of Theorem 1 holds after some manipulation.

3 Bootstrap covariance estimator

The bootstrap distribution is commonly used to estimate the covariance structure of $\widehat{\boldsymbol{\beta}}_n$ as well as other moments. In this section, we will concentrate on the bootstrap covariance estimator although similar results should apply for other bootstrap moment estimators.

Using the representation (4), we have

$$\begin{aligned} nE\left\{(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T\right\} &= E(\boldsymbol{W}_n \boldsymbol{W}_n^T) + E(\boldsymbol{W}_n \boldsymbol{\xi}_n^T) \\ &\quad + E(\boldsymbol{\xi}_n \boldsymbol{W}_n^T) + E(\boldsymbol{\xi}_n \boldsymbol{\xi}_n^T). \end{aligned}$$

Simple algebra gives us

$$\begin{aligned} E(\boldsymbol{W}_n \boldsymbol{W}_n^T) &= \frac{1}{4\lambda^2 n} C_0^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right) C_0^{-1} \\ &\to \frac{1}{4\lambda^2} C_0^{-1} \end{aligned} \tag{10}$$

Moreover, the convergence in distribution of $(\boldsymbol{W}_n, n^{1/4}\boldsymbol{\xi}_n)$ to $(\boldsymbol{W}, -\lambda^{-1}C_0^{-1}\boldsymbol{D}(\boldsymbol{W}))$ suggests that

$$n^{1/4} E(\boldsymbol{\xi}_n \boldsymbol{W}_n^T) \to \frac{1}{\lambda} C_0^{-1} E\left\{\boldsymbol{D}(\boldsymbol{W})\boldsymbol{W}^T\right\} = 0 \tag{11}$$

and

$$n^{1/2}E(\boldsymbol{\xi}_n\boldsymbol{\xi}_n^T) \to \frac{1}{\lambda^2}C_0^{-1}E\left\{\boldsymbol{D}(\boldsymbol{W})\boldsymbol{D}^T(\boldsymbol{W})\right\}C_0^{-1}. \qquad (12)$$

The convergence in (11) and (12) will hold given appropriate uniform integrability conditions on the sequence $\{n^{1/4}\boldsymbol{\xi}_n\}$, for example, if $E(\|n^{1/4}\boldsymbol{\xi}_n\|^{2+\delta})$ is bounded for some $\delta > 0$. It can be shown that uniform integrability of $\{n^{1/4}\boldsymbol{\xi}_n\}$ follows if $E[|\varepsilon_i|^\eta] < \infty$ for some $\eta > 0$.

It is straightforward to determine analogous results for the corresponding bootstrap quantities. Writing $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n) = \boldsymbol{W}_n^* + \boldsymbol{\xi}_n^*$, we have

$$E^*\left(\boldsymbol{W}_n^*\boldsymbol{W}_n^{*T}\right) \to \frac{1}{4\lambda^2}C_0^{-1} \qquad (13)$$

$$n^{1/4}E^*\left(\boldsymbol{\xi}_n^*\boldsymbol{W}_n^{*T}\right) \xrightarrow{d} \frac{1}{\lambda}C_0^{-1}E^*\left\{\boldsymbol{D}^T(\boldsymbol{W}^*)\boldsymbol{W}^*\right\} \neq \boldsymbol{0} \qquad (14)$$

$$n^{1/2}E^*\left(\boldsymbol{\xi}_n^*\boldsymbol{\xi}_n^{*T}\right) \xrightarrow{d} \frac{1}{\lambda^2}C_0^{-1}E\left[\{\boldsymbol{D}^*(\boldsymbol{W}^*) + \boldsymbol{D}(\boldsymbol{W}^*)\}\{\boldsymbol{D}^*(\boldsymbol{W}^*) + \boldsymbol{D}(\boldsymbol{W}^*)\}^T\right]C_0^{-1} \qquad (15)$$

(again assuming the appropriate uniform integrability conditions are satisfied). Note that (13), (14) and (15) imply that

$$n^{1/4}\left[n\,E^*\left\{(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)^T\right\} - \frac{1}{4\lambda^2}C_0^{-1}\right] \xrightarrow{d} -\frac{1}{\lambda}\left[C_0^{-1}E^*\left\{\boldsymbol{D}(\boldsymbol{W}^*)\boldsymbol{W}^{*T}\right\} + E^*\left\{\boldsymbol{W}^*\boldsymbol{D}^T(\boldsymbol{W}^*)\right\}C_0^{-1}\right] \qquad (16)$$

while (10), (11) and (12) imply that

$$n^{1/4}\left[n\,E\left\{(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T\right\} - \frac{1}{4\lambda^2}C_0^{-1}\right] \to 0.$$

The limiting random matrix in (16) is symmetric and has normally distributed entries.

