Asymptotics for L_1 regression estimators under general conditions

by

Keith Knight
Department of Statistics
University of Toronto

Technical Report No. 9716, September 24, (1997)

TECHNICAL REPORT SERIES

University of Toronto

Department of Statistics

Asymptotics for L_1 regression estimators under general conditions

Keith Knight
Department of Statistics
University of Toronto

Abstract: It is well-known that L_1-estimators of regression parameters are asymptotically Normal if the distribution function has a positive derivative at 0. In this paper, we derive the asymptotic distributions under more general conditions on the behaviour of the distribution function near 0. Second order or weak Bahadur-Kiefer representations are also derived.

1 Introduction

Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \tag{1}$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are unknown parameters and $\{\varepsilon_i\}$ are unobservable independent, identically distributed (i.i.d.) random variables each with median 0. For simplicity, we will assume that the x_{ki}'s are non-random although the results will typically hold for random x_{ki}'s. We will consider the asymptotic behaviour of L_1-estimators of $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$; that is, $\widehat{\beta}_0, \widehat{\beta}_1, \cdots \widehat{\beta}_p$ minimize the objective function

$$g_n(\boldsymbol{\phi}) = \sum_{i=1}^{n} |Y_i - \phi_0 - \phi_1 x_{1i} - \cdots - \phi_p x_{pi}|$$

over all $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_p)$.

The asymptotic behaviour of L_1-estimators in regression is well-known, at least in the case where the errors have a distribution function $F(t)$ which is differentiable at 0 with the derivative positive. In particular, if we denote this derivative by $\lambda = F'(0)$, we have $(X_n^T X_n)^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a $(p+1)$-variate Normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(4\lambda^2)^{-1} I$ provided that

$$\max_{1 \le i \le n} \boldsymbol{x}_i^T (X_n^T X_n)^{-1} \boldsymbol{x}_i \to 0 \quad \text{as } n \to \infty$$

where $\boldsymbol{x}_i^T = (1, x_{1i}, \cdots, x_{pi})$ and X_n is the $n \times (p+1)$ matrix whose i-th row is $\boldsymbol{x}_i^T$. (Note that $\boldsymbol{x}_i^T (X_n^T X_n)^{-1} \boldsymbol{x}_i$ $(i = 1, \cdots, n)$ are simply the diagonal elements of the so-called hat matrix

$X_n(X_n^T X_n)^{-1} X_n^T$.) If $n^{-1}(X_n^T X_n) \to C$ for some positive definite matrix C then it will follow that $\sqrt{n}(\widehat{\beta} - \beta)$ converges in distribution to the $(p+1)$-variate Normal distribution whose covariance matrix is $(4\lambda^2)^{-1} C^{-1}$. (See, for example, Bassett and Koenker (1978), Bloomfield and Steiger (1983), Bai *et al* (1990) and Pollard (1991) for various approaches to proving the asymptotic Normality.) Second order results are given by Arcones (1996a, 1996b), Babu (1989) and He and Shao (1996).

A natural question to ask is what happens when the distribution function does not have a positive derivative at 0. While these cases may seem pathological, they are, in fact, far from it. Indeed, while assuming the existence of a density seems reasonable, it is an assumption which is difficult to verify. In fact, previous work suggests that the asymptotic behaviour of L_1-estimators is very sensitive to this assumption. For i.i.d. observations, Smirnov (1952) identifies four possible types of limiting distributions for sample quantiles and characterizes their domains of attraction. Jurečková (1983) considers the asymptotic behaviour of M-estimators of location under non-regular conditions; her results include the L_1 estimator of location (namely the sample median) as a special case. On another front, Arcones (1994) considers the asymptotic behaviour of so-called L_p-median (that is, minimizers of $\sum_{i=1}^n |Y_i - \theta|^p$) for $0 < p \leq 1/2$ and shows that the convergence rate is slower than $O_p(n^{-1/2})$.

To consider the asymptotic behaviour of L_1-estimators, we will start by defining (for some sequence of constants a_n),

$$\Psi_n(t) = \int_0^t \sqrt{n}(F(s/a_n) - F(0))\, ds \tag{2}$$

which for each n is a convex function. If the limit of $\{\Psi_n(t)\}$ exists for each t, define

$$\Psi(t) = \lim_{n\to\infty} \Psi_n(t). \tag{3}$$

$\Psi(t)$ (if it exists) is a convex function taking values in $[0,\infty]$; note that $\Psi(t)$ may equal ∞ although typically $\Psi(t)$ will be finite. (See Examples 3 and 4 in section 3 for cases where $\Psi(t) = \infty$ for certain t.)

The exact form of Ψ in (3) can be more easily obtained by considering the limit of $\sqrt{n}(F(t/a_n) - F(0))$; if

$$\lim_{n\to\infty} \sqrt{n}(F(t/a_n) - F(0)) = \psi(t) \tag{4}$$

then typically

$$\Psi(t) = \int_0^t \psi(s)\, ds.$$

In the case where $F(x)$ is differentiable at $x = 0$ (with $F'(0) > 0$) then $a_n = \sqrt{n}$ and $\psi(t) = \lambda t$ where $\lambda = F'(0)$ and so $\Psi(t) = \lambda t^2/2$. More generally, condition (4) includes cases where F has one-sided derivatives at 0 ($\psi(t) = \lambda^+ t$ for $t > 0$ and $\psi(t) = \lambda^- t$ for $t < 0$ where $a_n = \sqrt{n}$) or is regularly varying in a neighbourhood of 0. These conditions are very similar to those given by Smirnov (1952). These assumptions are somewhat weaker than those used in Jurečková (1983) for the location case. In particular, notice that it is not necessary to assume that F is absolutely continuous (with respect to Lebesgue measure); in fact, F can contain discrete components.

We will show that (under suitable regularity conditions on the design) $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ converges in distribution. To do this, we will first modify the objective function g_n as follows:

$$Z_n(\boldsymbol{u}) = \frac{a_n}{\sqrt{n}} \sum_{i=1}^{n} \left[|\varepsilon_i - \boldsymbol{x}_i^T \boldsymbol{u}/a_n| - |\varepsilon_i| \right]. \tag{5}$$

It is easy to see that the vector $\widehat{\boldsymbol{u}}_n$ which minimizes Z_n is simply $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. If one now regards $\{Z_n\}$ as a sequence of random convex functions on R^{p+1} and if the finite dimensional distributions of $Z_n(\boldsymbol{u})$ converge in distribution to those of some function $Z(\boldsymbol{u})$ which has a unique minimum $\boldsymbol{U}$ then it will follow that

$$\boldsymbol{U}_n = a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \boldsymbol{U} = \operatorname{argmin}(Z)$$

as $n \to \infty$ (see Hjørt and Pollard, 1993; Geyer, 1996).

2 Limiting distributions

We will now formally state the regularity conditions needed to find the limiting distribution of the L_1-estimator.

(A1) $\{\varepsilon_i\}$ are i.i.d. random variables with median 0 with distribution function F continuous at 0.

