Asymptotic theory for M-estimators of boundaries

Keith Knight[1]

Department of Statistics, University of Toronto, Toronto, ON M5S 3G3

Summary

We consider some asymptotic distribution theory for M-estimators of the parameters of a linear model whose errors are non-negative; these estimators are the solutions of constrained optimization problems and their asymptotic theory is non-standard. Under weak conditions on the distribution of the errors and on the design, we show that a large class of estimators have the same asymptotic distributions in the case of i.i.d. errors; however, this invariance does not hold under non-i.i.d. errors.

Keywords: constrained optimization, epi-convergence, linear programming estimator, M-estimator, point processes.

[1]Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

Consider the linear regression model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + W_i \quad (i = 1, \cdots, n) \tag{1}$$

where $\boldsymbol{x}_i$ is a vector of covariates (of length p) whose first component is always 1, $\boldsymbol{\beta}$ is a vector of parameters and $W_1, \cdots, W_n$ are independent, identically distributed (i.i.d.) non-negative random variables whose essential infimum is 0. Thus $\boldsymbol{x}^T \boldsymbol{\beta}$ can be interpreted as the conditional minimum of the response Y given covariate values $\boldsymbol{x}$. (The assumption that the model (1) has an intercept is not always necessary in the sequel but will be assumed throughout as its inclusion reflects common practice.)

Suppose that the W_i's have common density

$$f(w) = \exp[-\rho(w)] \quad \text{for } w > 0$$

where $\rho(w) \to +\infty$ as $w \to \infty$. If ρ is assumed known and is lower semicontinuous (note that a lower semicontinuous version of ρ typically exists) then the maximum likelihood estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_n$, minimizes

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) \quad \text{subject to } Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi} \text{ for } i = 1, \cdots, n. \tag{2}$$

This type of estimator seems to have first been considered by Aigner & Chu (1968) for estimating the so-called "efficient frontier"; they considered $\rho(w) = w^2$ and $\rho(w) = w$. In a recent paper, Florens & Simar (2002) comment on the lack of development of statistical properties of these estimators. An estimator minimizing (2) seems to be sensible estimator of $\boldsymbol{\beta}$ generally for non-negative W_i's.

In fact, the asymptotics of $\widehat{\boldsymbol{\beta}}_n$ appear to have only been considered in the case where $\rho(w) = w$; in this case, $\widehat{\boldsymbol{\beta}}_n$ minimizes

$$-\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\phi} \quad \text{subject to } Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi} \text{ for } i = 1, \cdots, n, \tag{3}$$

which is a linear program. This estimator can also be viewed as a minimum regression quantile estimator as defined by Koenker & Bassett (1978). Limit theory for the estimator minimizing (3) can be derived under weak assumptions on the behaviour of the distribution of the W_i's near 0 and the behaviour of the empirical distribution of the $\boldsymbol{x}_i$'s; see, for example, Smith (1994), Portnoy & Jurečková (1999) and Knight (2001). A similar estimation method has been studied by (among others) Anděl (1989), An & Huang (1993) and Feigen & Resnick (1994) in the context of estimation in stationary autoregressive models with non-negative innovations; Feigen & Resnick

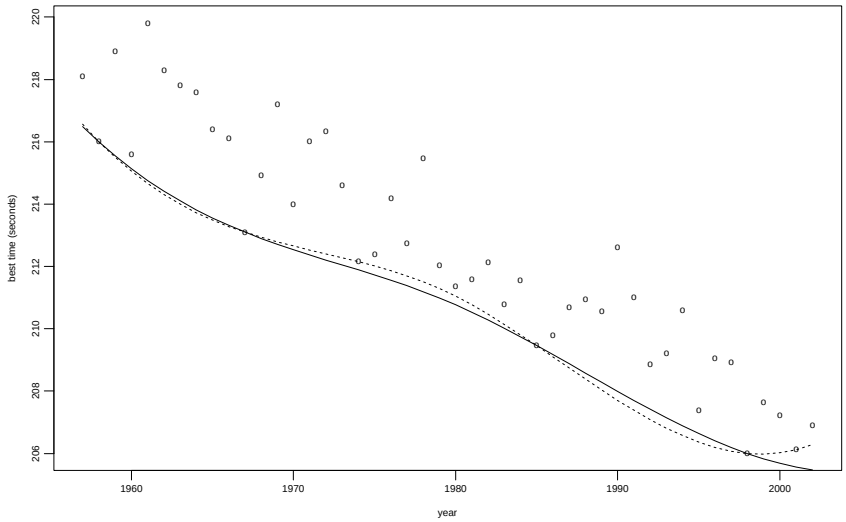

Figure 1: Yearly best men's outdoor 1500m times (in seconds) from 1957 to 2002 with estimated boundary lines using $\rho(w) = w$ (dotted) and $\rho(w) = w^2$ (solid).

(1994) derive limiting distributions of the estimators using an approach that relies heavily on point process arguments. For the general first order autoregressive model, Nielsen & Shephard (2003) derive the exact distribution of this estimator when the innovations have an exponential distribution.

In this paper, we will study the dependence of estimators minimizing (2) on the loss function ρ. Figure 1 shows the yearly best men's 1500m times from 1957 to 2002 with lower boundaries (which might be interpreted as the best possible time for a given year) estimated using $\rho(w) = w$ and $\rho(w) = w^2$; in both cases, we use a b-spline basis with three knots, which means that the parameter vector $\boldsymbol{\beta}$ has five elements (including an intercept). For these data, the two estimates are quite close although not identical; it is natural to ask whether this phenomenon occurs more generally. Note that the estimate for $\rho(w) = w$ is not strictly decreasing; depending on our interpretation of the lower boundary, it may be more natural to constrain the estimation so the estimate of the lower boundary is strictly decreasing.

In the i.i.d. setting (i.e. where $Y_i = \theta + W_i$), the analysis is straightforward to do. If $\rho(w)$ is increasing for $w \geq 0$ then the estimator is simply $\widehat{\theta}_n = \min(Y_1, \cdots, Y_n)$. More generally, suppose that $\rho(w)$ is convex and differentiable though not necessarily increasing for $w \geq 0$. Then the estimator

is again $\min(Y_1, \cdots, Y_n)$ unless there exists $\widehat{\theta}_n < \min(Y_1, \cdots, Y_n)$ such that

$$\sum_{i=1}^{n} \rho'(Y_i - \widehat{\theta}_n) = 0.$$

Suppose that $\min(Y_1, \cdots, Y_n) \xrightarrow{p} \theta$ and set $W_i = Y_i - \theta$. If $E[\rho'(W_i)] > 0$ then by convexity of ρ, it follows that $E[\rho'(W_i + t)] > 0$ for $t \geq 0$ and so

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} [\rho(W_i + t) - \rho(W_i)] &= \int_0^t \frac{1}{n}\sum_{i=1}^{n} \rho'(W_i + s)\, ds \\ &\xrightarrow{a.s.} \int_0^t E[\rho'(W_i + s)]\, ds \\ &> 0. \end{aligned}$$

¿From this we can conclude that $\widehat{\theta}_n$ is eventually equal to $\min(Y_1, \cdots, Y_n)$ if $E[\rho'(W_i)] > 0$.