Example 2.2. The sample median is the L_1 estimator of location in the i.i.d. case. If $\widehat{\mu}_n$ is the sample median of i.i.d. random variables

$Y_1, \cdots, Y_n$ and $\widehat{\mu}_n^*$ the median of a bootstrap sample from $Y_1, \cdots, Y_n$, we have (assuming the regularity conditions of the previous section),

$$\xi_n^* = \sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) - \frac{1}{2\lambda\sqrt{n}} \sum_{i=1}^{n} \{I(Y_i^* > \widehat{\mu}_n) - I(Y_i^* < \widehat{\mu}_n)\}$$

and

$$P^* \left(n^{1/4} \xi_n^* \le x\right) \xrightarrow{d} P^* \left[\frac{1}{\lambda}\{D^*(W^*) + D(W^*)\} \le x\right]$$

where D^* and D are independent Gaussian processes with $E[\{D^*(s) - D^*(t)\}^2] = E[\{D(s) - D(t)\}^2] = \lambda|s-t|$, and W^* is a zero mean Normal random variable with variance $1/(4\lambda^2)$.

For the bootstrap variance estimator, we have

$$\begin{aligned} n^{1/4} \left[n\, E^*\{(\widehat{\mu}_n^* - \widehat{\mu}_n)^2\} - \frac{1}{4\lambda^2}\right] &\xrightarrow{d} \\ \frac{2}{\lambda} \int_{-\infty}^{\infty} wD(w) \frac{2\lambda}{\sqrt{2\pi}} \exp(-2\lambda^2 w^2)\, dw& \end{aligned} \tag{17}$$

where the limiting random variable in (17) is Normal with mean 0 and variance

$$\begin{aligned} \sigma^2 &= \frac{4}{\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w\, v E\{D(w)D(v)\} \frac{4\lambda^2}{2\pi} \exp\{-2\lambda^2(w^2+v^2)\}\, dw\, dv \\ &= \frac{8}{\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{w\,v}{2} \left(|w| + |v| - |w-v|\right) \exp\{-2\lambda^2(w^2+v^2)\}\, dw\, dv \\ &= \frac{1}{4\lambda^4\sqrt{\pi}}. \end{aligned}$$

This agrees with Theorem 2.2 of Hall and Martin (1988).

Example 2.3. Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \cdots, n)$$

where $x_i = i/(n+1)$ and the ε_i's are i.i.d. with $F'(0) = \lambda > 0$. Then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, C_0^{-1}/(4\lambda^2))$$

where

$$C_0 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}.$$

We will focus on the bootstrap estimator of $\widehat{\beta}_{1n}$. Define the functions

$$\begin{aligned} \psi_{11}(a,b) &= \int_0^1 |a+bt|\,dt \\ \psi_{12}(a,b) &= \int_0^1 |a+bt|t\,dt \\ \psi_{22}(a,b) &= \int_0^1 |a+bt|t^2\,dt. \end{aligned}$$

Then $B(\boldsymbol{u}) = B(u_0, u_1)$ defined in (A6) becomes

$$B(\boldsymbol{u}) = \begin{pmatrix} \psi_{11}(u_0,u_1) & \psi_{12}(u_0,u_1) \\ \psi_{12}(u_0,u_1) & \psi_{22}(u_0,u_1) \end{pmatrix}.$$

Then

$$\begin{aligned} & n^{1/4}\left[n\,E^*\{(\widehat{\beta}^*_{1n} - \widehat{\beta}_{1n})^2\} - \frac{3}{\lambda^2}\right] \xrightarrow{d} \\ & \quad 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{e}^T C_0^{-1}\boldsymbol{D}(\boldsymbol{u})\boldsymbol{u}^T\boldsymbol{e} f(\boldsymbol{u})\,d\boldsymbol{u} \end{aligned} \tag{18}$$

where $E[\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}\{\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v})\}^T] = \lambda B(\boldsymbol{u}-\boldsymbol{v})$, $f(\boldsymbol{u})$ is a $\mathcal{N}(\boldsymbol{0}, C_0^{-1}/(4\lambda^2))$ density and $\boldsymbol{e} = (0,1)^T$. The variance of the limiting random variable in (18) is

