On the asymptotic distribution of the L_∞ estimator in linear regression

KEITH KNIGHT

Department of Statistical Sciences, University of Toronto Toronto, Ont. M5G 1Z5

E-mail: keith@utstat.toronto.edu

Abstract

The L_∞ (or Chebyshev) estimator minimizes the maximum absolute residual and is potentially useful in situations where the noise distribution is known to have bounded support. In this paper, we derive the asymptotic distribution of this estimator in cases where the noise distribution has bounded and unbounded support. We also discuss the lack of robustness and stability of the estimator and describe how to improve its robustness by convex regularization.

Key words: L_∞ estimator, Chebyshev norm, Poisson processes, linear programming, convex regularization.
JEL classification: C13, C21

1 Introduction

Consider the linear regression model

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

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 $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. random variables. (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.)

The L_∞ estimator (also known as the Chebyshev or minimax estimator) $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ is defined as the minimizer of the objective function

$$\max_{1 \le i \le n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}|. \tag{2}$$

Intuitively, $\widehat{\boldsymbol{\beta}}_n$ will be a good estimator of $\boldsymbol{\beta}$ in cases where $\{\varepsilon_i\}$ have bounded support with non-trivial probability mass near the boundaries of the support. In cases where the support is unbounded or where this is negligible probability mass near the boundary of the support, $\widehat{\boldsymbol{\beta}}_n$ will have a very slow convergence rate or may even be inconsistent.

Although it is not extensively used in practice, the L_∞ estimator definitely has a niche, especially in certain applications in the physical and environmental sciences; see, for example, the papers by James (1983a, 1983b), Brenner (2002), Zolghadri and Henry (2004), Bertsch *et al.* (2005), and Qi (2015). In finance, Jaschke (1998) uses L_∞ estimation in the context of computing arbitrage bounds where the minimax absolute error should be approximately equal to transaction costs; see also Jaschke and Küchler (2001). There is also a considerable literature in signal processing and systems engineering on estimation with bounded noise (Milanese and Belforte, 1982; Mäkilä, 1991; Tse *et al.*, 1993; Akçay *et al.*, 1996; Beck and Eldar, 2007). L_∞-norm minimization arises in the context of near lossless compression of images (Alecu *et al.*, 2006) as well as estimation in inverse problems with quantization errors (Clason, 2012; Dou *et al.*, 2013). The least median of squares (LMS) estimator (Rousseeuw, 1984) as well as the least quartile difference (LQD) estimator (Croux *et al.*, 1994) can be viewed within the L_∞ framework; both the LMS and the LQD estimators can be shown to be a L_∞ estimator of some (random) half sample. More recently, Castillo *et al.* (2009) consider combining least squares, L_1, and L_∞ estimators while Berenguer-Rico *et al.* (2019) consider models in which the LMS estimator is a maximum likelihood estimator. Du *et al.* (2019) consider L_∞ estimation with a LASSO (Tibshirani, 1996) penalty. Hayashi and Yoshida (2020) consider subspace proximity testing using the L_∞ norm.

A cautionary note: This paper should not be viewed as a methodological paper and as such, does not address inferential issues; this paper merely considers asymptotic properties of L_∞ estimation. However, in situations where it is appropriate (for example, image processing), this paper does provide (in section 5) some methodology for mitigating the negative aspects of L_∞ estimation.

2 Preliminaries

Analysis of the L_∞ estimator is greatly facilitated by viewing it as the solution of the following linear program:

$$\begin{aligned} \text{minimize } \gamma \quad \text{subject to} \quad & \boldsymbol{x}_i^T \boldsymbol{\phi} + \gamma \geq Y_i \quad (i = 1, \cdots, n) \\ \text{and} \quad & \boldsymbol{x}_i^T \boldsymbol{\phi} - \gamma \leq Y_i \quad (i = 1, \cdots, n) \end{aligned} \tag{3}$$

The solution of the linear program (3) can, in fact, be viewed as a regression quantile (Koenker, 2005) solution for the quantile $\tau = 1$ (that is, the conditional maximum) for the augmented data $\{(\boldsymbol{x}_i^*, Y_i^*) : i = 1, \cdots, 2n\}$ where

$$Y_i^* = \begin{cases} Y_i & \text{for } i = 1, \cdots, n \\ -Y_{i-n} & \text{for } i = n+1, \cdots, 2n \end{cases}$$

and

$$\boldsymbol{x}_i^* = \begin{cases} (1, \boldsymbol{x}_i^T)^T & \text{for } i = 1, \cdots, n \\ (1, -\boldsymbol{x}_{i-n}^T)^T & \text{for } i = n+1, \cdots, 2n. \end{cases}$$

This suggests that the properties of the L_∞ estimator are similar to those of the estimator of the extreme regression quantile.

Alternatively, the L_∞ estimator can be computed using Lawson's algorithm (Lawson, 1961; Cline, 1972), which is an iteratively reweighted least squares (IRLS) algorithm for computing L_∞ estimates. (In fact, Daubechies *et al.* (2010) suggest that Lawson's algorithm may be the original IRLS algorithm.) This algorithm generates a sequence of estimates $\widehat{\boldsymbol{\beta}}^{(k)}$ (for $k = 0, 1, 2, \cdots$) by minimizing

$$\sum_{i=1}^{n} w_i^{(k)} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})^2$$

where the weights $\{w_i^{(k)}\}$ are defined iteratively by

$$w_i^{(k)} = \frac{w_i^{(k-1)} \left| Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}^{(k-1)} \right|}{\sum_{j=1}^{n} w_j^{(k-1)} \left| Y_j - \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}^{(k-1)} \right|}$$

with $w_i^{(0)} = 1$ for all i. Although Lawson's algorithm (as given above) is not guaranteed to converge, it can be modified so that convergence is guaranteed, in which case, convergence is linear; in practice, convergence of Lawson's algorithm is painfully slow.

It follows from the linear programming representation of the L_∞ estimator that the maximum absolute residual will be attained at $p + 1$ points where p is the dimension of $\boldsymbol{x}_i$. It is this property that, to a large extent, makes the L_∞ estimator unattractive in practice beyond some special cases. The following example illustrates this main pitfall, namely the misfitting of some or all of the data in order to attain a minimum uniform error.

EXAMPLE 1. We will consider the well-known data from a simulated motorcycle crash as presented by Silverman (1985). These data, which are described in detail by Schmidt *et al.* (1981), consist of 133 accelerometer readings taken over time. If A_i is acceleration measured at time t_i, we assume the model

$$\begin{aligned} A_i &= \beta_0 + \sum_{j=1}^{15} \beta_j \phi_j(t_i) + \varepsilon_i \\ &= g(t_i) + \varepsilon_i \quad \text{for } i = 1, \cdots, 133 \end{aligned}$$

where the functions $\phi_1, \cdots, \phi_{15}$ are B-spline functions. Figure 1 shows the least squares and L_∞ estimates of g while Figure 2 shows the least squares estimate as well as an estimate computed using a variation of Lawson's algorithm with 1000 iterations; the maximum absolute residual for the Lawson estimate (62.89166) is only slightly greater than that of the L_∞ estimate (62.89160). Clearly, the L_∞ estimate misfits the data especially for smaller values of t while the Lawson and least squares estimates are very similar for these values of t. For larger values of t, the L_∞ and Lawson estimates are essentially equal and quite different from the least squares estimate; for larger values of $\{t_i\}$, the noise in the corresponding $\{A_i\}$ makes it less clear what the correct form of g should be although the least squares estimate is certainly aesthestically more pleasing. A similar phenomenon for L_∞ image compression is noted by Chuah *at al.* (2013). The lack of stability (or lack of robustness) of the L_∞ estimator will be discussed in section 4 and methods for stabilizing the L_∞ estimator will be discussed in section 5.