The purpose of this paper is to extend the equivalence in an asymptotic sense to the regression case under general conditions on the $\boldsymbol{x}_i$'s and the distribution of the W_i's; in particular, we will not assume any relationship between the density of the W_i's and the loss function ρ. We will also show that the asymptotic equivalence does not necessarily hold for non-i.i.d. errors.

2 Asymptotics

As in Knight (2001), the key tools used in deriving the limiting distribution of $\widehat{\boldsymbol{\beta}}_n$ minimizing (2) are epi-convergence in distribution (Pflug 1994, Pflug 1995, Geyer 1994, Geyer 1996, Knight 1999) and point process convergence for extreme values (Kallenberg 1983, Leadbetter *et al.* 1983). Point processes defined on a space can be characterized by random measures that count the (random) number of points lying in subsets of the space; point process convergence is characterized by the weak convergence of integrals of bounded continuous functions with compact support with respect to the random measures (Kallenberg 1983). Under appropriate regularity conditions (described below), the configuration of points $\{(\boldsymbol{x}_i, Y_i)\}$ generated from (1) lying in a neighbourhood of the plane $\boldsymbol{x}^T\boldsymbol{\beta}$ can be approximated (in a distributional sense) by a Poisson process when n is large and the asymptotic behaviour of $\widehat{\boldsymbol{\beta}}_n$ (perhaps not surprisingly) turns out to depend on this Poisson process.

Epi-convergence in distribution gives us an elegant way of proving convergence in distribution of "argmin" (and "argmax") estimators, and is particularly useful for constrained estimation procedures. A sequence of random lower semicontinuous functions $\{Z_n\}$ epi-converges in distribution to

Z ($Z_n \xrightarrow{e-d} Z$) if for any closed rectangles $R_1, \cdots, R_k$ with open interiors $R_1^o, \cdots, R_k^o$ and any real numbers $a_1, \cdots, a_k$,

$$\begin{aligned} & P\left\{\inf_{\mathbf{u}\in R_1} Z(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z(\boldsymbol{u}) > a_k\right\} \\ \leq & \liminf_{n\to\infty} P\left\{\inf_{\mathbf{u}\in R_1} Z_n(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z_n(\boldsymbol{u}) > a_k\right\} \\ \leq & \limsup_{n\to\infty} P\left\{\inf_{\mathbf{u}\in R_1^o} Z_n(\boldsymbol{u}) \geq a_1, \cdots, \inf_{\mathbf{u}\in R_k^o} Z_n(\boldsymbol{u}) \geq a_k\right\} \\ \leq & P\left\{\inf_{\mathbf{u}\in R_1^o} Z(\boldsymbol{u}) \geq a_1, \cdots, \inf_{\mathbf{u}\in R_k^o} Z(\boldsymbol{u}) \geq a_k\right\}. \end{aligned}$$

For an extended real-valued lower-semicontinuous function g, define

$$\begin{aligned} \operatorname{argmin}(g) &= \left\{\boldsymbol{u}_0 : g(\boldsymbol{u}_0) = \inf_{\mathbf{u}} g(\boldsymbol{u})\right\} \\ \epsilon - \operatorname{argmin}(g) &= \left\{\boldsymbol{u}_0 : g(\boldsymbol{u}_0) \leq \inf_{\mathbf{u}} g(\boldsymbol{u}) + \epsilon\right\}. \end{aligned}$$

Suppose that $\boldsymbol{U}_n \in \operatorname{argmin}(Z_n)$ where $Z_n \xrightarrow{e-d} Z$ and $\boldsymbol{U}_n = O_p(1)$; then $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U} = \operatorname{argmin}(Z)$ provided that $\operatorname{argmin}(Z)$ is (with probability 1) a singleton. (The condition that $\boldsymbol{U}_n \in \operatorname{argmin}(Z_n)$ can be weakened to $\boldsymbol{U}_n \in \epsilon_n - \operatorname{argmin}(Z_n)$ where $\epsilon_n \xrightarrow{p} 0$.) If the Z_n's are convex (as will be the case here) then epi-convergence is quite simple to prove; finite dimensional convergence in distribution of Z_n to Z ($Z_n \xrightarrow{f-d} Z$) is sufficient for epi-convergence in distribution provided that Z is finite on an open set. (In fact, it is sufficient to prove this finite dimensional convergence on a countable dense subset.) Moreover, if $\operatorname{argmin}(Z)$ is a singleton then $\boldsymbol{U}_n = O_p(1)$ is implied by $Z_n \xrightarrow{e-d} Z$.

In order to consider the asymptotics of estimators minimizing (2), we need to make some mild assumptions. We will assume that ρ in (2) is a convex function with

$$\rho(w) = \int_0^w \psi(t)\, dt \tag{4}$$

for some non-decreasing function ψ satisfying

$$|\psi(w+t) - \psi(w)| \leq M(w)|t|^\delta \tag{5}$$

for $w > 0$ and $|t| \leq \epsilon$ where $\delta > 0$. In addition, we will make the following assumptions about the design and the distributions of the W_i's:

(A0) For ψ defined in (4), $E[\psi(W_i)] > 0$ and $E[\psi^2(W_i)] < \infty$.

(A1) For some sequence $a_n \to \infty$, we have

$$|nP\left(a_n W_i \le t\right) - t^\alpha| \le \tau_n t^\alpha$$

where $\alpha > 0$ and $\tau_n \to 0$.

(A2) There exists a sequence of matrices $\{C_n\}$ and a probability measure μ on R^p such that for each set B with $\mu(\partial B) = 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(C_n^{-1}\boldsymbol{x}_i \in B) = \mu(B).$$

(A3) $\int \|\boldsymbol{x}\|\, \mu(d\boldsymbol{x}) < \infty$ with

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} C_n^{-1}\boldsymbol{x}_i &= \int \boldsymbol{x}\, \mu(d\boldsymbol{x}) = \boldsymbol{\gamma}, \\ \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \|C_n^{-1}\boldsymbol{x}_i\|^2 &= 0. \end{aligned}$$

(A4) $\mu(D_\gamma) = 0$ where

$$D_\gamma = \left\{\boldsymbol{x} : \boldsymbol{x}^T\boldsymbol{c} = 0 \text{ for some } \boldsymbol{c} \ne \boldsymbol{0} \text{ with } \boldsymbol{\gamma}^T\boldsymbol{c} = 0\right\}$$

where $\boldsymbol{\gamma}$ is defined in (A3).