$$\begin{aligned} \sigma^2 &= 2\lambda^2 \int\int\int\int \left(\boldsymbol{e}^T C_0^{-1}\{B(\boldsymbol{u}) + B(\boldsymbol{v}) - B(\boldsymbol{u}-\boldsymbol{v})\} C_0^{-1}\boldsymbol{e}\right) \\ & \quad (\boldsymbol{u}^T\boldsymbol{e})(\boldsymbol{v}^T\boldsymbol{e}) f(\boldsymbol{u}) f(\boldsymbol{v})\,d\boldsymbol{u}d\boldsymbol{v} \\ &= \frac{21.02}{\lambda^4} \end{aligned} \tag{19}$$

where the numerator above has been evaluated via Monte Carlo integration (a 95% confidence interval for the numerator is 21.02 ± 0.01).

To investigate how well the asymptotic variance approximates the variance of $E^*\{(\widehat{\beta}^*_{1n} - \widehat{\beta}_{1n})^2\}$ for finite samples, we simulated 1000 samples for various n from the simple linear regression model with

n	20	50	100	200	500	1000	2000	5000	10000
est'd σ_n^2	322.5	359.0	390.0	460.7	493.0	585.6	616.7	663.2	870.7

Table 1: Monte Carlo estimates of σ_n^2 in (20) for various values of n; the standard errors of the estimated σ_n^2's range from approximately 16 (for $n = 20$) to approximately 40 (for $n = 10000$).

standard Normal errors; for each sample, we approximated $E^*\{(\widehat{\beta}^*_{1n} - \widehat{\beta}_{1n})^2\}$ using 200 bootstrap samples. For each n, the 1000 samples were used to approximate

$$\sigma_n^2 = n^{1/2} E\left(\left[n\, E^*\{(\widehat{\beta}^*_{1n} - \widehat{\beta}_{1n})^2\} - \frac{3}{\lambda^2}\right]^2\right) \tag{20}$$

where $\lambda = (2\pi)^{-1/2}$. The asymptotic variance σ^2 given by equation (19) is 829.8. As can be seen from Table 1, the convergence of σ_n^2 to its asymptotic value is quite slow.

4 Adaptive regression

Dodge and Jurečkova (1992) propose estimating $\boldsymbol{\beta}$ in (1) by minimizing the objective function

$$\delta \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}| + (1-\delta) \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})^2 \tag{21}$$

for some $\delta \in [0, 1]$; they describe a method for adaptively choosing δ based on minimizing an estimate of the asymptotic variance.

Assuming that the ε_i's have mean and median 0 with $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and that conditions (A1) and (A2) hold, then for any $\delta \in [0, 1]$, the estimator $\widehat{\boldsymbol{\beta}}_n(\delta)$ minimizing (21) satisfies

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n(\delta) - \boldsymbol{\beta}) \quad &= \quad \frac{1}{\{2\delta\lambda + 2(1-\delta)\}\sqrt{n}} \\ & \qquad C_0^{-1} \sum_{i=1}^{n} \{\delta \,\mathrm{sgn}(\varepsilon_i) + 2(1-\delta)\varepsilon_i\}\, \boldsymbol{x}_i + \boldsymbol{\xi}_n(\delta)\end{aligned}$$

$$\xrightarrow{d} \mathcal{N}(\mathbf{0}, \omega(\delta) C_0^{-1})$$

where $\boldsymbol{\xi}_n(\delta) = O_p(n^{-1/4})$ and

$$\omega(\delta) = \frac{\delta^2 + 4(1-\delta)^2\sigma^2 + 4\delta(1-\delta)E(|\varepsilon_i|)}{\{2\delta\lambda + 2(1-\delta)\}^2}. \tag{22}$$

Under suitable regularity conditions, we have

$$n^{1/4}\boldsymbol{\xi}_n(\delta) \xrightarrow{d} \frac{\delta}{1+\delta(\lambda-1)} C_0^{-1}\boldsymbol{D}(\boldsymbol{W}(\delta))$$

where $\boldsymbol{D}$ is a Gaussian process with covariance structure given in (6) and $\boldsymbol{W}(\delta) \sim \mathcal{N}(\mathbf{0}, \omega(\delta)C_0^{-1})$; in fact, $\boldsymbol{W}(\delta)$ is a zero mean Gaussian process with covariance structure

$$E\left\{\boldsymbol{W}(\delta_1)\boldsymbol{W}(\delta_2)^T\right\} = \\ \frac{\delta_1\delta_2 + 4(1-\delta_1)(1-\delta_2)\sigma^2 + 2E[|\varepsilon_i|]\left\{\delta_1(1-\delta_2)+\delta_2(1-\delta_1)\right\}}{\{2\delta_1\lambda + 2(1-\delta_1)\}\{2\delta_2\lambda + 2(1-\delta_2)\}} C_0^{-1}.$$