Some insight into the asymptotic behaviour of the L_∞ estimator can be obtained by thinking of $\widehat{\boldsymbol{\beta}}_n$ minimizing (2) as the limit L_r estimators $\widehat{\boldsymbol{\beta}}_n^{(r)}$ minimizing

$$\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}|^r$$

taking $r \to \infty$. To simplify the computations, assume that the noise $\{\varepsilon_i\}$ have common density

$$f_\alpha(t) = \frac{\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma^2(\alpha)}(1-t^2)^{\alpha-1} \quad \text{for } |t| \le 1 \tag{4}$$

and that $\{\boldsymbol{x}_i\}$ satisfy

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T &\to C \\ \frac{1}{n}\max_{1\le i\le n} \boldsymbol{x}_i^T \boldsymbol{x}_i &\to 0 \end{aligned}$$

where C is positive definite. The assumptions on $\{\boldsymbol{x}_i\}$ are standard for asymptotic normality while the parameter α in the noise density f_α describes the concentration of probability mass near the endpoints ± 1. Under these assumptions, for each fixed $r \ge 1$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(r)} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2(r,\alpha) C^{-1})$$

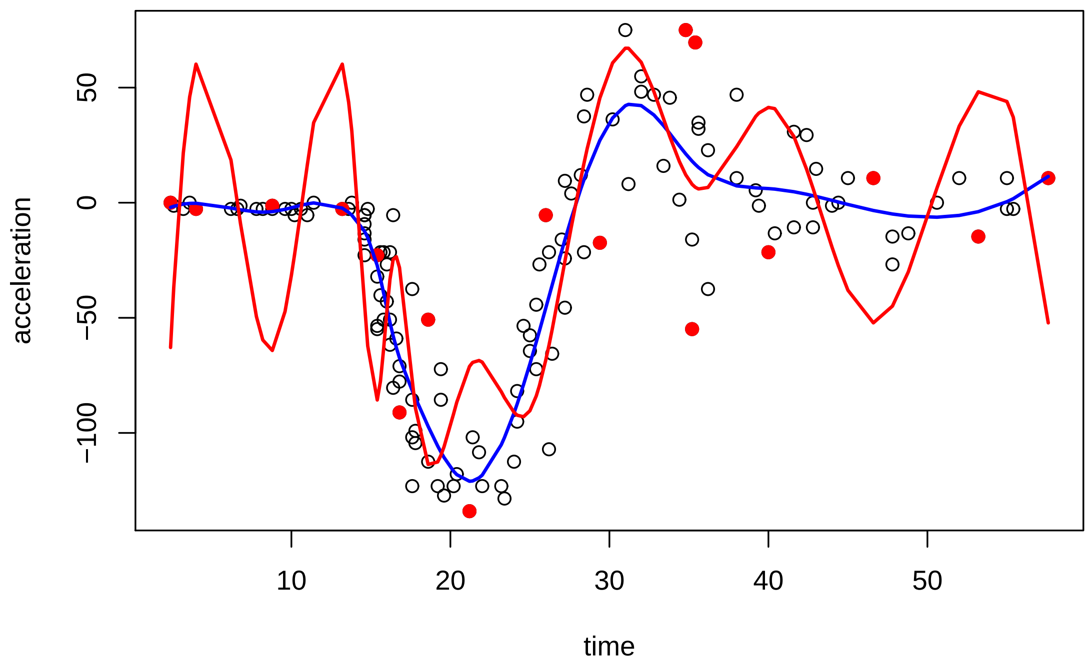

Figure 1: Motorcycle data with least squares (blue) and L_∞ (red) estimates of g; the red points are those where the maximum absolute residual for the L_∞ estimate is attained.

where

$$\sigma^2(r,\alpha) = \frac{2^{2\alpha-3}\Gamma(r-1/2)\Gamma(\alpha)\Gamma^2(\alpha+r/2-1/2)}{\Gamma(2\alpha)\Gamma(\alpha+r-1/2)\Gamma^2(r/2+1/2)}.$$

(Lai and Lee (2005) give a comprehensive survey of the asymptotics of L_r estimators in regression.) For large values of r, $\sigma^2(r,\alpha)$ behaves like a multiple of $r^{\alpha-2}$; taking the limit of $\sigma^2(r,\alpha)$ as $r\to\infty$ for each fixed $\alpha>0$, we get

$$\lim_{r\to\infty}\sigma^2(r,\alpha) = \begin{cases} 0 & \text{if } \alpha<2 \\ 1/12 & \text{if } \alpha=2 \\ \infty & \text{if } \alpha>2. \end{cases}$$

(If $\{\varepsilon_i\}$ are normally distributed then the corresponding asymptotic variance behaves like a multiple of $2^r/r$ as $r\to\infty$.) These heuristics (which take the limits as r and n tend to infinity in the wrong order) suggest that the convergence rate of the L_∞ estimator increases with the concentration of probability mass near the endpoints of the noise distribution. It is interesting to note that in the location case Rider (1957) (see also Harter (1975a, 1975b) and Sposito (1990)) recommends as a rule-of-thumb using the sample midrange as a estimator of location for distributions with kurtosis smaller than 2.2; the density f_α in (4) has kurtosis

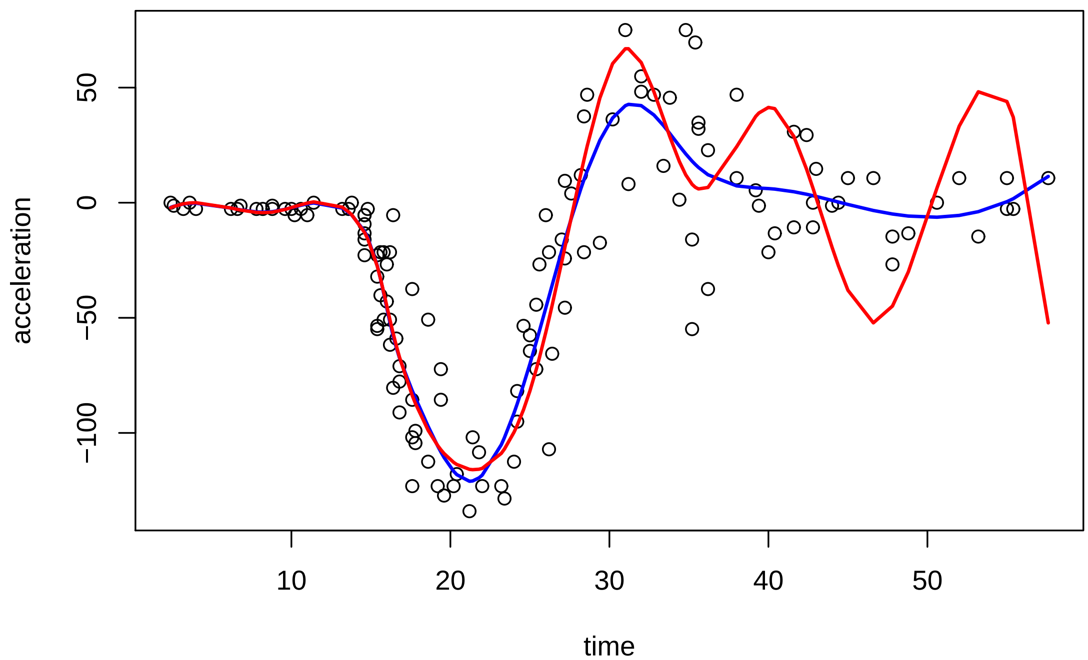

Figure 2: Motorcycle data with least squares (blue) and Lawson algorithm (red) estimates of g.

$3(1 + 2\alpha)/(3 + 2\alpha)$. Thus the limit of $\sigma^2(r, \alpha)$ (as $r \to \infty$) is 0 for $\alpha < 2$ or for densities f_α with kurtosis smaller than $15/7 \approx 2.14$. We will see that the asymptotics of the Chebshev estimator do depend strongly on the parameter α in (4) where α describes the concentration of probability mass near the boundaries ± 1 of the support of f_α in (4).