(A5) The (closed) set

$$K = \left\{\boldsymbol{u} : \int (\boldsymbol{u}^T\boldsymbol{x})_+^\alpha\, \mu(d\boldsymbol{x}) < \infty\right\}$$

has an open interior and for each $\boldsymbol{u} \in \text{int}(K)$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i)_+^\alpha &= \int (\boldsymbol{u}^T\boldsymbol{x})_+^\alpha\, \mu(d\boldsymbol{x}) \\ \lim_{n\to\infty} \frac{1}{n} \max_{1\le i\le n} (\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i)_+^\alpha &= 0 \end{aligned}$$

where $x_+ = \max(x, 0)$ denotes the positive part of x.

(A6) $E[M(W_i)] < \infty$ and

$$\lim_{n\to\infty} \frac{1}{a_n^\delta} \max_{1\le i\le n} \|C_n^{-1}\boldsymbol{x}_i\|^{\delta+1} = 0$$

where $M(\cdot)$ and δ are defined as in (5).

It is worth commenting at this point on the *raison d'être* of these conditions. The first part of condition (A0) is essentially necessary for consistency; if $E[\psi(W_i)] < 0$ then $\widehat{\boldsymbol{\beta}}_n$ will not converge to $\boldsymbol{\beta}$. Condition (A1) generalizes the condition on the density of the W_i's assumed in Smith (1994) and implies that the W_i's are in the domain of attraction of a type III extreme value distribution. Condition (A2) is effectively a weak convergence condition for the empirical distribution of the $\boldsymbol{x}_i$'s; if the $\boldsymbol{x}_i$'s are a random sample from some distribution then we would have $C_n = I$ and μ equal to the underlying probability measure of the $\boldsymbol{x}_i$'s. Even for fixed designs, (A2) is a reasonable condition although C_n need not equal I (although it is typically a diagonal matrix). For example, if $\boldsymbol{x}_i = (1, i, i^2)^T$ for $i = 1, \cdots, n$ then the diagonal elements of C_n are $(1, n, n^2)$ and μ is the probability measure of the random vector $(1, U, U^2)$ where U is uniformly distributed on $[0, 1]$. More importantly, (A2) implies similar weak convergence results about the empirical distribution of $\boldsymbol{u}^T C_n^{-1} \boldsymbol{x}_i$ $(i = 1, \cdots, n)$ for a given $\boldsymbol{u}$ (or finite number of $\boldsymbol{u}$'s). Moreover, if $C_n^{-1}\boldsymbol{x}_i$ is bounded then condition (A1) can be replaced by

$$nP\left(a_n W_i \leq t\right) \to t^\alpha$$

for each $t > 0$. Conditions (A3)–(A5) are used to facilitate the proof of epi-convergence in distribution of an appropriate sequence of objective functions; for example, (A4) will imply that the limiting objective function has a unique minimizer (with probability 1) while (A5) will imply that the limiting objective function is finite on a open set and so finite dimensional weak convergence will imply epi-convergence in distribution. (In fact, condition (A5) is not necessary and is included only to simplify the proof.) Condition (A6) together with condition (A3) allows us to approximate the finite part of the objective function by a linear function.

Note that conditions (A3), (A5), and (A6) are essentially moment conditions on the $\boldsymbol{x}_i$'s (or, more precisely, on the $C_n^{-1}\boldsymbol{x}_i$'s); depending on the value of α, one of these conditions may imply all or part of the others. The conditions as stated are certainly far from minimal and can be weakened

THEOREM 1. Assume the model (1) and suppose that $\widehat{\boldsymbol{\beta}}_n$ is minimizes of (2) where ρ is convex and satisfies (4) and (5). If conditions (A0)–(A6) hold then

$$a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ is the solution of the linear programming problem:

$$\text{maximize} \quad \boldsymbol{u}^T \boldsymbol{\gamma} \quad \text{subject to } \Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i \text{ for } i = 1, 2, 3, \cdots$$

where

(i) $\Gamma_i = (E_1 + \cdots + E_i)^{1/\alpha}$ for unit mean i.i.d. exponential random variables $E_1, E_2, \cdots$;

(ii) $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots$ are i.i.d. with distribution $P(\boldsymbol{X}_i \in A) = \mu(A)$;

(iii) the $\boldsymbol{X}_i$'s are independent of the E_i's (and hence of the Γ_i's).

Proof. The proof follows much along the lines of the proof of Theorem 1 of Knight (2001). First of all, note that $\boldsymbol{U}_n = a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is the solution to the linear programming problem:

$$\begin{aligned} \text{minimize} \quad & \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n) - \rho(W_i)\right] \\ \text{subject to} \quad & a_n W_i \geq \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i \quad \text{for } i = 1, \cdots, n. \end{aligned}$$

Defining $\phi_n(\boldsymbol{u})$ to be 0 when the constraints above are all satisfied and $+\infty$ otherwise, $\boldsymbol{U}_n$ minimizes

$$Z_n(\boldsymbol{u}) = \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n) - \rho(W_i)\right] + \phi_n(\boldsymbol{u}). \tag{6}$$

Z_n is a convex function for each n and so to prove that $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$, it suffices to show that $Z_n \xrightarrow{e-d}$ some Z where $\boldsymbol{U} = \text{argmin}(Z)$; we will show that

$$Z(\boldsymbol{u}) = -E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + \phi(\boldsymbol{u}) \tag{7}$$

where $\phi(\boldsymbol{u}) = 0$ if $\Gamma_i \geq \boldsymbol{u}^T\boldsymbol{X}_i$ for all i and $\phi(\boldsymbol{u}) = +\infty$ otherwise.

Using the integral representation (4) for ρ and condition (5), we obtain

$$\begin{aligned} & \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n) - \rho(W_i)\right] \\ = \quad & -\frac{1}{n}\sum_{i=1}^{n}\psi(W_i)\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i + o_p(1) \\ = \quad & -E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + o_p(1) \end{aligned}$$

using condition (A3) to establish the weak law of large numbers and condition (A6) to establish the asymptotic linearity. ¿From the convexity of Z_n, $Z_n \xrightarrow{e-d} Z$ follows from $Z_n \xrightarrow{f-d} Z$ provided that Z is finite on an open set with probability 1; the latter follows since $\Gamma_i \sim i^{1/\alpha}$ (with probability 1) as $i \to \infty$ and so by the first Borel-Cantelli lemma $P(\boldsymbol{u}^T\boldsymbol{X}_i > \Gamma_i \text{ infinitely often}) = 0$ for any $\boldsymbol{u} \in K$ (since $E[(\boldsymbol{u}^T\boldsymbol{X}_i)_+^\alpha] < \infty$ on K); for $\boldsymbol{u} \notin K$, we also have $P(\boldsymbol{u}^T\boldsymbol{X}_i > \Gamma_i \text{ infinitely often}) = 1$ (since $E[(\boldsymbol{u}^T\boldsymbol{X}_i)_+^\alpha] = \infty$) by the second Borel-Cantelli lemma. Thus for a given $\boldsymbol{u} \in K$, at most a finite number of constraints are violated, the rest being trivially satisfied. Since $\boldsymbol{u} \in K$ implies that $t\boldsymbol{u} \in K$ for $t > 0$, taking t sufficiently small guarantees that all

the constraints are satisfied. Since int(K) is open (by condition (A5)), it is possible (with probability 1) to find a finite number of points in K such that all the constraints are satisfied and the convex hull of these points contains an open set. Since Z is finite at these points, it is necessarily finite on the convex hull (since Z is convex).