(A2) For some positive definite matrix C,

$$\lim_{n\to\infty} \frac{1}{n} X_n^T X_n = C.$$

(A3) For each $\boldsymbol{u}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi_n(\boldsymbol{u}^T \boldsymbol{x}_i) = \tau(\boldsymbol{u})$$

for some convex function $\tau(\boldsymbol{u})$ taking values in $[0,\infty]$ where $\{\Psi_n(t)\}$ is defined as in (2) (for some sequence $\{a_n\}$).

At this point, it is worth making a few comments on the regularity conditions. Condition (A2) is standard and implies, for example, that

$$\frac{1}{n} \max_{1\le i\le n} \boldsymbol{x}_i^T \boldsymbol{x}_i \to 0.$$

(A3) is similar in spirit to (A2); it is essentially another moment condition for the $\boldsymbol{x}_i$'s. If $\Psi(t)$ (defined in (3)) is finite for all t then $\tau(\boldsymbol{u})$ in (A3) can sometimes be evaluated as

$$\tau(\boldsymbol{u}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi(\boldsymbol{u}^T \boldsymbol{x}_i)$$

(assuming the convergence of Ψ_n to Ψ is sufficiently uniform). If this is the case and (4) holds with $\psi(t) = \lambda t$ then $\Psi(t) = \lambda t^2/2$ and so (A3) is implied by (A2) with

$$\tau(\boldsymbol{u}) = \frac{\lambda}{2}\boldsymbol{u}^T C \boldsymbol{u}.$$

Note that the conditions (A2) and (A3) rule out certain designs for which the moment conditions are not appropriate; for example, consider $\boldsymbol{x}_i^T = (1, i)$. In such cases, it may be possible to reformulate conditions (A2) and (A3) so that a non-degenerate limiting distribution exists.

THEOREM 1. Assume the model (1) for $Y_1, Y_2, \cdots$ and define $Z_n(\cdot)$ as in (5). If assumptions (A1), (A2) and (A3) hold then for any $(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k)$,

$$(Z_n(\boldsymbol{u}_1), \cdots, Z_n(\boldsymbol{u}_k)) \to_d (Z(\boldsymbol{u}_1), \cdots, Z(\boldsymbol{u}_k))$$

where

$$Z(\boldsymbol{u}) = -\boldsymbol{u}^T \boldsymbol{W} + 2\tau(\boldsymbol{u})$$

with $\boldsymbol{W}$ a $(p+1)$-variate Normal random vector with mean vector $\boldsymbol{0}$ and covariance matrix C.

(The $-$ in front of $\boldsymbol{u}^T\boldsymbol{W}$ is obviously unnecessary but will be useful in the sequel.)

COROLLARY 2. Under the hypotheses of Theorem 1, if $Z(\boldsymbol{u})$ has unique minimum (with probability 1) then

$$a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(Z).$$

Proof of Theorem 1. We will use the identity

$$|x-y| - |x| = -y[I(x>0) - I(x<0)] + 2\int_0^y [I(x \le s) - I(x \le 0)]\, ds$$

which is valid for $x \neq 0$. ($I(A)$ is the indicator function of the set A.) Now

$$Z_n(\boldsymbol{u}) = Z_n^{(1)}(\boldsymbol{u}) + Z_n^{(2)}(\boldsymbol{u})$$

where

$$\begin{aligned} Z_n^{(1)}(\boldsymbol{u}) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{u}[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)] \\ \text{and} \quad Z_n^{(2)}(\boldsymbol{u}) &= \frac{2a_n}{\sqrt{n}}\sum_{i=1}^n \int_0^{v_{ni}} [I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]\, ds \\ &= \sum_{i=1}^n Z_{ni}^{(2)}(\boldsymbol{u}) \end{aligned}$$

(with $v_{ni} = \boldsymbol{x}_i^T\boldsymbol{u}/a_n$). By the Lindeberg-Feller Central Limit Theorem, for each $\boldsymbol{u}$,

$$Z_n^{(1)}(\boldsymbol{u}) \to_d -\boldsymbol{u}^T\boldsymbol{W}$$

(using the fact that $n^{-1}(X_n^T X_n)$ converges to C) and the convergence in distribution holds for any finite collection of $\boldsymbol{u}$'s. For $Z_n^{(2)}(\boldsymbol{u})$, we have

$$Z_n^{(2)}(\boldsymbol{u}) = \sum_{i=1}^{n} E(Z_{ni}^{(2)}(\boldsymbol{u})) + \sum_{i=1}^{n}(Z_{ni}^{(2)}(\boldsymbol{u}) - E(Z_{ni}^{(2)}(\boldsymbol{u}))).$$

Letting $v_i = \boldsymbol{x}_i^T \boldsymbol{u} = a_n v_{ni}$, it follows that

$$\begin{aligned}
\sum_{i=1}^{n} E(Z_{ni}^{(2)}(\boldsymbol{u})) &= \frac{2a_n}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{v_{ni}} (F(s) - F(0))\, ds \\
&= \frac{2}{n} \sum_{i=1}^{n} \int_0^{v_i} \sqrt{n}(F(s/a_n) - F(0))\, ds \\
&= \frac{2}{n} \sum_{i=1}^{n} \Psi_n(\boldsymbol{u}^T \boldsymbol{x}_i) \\
&\to 2\tau(\boldsymbol{u}).
\end{aligned}$$

For the remainder term in $Z_n^{(2)}(\boldsymbol{u})$, we have

$$\begin{aligned}
\mathrm{Var}(Z_n^{(2)}(\boldsymbol{u})) &= \sum_{i=1}^{n} E[(Z_{ni}^{(2)}(\boldsymbol{u}) - E(Z_{ni}^{(2)}(\boldsymbol{u})))^2] \\
&\leq \frac{2}{\sqrt{n}} \max_{1 \leq i \leq n} |\boldsymbol{x}_i^T \boldsymbol{u}| \sum_{i=1}^{n} E(Z_{ni}^{(2)}(\boldsymbol{u})) \\
&= \frac{2}{\sqrt{n}} \max_{1 \leq i \leq n} |\boldsymbol{x}_i^T \boldsymbol{u}| E(Z_n^{(2)}(\boldsymbol{u})).
\end{aligned}$$