There has been some research on the asymptotics of the L_∞ estimator although most this has been restricted to special cases. For example, Aoki (1965) and Broffitt (1974) consider the distribution of the midrange (which is the L_∞ estimator in the location model), Akçay and At (2006) consider the distribution of the L_∞ estimator in a simple scalar regression model. Schechtmann and Schechtmann (1986) consider maximum likelihood estimation for a simple linear regression model with uniformly distributed errors, which is similar in spirit to L_∞ estimation.

In the next section, we will use the approach of Knight (2001) and Chernozhukov (2005) (which exploits the linear programming representation (3)) to derive limiting distributions for the L_∞ estimator.

3 Asymptotics for i.i.d. noise

In this section, we will assume that the noise $\{\varepsilon_i\}$ are i.i.d. and bounded in absolute value. In particular, we will assume that there exists a typically unknown parameter γ_0 such that

$$P(-\gamma_0 < \varepsilon_i < \gamma_0) = 1 \quad \text{for } i = 1, \cdots, n$$

and that $\{\varepsilon_i\}$ is "boundary visiting" in the sense that for any $\delta > 0$,

$$P(-\gamma_0 + \delta < \varepsilon_i < \gamma_0 - \delta) < 1 \quad \text{for } i = 1, \cdots, n.$$

The asymptotics of the L_∞ estimator depend on the behaviour of the distribution function of $\{\varepsilon_i\}$ close to the endpoints $\pm\gamma_0$.

If F is the distribution function of $\{\varepsilon_i\}$, we define a function G by

$$G(t) = \begin{cases} F(\gamma_0 + t) - 1 & \text{for } -\gamma_0 \leq t < 0 \\ F(-\gamma_0 + t) & \text{for } 0 \leq t \leq \gamma_0. \end{cases} \tag{5}$$

Clearly, G is a non-decreasing function. To obtain non-degenerate limiting distributions, we will assume that for some sequence of constants $\{a_n\}$ with $a_n \to \infty$, we have

$$nG(t/a_n) \to \psi(t) \quad \text{as } n \to \infty \tag{6}$$

where ψ is a non-decreasing function of the form

$$\psi(t) = \begin{cases} \kappa t^\alpha & \text{for } t \geq 0, \\ -(1-\kappa)(-t)^\alpha & \text{for } t < 0 \end{cases} \tag{7}$$

where κ and α are both positive numbers. In the "standard" case where we assume $\{\varepsilon_i\}$ to have a continuous density f on $[-\gamma_0, \gamma_0]$ with $f(-\gamma_0) = \lambda^- > 0$ and $f(\gamma_0) = \lambda^+ > 0$, then $a_n = (\lambda^+ + \lambda^-)n$ and

$$\psi(t) = \begin{cases} \kappa t & \text{for } t \geq 0 \\ (1-\kappa)t & \text{for } t < 0 \end{cases}$$

so that $\alpha = 1$ and $\kappa = \lambda^-/(\lambda^+ + \lambda^-)$. More generally, $a_n = n^{1/\alpha}L(n)$ where L is a slowly varying (at infinity) function. For the family of densities $\{f_\alpha(t)\}$ defined in (4), the parameter α corresponds to the parameter α in (7) with $\kappa = 1/2$ and $a_n = k(\alpha)n^{1/\alpha}$ where $k(\alpha)$ does not depend on n.

Given these assumptions on the distribution of $\{\varepsilon_i\}$, it is straightforward to derive a point process convergence result for the number of $\{\varepsilon_i\}$ lying within $O(a_n^{-1})$ of $\pm\gamma_0$. Define

$$\tau_n(x) = \begin{cases} a_n(x - \gamma_0) & \text{for } 0 \leq x \leq \gamma_0, \\ a_n(x + \gamma_0) & \text{for } -\gamma_0 \leq x < 0. \end{cases} \tag{8}$$

Then the point process

$$\sum_{i=1}^{n} I\left\{\tau_n(\varepsilon_i) \in A\right\}$$

converges to a Poisson process whose mean measure is given by

$$\int_A \psi'(t)\, dt.$$

We will make the following assumptions about the noise $\{\varepsilon_i\}$ and the design $\{\boldsymbol{x}_i\}$:

(A1) $\{\varepsilon_i\}$ are i.i.d. random variables on $[-\gamma_0, \gamma_0]$ with distribution function F where G defined in (5) satisfies (6) for some sequence $\{a_n\}$ and some non-decreasing function ψ of the form (7) where $\alpha > 0$ and $0 < \kappa < 1$.

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

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

where μ is not concentrated on a lower dimensional subspace.

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. 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). Under conditions (A1) and (A2), it is easy to verify that the point process

$$M_n(A \times B) = \sum_{i=1}^{n} I\left\{\tau_n(\varepsilon_i) \in A, C_n^{-1}\boldsymbol{x}_i \in B\right\} \tag{9}$$

converges in distribution with respect to the vague topology on measure (Kallenberg, 1983) to a Poisson process M whose mean measure is given by

$$E[M(A \times B)] = \left\{\int_A \psi'(t)\, dt\right\} \mu(B).$$

The convergence in distribution of M_n to M is equivalent to the following: For any bounded continuous function g with compact support, we have

$$\sum_{i=1}^{n} g(\tau_n(\varepsilon_i), C_n^{-1}\boldsymbol{x}_i) \xrightarrow{d} \int g(w, \boldsymbol{x})\, M(dw \times d\boldsymbol{x}).$$

In the case where the support of $\{C_n^{-1}\boldsymbol{x}_i\}$ is bounded, conditions (A1) and (A2) are sufficient to establish the limiting distribution of $a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. However, in the case of unbounded $\{C_n^{-1}\boldsymbol{x}_i\}$, we need to make stronger assumptions both about the function $G(t)$ as well as the behaviour of $\{C_n^{-1}\boldsymbol{x}_i\}$.

(A3) The function G defined in (5) satisfies

$$nG(t/a_n) = \psi(t)\{1 + r_n(t)\}$$

where for each $\boldsymbol{u}$,

$$\max_{1\le i\le n} |r_n(\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i)| \to 0.$$

(A4) For all $\boldsymbol{u}$,

$$\int |\boldsymbol{u}^T\boldsymbol{x}|^\alpha \,\mu(d\boldsymbol{x}) < \infty.$$

Moreover, for each $\boldsymbol{u}$,

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

Clearly, conditions (A3) and (A4) are redundant given (A1) and (A2) if $\{C_n^{-1}\boldsymbol{x}_i\}$ is bounded. If, for example, $G(t) = t^\alpha\{1 + kt^\delta\}$ for some $\delta > 0$ then $a_n = n^{1/\alpha}$ and so condition (A3) becomes

$$\max_{1\le i\le n} \frac{|\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i|^\delta}{n^{\delta/\alpha}} \to 0$$

which is equivalent to

$$\frac{1}{n}\max_{1\le i\le n} |\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i|^\alpha \to 0$$

as required in (A4).

The key tools used in deriving the limiting distributions are epi-convergence in distribution (Pflug, 1994, 1995; Geyer, 1994, 1996; Knight, 1999, 2001; Chernozhukov, 2005; Chernozhukov and Hong, 2004) and point process convergence for extreme values (Kallenberg, 1983; Leadbetter *et al*, 1983). 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 $a_1, \cdots, a_k$,

$$\begin{aligned} &P\left\{\inf_{\mathbf{u}\in R_1} Z(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z(\boldsymbol{u}) > a_k\right\} \\ &\quad\le \liminf_{n\to\infty} P\left\{\inf_{\mathbf{u}\in R_1} Z_n(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z_n(\boldsymbol{u}) > a_k\right\} \end{aligned}$$

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

See the monograph by Molchanov (2005) for more details. Epi-convergence in distribution gives us an elegant way of proving convergence in distribution of "argmin" (and "argmax") estimators: suppose that $\boldsymbol{U}_n = \arg\min(Z_n)$ where $Z_n \xrightarrow{e-d} Z$ and $\boldsymbol{U}_n = O_p(1)$; then $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U} = \arg\min(Z)$ provided that the latter argmin is unique. 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 $\arg\min(Z)$ is unique then $\boldsymbol{U}_n = O_p(1)$ is implied by $Z_n \xrightarrow{e-d} Z$.