To show the finite dimensional weak convergence of ϕ_n, we first define the following point process (random measure) on R^{p+1}:

$$\nu_n(A \times B) = \sum_{i=1}^{n} I(a_n W_i \in A, C_n^{-1} \boldsymbol{x}_i \in B).$$

It is easy to verify that ν_n tends in distribution with respect to the vague topology (Kallenberg 1983) to a Poisson process (random measure) ν whose mean measure is

$$E[\nu(A \times B)] = \mu(B) \int_{A \cap (0,\infty)} \alpha x^{\alpha-1}\, dx.$$

We can represent the points of this Poisson process by $\{(\Gamma_i, \boldsymbol{X}_i) : i \geq i\}$ where the Γ_i's and $\boldsymbol{X}_i$'s are as defined above. Thus it suffices to show that

$$P\left[\phi_n(\boldsymbol{u}_1) = 0, \cdots, \phi_n(\boldsymbol{u}_k) = 0\right] \to P\left[\phi(\boldsymbol{u}_1) = 0, \cdots, \phi(\boldsymbol{u}_k) = 0\right]$$

where $\phi(\boldsymbol{u}) = 0$ if $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i$ for all i and ∞ otherwise. Exploiting the convergence in distribution of ν_n to the Poisson random measure ν, we have

$$\begin{aligned}
&P\left[\phi_n(\boldsymbol{u}_1) = 0, \cdots, \phi_n(\boldsymbol{u}_k) = 0\right] \\
&= \quad P\left\{\sum_{i=1}^{n} I\left[0 \leq a_n W_i < \max_{1 \leq j \leq k} (\boldsymbol{u}_j^T C_n^{-1} \boldsymbol{x}_i)_+\right] = 0\right\} \\
&\to \quad \exp\left[-\int \max_{1 \leq j \leq k} (\boldsymbol{u}_j^T \boldsymbol{x})_+^{\alpha}\, \mu(d\boldsymbol{x})\right] \\
&= \quad P\left[\phi(\boldsymbol{u}_1) = 0, \cdots, \phi(\boldsymbol{u}_k) = 0\right].
\end{aligned}$$

Hence for Z_n given in (6), we have $Z_n \xrightarrow{f-d} Z$ where Z is defined in (7). Finally, to show that Z has a unique minimizer (with probability 1), we note that if $\boldsymbol{U}$ minimizes Z then for some indices $i_1 < i_2 < \cdots < i_p$, we have $\boldsymbol{U}^T \boldsymbol{X}_{i_k} = \Gamma_{i_k}$ with $\Gamma_j > \boldsymbol{U}^T \boldsymbol{X}_j$ for $j \notin \{i_1, i_2, \cdots, i_p\}$. If $\boldsymbol{U}$ and $\boldsymbol{U}^*$ both minimize Z then $\boldsymbol{U}^* = \boldsymbol{U} + t\boldsymbol{c}$ for some vector $\boldsymbol{c}$ with $\boldsymbol{c}^T \boldsymbol{\gamma} = 0$ and so $t\boldsymbol{c}^T \boldsymbol{X}_{i_k} = 0$ for $k = 1, \cdots, p$. However, condition (A4) says that $P(\boldsymbol{c}^T \boldsymbol{X}_i = 0) = 0$ (when $\boldsymbol{c}^T \boldsymbol{\gamma} = 0$) and so Z is a unique minimizer (with probability 1). □

As mentioned above, the conclusion of Theorem 1 holds even if the set K defined in condition (A5) does not have an open interior. In this case, the limiting objective function Z will not be finite on an open set and so

$Z_n \xrightarrow{e-d} Z$ will not follow immediately from $Z_n \xrightarrow{f-d} Z$. However, the epi-convergence in distribution will still hold; we need to establish that the sequence of functions $\{\phi_n\}$ (which describe the constraints) are stochastically equi-lower-semicontinuous (Knight 1999). For this, we need to show that for any bounded set B and $\delta > 0$, there exist points $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m$ in B and open neighbourhoods $V(\boldsymbol{u}_1), \cdots, V(\boldsymbol{u}_m)$ of these points such that

$$B \subset \bigcup_{i=1}^{m} V(\boldsymbol{u}_i)$$

(that is, B is covered by the neighbourhoods) and

$$\limsup_{n\to\infty} P\left\{\bigcup_{i=1}^{m}\left[\inf_{u\in V(u_i)} \phi_n(\boldsymbol{u}) = 0, \phi_n(\boldsymbol{u}_i) = \infty\right]\right\} < \delta.$$

This turns out to be reasonably straightforward to show since

$$\begin{aligned} & P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in V(\mathbf{u}_i)} \phi_n(\boldsymbol{u}) = 0, \phi_n(\boldsymbol{u}_i) = \infty\right]\right\} \\ \leq \quad & P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in V(\mathbf{u}_i)} \phi_n(\boldsymbol{u}) = 0\right]\right\} - P\left\{\bigcup_{i=1}^{m}\left[\phi_n(\boldsymbol{u}_i) = 0\right]\right\}. \end{aligned}$$

The right hand side above can be made arbitrarily small by making the neighbourhoods uniformly small.

In the case where $\rho(w) = w$, Smith (1994) as well as Portnoy & Jurečková (1999) determine the limiting distribution by finding the limiting density of $\boldsymbol{U}_n = \text{argmin}(Z_n)$; however, they need to assume a specific form for the density of the W_i's, from which the density of $\boldsymbol{U}_n$ can be approximated. The conclusion of Theorem 1 holds under a weak assumption (condition (A1)) about the distribution of the W_i's, which in particular does not imply the existence of a density function. Chernozhukov (2000) also uses an epi-convergence approach to study the asymptotic behaviour of "near extreme" regression quantile estimators.