Thus if $\tau(\boldsymbol{u}) < \infty$,

$$Z_n^{(2)}(\boldsymbol{u}) - E(Z_n^{(2)}(\boldsymbol{u})) \to_p 0 \quad \text{as } n \to \infty$$

and so $Z_n^{(2)}(\boldsymbol{u}) \to_p 2\tau(\boldsymbol{u})$. If $\tau(\boldsymbol{u}) = \infty$ then

$$\begin{aligned}
P\left(\left|Z_n^{(2)}(\boldsymbol{u}) - E(Z_n^{(2)}(\boldsymbol{u}))\right| > \epsilon E(Z_n^{(2)}(\boldsymbol{u}))\right) &\leq \frac{\mathrm{Var}(Z_n^{(2)}(\boldsymbol{u}))}{\epsilon^2 E(Z_n^{(2)}(\boldsymbol{u}))^2} \\
&\leq 2\frac{\max_{1 \leq i \leq n} |\boldsymbol{x}_i^T \boldsymbol{u}|/\sqrt{n}}{\epsilon^2 E(Z_n^{(2)}(\boldsymbol{u}))} \\
&\to 0
\end{aligned}$$

which implies that $Z_n^{(2)}(\boldsymbol{u}) \to_p \infty = \tau(\boldsymbol{u})$. Thus we have

$$Z_n(\boldsymbol{u}) \to_d -\boldsymbol{u}^T \boldsymbol{W} + 2\tau(\boldsymbol{u}) = Z(\boldsymbol{u})$$

and the finite dimensional convergence holds trivially. □

Proof of Corollary 2. If $Z(\boldsymbol{u})$ has a unique minimum then by the convexity of the Z_n's, it follows that

$$\mathrm{argmin}(Z_n) = a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{argmin}(Z)$$

as $n \to \infty$. (See, for example, Geyer (1996).) □

In general, argmin(Z) will have a multivariate Normal distribution if, and only if, τ (and hence Z) is a quadratic function. Moreover, τ will be quadratic if, and only if, the function ψ defined in (4) is linear. When τ is differentiable (with gradient $\nabla\tau$) then

$$a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \boldsymbol{U}$$

where $\boldsymbol{U}$ satisfies the equation

$$2\nabla\tau(\boldsymbol{U}) = \boldsymbol{W}.$$

Under appropriate regularity conditions, we can evaluate $\nabla\tau$ as

$$\nabla\tau(\boldsymbol{u}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \psi(\boldsymbol{u}^T \boldsymbol{x}_i)$$

where ψ is defined in (4).

Cases where $\tau(\boldsymbol{u})$ is infinite for $\boldsymbol{u}$ outside some compact set K can occur in many ways. For example, suppose that F is absolutely continuous with density f where

$$\lim_{x\to 0-} f(x) = \infty \quad \text{and} \quad \lim_{x\to 0+} f(x) = \lambda > 0.$$

In this case, we have $a_n = \sqrt{n}$ and

$$\Psi(t) = \begin{cases} \infty & \text{if } t < 0 \\ \lambda t^2/2 & \text{if } t \geq 0 \end{cases}$$

Therefore, $\tau(\boldsymbol{u})$ will take infinite values. This case is considered further in Example 3 below; see also Example 4.

The results given in this section can be extended in numerous directions. For example, we can obtain similar results for so-called "regression quantiles" (Koenker and Bassett, 1978) by replacing $|x|$ by the function $\rho_q(x) = |x| - (2q-1)x$ for some $0 < q < 1$. Similarly, we could consider regression M-estimators with discontinuous "ψ" functions similar to Jurečková (1983).

3 Examples

The limiting distributions for $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ are, in general, quite complicated (but not impossible) to determine in closed-form.

In Examples 1 and 2 below, we will assume that the ε_i's have a distribution function satisfying

$$F(x) - F(0) = \lambda \text{sgn}(x)|x|^{\alpha} L(|x|) \tag{6}$$

for x in a neighbourhood of 0 where $\alpha > 0$ and L is a slowly varying function at 0. ($\text{sgn}(x) = 1$ if x is positive and -1 if x is negative.) In this case, we can take

$$a_n = n^{1/(2\alpha)} L^*(n) \quad \text{and} \quad \psi(t) = \lambda \text{sgn}(t)|t|^{\alpha}$$

where L^* is a slowly varying function at infinity. For example, if $F(x) - F(0) = \lambda x \ln(|x|^{-1})$ (for x close to 0) then we can take $a_n = \sqrt{n}\ln(n)/2$ and $\psi(t) = \lambda t$. Moreover, for distributions satisfying (6), we have

$$\Psi(t) = \frac{\lambda}{\alpha+1}|t|^{\alpha+1}.$$

Thus for larger α, condition (A3) is a more stringent "moment" condition on the $\boldsymbol{x}_i$'s.

EXAMPLE 1. Suppose that $Y_1, Y_2, \cdots$ are i.i.d. random variables with $Y_i = \mu + \varepsilon_i$ where μ is the median of the distribution of the Y_i's. The sample median $\widehat{\mu}_n$ minimizes

$$g_n(\theta) = \sum_{i=1}^{n} |Y_i - \theta|.$$

If we assume that the distribution function of the ε_i's satisfies (6) for some $\alpha > 0$, then the limit of $Z_n(u)$ as defined in (5) is

$$Z(u) = -uW + \frac{2\lambda}{\alpha+1}|u|^{\alpha+1}$$

where W is Normal with mean 0 and variance 1. $Z(u)$ is minimized at U which satisfies

$$2\lambda|U|^{\alpha}\mathrm{sgn}(U) = W$$

and so $U = \mathrm{sgn}(W)|W/(2\lambda)|^{1/\alpha}$; the density of U (and hence the limiting density of $a_n(\widehat{\mu}_n - \mu)$) is

$$f(x) = \frac{\lambda\alpha\sqrt{2}}{\sqrt{\pi}}|x|^{\alpha-1}\exp\left(-2\lambda^2|x|^{2\alpha}\right).$$

(See also Smirnov (1952) and Jurečková (1983, Corollary 1).) Note that the density has a singularity at 0 if $\alpha < 1$ and that the density is bimodal if $\alpha > 1$. ◇

EXAMPLE 2. Suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i's are i.i.d. random variables whose distribution satisfies (6) for some $\alpha > 0$. Suppose also that half of the x_i's are 1 and the other half are -1. Then

$$\frac{1}{n}\sum_{i=1}^{n}|u_0 + u_1 x_i|^{\alpha+1} \to \frac{1}{2}\left(|u_0 - u_1|^{\alpha+1} + |u_0 + u_1|^{\alpha+1}\right).$$

The limit of $Z_n(u_0, u_1)$ is

$$Z(u_0, u_1) = -(u_0 W_0 + u_1 W_1) + \frac{\lambda}{\alpha+1}\left(|u_0 - u_1|^{\alpha+1} + |u_0 + u_1|^{\alpha+1}\right)$$

where W_0 and W_1 are independent Normal random variables each with mean 0 and variance 1. If U_0 and U_1 minimize $Z(u_0, u_1)$ then they satisfy the equations

$$\begin{aligned}
\lambda\left[|U_0 + U_1|^{\alpha}\mathrm{sgn}(U_0 + U_1) + |U_0 - U_1|^{\alpha}\mathrm{sgn}(U_0 - U_1)\right] &= W_0 \\
\lambda\left[|U_0 + U_1|^{\alpha}\mathrm{sgn}(U_0 + U_1) - |U_0 - U_1|^{\alpha}\mathrm{sgn}(U_0 - U_1)\right] &= W_1.
\end{aligned}$$

The joint density of (U_0, U_1) is then

$$f(u_0, u_1) = \frac{2\lambda^2\alpha^2}{\pi}|u_0^2 - u_1^2|^{\alpha-1}\exp\left[-\lambda^2\left(|u_0 + u_1|^{2\alpha} + |u_0 - u_1|^{2\alpha}\right)\right]$$