THEOREM 1. *Assume the model (1) and conditions (A1) through (A4). Then if $\widehat{\boldsymbol{\beta}}_n$ is the solution of (3),*

$$
\left(a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n(\widehat{\gamma}_n - \gamma_0)\right) \xrightarrow{d} (\boldsymbol{U}, V)
$$

where $(\boldsymbol{U}, V)$ minimizes v subject to the constraints

$$
\begin{aligned}
\Gamma_i &\geq \boldsymbol{u}^T \boldsymbol{X}_i - v \quad \text{for } i = 1, 2, 3, \cdots \\
\Gamma_i &\leq \boldsymbol{u}^T \boldsymbol{X}_i + v \quad \text{for } i = -1, -2, -3, \cdots
\end{aligned}
$$

where

(i) $\Gamma_i = \psi^{-1}(E_1 + \cdots + E_i)$ *for* $i \geq 1$ *and* $\Gamma_i = \psi^{-1}(-E_{-1} - E_{-2} - \cdots - E_i)$ *for* $i \leq -1$ *where* $\{E_i : |i| \geq 1\}$ *is a sequence of i.i.d. unit mean exponential random variables.*

(ii) $\{\boldsymbol{X}_i : |i| \geq 1\}$ *is a sequence of i.i.d. random vectors whose distribution is* μ.

(iii) $\{\boldsymbol{X}_i\}$ *is independent of* $\{E_i\}$.

Proof. Write $\boldsymbol{U}_n = a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ and $V_n = a_n(\widehat{\gamma}_n - \gamma_0)$ and note that $(\boldsymbol{U}_n, V_n)$ minimizes v subject to

$$
\begin{aligned}
a_n(\varepsilon_i - \gamma_0) - v &\leq \boldsymbol{u}^T C_n^{-1} \boldsymbol{x}_i \\
&\leq a_n(\varepsilon_i + \gamma_0) + v \quad \text{for } i = 1, \cdots, n
\end{aligned} \tag{10}
$$

Define $\varphi_n(\boldsymbol{u}, v)$ to be 0 when the constraints in (10) are all satisfied and $+\infty$ otherwise. Then $(\boldsymbol{U}_n, V_n)$ minimizes

$$
Z_n(\boldsymbol{u}, v) = v + \varphi_n(\boldsymbol{u}, v) \tag{11}
$$

where Z_n is a convex function. Thus it suffices to show that $Z_n \xrightarrow{e-d} Z$ where Z has unique minimizer with probability 1 and for this, it suffices to show finite dimensional convergence of Z_n to Z provided that Z is finite on an open set; this latter fact is true since $\Gamma_i \sim i^{1/\alpha}$ (with probability 1) as $i \to \infty$ and so by the first Borel-Cantelli lemma

$$\begin{aligned} P(\boldsymbol{u}^T \boldsymbol{X}_i - v > \Gamma_i \text{ for infinitely many } i \geq 1) &= 0 \\ \text{and} \quad P(\boldsymbol{u}^T \boldsymbol{X}_i + v < \Gamma_i \text{ for infinitely many } i \leq -1) &= 0 \end{aligned}$$

for all $(\boldsymbol{u}, v)$ since $E[|\boldsymbol{u}^T \boldsymbol{X}_i|^\alpha] < \infty$. Thus for a given $(\boldsymbol{u}, v)$, at most a finite number of constraints are violated, the rest being trivially satisfied. Thus for a given $(\boldsymbol{u}, v)$, there exists some $t > 0$ such that all the constraints are satisfied for $(t\boldsymbol{u}, tv)$. We then take a finite number of such points whose convex hull contains an open set; since Z is finite at each point, it will be finite on the open set by convexity.

To show finite dimensional weak convergence of φ_n, it suffices to show that

$$P\left[\varphi_n(\boldsymbol{u}_1, v_1) = 0, \cdots, \varphi_n(\boldsymbol{u}_k, v_k) = 0\right] \to P\left[\varphi(\boldsymbol{u}_1, v_1) = 0, \cdots, \varphi(\boldsymbol{u}_k, v_k) = 0\right]$$

where $\varphi(\boldsymbol{u}, v) = 0$ if $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i - v$ for all $i \geq 1$ and $\Gamma_i \leq \boldsymbol{u}^T \boldsymbol{X}_i + v$ for all $i \leq -1$ and $\varphi(\boldsymbol{u}, v) = +\infty$ otherwise. Exploiting the convergence in distribution of ν_n to the Poisson random measure ν, we have

$$\begin{aligned} & P\left\{\varphi_n(\boldsymbol{u}_1, v_1) = 0, \cdots, \varphi_n(\boldsymbol{u}_k, v_k) = 0\right\} \\ = & \prod_{i=1}^n P\left\{-\gamma_0 + a_n^{-1} \max_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T C_n^{-1} \boldsymbol{x}_i - v_j\right) \leq \varepsilon_i \leq \gamma_0 + a_n^{-1} \min_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T C_n^{-1} \boldsymbol{x}_i + v_j\right)\right\} \\ = & \prod_{i=1}^n \left\{1 - G\left(a_n^{-1} \max_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T C_n^{-1} \boldsymbol{x}_i - v_j\right)_+\right) - G\left(a_n^{-1} \min_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T C_n^{-1} \boldsymbol{x}_i + v_j\right)_-\right)\right\} \\ \to & \exp\left[-\int \left\{\psi\left(\max_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T \boldsymbol{x} - v_j\right)_+\right) + \psi\left(\min_{1 \leq j \leq k} \left(\boldsymbol{u}_j^T \boldsymbol{x} + v_j\right)_-\right)\right\} \mu(d\boldsymbol{x})\right] \\ = & P\left\{\varphi(\boldsymbol{u}_1, v_1) = 0, \cdots, \varphi(\boldsymbol{u}_k, v_k) = 0\right\}. \end{aligned}$$

Hence for Z_n given in (11), we have $Z_n \xrightarrow{f-d} Z$ where

$$Z(\boldsymbol{u}) = v + \varphi(\boldsymbol{u}, v).$$

Finally, to show that Z has a unique minimizer (with probability 1) note that Z is minimized at a basic solution $(\boldsymbol{u}, v)$ satisfying for some $i_1, \cdots, i_{p+1}$

$$\begin{aligned} \boldsymbol{u}^T \boldsymbol{X}_{i_k} - v &= \Gamma_{i_k} \quad \text{for } i_k > 0, \\ \boldsymbol{u}^T \boldsymbol{X}_{i_k} + v &= \Gamma_{i_k} \quad \text{for } i_k < 0. \end{aligned}$$

Absolute continuity of the distribution of $\{\Gamma_i\}$ guarantees that with probability 1, no two basic solutions will be equal. $\square$

By combining the constraints for negative and positive values of $\{\Gamma_i\}$, we can also represent $(\boldsymbol{U}, V)$ as the minimizer of v subject to

$$\Gamma_i' \geq \boldsymbol{u}^T \boldsymbol{X}_i' - v \quad \text{for } i = 1, 2, \cdots$$

where $\Gamma_i' = (E_1 + \cdots + E_i)^{1/\alpha}$ with $\{E_i\}$ i.i.d. unit mean exponential random variables and $\{\boldsymbol{X}_i'\}$ i.i.d. random vectors with probability measure $\bar{\mu}$ defined in terms of μ by

$$\bar{\mu}(B) = \kappa\mu(B) + (1-\kappa)\mu(-B). \tag{12}$$

Using this representation, we can easily determine the density of $(\boldsymbol{U}, V)$; using properties of the dual problem, it follows that the density of $(\boldsymbol{U}, V)$ is

$$g(\boldsymbol{u}, v) = \frac{\alpha^{p+1}}{(p+1)!} \exp\left[-\int (\boldsymbol{u}^T\boldsymbol{x} - v)_+^{\alpha}\, \bar{\mu}(d\boldsymbol{x})\right] \int \cdots \int |D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1})| \prod_{i=1}^{p+1} \left\{(\boldsymbol{u}^T\boldsymbol{x}_i - v)_+^{\alpha-1}\, \bar{\mu}(d\boldsymbol{x}_i)\right\} \tag{13}$$

where $s_+ = sI(s > 0)$ is the positive part of s, $\bar{\mu}$ is the measure defined in (12) and $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1})$ is the determinant of the matrix

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{p+1} \end{pmatrix} \tag{14}$$

if $\mathbf{0}$ lies in the convex hull of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}$ and $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}) = 0$ otherwise.