In the case where the set K defined in (A5) satisfies $K = \text{cl}(\text{int}(K))$, we can determine the limiting joint density, that is, the density of $\boldsymbol{U} = \text{argmin}(Z)$. Using the Poisson process representation of Z, it follows that the density of $\boldsymbol{U}$ is

$$g(\boldsymbol{u}) = \kappa(\boldsymbol{u}; \alpha, p, \mu) \int \cdots \int |D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)| \prod_{i=1}^{p} \left\{(\boldsymbol{u}^T \boldsymbol{x}_i)_+^{\alpha-1} \, \mu(d\boldsymbol{x}_i)\right\} \quad (8)$$

where

$$\kappa(\boldsymbol{u}; \alpha, p, \mu) = \frac{\alpha^p}{p!} \exp\left[-\int (\boldsymbol{u}^T \boldsymbol{x})_+^{\alpha} \, \mu(d\boldsymbol{x})\right]$$

and $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$ is the determinant of the matrix with columns $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ if $\boldsymbol{\gamma}$ lies in the convex hull of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ and $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) = 0$ otherwise. (If there is no intercept in the model (1) then $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$ is the determinant if $\boldsymbol{\gamma} = \sum_{j=1}^{p} t_j \boldsymbol{x}_j$ for non-negative t_j's with $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) = 0$ otherwise.) The density $g(\boldsymbol{u})$ is not easy to evaluate in closed-form (except in special cases) but can be approximated quite easily using Monte Carlo techniques (by sampling from the probability measure μ). However, it seems that this density does not provide as much insight into the limiting distribution as does the representation of $\boldsymbol{U}$ as the solution of a linear programming problem.

Other estimation problems in which the limiting objective function is related to a Poisson process are considered by Pflug (1994). Theorem 1 implies that we obtain the same limiting distribution for any convex ρ satisfying some mild regularity conditions so that all such estimators differ by $o_p(a_n^{-1} C_n^{-1})$. However, an examination of the proof of Theorem 1 suggests that this asymptotic equivalence is a consequence of the i.i.d. assumption on the W_i's.

Suppose instead we assume that the W_i's in (1) are independent with the distribution of W_i depending on $\boldsymbol{x}$ such that

$$|nP(a_n W_i \leq t | \boldsymbol{x}) - \lambda(\boldsymbol{x}) t^{\alpha}| \leq \tau_n(\boldsymbol{x}) t^{\alpha}$$

where

$$\max_{1 \leq i \leq n} |\tau_n(\boldsymbol{x}_i)| \to 0.$$

Under condition (A2) on the $\boldsymbol{x}_i$'s, it then follows that the point process

$$\nu_n(A \times B) = \sum_{i=1}^{n} I(a_n W_i \in A, C_n^{-1} \boldsymbol{x}_i \in B)$$

converges in distribution to a point process ν whose mean measure is given by

$$E[\nu(A \times B)] = \int_A \int_B \alpha \lambda(\boldsymbol{x}) t^{\alpha - 1} \, \mu(d\boldsymbol{x}) \, dt.$$

The points of ν can be represented by $\{(\Gamma_i / \lambda(\boldsymbol{X}_i), \boldsymbol{X}_i) : i = 1, 2, \cdots\}$ where the Γ_i's and $\boldsymbol{X}_i$'s are defined as in Theorem 1. Assuming that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \psi(W_i) \boldsymbol{u}^T C_n^{-1} \boldsymbol{x}_i \quad &\xrightarrow{p} \quad \int E(\psi(W) | \boldsymbol{x}) \boldsymbol{u}^T \boldsymbol{x} \, \mu(d\boldsymbol{x}) \\ &= \quad \boldsymbol{u}^T \boldsymbol{\gamma}_{\psi} \end{aligned}$$

it will follow (under appropriate modifications of the regularity conditions) that

$$a_n C_n (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ maximizes $\boldsymbol{u}^T \boldsymbol{\gamma}_{\psi}$ subject to $\Gamma_i \geq \lambda(\boldsymbol{X}_i) \boldsymbol{u}^T \boldsymbol{X}_i$ for all i. Note that $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\gamma}_{\psi}, \nu)$ where the point process ν does not depend on the loss function ρ (nor its "derivative" ψ).

We can extend (8) to obtain the density of $\boldsymbol{U}$ in this case:

$$g(\boldsymbol{u}) = \kappa_\lambda(\boldsymbol{u};\alpha,p,\mu)\int\cdots\int |D_\lambda(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)| \prod_{i=1}^{p}\left\{[\lambda(\boldsymbol{x}_i)\boldsymbol{u}^T\boldsymbol{x}_i]_+^{\alpha-1}\,\mu(d\boldsymbol{x}_i)\right\} \tag{9}$$

where

$$\kappa_\lambda(\boldsymbol{u};\alpha,p,\mu) = \frac{\alpha^p}{p!}\exp\left[-\int(\lambda(\boldsymbol{x})\boldsymbol{u}^T\boldsymbol{x})_+^\alpha\,\mu(d\boldsymbol{x})\right]$$

and $D_\lambda(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)$ is the determinant of the matrix whose columns are $\lambda(\boldsymbol{x}_1)\boldsymbol{x}_1,\cdots,\lambda(\boldsymbol{x}_p)\boldsymbol{x}_p$ if

$$\boldsymbol{\gamma}_\psi = \sum_{j=1}^{p} t_j\lambda(\boldsymbol{x}_j)\boldsymbol{x}_j$$

for some non-negative $t_1,\cdots,t_p$ and $D_\lambda(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)$ is 0 otherwise.

EXAMPLE 1. Consider the simple regression model

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

where $W_1,\cdots,W_n$ are independent (but identically distributed) random variables with the distribution of W_i depending on x_i, and we will assume that the x_i's are uniformly distributed on the interval $[-1,1]$, which implies that μ is a uniform distribution on $[-1,1]$. For a given loss function ρ (with "derivative" ψ), the vector $\boldsymbol{\gamma}_\psi$ is simply

$$\boldsymbol{\gamma}_\psi = \frac{1}{2}\int_{-1}^{1} E[\psi(W)|x]\begin{pmatrix}1\\x\end{pmatrix}dx = \left\{\int_{-1}^{1} E[\psi(W)|x]\,dx\right\}\begin{pmatrix}1\\c_\psi\end{pmatrix}$$

where $-1 < c_\psi < 1$; note that for $\rho(x) = x$, $c_\psi = 0$. For simplicity, we will take $\alpha = 1$ and set $\lambda(x) = 1$ (which is possible even when the W_i's are not identically distributed). Thus for a given ρ (and corresponding ψ), we have

$$n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U} = \begin{pmatrix}U_0\\U_1\end{pmatrix}$$

where $\boldsymbol{U}$ maximizes $u_0 + c_\psi u_1$ subject to $\Gamma_i \geq u_0 + u_1 X_i$ for $i \geq 1$ where the Γ_i's are partial sums of i.i.d. unit mean exponential random variables and the X_i's are i.i.d. uniform random variables on $[-1,1]$. Thus the limiting distribution depends only on the constant c_ψ (which depends on ψ and the dependence between the W_i's and the x_i's). Figures 2 to 5 show contour plots of the joint density of $\boldsymbol{U}$ (using (9)) for $c_\psi = 0, 0.25, 0.5, 0.75$. In all cases, the distribution of U_0 (intercept) is concentrated on the positive part of the real line. As c_ψ increases, more probability mass is shifted to the positive part of the distribution of U_1, that is, the bias of the slope estimator becomes more positive as c_ψ increases; likewise, the bias becomes more negative as c_ψ decreases from 0 to -1.