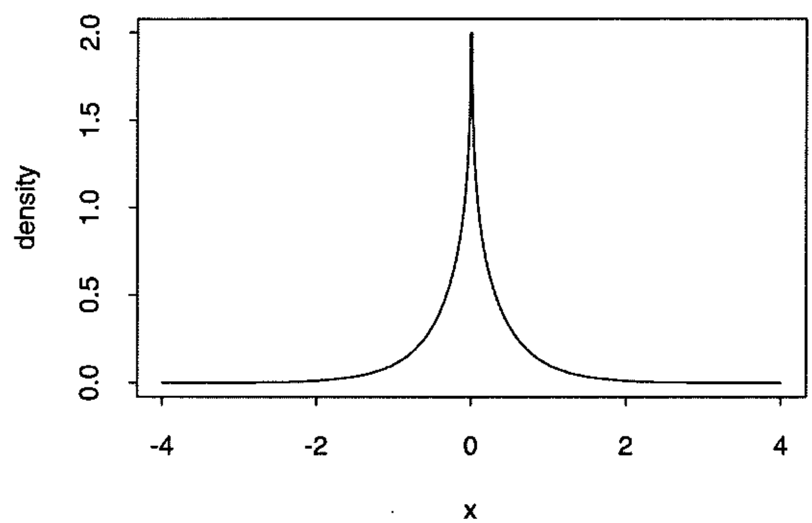

Figure 1: Asymptotic marginal density for $\alpha = 0.5$

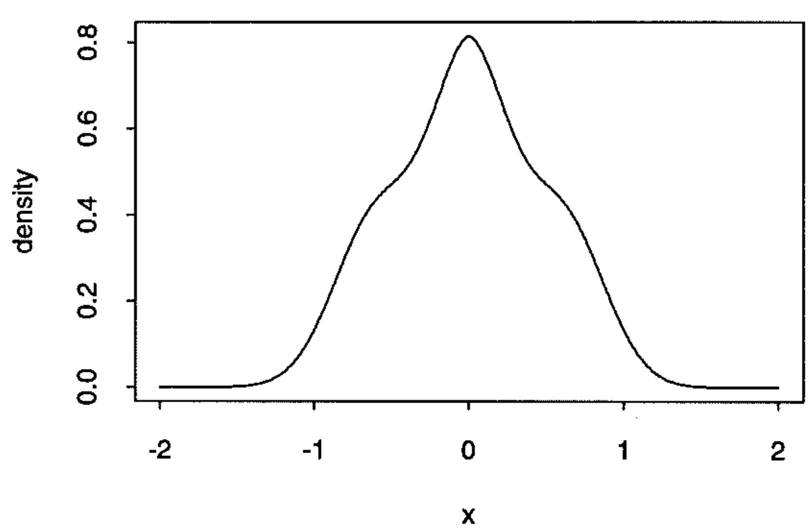

Figure 2: Asymptotic marginal density for $\alpha = 1.5$

Thus $a_n(\widehat{\beta}_{n0} - \beta_0)$ and $a_n(\widehat{\beta}_{n1} - \beta_1)$ are asymptotically independent if, and only if, $\alpha = 1$. The asymptotic marginal densities are, in this example, identical; if $\alpha \leq 1/2$, the marginal densities will have a singularity at 0. Figures 1 and 2 give these limiting densities for $\alpha = 1/2$ and $\alpha = 3/2$ (with $\lambda = 1$). (Note that the limiting joint distribution could also be deduced from the limiting distribution in Example 1.) It is also possible to see that as $\alpha \to \infty$, the asymptotic distribution concentrates around the points $(0,1)$, $(0,-1)$, $(1,0)$ and $(-1,0)$ with equal probability. ◇

EXAMPLE 3. Suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the x_i's are uniformly distributed over the interval $[-1,1]$. In this case, we have

$$\frac{1}{n} X_n^T X_n \to C = \begin{pmatrix} 1 & 0 \\ 0 & 1/3 \end{pmatrix}$$

We will assume that the distribution function F of the ε_i's satisfies

$$\lim_{n\to\infty} \sqrt{n}(F(t/\sqrt{n}) - F(0)) = \begin{cases} -\infty & \text{if } t < 0 \\ \lambda t & \text{if } t \geq 0 \end{cases}$$

and so

$$\Psi(t) = \begin{cases} \infty & \text{if } t < 0 \\ \lambda t^2/2 & \text{if } t \geq 0 \end{cases}$$

Thus, given that the x_i's are contained in $[-1,1]$, we have

$$\tau(u_0, u_1) = \begin{cases} \lambda(u_0^2/2 + u_1^2/6) & \text{if } u_0 \geq |u_1| \\ \infty & \text{otherwise} \end{cases}$$

Then letting W_0 and W_1 be independent 0 mean Normal random variables with variances 1 and 1/3 respectively, it follows that

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_{0n} - \beta_0 \\ \widehat{\beta}_{1n} - \beta_1 \end{pmatrix} \to_d \begin{pmatrix} U_0 \\ U_1 \end{pmatrix}$$

where

$$U_0 = \begin{cases} W_0/(2\lambda) & \text{if } W_0 \geq 3|W_1| \\ 3(W_0 + W_1)/(8\lambda) & \text{if } 0 < W_0 + W_1 < 4W_1 \\ 3(W_0 - W_1)/(8\lambda) & \text{if } 0 < W_0 - W_1 < -4W_1 \\ 0 & \text{if } W_0 \leq -|W_1| \end{cases}$$

and

$$U_1 = \begin{cases} 3W_1/(2\lambda) & \text{if } W_0 \geq 3|W_1| \\ 3(W_0 + W_1)/(8\lambda) & \text{if } 0 < W_0 + W_1 < 4W_1 \\ 3(W_1 - W_0)/(8\lambda) & \text{if } 0 < W_0 - W_1 < -4W_1 \\ 0 & \text{if } W_0 \leq -|W_1| \end{cases}$$

Note that much of the limit distribution is concentrated on the set $B = \{(u_0, u_1) : u_0 = |u_1|\}$; a straightforward computation gives

$$P\left[(U_0, U_1) \in B\right] = P\left[W_0 < 3|W_1|\right] = \frac{1}{\pi\sqrt{3}} \int_0^\infty \int_{-\infty}^x \exp\left(-\frac{1}{6}(3t^2 + x^2)\right) dt\, dx = \frac{5}{6}.$$

Likewise, $P(U_0 = U_1 = 0) = P(W_0 \leq -|W_1|) = 1/6$. $\diamond$

EXAMPLE 4. As in Example 3, suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the x_i's are uniformly distributed on the interval $[-1, 1]$. Suppose that the density of the ε_i's is

$$f(x) = \frac{1}{2x^2} \exp\left(-|x|^{-1}\right)$$

so that its distribution function is

$$F(x) = \frac{1}{2}\left(1 + \operatorname{sgn}(x) \exp\left(-|x|^{-1}\right)\right). \tag{7}$$

It is now easy to see that

$$\sqrt{n}(F(t/\ln(n)) - F(0)) \to \psi(t) = \begin{cases} 0 & \text{for } |t| < 2 \\ \operatorname{sgn}(t)1/2 & \text{for } |t| = 2 \\ \operatorname{sgn}(t)\infty & \text{for } |t| > 2 \end{cases}$$

so that

$$\Psi_n(t) \to \Psi(t) = \begin{cases} 0 & \text{for } |t| \leq 2 \\ \infty & \text{for } t > 2 \end{cases}.$$