The limiting density $g(\boldsymbol{u}, v)$ given in (13) 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 $\bar{\mu}$). However, it seems that this density does not provide as much insight into the limiting distribution as does the representation of $(\boldsymbol{U}, V)$ as the solution of a linear program.

EXAMPLE 2. Consider the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, \cdots, n$$

where $\{\varepsilon_i\}$ are i.i.d. with distribution F where G defined in (5) satisfies (6) and (7) for $\kappa = 1/2$ and $\alpha = 1$. We will consider the limiting distributions of $a_n(\widehat{\beta}_{1n} - \beta_1)$ in the cases where the limiting measure for $\{x_i\}$ is

(a) normal with mean 0 and variance 1;

(b) Laplace (double exponential) with mean 0 and variance 1;

(c) Rademacher with mass 1/2 at ± 1.

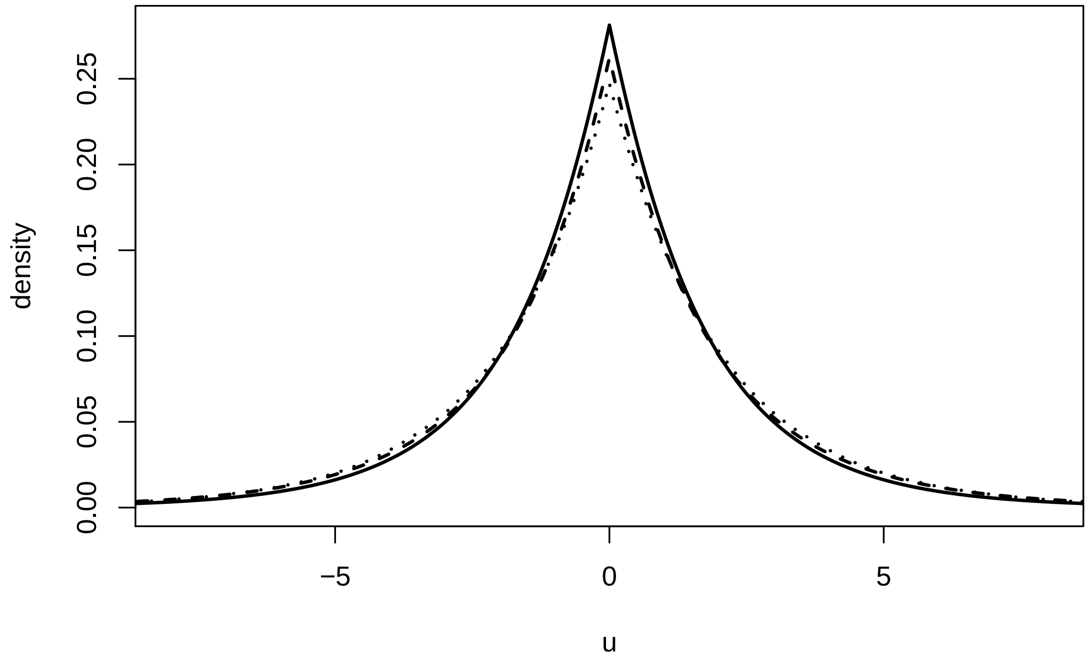

Figure 3: Limiting densities of $a_n(\widehat{\beta}_{1n} - \beta_1)$ for the three limiting design measures considered in Example 2. The solid line corresponds to the normal measure, the dashed line to the Laplace measure, and the dotted line the Rademacher measure.

The means and variances of these three distributions are the same; thus the limiting distributions of least squares, regression quantile (for fixed quantiles between 0 and 1), and many other estimators of $\boldsymbol{\beta}$ would be the same for the two designs. Figure 3 gives the limiting densities for the three limiting measures; the differences among the three limiting densities are quite small.

There are several possible variations on Theorem 1. For example, we may relax the assumption that $E[|\boldsymbol{u}^T \boldsymbol{X}_i|^\alpha]$ is finite for all $\boldsymbol{u}$ to allow $E[(\boldsymbol{u}^T \boldsymbol{X}_i)_+^\alpha]$ or $E[(\boldsymbol{u}^T \boldsymbol{X}_i)_-^\alpha]$ (or both) to be infinite for some $\boldsymbol{u}$. In such cases, the conclusion of Theorem 1 will still hold although the limiting distribution may put all its mass at 0 for certain components of $\boldsymbol{U}$. We can also extend Theorem 1 to the case where the behaviour of the distribution of $\{\varepsilon_i\}$ differs at the two endpoints $\pm\gamma_0$. For example, suppose that for some sequence $\{a_n\}$,

$$nP\left\{a_n(\gamma_0 - \varepsilon_i) \leq x\right\} \to x^\alpha$$

while

$$nP\left\{a_n(\varepsilon_i + \gamma_0) \leq x\right\} \to \infty$$

for any $x > 0$. This implies that

$$a_n \left(\min_{1 \leq i \leq n} \varepsilon_i + \gamma_0 \right) \xrightarrow{p} 0$$

and (more importantly) that the number of $\{\varepsilon_i\}$ falling within $O(a_n^{-1})$ of $-\gamma_0$ is unbounded as $n \to \infty$. In this case, we can effectively set $\Gamma_1, \Gamma_2, \cdots$ in Theorem 1 to 0, which means that the set of $(\boldsymbol{u}, v)$ satisfying the constraints $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i - v$ for $i \geq 1$ is

$$A = \left\{ (\boldsymbol{u}, v) : \mu\{\boldsymbol{x} : \boldsymbol{u}^T \boldsymbol{x} - v \leq 0\} = 1 \right\}.$$

Note that A always contains $\mathbf{0}$. This suggests that $(\boldsymbol{U}_n, V_n)$ converges in distribution to the minimizer of v subject to

$$\begin{aligned} \Gamma_i &\geq \boldsymbol{u}^T \boldsymbol{X}_i - v \quad \text{for } i \geq 1 \\ \text{and} \quad (\boldsymbol{u}, v) &\in A \end{aligned}$$

where $\Gamma_i = (E_1 + \cdots + E_i)^{1/\alpha}$ for i.i.d. unit mean exponential random variables $\{E_i\}$. Similarly, it is possible to extend Theorem 1 to the case where $\alpha = \infty$ in (7) (so that $\psi(t) = 0$ for $|t| < 1$ and $\psi(t) = \operatorname{sgn}(t)\infty$ for $|t| > 1$), in which case $a_n = L(n)$ where L is a slowly varying function. Assuming that $\{C_n^{-1} \boldsymbol{x}_i\}$ is bounded, Theorem 1 holds with $\Gamma_i = 1$ for $i \geq 1$ and $\Gamma_i = -1$ for $i \leq -1$.

We can also consider the asymptotics for i.i.d. noise that are unbounded but whose distributions have light tails. Condition (A1) on the distribution of $\{\varepsilon_i\}$ implies that both maxima and minima of $\varepsilon_1, \cdots, \varepsilon_n$ (suitably normalized) converge in distribution to the same type-III (or Weibull) distribution. We can also generalize to distributions whose extremes (minima and maxima) converge in distribution to the other two extreme value distributions. We will limit our discussion here to the case where the limiting distribution of the extremes is a type-I (or Gumbel) distribution; distributions in this domain of attraction have light (exponential) tails in contrast with distributions in the domain of attaction of a type-II (or Frechet) distribution, which have heavier tails. The extension to type-II distributions is straightforward given appropriate modifications to the regularity conditions.