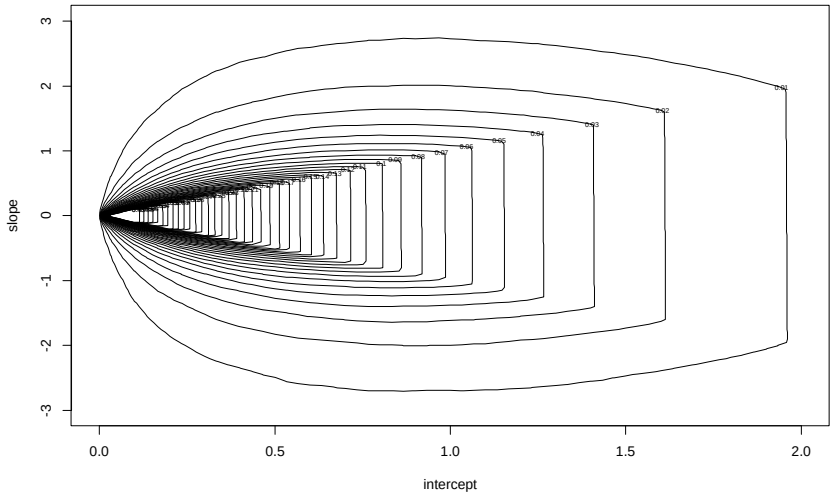

Figure 2: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0$; the interval between adjacent contours is 0.01.

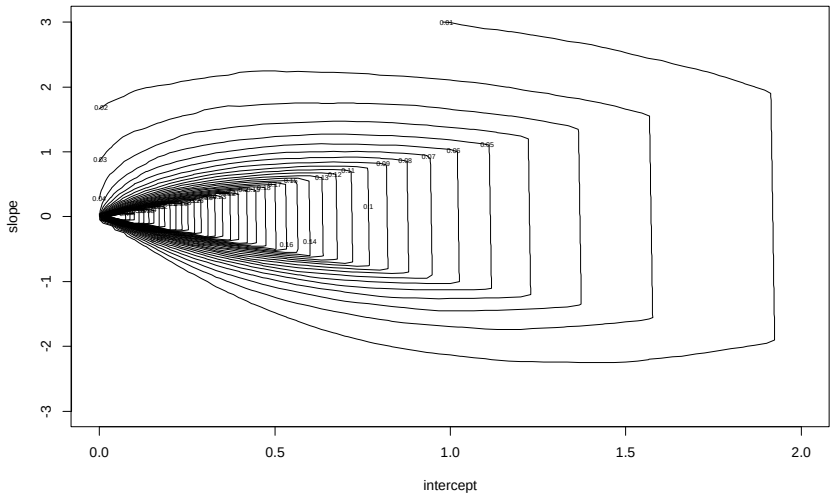

Figure 3: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.25$; the interval between adjacent contours is 0.01.

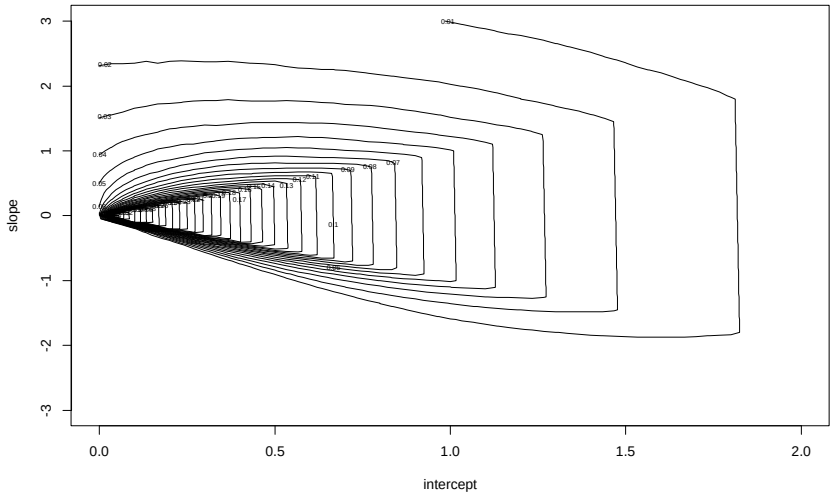

Figure 4: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.5$; the interval between adjacent contours is 0.01.

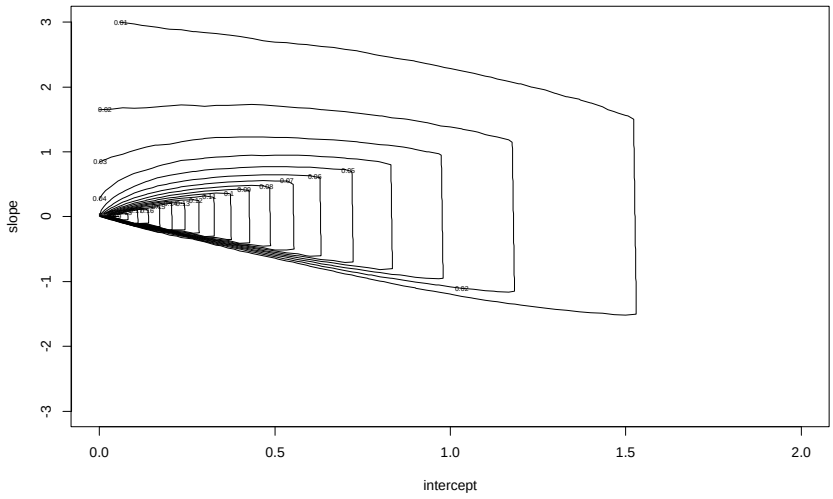

Figure 5: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.75$; the interval between adjacent contours is 0.01.

3 Barrier regularization

Estimators minimizing (2) are inherently biased upwards since necessarily we have

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n) < \sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

and so $\boldsymbol{x}^T \widehat{\boldsymbol{\beta}}_n$ tends to be systematically smaller than $\boldsymbol{x}^T \boldsymbol{\beta}$ (since ρ is "on average" increasing under condition (A0)). In general, reducing bias is a tricky proposition since such a reduction often leads to an increase in variance. In this problem, the bias typically manifests itself in the estimation of the intercept and so one might consider reducing bias simply by adjusting (downwards) the intercept estimator.