Defining Z_n as in (5) with $a_n = \ln(n)$, it is easy to determine that

$$Z_n(u_0, u_1) \to_d Z(u_0, u_1) = u_0 W_0 + u_1 W_1 + \tau(u_0, u_1)$$

where W_0, W_1 are independent Normal random variables with mean 0 (variances 1 and 1/3 respectively) and

$$\tau(u_0, u_1) = \begin{cases} 0 & \text{if } -2 \leq u_0 + u_1 \leq 2 \text{ and } -2 \leq u_0 - u_1 \leq 2 \\ \infty & \text{otherwise} \end{cases}.$$

(We have

$$\frac{1}{n}\sum_{i=1}^n \Psi_n(\boldsymbol{u}^T \boldsymbol{x}_i) \to \tau(\boldsymbol{u})$$

since the $\boldsymbol{x}_i$'s are contained in a compact set.) Thus determining the limiting distribution of $\ln(n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ depends on minimizing $u_0 W_0 + u_1 W_1$ over the region

$$A = \{(u_0, u_1) : -2 \leq u_0 + u_1 \leq 2 \text{ and } -2 \leq u_0 - u_1 \leq 2\}.$$

Since $P(W_0 = W_1) = P(W_0 = -W_1) = 0$, $Z(u_0, u_1)$ has an almost sure unique minimum; the minimizer will be one of the 4 corners of the region A with probability 1/4 for each corner. (Likewise,

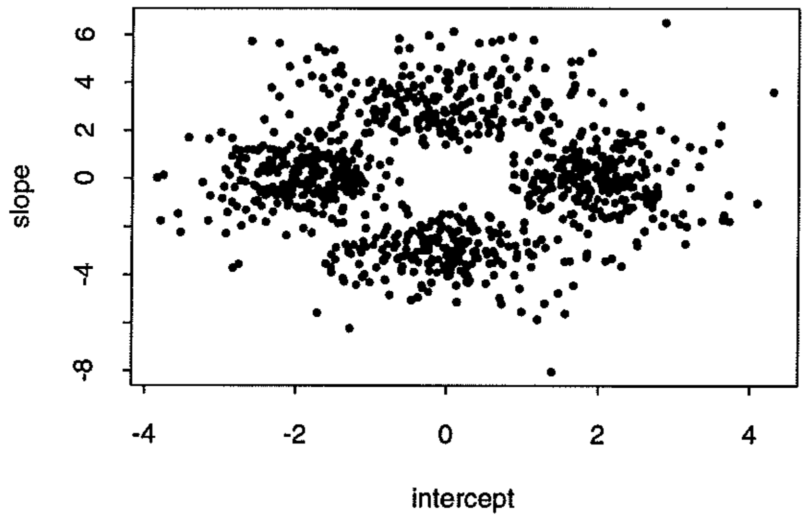

Figure 3: Plot of $\ln(n)(\widehat{\beta}_{n1} - \beta_1)$ versus $\ln(n)(\widehat{\beta}_{n0} - \beta_0)$ for $n = 100$

if $Y_i = \mu + \varepsilon_i$ where the ε_i's have the distribution function (7) then it is easy to show that for the sample median $\widehat{\mu}_n$ of $Y_1, \cdots, Y_n$, we have

$$\ln(n)(\widehat{\mu}_n - \mu) \to_d U$$

where $P(U = 2) = P(U = -2) = 1/2$; this is one of the four types of limiting distribution for the sample median given by Smirnov (1952).)

The convergence to the limiting distribution is very slow (as might be expected). Figures 3 and 4 show plots of $\ln(n)(\widehat{\beta}_{n1} - \beta_1)$ versus $\ln(n)(\widehat{\beta}_{n0} - \beta_0)$ for $n = 100$ and $n = 100000$ based on 1000 simulations; note the tendency for the points to concentrate around the corners of A. However, even for the extremely large sample size, there is little evidence that the limiting distribution would be a good approximation to the true distribution.

More generally, when $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ for i.i.d. ε_i's with distribution function (7) then

$$Z_n(\boldsymbol{u}) \to_d Z(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{W} + \tau(\boldsymbol{u})$$

where $\boldsymbol{W}$ is $(p+1)$-variate Normal with mean $\mathbf{0}$ and covariance matrix

$$C = \lim_{n\to\infty} \frac{1}{n} X_n^T X_n$$

and

$$\tau(\boldsymbol{u}) = \begin{cases} 0 & \text{if } \boldsymbol{u} \in A \\ \infty & \text{otherwise} \end{cases}$$

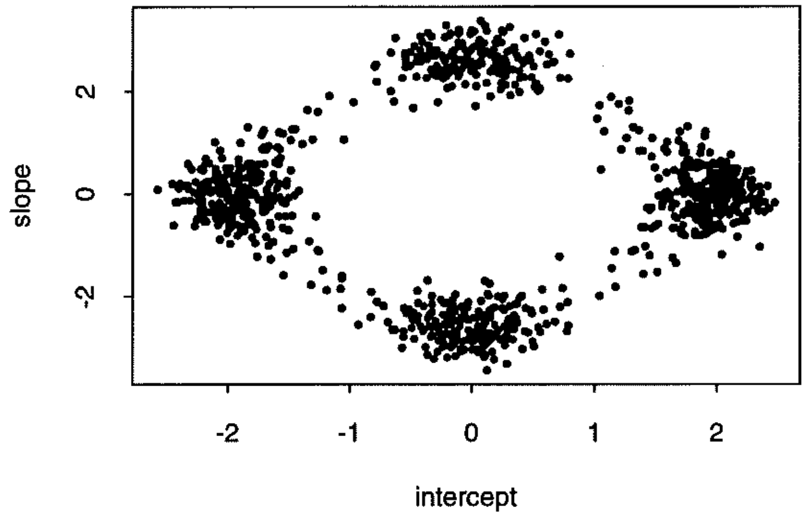

Figure 4: Plot of $\ln(n)(\widehat{\beta}_{n1} - \beta_1)$ versus $\ln(n)(\widehat{\beta}_{n0} - \beta_0)$ for $n = 100000$

with

$$A = \left\{ \boldsymbol{u} : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(|\boldsymbol{u}^T \boldsymbol{x}_i| \leq 2) = 1 \right\}.$$

Note that A is always non-empty in that it will always contain $\mathbf{0}$. Moreover, in order to obtain a non-degenerate limit distribution, the $\boldsymbol{x}_i$'s need to be "essentially" bounded in the sense that a negligible fraction lies outside a bounded set. $\diamond$

4 Second-order results

The proof of Theorem 1 implies that we can approximate the function Z_n by

$$Z_n^*(\boldsymbol{u}) = -\boldsymbol{u}^T \boldsymbol{W}_n + 2\tau(\boldsymbol{u})$$

where

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i [I(\varepsilon_i > 0) - I(\varepsilon_i < 0)].$$

The convexity of Z_n and Z_n^* implies that we can approximate $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \text{argmin}(Z_n)$ by $\text{argmin}(Z_n^*)$.