Conditions (A1), (A3), and (A4) are generalized in the following way:

(A1′) For some sequences of constants $\{a_n\}$ and $\{b_n\}$

$$\begin{aligned} \lim_{n \to \infty} nP\left\{a_n \varepsilon_i > t + b_n\right\} &= \kappa \exp(-t) \\ \lim_{n \to \infty} nP\left\{a_n \varepsilon_i < -(t + b_n)\right\} &= (1 - \kappa) \exp(-t) \end{aligned}$$

for all t.

(A3′) For all $t > 0$,

$$nP\{a_n|\varepsilon_i| > t + b_n\} = \exp(-t)\{1 + r_n(t)\}$$

where for each $\boldsymbol{u}$,

$$\max_{1\le i\le n} |r_n(\boldsymbol{u}^T C_n^{-1}\boldsymbol{x}_i)| \to 0.$$

(A4′) For all $\boldsymbol{u}$,

$$\int \exp(\boldsymbol{u}^T\boldsymbol{x})\,\mu(d\boldsymbol{x}) < \infty.$$

Moreover, for each $\boldsymbol{u}$,

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

Condition (A1′) is satisfied (for example) by both the normal and Laplace (double exponential) distributions; for the normal distribution (with mean 0 and variance 1), we have $a_n = \sqrt{2\ln(n)}$ and $b_n = 2\ln(n) - \ln(4\pi\ln(n))/2$ while for an Exponential distribution with mean 1, we have $a_n = 1$ and $b_n = \ln(n)$.

THEOREM 2. *Assume the model (1) and conditions (A1′), (A2), (A3′), and (A4′). Then if $\widehat{\boldsymbol{\beta}}_n$ is the solution of (3),*

$$\left(a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n\widehat{\gamma}_n - b_n\right) \xrightarrow{d} (\boldsymbol{U}, V)$$

where $(\boldsymbol{U}, V)$ is the solution of the linear program:

$$\textit{minimize } v \textit{ subject to } \Gamma_i' \ge \boldsymbol{u}^T\boldsymbol{X}_i' - v \quad \textit{for } i = 1, 2, 3, \cdots$$

where

(i) $\Gamma_i' = -\ln(E_1 + \cdots + E_i)$ for $i \ge 1$ where $\{E_i\}$ is a sequence of i.i.d. unit mean exponential random variables.

(ii) $\{\boldsymbol{X}_i'\}$ is a sequence of i.i.d. random vectors whose distribution is $\bar{\mu}$.

(iii) $\{\boldsymbol{X}_i'\}$ is independent of $\{E_i\}$.

Proof. The proof of Theorem 2 is essentially the same as that of Theorem 1 with only minor modifications. The moment conditions for μ guarantee that the limiting objective function is finite on an open set. □

Using the same considerations as before, we can determine the limiting density under the conditions of Theorem 2. The density of $(\boldsymbol{U}, V)$ is

$$g(\boldsymbol{u}, v) = \frac{1}{(p+1)!} \exp\left[-\int \exp(v - \boldsymbol{u}^T\boldsymbol{x})\,\bar{\mu}(d\boldsymbol{x})\right] \int \cdots \int |D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1})| \prod_{i=1}^{p+1} \left\{\exp(v - \boldsymbol{u}^T\boldsymbol{x}_i)\,\bar{\mu}(d\boldsymbol{x}_i)\right\} \tag{15}$$

where $\bar{\mu}$ is defined as in (15) and $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1})$ is the determinant of the matrix (14) if $\mathbf{0}$ lies in the convex hull of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}$ with $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}) = 0$ otherwise.

4 Robustness issues

To this point, we have ignored the question of what the L_∞ estimator is actually estimating; this has not been an issue since we have assumed i.i.d. noise. However, more generally, this question is somewhat more complicated.

We will assume that, as a function of $\boldsymbol{x}$, the response is bounded between two functions $g^-(\boldsymbol{x})$ and $g^+(\boldsymbol{x})$, which represent the essential infimum and supremum of the response given a covariate value $\boldsymbol{x}$. Thus

$$g^-(\boldsymbol{x}_i) \le Y_i \le g^+(\boldsymbol{x}_i) \quad (i = 1, \cdots, n).$$

Thus, under mild conditions, the L_∞ estimator of $\boldsymbol{\beta}$ will converge in probability to the value of $\boldsymbol{\phi}$ that minimizes γ subject to the constraints

$$\begin{aligned} g^+(\boldsymbol{x}) &\le \boldsymbol{x}^T\boldsymbol{\phi} + \gamma \quad \text{for all } \boldsymbol{x} \\ g^-(\boldsymbol{x}) &\ge \boldsymbol{x}^T\boldsymbol{\phi} - \gamma \quad \text{for all } \boldsymbol{x} \end{aligned}$$

provided that this minimizer is unique. We will have a unique minimizer $(\boldsymbol{\beta}, \gamma_0)$ if there exist positive constants $\lambda_1, \cdots, \lambda_{p+1}$ with $\lambda_1 + \cdots + \lambda_{p+1} = 1$ and points $\boldsymbol{x}_1^\circ, \cdots, \boldsymbol{x}_{p+1}^\circ$ such that

$$\sum_{j=1}^{p^\circ} \lambda_j \boldsymbol{x}_j^\circ - \sum_{j=p^\circ+1}^{p+1} \lambda_j \boldsymbol{x}_j^\circ = \mathbf{0}$$

(where $1 \le p^\circ \le p$) and

$$\begin{aligned} g^+(\boldsymbol{x}_j) &= \boldsymbol{\beta}^T\boldsymbol{x}_j^\circ + \gamma_0 \quad \text{for } j = 1, \cdots, p^\circ \\ g^-(\boldsymbol{x}_j) &= \boldsymbol{\beta}^T\boldsymbol{x}_j^\circ - \gamma_0 \quad \text{for } j = p^\circ + 1, \cdots, p+1 \end{aligned}$$

with

$$\begin{aligned} g^+(\boldsymbol{x}) &\le \boldsymbol{x}^T\boldsymbol{\beta} + \gamma_0 \quad \text{for all } \boldsymbol{x} \\ g^-(\boldsymbol{x}) &\ge \boldsymbol{x}^T\boldsymbol{\beta} - \gamma_0 \quad \text{for all } \boldsymbol{x}. \end{aligned}$$

In general though, uniqueness of $\boldsymbol{\beta}$ requires fairly strong assumptions on the model, both in terms of the noise structure as well as any possible model misspecification.

In view of the above, it seems that Theorem 1 masks the lack of robustness of L_∞ estimation. The hypotheses of Theorem 1 assume that the distributions of $\{\varepsilon_i\}$ are homogeneous over the covariates $\{\boldsymbol{x}_i\}$; that is, each ε_i is potentially boundary visiting with no distributional dependence on $\boldsymbol{x}_i$ (at least in neighbourhoods of the boundaries $\pm\gamma_0$). However, the proof of Theorem 1 shows that the basic result depends mainly on the convergence of the point process M_n defined in (9); we can weaken the conditions on $\{\varepsilon_i\}$ and $\{\boldsymbol{x}_i\}$ (or more generally $\{C_n^{-1}\boldsymbol{x}_i\}$ while retaining the weak convergence of $\{M_n\}$ to some limiting point process M, which will give the constraints in the limiting linear program. For example, suppose that F_i is the distribution function of ε_i with G_i defined analogously to G in (5). If we assume that

$$nG_i(t/a_n) \to \begin{cases} \lambda^+(\boldsymbol{x}_i)t^\alpha & \text{for } t > 0 \\ -\lambda^-(\boldsymbol{x}_i)(-t)^\alpha & \text{for } t < 0. \end{cases}$$

where $\lambda^+(\boldsymbol{x}_i), \lambda^-(\boldsymbol{x}_i) \geq 0$ then (assuming appropriate additional regularity conditions) the point process M_n defined in (9) (setting $C_n = I$ for simplicity) converges weakly to M whose points are represented by

$$\begin{array}{ll} (\lambda^+(\boldsymbol{X}_i)^{-1/\alpha}\Gamma_i, \boldsymbol{X}_i) & \text{for } i \geq 1, \\ (\lambda^-(\boldsymbol{X}_i)^{-1/\alpha}\Gamma_i, \boldsymbol{X}_i) & \text{for } i \leq 1 \end{array}$$

where $\{\Gamma_i\}$ and $\{\boldsymbol{X}_i\}$ are as defined in Theorem 1. In the case where $\lambda^+(\boldsymbol{X}_i) = 0$ or $\lambda^-(\boldsymbol{X}_i) = 0$, we set $\lambda^\pm(\boldsymbol{X}_i)\Gamma_i = \pm\infty$ and hence we can remove the corresponding constraints from the limiting linear program; Theorem 1 will hold provided that the sets $\{\boldsymbol{x} : \lambda^+(\boldsymbol{x}) > 0\}$ and $\{\boldsymbol{x} : \lambda^-(\boldsymbol{x}) > 0\}$ both have positive μ-measure.