An alternative approach to reducing bias is to replace the constraints $Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi}$ $(i = 1, \cdots, n)$ in (2) by a "barrier" function that pushes the estimator away from the boundary of the constraint region. Specifically, we will define $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ to minimize

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) + \epsilon \sum_{i=1}^{n} \tau(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) \quad \text{subject to } Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi} \text{ for all } i \quad (10)$$

where ϵ is a positive constant and the barrier function $\tau(w)$ is a convex function on $(0, \infty)$ satisfying

$$\lim_{w \downarrow 0} \tau(w) = +\infty,$$

for example, $\tau(w) = w^{-r}$ for $r > 0$ or $\tau(w) = -\ln(w)$. It is easy to see that, for any $\epsilon > 0$, the minimizer of (10) will lie in the interior of the set $\{\boldsymbol{\phi} : Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi} \text{ for } i = 1, \cdots, n\}$ and so if $\rho(w)$ and $\tau(w)$ are differentiable for $w > 0$, it follows that $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ satisfies

$$\sum_{i=1}^{n} \left[\rho'(Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n(\epsilon)) + \epsilon \tau'(Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n(\epsilon)) \right] \boldsymbol{x}_i = \mathbf{0}.$$

More importantly, by choosing $\epsilon = \epsilon_n$ appropriately, we may be able to reduce the bias of $\widehat{\boldsymbol{\beta}}_n(\epsilon_n)$ while retaining many of the otherwise attractive properties possessed by $\widehat{\boldsymbol{\beta}}_n$.

There is a connection between the estimators minimizing (2) and (10). If $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ minimizes (10) and $\widehat{\boldsymbol{\beta}}_n$ minimizes (2) then

$$\lim_{\epsilon \downarrow 0} \widehat{\boldsymbol{\beta}}_n(\epsilon) = \widehat{\boldsymbol{\beta}}_n.$$

(This follows since the objective function implied by (10) epi-converges to the objective function implied by (2) as $\epsilon \downarrow 0$ for each fixed n.) This observation

turns out to be useful in the computation of $\widehat{\boldsymbol{\beta}}_n$ minimizing (2). For each $\epsilon > 0$, (10) can be minimized using "standard" optimization techniques (for example, Newton and quasi-Newton methods) and so we can obtain an arbitrarily good approximation to $\widehat{\boldsymbol{\beta}}_n$ minimizing (2) by computing a sequence of minimizers of (10), $\widehat{\boldsymbol{\beta}}_n(\epsilon_k)$ with $\epsilon_k \downarrow 0$. Such numerical methods for solving constrained optimization problems are commonly referred to as barrier or interior point methods; some theory can be found in Fiacco & McCormick (1990).

By taking $\tau(w) = w^{-r}$ for r sufficiently large, we obtain the following analogue of Theorem 1.

THEOREM 2. Assume the model (1) and suppose that $\widehat{\boldsymbol{\beta}}_n(\epsilon_n)$ minimizes (10) (with $\epsilon = \epsilon_n$) where ρ is convex and satisfies (4) and (5). If conditions (A0)–(A6) hold and $\tau(w) = w^{-r}$ where $r > \alpha$ and

$$\lim_{n\to\infty} \frac{a_n^{r+1}}{n}\epsilon_n = \epsilon_0$$

then

$$a_n C_n(\widehat{\boldsymbol{\beta}}_n(\epsilon_n) - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ minimizes

$$-E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + \epsilon_0 \sum_{i=1}^{\infty} (\Gamma_i - \boldsymbol{u}^T\boldsymbol{X}_i)^{-r}$$

subject to $\Gamma_i \geq \boldsymbol{u}^T\boldsymbol{X}_i$ for all i with $\{\Gamma_i\}$ and $\{\boldsymbol{X}_i\}$ defined as in Theorem 1.

Proof. The proof follows along the same lines as the proof of Theorem 1. We redefine Z_n in (6) by

$$\begin{aligned} Z_n(\boldsymbol{u}) &= \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n) - \rho(W_i)\right] \\ &\quad + \frac{a_n}{n}\epsilon_n \sum_{i=1}^{n}\left(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n\right)^{-r} \\ &= \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i/a_n) - \rho(W_i)\right] \\ &\quad + \frac{a_n^{r+1}}{n}\epsilon_n \sum_{i=1}^{n}\left(a_n W_i - \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i\right)^{-r} \\ &= Z_n^{(1)}(\boldsymbol{u}) + Z_n^{(2)}(\boldsymbol{u}) \end{aligned}$$

provided that $a_n W_i \geq \boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i$ for all i with $Z_n(\boldsymbol{u}) = +\infty$ otherwise. The

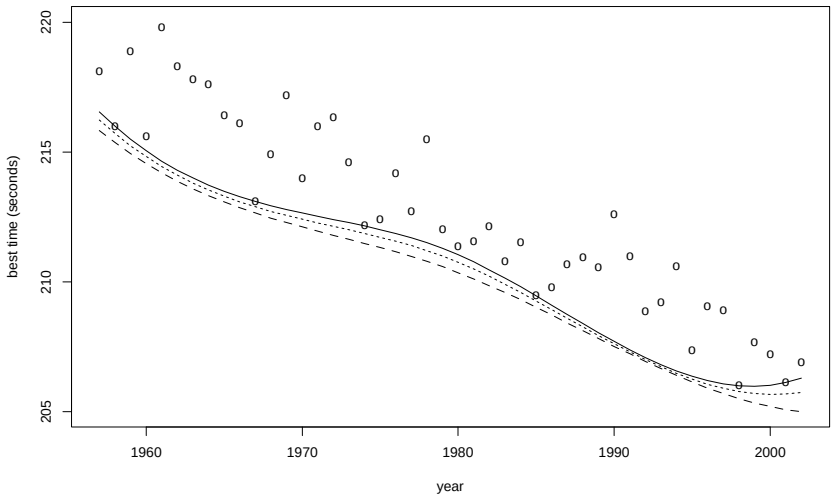

Figure 6: Estimated boundary lines for the 1500m data using $\rho(w) = w$ and $\tau(w) = w^{-2}$ for $\epsilon = 0.05$ (dotted) and $\epsilon = 0.5$ (dashed). The solid line is estimate given in Figure 1 and corresponds to the limit as $\epsilon \downarrow 0$.

only technical complication lies in showing that $Z_n^{(2)} \xrightarrow{f-d} Z^{(2)}$ where

$$Z^{(2)}(\boldsymbol{u}) = \epsilon_0 \sum_{i=1}^{\infty} (\Gamma_i - \boldsymbol{u}^T \boldsymbol{X}_i)^{-r}$$

when $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i$ for all i with $Z^{(2)}(\boldsymbol{u}) = +\infty$ otherwise; this can be done by truncating the barrier function $\tau(w) = w^{-r}$ to make it bounded with compact support and then using Slutsky-type arguments to take care of the difference. □

The assumption that $r > \alpha$ is inconvenient but seems to be necessary in order to obtain non-degenerate asymptotic results, at least, with the "right" rate of convergence; if $\tau(w) \to \infty$ too slowly as $w \downarrow 0$ then typically we will obtain a slower convergence rate for the resulting estimators. In particular, it rules out the barrier function $\tau(w) = -\ln(w)$, which is quite useful for numerical computation.