When τ is differentiable with gradient $\nabla\tau$, $\text{argmin}(Z_n^*) = (\nabla\tau)^{-1}(\boldsymbol{W}_n/2)$ where $(\nabla\tau)^{-1}$ is the inverse of $\nabla\tau$. In this case, it is possible to study asymptotic behaviour of the approximation error $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) - (\nabla\tau)^{-1}(\boldsymbol{W}_n/2)$ by examining the asymptotic behaviour of $Z_n(\boldsymbol{u}) - Z_n^*(\boldsymbol{u})$. We will show below that (under appropriate regularity conditions)

$$n^{1/4}(Z_n - Z_n^*) \to_d V$$

where the limiting process V has a gradient process $\boldsymbol{D}$ which is a R^{p+1}-valued Gaussian process defined on R^{p+1}. Using Lemma A (stated in the appendix), it follows that

$$n^{1/4}\left(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) - (\nabla\tau)^{-1}(\boldsymbol{W}_n/2)\right) \to_d -H^{-1}(\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U})$$

where H is the Hessian of τ and $\boldsymbol{U} = (\nabla\tau)^{-1}(\boldsymbol{W}/2)$. Under the classical assumption that $F'(0) = \lambda > 0$, such "weak Bahadur-Kiefer representations" (Bahadur, 1966; Kiefer, 1967) are given by Arcones (1996a, 1996b).

What are the regularity conditions needed to make the previous discussion rigorous? We will assume that F satisfies (4) for some strictly increasing continuous function ψ; more precisely, let

$$\sqrt{n}(F(t/a_n) - F(0)) = \psi(t) + r_n(t). \tag{8}$$

We will also need the following regularity conditions.

(A4) For each compact set K,

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

for r_n defined in (8).

(A5) For each compact set K,

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

for ψ defined in (8).

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

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

In Theorem 3 below, we will talk about convergence in distribution of sequences of real- and R^{p+1}-valued locally bounded functions; we will use the topology of uniform convergence on compact sets. The limiting functions are always continuous and hence belong to a separable subset of the space of locally bounded functions; thus we can employ almost sure representations (van der Vaart and Wellner, 1996) whenever necessary. It should also be noted that the finite dimensional convergence of Z_n to Z in Theorem 2 implies $Z_n \to_d Z$ when Z is continuous; this follows from the convexity of the Z_n's (Geyer, 1996).

THEOREM 3. Assume that Theorem 1 holds where Z is continuous with unique minimizer $\boldsymbol{U}$ and, in addition, assume conditions (A4), (A5) and (A6). Suppose that, with probability 1,

$$\|\nabla\tau(\boldsymbol{u}) - \nabla\tau(\boldsymbol{v})\| \leq k\|\boldsymbol{u}-\boldsymbol{v}\|^{\alpha}$$

for all $\boldsymbol{u}$, $\boldsymbol{v}$ in a neighbourhood of $\boldsymbol{U}$ where $\alpha, k > 0$ may depend on $\boldsymbol{U}$. Then

$$n^{1/4}\left(\nabla\tau(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})) - \frac{\boldsymbol{W}_n}{2}\right) \to_d -\boldsymbol{D}(\boldsymbol{U})$$

where $\boldsymbol{D}$ is a zero-mean Gaussian process (independent of Z and $\boldsymbol{U}$) with $\boldsymbol{D}(\boldsymbol{0}) = \boldsymbol{0}$ and $E[(\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v}))(\boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{D}(\boldsymbol{v}))^T] = B(\boldsymbol{u}, \boldsymbol{v})$.

Moreover, if τ is twice continuously differentiable with invertible Hessian H (except on a set A with $P(\boldsymbol{U} \in A) = 0$) then

$$n^{1/4}\left(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) - (\nabla\tau)^{-1}(\boldsymbol{W}_n/2)\right) \to_d -H^{-1}(\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U}).$$

Proof. First of all, define $V_n(\boldsymbol{u}) = n^{1/4}(Z_n(\boldsymbol{u}) - Z_n^*(\boldsymbol{u}))$. It is easy to see that

$$V_n(\boldsymbol{u} + t\boldsymbol{w}) - V_n(\boldsymbol{u}) = 2\int_0^t \boldsymbol{w}^T \boldsymbol{D}_n(\boldsymbol{u} + s\boldsymbol{w})\, ds$$

where

$$\begin{aligned} \boldsymbol{D}_n(\boldsymbol{u}) &= n^{-3/4}\sum_{i=1}^n \left[\boldsymbol{x}_i\sqrt{n}(I(\varepsilon_i \leq \boldsymbol{u}^T\boldsymbol{x}_i/a_n) - I(\varepsilon_i \leq 0)) - \nabla\tau(\boldsymbol{u})\right] \\ &= n^{-3/4}\sum_{i=1}^n \boldsymbol{x}_i\left[\sqrt{n}(I(\varepsilon_i \leq \boldsymbol{u}^T\boldsymbol{x}_i/a_n) - I(\varepsilon_i \leq 0)) - \psi(\boldsymbol{u}^T\boldsymbol{x}_i)\right] + o_p(1) \\ &= n^{-1/4}\sum_{i=1}^n \boldsymbol{x}_i\left[(I(\varepsilon_i \leq \boldsymbol{u}^T\boldsymbol{x}_i/a_n) - I(\varepsilon_i \leq 0)) - (F(\boldsymbol{u}^T\boldsymbol{x}_i/a_n) - F(0))\right] + o_p(1) \end{aligned}$$

where the $o_p(1)$ remainder terms are uniform over compact sets. Applying, for example, the Lindeberg Central Limit Theorem, it follows now that the finite dimensional distributions of $\boldsymbol{D}_n$ converge weakly to those of $\boldsymbol{D}$. Applying Theorem 2.11.11 of van der Vaart and Wellner (1996), it follows that $\boldsymbol{D}_n \to_d \boldsymbol{D}$ on the space of locally bounded functions. It is also easy to verify that $\boldsymbol{D}$ is independent of Z. Choosing almost surely convergent sequences (which must exist since $\boldsymbol{D}$ and Z are continuous) and applying Lemma A, the conclusion follows. □

Note that the proof of Theorem 3 actually implies the somewhat stronger result

$$\left(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), n^{1/4}\left(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) - (\nabla\tau)^{-1}(\boldsymbol{W}_n/2)\right)\right) \to_d \left(\boldsymbol{U}, -H^{-1}(\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U})\right).$$

It is also straightforward to find the density of $-H^{-1}(\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U})$. Let $B(\boldsymbol{u}) = B(\boldsymbol{u}, \boldsymbol{0})$; then the density of $\boldsymbol{D}(\boldsymbol{u})$ is

$$f(\boldsymbol{x}; \boldsymbol{u}) = \frac{1}{(2\pi)^{(p+1)/2}|B(\boldsymbol{u})|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T B^{-1}(\boldsymbol{u})\boldsymbol{x}\right)$$

provided, of course, $B^{-1}(\boldsymbol{u})$ exists. If $B^{-1}(\boldsymbol{u})$ exists for almost all $\boldsymbol{u}$, the density of $-H^{-1}(\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U})$ is

$$g(\boldsymbol{x}) = \int \frac{2^{(p+1)/2}|H(\boldsymbol{u})|^2}{\pi^{(p+1)/2}|C|^{1/2}} \exp\left(-2(\nabla\tau(\boldsymbol{u}))^T C^{-1}(\nabla\tau(\boldsymbol{u}))\right) f(H(\boldsymbol{u})\boldsymbol{x};\boldsymbol{u})\, d\boldsymbol{u}$$

where the integration is over R^{p+1}.