A more interesting case from a robustness point of view occurs when the noise $\{\varepsilon_i\}$ are boundary visiting only in the neighbourhoods of a finite number of points $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_q^*$; for example, this seems to be a good approximation for the motorcycle data considered in Example 1. (We assume that the matrix whose columns are $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_q^*$ has full rank.) Define disjoint neighbourhoods $B_1^*, \cdots, B_q^*$ of $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_q^*$, respectively, and assume that for some sequence $\{a_n\}$ the point processes

$$M_{nj}^*(A) = \sum_{\mathbf{x}_i \in B_j^*} I\{\tau_n(\varepsilon_i) \in A\} \quad (j = 1, \cdots, q) \tag{16}$$

(where τ_n is defined as in (8)) converge in distribution to independent Poisson processes $M_1^*, \cdots, M_q^*$ whose mean measures are given by

$$E[M_j^*(dt)] = \begin{cases} \lambda_j^+ \alpha t^{\alpha-1}\, dt & \text{for } t > 0 \\ \lambda_j^- \alpha(-t)^{\alpha-1}\, dt & \text{for } t < 0, \end{cases} \qquad (j = 1, \cdots, q) \tag{17}$$

where $\lambda_1^+, \cdots, \lambda_q^+, \lambda_1^-, \cdots, \lambda_q^- \geq 0$ with at least one strictly positive. We also assume that

$$\sum_{\mathbf{x}_i \notin \cup_j B_j^*} I\{\tau_n(\varepsilon_i) \in A\} \xrightarrow{p} 0$$

for all A. Define $M_n(A \times B)$ as in (9) setting $C_n = I$ (for simplicity). Then M_n converges in distribution to a Poisson process M^* whose mean measure is given by

$$E[M^*(dt \times B)] = \sum_{j=1}^{q} E[M_j^*(dt)] I(\boldsymbol{x}_j^* \in B).$$

In order to extend Theorem 1, we require at least $p+1$ of $\lambda_1^+, \cdots, \lambda_q^+, \lambda_1^-, \cdots, \lambda_q^-$ to be positive with at least one positive elements in each of the sets $\{\lambda_j^+\}$ and $\{\lambda_j^-\}$; in this case,

$$\left(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n(\widehat{\gamma}_n - \gamma_0)\right) \xrightarrow{d} (\boldsymbol{U}, V),$$

which is the minimizer of v subject to

$$\Gamma_i' \geq \boldsymbol{u}^T \boldsymbol{X}_i' - v \quad \text{for } i = 1, 2, \cdots$$

where

$$\Gamma_i' = \Lambda^{-1/\alpha}(E_1 + \cdots + E_i)^{1/\alpha} \tag{18}$$

with

$$\Lambda = \sum_{j=1}^{q} (\lambda_j^+ + \lambda_j^-)$$

and $\{E_i\}$ i.i.d. unit mean exponential random variables; $\{\boldsymbol{X}_i'\}$ are i.i.d. random vectors with probability measure $\bar{\nu}$ defined by

$$\bar{\nu}(B) = \frac{1}{\Lambda} \sum_{j=1}^{q} \left\{ \lambda_j^+ I(\boldsymbol{x}_j^* \in B) + \lambda_j^- I(\boldsymbol{x}_j^* \in -B) \right\}. \tag{19}$$

The positivity condition on $\{\lambda_j^+\}$, $\{\lambda_j^-\}$ implies that the measure $\bar{\nu}$ puts probability mass 0 on any lower dimensional subspace. In the case where fewer than $p+1$ of $\lambda_1^+, \cdots, \lambda_q^+, \lambda_1^-, \cdots, \lambda_q^-$ are positive, we will not necessarily have $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1)$ although we may have $a_n(\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_n - \boldsymbol{c}^T \boldsymbol{\beta}) = O_p(1)$ for $\boldsymbol{c}$ belonging to some lower dimensional subspace. Note, however, that $a_n(\widehat{\gamma}_n - \gamma_0) = O_p(1)$ if $\mathbf{0}$ lies in the convex hull of some $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s$ in the support of $\bar{\nu}$. In section 5, we will discuss a modification of L_∞ estimation that may, under these conditions, give estimators of $\boldsymbol{\beta}$ with $O_p(a_n^{-1})$ convergence when this fails for the L_∞ estimator.

EXAMPLE 3. Consider the simple linear regression model

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

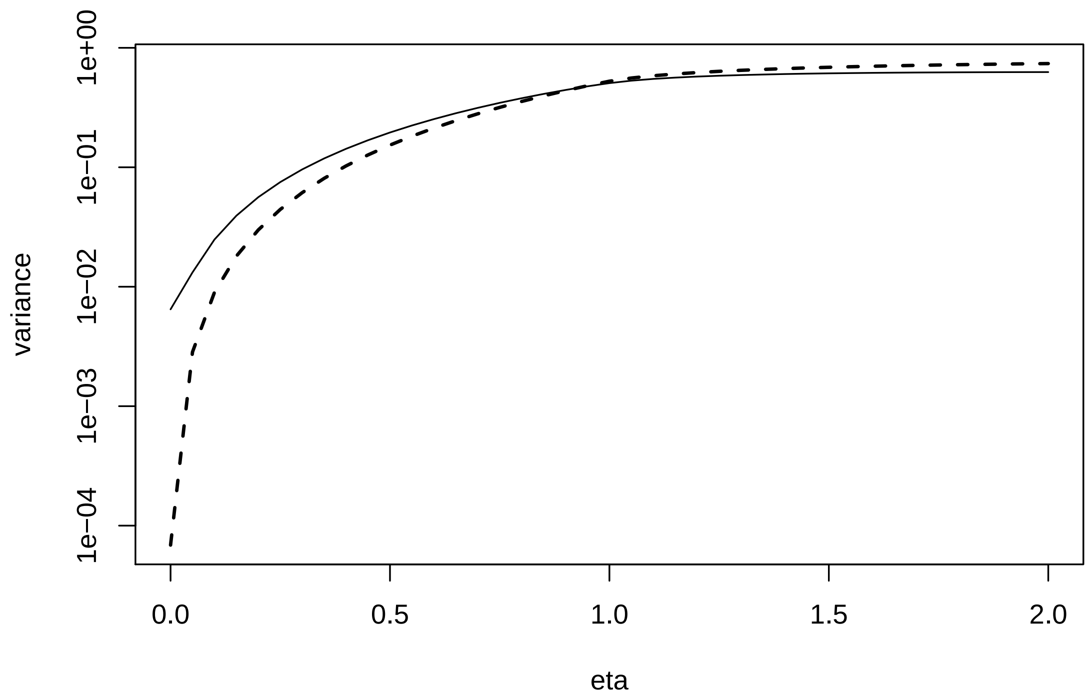

Figure 4: Variances (on a logarithmic scale) of the L_∞ estimator of β_1 in Example 3 for $n = 100$ (solid line) and $n = 1000$ (dashed line) for $0 \le \eta \le 2$.