Define δ_0 as the value of $\delta \in [0,1]$ that minimizes $\omega(\delta)$ defined in (22); if $\{\widehat{\delta}_n\}$ satisfies $\widehat{\delta}_n \xrightarrow{p} \delta_0$ then it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n(\widehat{\delta}_n) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \omega(\delta_0)C_0^{-1}).$$

(On the other hand, if $\widehat{\boldsymbol{\beta}}_n(\delta) \xrightarrow{p} \boldsymbol{\beta}(\delta)$, which would occur, for example, if the mean and median of the ε_i's were different, then the convergence rate of $\widehat{\boldsymbol{\beta}}_n(\widehat{\delta}_n)$ to $\boldsymbol{\beta}(\delta_0)$ for some δ_0 could depend on the convergence rate of $\widehat{\delta}_n$ to δ_0.) The estimator proposed by Dodge and Jurečkova (1992) has $\widehat{\delta}_n = \delta_0 + O_p(n^{-1/4})$ and it is possible to construct estimators with $\widehat{\delta}_n = \delta_0 + O_p(n^{-2/5})$. Another approach to choosing δ might be to minimize the bootstrap estimate of $\mathrm{Var}\{\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_n(\delta)\}$ over $\delta \in [0,1]$. Mimicking the argument used in the preceding section, we get

$$n^{1/4}\left[n\,E^*\left\{(\widehat{\boldsymbol{\beta}}_n^*(\delta) - \widehat{\boldsymbol{\beta}}_n(\delta))(\widehat{\boldsymbol{\beta}}_n^*(\delta) - \widehat{\boldsymbol{\beta}}_n(\delta))^T\right\} - \omega(\delta)C_0^{-1}\right] \xrightarrow{d} \\ \frac{\delta}{1+\delta(\lambda-1)}\left[C_0^{-1}E^*\left\{\boldsymbol{D}(\boldsymbol{W}^*(\delta))\boldsymbol{W}^*(\delta)^T\right\} + E^*\left\{\boldsymbol{W}^*(\delta)\boldsymbol{D}^T(\boldsymbol{W}^*(\delta))\right\}C_0^{-1}\right] \\ = J(\delta)$$

where $\boldsymbol{W}^*(\delta) \sim \mathcal{N}(\mathbf{0}, \omega(\delta) C_0^{-1})$ (and is independent of $\boldsymbol{D}$) and the expectations are taken with respect to the conditional distribution of $\boldsymbol{W}^*(\delta)$ given $\boldsymbol{D}$. The matrix $J(\delta)$ has Normally distributed entries for each δ and is differentiable in δ with probability 1. If $\delta_0 \in (0,1)$ and $\widehat{\delta}_n(\boldsymbol{a})$ minimizes (over δ) the bootstrap quantity $E^*[\{\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_n^*(\delta) - \boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_n(\delta)\}^2]$ then the result above suggests that

$$n^{1/4}\{\widehat{\delta}_n(\boldsymbol{a}) - \delta_0\} \xrightarrow{d} V(\boldsymbol{a}).$$

Note that the limiting random above may depend on the vector $\boldsymbol{a}$ even though δ_0 does not.

Note that the bootstrap procedure described above for estimating δ is not practical as the computational cost of the minimization of the bootstrap variance over the interval $[0,1]$ is very high. However, the result above does suggest that any estimation of δ based on minimizing bootstrap variance over a suitably fine grid is likely to be unstable. This is not a great concern if $\widehat{\boldsymbol{\beta}}_n(\delta)$ minimizing (21) is more or less constant for $\delta \in [0,1]$; if this is not the case then a more stable method of choosing δ may be desirable.

In the context of robust filtering in spatial analysis, Josselin and Ladiray (2002) propose an estimator (which they call the "meadian") that is an adaptive linear combination (indexed by a single parameter δ) of the sample mean and sample median; their choice of δ minimizes the bootstrap variance. (A similar estimator was proposed by Chan and He (1994).) This estimator of δ suffers from the same problem as above, namely the slow convergence rate of the estimator $\widehat{\delta}_n$ to its optimal value δ_0; as before, we have $\widehat{\delta}_n = \delta_0 + O_p(n^{-1/4})$.