EXAMPLE 5. Consider the model described in Example 2 where the ε_i's have distribution F satisfying (6) in a neighbourhood of 0 for $\alpha = 1$ and some $\lambda > 0$. Using the notation developed above, we have $C = I$, $H(u_0, u_1) = \lambda I/2$,

$$\tau(u_0, u_1) = \frac{\lambda}{4}\left(|u_0 - u_1|^2 + |u_0 + u_1|^2\right) = \frac{\lambda}{2}(u_0^2 + u_1^2)$$

and

$$B(u_0, u_1) = \frac{\lambda}{2}\begin{pmatrix} |u_0 + u_1| + |u_0 - u_1| & |u_0 + u_1| - |u_0 - u_1| \\ |u_0 + u_1| - |u_0 - u_1| & |u_0 + u_1| + |u_0 - u_1| \end{pmatrix}.$$

Then

$$f(x_0, x_1; u_0, u_1) = \frac{1}{2\pi\lambda|u_0^2 - u_1^2|^{1/2}} \exp\left[-\left(\frac{(x_0^2 + x_1^2)(|u_0 + u_1| + |u_0 - u_1|) - 2x_0x_1(|u_0 + u_1| - |u_0 - u_1|)}{4\lambda|u_0^2 - u_1^2|}\right)\right]$$

and

$$g(x_0, x_1) = \frac{\lambda^4}{8\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-2(u_0^2 + u_1^2)\right] f(x_0, x_1; u_0, u_1)\, du_0\, du_1.$$

By symmetry, the marginal densities of $2\boldsymbol{D}(U_0, U_1)/\lambda$ are the same; the characteristic function for this marginal density is

$$\phi(t) = 4\exp\left(\frac{t^4}{2\lambda^4}\right)\left[1 - \Phi\left(\frac{t^2}{\sqrt{2}\lambda^2}\right)\right]^2$$

where Φ is the standard Normal distribution function. This marginal density is given in Figure 5 for $\lambda = 1$. (The marginal density is perhaps most easily obtained by integrating the characteristic function.) Note that, for this problem, the components of $2\boldsymbol{D}(U_0, U_1)/\lambda$ are not independent even though U_0 and U_1 are. ◇

5 Almost sure representations

Lemma A can also be used to obtain "strong" Bahadur-Kiefer representations. We will just sketch the argument here since more complete developments are given by He and Shao (1996) as well as Arcones (1996b) for the case where $F'(0) = \lambda > 0$. All of the claims made below hold subject to appropriate regularity conditions; these conditions are somewhat stronger than those given in section 4.

Let $m_n = n/\ln\ln(n)$ and define b_n so that

$$\sqrt{m_n}(F(t/b_n) - F(0)) \to \psi(t)$$

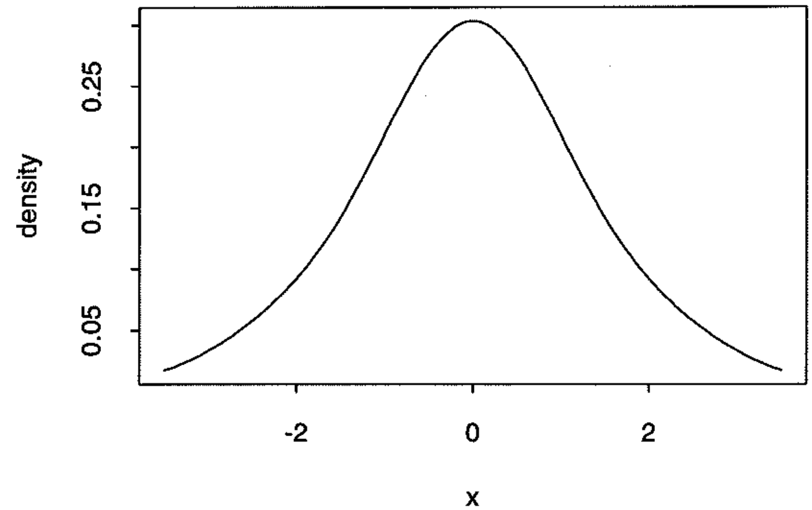

Figure 5: Asymptotic marginal density ($\lambda = 1$) in Example 5

for a strictly increasing function ψ. (Note that we could take $b_n = a_{[m_n]}$ where $[\cdot]$ is the greatest integer function.) Now redefine Z_n to be

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

Then under appropriate conditions, we have

$$Z_n(\boldsymbol{u}) = -\boldsymbol{u}^T \boldsymbol{W}_n + 2\tau(\boldsymbol{u}) + o(1) \quad \text{with probability 1}$$

where $\boldsymbol{W}_n$ is now defined to be

$$\boldsymbol{W}_n = \frac{1}{(n \ln\ln(n))^{1/2}} \sum_{i=1}^{n} \boldsymbol{u}^T \boldsymbol{x}_i [I(\varepsilon_i > 0) - I(\varepsilon_i < 0)] \tag{9}$$

where the $o(1)$ term is uniform over compact sets. By a law of the iterated logarithm, the set of (almost sure) limit points of $\{\boldsymbol{W}_n\}$ is

$$K_W = \{\boldsymbol{w} : \boldsymbol{w}^T C^{-1} \boldsymbol{w} \le 2\}$$

and so the set of limit points of $\{b_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\}$ is

$$K_U = \{(\nabla \tau)^{-1}(\boldsymbol{w}/2) : \boldsymbol{w} \in K_W\}.$$

Next redefine $Z_n^*(\boldsymbol{u}) = -\boldsymbol{u}^T \boldsymbol{W}_n + 2\tau(\boldsymbol{u})$ (for $\boldsymbol{W}_n$ in (9)) and let $V_n(\boldsymbol{u}) = m_n^{1/4}(Z_n(\boldsymbol{u}) - Z_n^*(\boldsymbol{u}))$. We then have

$$V_n(\boldsymbol{u} + t\boldsymbol{w}) - V_n(\boldsymbol{u}) = 2 \int_0^t \boldsymbol{w}^T \boldsymbol{D}_n(\boldsymbol{u} + s\boldsymbol{w})\, ds$$

where (with probability 1)

$$D_n(u) = \frac{1}{(n \ln\ln(n))^{1/4}} \sum_{i=1}^{n} x_i \left[(I(\varepsilon_i \le u^T x_i / b_n) - I(\varepsilon_i \le 0)) - (F(u^T x_i / b_n) - F(0)) \right] + o(1)$$

and the $o(1)$ remainder term is uniform over compact sets. If $\{(W_n, D_n)\}$ has limit set $\mathcal{K}$ then by Lemma A, the limit set of $m_n^{1/4}(\nabla\tau(b_n(\widehat{\beta}_n - \beta)) - W_n/2)$ is

$$K_R = \left\{ -d((\nabla\tau)^{-1}(w/2)) : (w, d) \in \mathcal{K} \right\}.$$

Acknowledgements. I would like to thank Miguel Arcones for sending me preprints of his work. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Arcones, M.A. (1994) Distributional convergence of M-estimators under unusual rates. *Statistics and Probability Letters.* **21**, 271-280.