Figure 6 shows the estimated boundaries for the 1500m data discussed in section 1 using $\rho(w) = w$ and $\tau(w) = w^{-2}$ in (10) with $\epsilon = 0.05$ and $\epsilon = 0.5$. The choice of ϵ for a given value of r is an open question; however, for these data, the estimates seem somewhat insensitive to the value of ϵ.

4 Final comments

Models such as (1) fit into framework considered by Chernozhukov & Hong (2002), Donald & Paarsch (2002), and Hirano & Porter (2003), who consider asymptotic theory for estimation in models with parameter-dependent support. Unlike classical statistical models (where the support is independent of the parameters), maximum likelihood estimation does not have any particular asymptotic optimality. Both Chernozhukov & Hong (2002) and Hirano & Porter (2003) consider the asymptotics of Bayes estimators for a given loss function and prior distribution on the parameter space. Such estimators have the advantage of being admissible (with respect to loss function) and have asymptotic distributions that are independent of the prior distribution. Of course, these admissibility results are dependent on the model being correctly specified although one might expect Bayes estimators to be useful more generally.

It is also possible to extend the results to estimators $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n)$ minimizing

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}; \boldsymbol{\zeta}) \quad \text{subject to } Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi} \text{ for } i = 1, \cdots, n$$

where $\rho(w; \boldsymbol{\zeta})$ is a three times differentiable (or otherwise sufficiently smooth) function in $\boldsymbol{\zeta}$; the support of the response depends on $\boldsymbol{\beta}$ but not on the "nuisance" parameter $\boldsymbol{\theta}$. We assume that for some matrices $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$, we have

$$\begin{aligned} E\left[\nabla_\zeta \rho(W_i; \boldsymbol{\theta})\right] &= 0, \\ E\left[\nabla_\zeta \rho(W_i; \boldsymbol{\theta}) \nabla_\zeta^T \rho(W_i; \boldsymbol{\theta})\right] &= A(\boldsymbol{\theta}), \\ \text{and} \quad E\left[\nabla_{\zeta\zeta} \rho(W_i; \boldsymbol{\theta})\right] &= B(\boldsymbol{\theta}) \end{aligned}$$

where ∇_ζ and $\nabla_{\zeta\zeta}$ are, respectively, the gradient and Hessian operators with respect to $\boldsymbol{\zeta}$. Then under additional regularity conditions (including, for example, appropriate modifications of (A0)–(A6)), we have the same limiting behaviour for $a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ as given in Theorem 1; moreover,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, B^{-1}(\boldsymbol{\theta}) A(\boldsymbol{\theta}) B^{-1}(\boldsymbol{\theta})\right)$$

with the two limiting distributions being independent.

We can also consider non-parametric estimation of boundaries by fitting parametric models (for example, polynomials) locally in the neighbourhood of a given point; the asymptotic behaviour of such non-parametric estimators can be determined using the theory discussed in sections 2 and 3 with appropriate modifications. An alternative non-parametric approach to boundary estimation is given by Bouchard *et al.* (2003). This approach defines the boundary as a linear combination of kernel functions with non-negative

weights estimated as the solution of a linear programming problem. In the context of production frontier estimation, a good survey of non-parametric estimation methods can be found in Florens & Simar (2002).

References

Aigner, D.J. & Chu, S.F. (1968) On estimating the industry production function. *American Economic Review.* **58**, 826-839.

An, H.Z. & Huang, F.C. (1993) Estimation for regressive and autoregressive models with nonnegative residual errors. *Journal of Time Series Analysis.* **14**, 179-191.

Anděl, J. (1989) Nonnegative autoregressive processes. *Journal of Time Series Analysis.* **10**, 1-11.

Bouchard, G., Girard, S., Iouditski, A. & Nazin, A. (2003) Linear programming problems for frontier estimation. *Rapport de Recherche INRIA RR–4717.*

Chernozhukov, V. (2000) Conditional extremes and near extremes: estimation, inference and economic applications. Ph.D. thesis, Department of Economics, Stanford University.

Chernozhukov, V. & Hong, H. (2002) Likelihood inference in a class of nonregular econometric models. *MIT Department of Economics Working Paper 02-05.*

Donald, S.G. & Paarsch, H.J. (2002) Superconsistent estimation and inference in structural econometric models using extreme order statistics. *Journal of Econometrics.* **109**, 305-340.

Feigen, P.D. & Resnick, S.I. (1994) Limit distributions for linear programming time series estimators. *Stochastic Processes and their Applications.* **51**, 135-166.

Fiacco, A.V. & McCormick, G.P. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* SIAM, Philadelphia.

Florens, J.-P. & Simar, L. (2002) Parametric approximations of nonparametric frontiers. (unpublished manuscript)

Geyer, C.J. (1994) On the asymptotics of constrained M-estimation. *Annals of Statistics.* **22**, 1993-2010.

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

Hirano, K. & Porter, J. (2003) Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica.* **71**, forthcoming.

Kallenberg, O. (1983) *Random Measures.* (third edition) Akademie-Verlag, Berlin.

Knight, K. (1999) Epi-convergence in distribution and stochastic equisemicontinuity. (unpublished manuscript)

Knight, K. (2001) Limiting distributions of linear programming estimators. *Extremes.* **4**, 87-104.

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

Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

Nielsen, B. & Shephard, N. (2003) Likelihood analysis of a first order autoregressive model with exponential innovations. *Journal of Time Series Analysis.* **24**, 337-344.

Pflug, G.Ch. (1994) On an argmax-distribution connected to the Poisson process, in *Asymptotic Statistics* (P. Mandl & M. Hušková, eds) 123-130, Physica-Verlag, Heidelberg.

Pflug, G. Ch. (1995) Asymptotic stochastic programs. *Mathematics of Operations Research.* **20**, 769-789.

Portnoy, S. & Jurečková, J. (1999) On extreme regression quantiles. *Extremes.* **2**, 227-243.

Smith, R.L. (1994) Nonregular regression. *Biometrika.* **81**, 173-183.