5 Final comments

In this paper, we have shown that it is possible to derive second order properties of the bootstrap of L_1 (and related) estimators in regression by exploiting properties of the corresponding objective functions. The slow convergence rate ($O_p(n^{-1/4})$ versus $O_p(n^{-1/2})$ for "smoother" estimators) is a consequence of the lack of differentiability in the function $|x|$ at $x = 0$; replacing $|x|$ by an approximation that is smooth around $x = 0$ may improve the situation but may also create other problems.

The second order properties of the bootstrap are applicable in examining the properties of an estimator produced by "bagging" (Breiman, 1996) or by "subagging" (Bühlmann and Yu, 2002) L_1 estimators. These methods use averages of estimators from bootstrap samples (bagging) and without replacement subsamples (subagging) to create a new estimator, which for many "non-smooth" estimators have superior properties compared to the original estimator. In the case of L_1 estimation, there is some evidence that bagging and subagging improve the performance of the L_1 estimator; defining $\widetilde{\boldsymbol{\beta}}_n$ to a bagged or subagged estimator of $\boldsymbol{\beta}$, we have (analogous to (4))

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \boldsymbol{W}_n + \boldsymbol{\xi}'_n$$

where $\boldsymbol{W}_n$ is exactly the same as in (4) and $n^{1/4}\boldsymbol{\xi}'_n \xrightarrow{d} \boldsymbol{\xi}'_0$. For bagging, we have

$$\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{\xi}'_0) \leq \mathrm{Var}(\boldsymbol{a}^T\boldsymbol{\xi}_0)$$

for all $\boldsymbol{a}$ where $\boldsymbol{\xi}_0 = C_0^{-1}\boldsymbol{D}(\boldsymbol{W})/\lambda$ is the limiting distribution in (5) with similar results holding for subagged estimators as well. However, much more work needs to be done in this area.

Acknowledgment

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Arcones, M. A. (1998), Second order representations of the least absolute deviation regression estimator. Annals of the Institute of Statistical Mathematics, **50**, 87-117.

Bahadur, R. R. (1966), A note on quantiles in large samples. Annals of Mathematical Statistics, **37**, 577-580.

Bassett, G. and Koenker, R. (1978), Asymptotic theory of least absolute error regression. Journal of the American Statistical Association, **73**, 618-622.

Breiman, L. (1996), Bagging predictors. Machine Learning, **24**, 123-140.

Bühlman, P. and Yu, B. (2002), Analyzing bagging. Annals of Statistics, **30**, 927-961.

Chan, Y. M. and He, X. (1994), A simple and competitive estimator of location. Statistics and Probability Letters, **19**, 137-142.

Dodge, Y. and Jurečková, J. (1992), A class of estimators based on adaptive convex combinations of two estimation procedures. In L_1-Statistical Analysis and Related Methods, Ed Dodge, Y. pp 31-45.

Efron, B. (1979), Bootstrap methods: another look at the jackknife. Annals of Statistics, **7**, 1-26.

De Angelis, D., Hall, P. and Young, G. A. (1993), Analytical and bootstrap approximations to estimator distributions in L^1 regression. Journal of the American Statistical Association, **88**, 1310-1316.

Geyer, C. J. (1996), On the asymptotics of convex stochastic optimization. Unpublished manuscript.

Hall, P. and Martin, M. A. (1988), Exact convergence rate of bootstrap quantile variance estimator. Probability Theory and Related Fields, **80**, 261-288.

Hjørt, N. L. and Pollard, D. (1993), Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Josselin, D. and Ladiray, D. (2002), Combining L_1 and L_2 norms for a more robust spatial analysis: the "meadian attitude". Unpublished manuscript.

Kallenberg, O. (1976), Random Measures. Berlin: Akademie Verlag.

Kiefer, J. (1967), On Bahadur's representation of sample quantiles. Annals of Mathematical Statistics, **38**, 1323-1342.

Knight, K. (1997), Asymptotics for L_1 regression estimators under general conditions. Unpublished manuscript.

Knight, K. (1998), Limiting distributions for L_1 regression estimators under general conditions. Annals of Statistics, **26**, 755-770.

Knight, K. (1999), Asymptotics for L_1-estimators of regression parameters under heteroscedasticity. Canadian Journal of Statistics, **27**, 497-507.

Koenker, R. and Bassett, G. (1978), Regression quantiles. Econometrica, **46**, 33-50.

Pollard, D. (1991), Asymptotics for least absolute deviation regression estimators. Econometric Theory, **7**, 186-199.