[2] Arcones, M.A. (1996a) The Bahadur-Kiefer representation of L_p regression estimators. *Econometric Theory.* **12**, 257-283.

[3] Arcones, M.A. (1996b) Second order representations of the least absolute deviation regression estimator. (unpublished manuscript)

[4] Babu, G.J. (1989) Strong representation for LAD estimators in linear models. *Probability Theory and Related Fields.* **83**, 547-558.

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

[6] Bai, Z.D., Chen, X.R., Wu, Y. and Zhao, L.C. (1990) Asymptotic normality of minimum L_1 norm estimates in linear models. *Chinese Sciences A.* **33**, 449-463.

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

[8] Bloomfield, P. and Steiger, W.L. (1983) *Least Absolute Deviations: Theory, Applications and Algorithms.* Boston: Birkhäuser.

[9] Geyer, C.J. (1996) On the asymptotics of convex stochastic optimization. (unpublished manuscript)

[10] He, X. and Shao, Q.-M. (1996) General Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Annals of Statistics.* **24**, 2608-2630.

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

[12] Jurečková, J. (1983) Asymptotic behavior of M-estimators of location in nonregular cases. *Statistics and Decisions.* **1**, 323-340.

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

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

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

[16] Smirnov, N.V. (1952) Limit distributions for the terms of a variational series. *American Mathematical Society Translation Series (1).* **11**, 82-143.

[17] van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes with Applications to Statistics.* New York: Springer.

Appendix

LEMMA A. Define

$$g_n(\boldsymbol{u}) = -\boldsymbol{x}_n^T\boldsymbol{u} + \rho_n(\boldsymbol{u}) \qquad h_n(\boldsymbol{u}) = -\boldsymbol{x}_n^T\boldsymbol{u} + \rho(\boldsymbol{u})$$

and let $\boldsymbol{u}_n = \operatorname{argmin}(g_n)$, $\boldsymbol{v}_n = \operatorname{argmin}(h_n)$. Suppose that

(i) $\boldsymbol{x}_n \to \boldsymbol{x}_0$;

(ii) $\boldsymbol{u}_n - \boldsymbol{v}_n \to 0$;

(iii) for any t, $\boldsymbol{u}$ and $\boldsymbol{w}$,

$$\rho_n(\boldsymbol{u} + t\boldsymbol{w}) - \rho_n(\boldsymbol{u}) = \int_0^t \boldsymbol{w}^T\boldsymbol{\psi}_n(\boldsymbol{u} + s\boldsymbol{w})\,ds$$

and

$$\rho(\boldsymbol{u} + t\boldsymbol{w}) - \rho(\boldsymbol{u}) = \int_0^t \boldsymbol{w}^T\boldsymbol{\psi}(\boldsymbol{u} + s\boldsymbol{w})\,ds$$

for some functions $\{\boldsymbol{\psi}_n\}$ and $\boldsymbol{\psi}$ where $\boldsymbol{\psi}$ is one-to-one.

(iv) $\boldsymbol{v}_0 = \boldsymbol{\psi}^{-1}(\boldsymbol{x}_0)$ exists and for some $\alpha > 0$

$$\|\boldsymbol{\psi}(\boldsymbol{u}) - \boldsymbol{\psi}(\boldsymbol{v})\| \le k\|\boldsymbol{u} - \boldsymbol{v}\|^\alpha$$

for all $\boldsymbol{u}$, $\boldsymbol{v}$ in a neighbourhood of $\boldsymbol{v}_0$;

(v) For some sequence $\{b_n\}$ with $b_n \to \infty$ and any compact set K,

$$\sup_{u \in K} \|b_n(\psi_n(u) - \psi(u)) - d_0(u)\| \to 0$$

where d_0 is a continuous function.

Then

$$b_n(\psi(u_n) - \psi(v_n)) \to -d_0(v_0)$$

where $v_0 = \psi^{-1}(x_0)$. Furthermore, if ρ is twice continuously differentiable at v_0 and its Hessian matrix is invertible at v_0 then

$$b_n(u_n - v_n) \to -H^{-1}(v_0)d_0(v_0).$$

Proof. First of all, note that $x_n = \psi(v_n)$. Since u_n minimizes g_n, we have

$$g_n(u_n) \leq g_n(u_n + tw/b_n^{1/\alpha})$$

for any t and any unit vector w. Thus

$$\begin{aligned}
0 &\leq -\frac{t}{b_n^{1/\alpha}} w^T \psi(v_n) + \int_0^{t/b_n^{1/\alpha}} \psi_n(u_n + sw)\, ds \\
&= \frac{t}{b_n^{1/\alpha}} w^T(\psi(u_n) - \psi(v_n)) + \frac{1}{b_n^{1/\alpha}} \int_0^t w^T \left(\psi_n(u_n + sw/b_n^{1/\alpha}) - \psi(u_n + sw/b_n^{1/\alpha})\right) ds \\
&\quad + \frac{1}{b_n^{1/\alpha}} \int_0^t w^T \left(\psi(u_n + sw/b_n^{1/\alpha}) - \psi(u_n)\right) ds \\
&\leq \frac{t}{b_n^{1/\alpha}} w^T(\psi(u_n) - \psi(v_n)) + \frac{1}{b_n^{1/\alpha}} \int_0^t w^T \left(\psi_n(u_n + sw/b_n^{1/\alpha}) - \psi(u_n + sw/b_n^{1/\alpha})\right) ds \\
&\quad + \frac{k'}{b_n^{1+1/\alpha}} |t|^{1+\alpha}.
\end{aligned}$$

Now multiplying both sides of the inequality above by $b_n^{1+1/\alpha}$, we have

$$\epsilon_n(t; w) \leq w^T \left(b_n(\psi(u_n) - \psi(v_n)) + d_0(v_0)\right) t + k'|t|^{1+\alpha}$$

where

$$\epsilon_n(t; w) = -\int_0^t w^T \left[b_n(\psi_n(u_n + sw/b_n^{1/\alpha}) - \psi(u_n + sw/b_n^{1/\alpha})) - d_0(v_0)\right] ds.$$

Since $u_n + sw/b_n^{1/\alpha} \to v_0$ uniformly over s in a compact set and unit vectors w, it follows that

$$\sup_{t \in K, w} |\epsilon_n(t; w)| \to 0$$

for any compact set K. Thus

$$w^T [b_n(\psi(u_n) - \psi(v_n)) + d_0(v_0)] \to 0$$

for any unit vector w which completes the proof. □