where $x_i = i/n$ (so that $\{x_i\}$ are asymptotically uniformly distributed on $[0,1]$) and $\{\varepsilon_i\}$ are independent random variables with ε_i uniformly distributed on $[-x_i^\eta, x_i^\eta]$ for some $\eta \ge 0$. When $\eta = 0$, Theorem 1 holds with $\alpha = 1$ with $n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$ where $\boldsymbol{U}$ is defined as in Theorem 1. The situation becomes a little more complicated when $\eta > 0$. In this case, we do not have unique identification of $\boldsymbol{\beta}$; for a given (β_0, β_1), the maximum absolute error $\gamma_0 = 1$ can be attained for any $(\beta_0 - \phi, \beta_1 + \phi)$ satisfying $|\phi| \le \min(\eta, 1)$. If we take $\phi = 0$ then $x_1^* = 1$ and the point process M_{n1}^* defined in (16) converges to a Poisson process with $a_n = n^{1/2}$ and $\lambda_1^+ = \lambda_1^- = (4\eta)^{-1}$; however, because of non-uniqueness in the identification of $\boldsymbol{\beta}$, the extension of Theorem 1 does not hold. (In fact, the L_∞ estimator is not consistent in this case.) Figure 4 gives the variances of the L_∞ estimator of β_1 for $n = 100$ and $n = 1000$ (based on 100000 Monte Carlo replications) for values of η between 0 and 2. This plot illustrates how unstable the L_∞ estimator is, particularly for larger values of η. In particular, one might expect the distribution of the L_∞ estimator of β_1 to be mostly concentrated on the interval $[-\min(\eta, 1), \min(\eta, 1)]$; the variance of a uniform distribution on $[-\eta, \eta]$ is $\eta^2/3$ and so the distributions of the estimators have a greater larger variance than a uniform distribution on $[-\min(\eta,1),\min(\eta,1)]$. A natural question to ask is whether we can modify L_∞ estimation appropriately (that is, retaining the spirit of minimizing, at least approximately, the maximum absolute residual) to obtain a $n^{1/2}$-consistent estimator. For example, the L_r-estimator $\widehat{\boldsymbol{\beta}}_n^{(r)}$ is $n^{1/2}$-consistent and asympotically normal with mean $\mathbf{0}$ and variance-covariance matrix $D(r,\eta) = D_2^{-1}(r,\eta)D_1(r,\eta)D_2^{-1}(r,\eta)$ where

$$D_1(r,\eta) = \frac{1}{2r-1}\begin{pmatrix} \{2\eta(r-1)+1\}^{-1} & \{2\eta(r-1)+2\}^{-1} \\ \{2\eta(r-1)+2\}^{-1} & \{2\eta(r-1)+3\}^{-1} \end{pmatrix}$$

and

$$D_2(r,\eta) = \begin{pmatrix} \{\eta(r-2)+1\}^{-1} & \{\eta(r-2)+2\}^{-1} \\ \{\eta(r-2)+2\}^{-1} & \{\eta(r-2)+3\}^{-1} \end{pmatrix}.$$

The L_r-estimator approximates the L_∞ estimator for large values of r; however, some tedious calculations reveal that the larger eigenvalue of $D(r,\eta)$ is approximately $\eta^3 r^2/4$ for large r while the smaller eigenvalue tends to $\eta/16$.

In the following section, we will discuss a simple method for stabilizing the L_∞ estimator, thereby improving its robustness.

5 Stabilizing L_∞ estimation

As noted above, the L_∞ estimator can be viewed as an extreme ($\tau = 1$) quantile regression estimator on augmented data. We can obtain extensions of the L_∞ estimator by defining ρ to be a non-decreasing, non-negative function on the positive real line and defining $(\widehat{\boldsymbol{\beta}}_n, \widehat{\gamma}_n)$ to minimize

$$\sum_{i=1}^{n}\left\{\rho(\gamma + \boldsymbol{x}_i^T\boldsymbol{\phi} - Y_i) + \rho(Y_i + \gamma - \boldsymbol{x}_i^T\boldsymbol{\phi})\right\} \tag{20}$$

subject to the constraints

$$Y_i \le \boldsymbol{x}_i^T\boldsymbol{\phi} + \gamma \quad \text{for } i = 1,\cdots,n, \tag{21}$$

$$Y_i \ge \boldsymbol{x}_i^T\boldsymbol{\phi} - \gamma \quad \text{for } i = 1,\cdots,n. \tag{22}$$

The L_∞ estimator corresponds to the case where $\rho(x) = x$ for $x \ge 0$. In the case where $\rho(x) = x^2$, (20) becomes

$$2n\left\{\gamma^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\phi})^2\right\} \tag{23}$$

so that we are minimizing a combination of the maximum absolute residual and the residual sum of squares subject to the constraints (21) and (22). In the absence of a better name, we will refer to a general estimator minimizing (20) subject to (21) and (22) as an M-Chebyshev estimator and the estimator minimizing (23) subject to (21) and (22) as the LS-Chebyshev estimator.

For the "nice" noise configurations studied earlier, the asymptotic theory for M-Chebyshev estimators follows quite simply from the asymptotics for the L_∞ estimator. For simplicity, assume that ρ is strictly convex with a sufficiently smooth derivative ρ'. Then under the assumptions of Theorem 1 (plus some additional regularity conditions), we would have that

$$\left(a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n(\widehat{\gamma}_n - \gamma_0)\right) \xrightarrow{d} (\boldsymbol{U}, V)$$

where $(\boldsymbol{U}, V)$ minimizes

$$\{E[\rho'(\gamma_0 - \varepsilon_i)] + E[\rho'(\varepsilon_i + \gamma_0)]\}\, v + \{E[\rho'(\gamma_0 - \varepsilon_i)] - E[\rho'(\varepsilon_i + \gamma_0)]\} \int \boldsymbol{x}^T \boldsymbol{u}\, \mu(d\boldsymbol{x})$$

subject to the constraints

$$\Gamma_i' \geq \boldsymbol{u}^T \boldsymbol{X}_i' - v \quad \text{for } i = 1, 2, 3, \cdots$$

where $\{\Gamma_i'\}$ is defined as in (18) and $\{\boldsymbol{X}_i'\}$ are i.i.d. random vectors (independent of of $\{\Gamma_i'\}$) with distribution $\bar{\mu}$ defined in (19). The limiting distribution of the general estimator will thus be the same as that of the L_∞ estimator if

$$E[\rho'(\gamma_0 - \varepsilon_i)] = E[\rho'(\varepsilon_i + \gamma_0)],$$

which would occur, for example, if the distribution of $\{\varepsilon_i\}$ were symmetric around 0. The density of $(\boldsymbol{U}, V)$ has the same form as the density (13) except that $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1})$ is now defined to be the determinant of the matrix (14) if the convex hull of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}$ contains the vector

$$\frac{E[\rho'(\gamma_0 - \varepsilon_i)] - E[\rho'(\varepsilon_i + \gamma_0)]}{E[\rho'(\gamma_0 - \varepsilon_i)] + E[\rho'(\varepsilon_i + \gamma_0)]} \int \boldsymbol{x}\, \mu(d\boldsymbol{x})$$

with $D(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p+1}) = 0$ otherwise.

The previous result is somewhat disappointing since it essentially implies a first order asymptotic equivalence between L_∞ and M-Chebyshev estimation under i.i.d. noise. However, taking $\rho(x)$ in (20) to be strictly convex for $x > 0$ can lead to more stable estimation in cases where L_∞ estimation is unstable, namely when the noise is heterogeneous (which may lead to L_∞ estimation being inconsistent as discussed in section 4). To illustrate this, we will consider the more general setup discussed in section 4 (again setting $C_n = I$ for simplicity). In section 4, we assumed independent but non-identically distributed noise $\{\varepsilon_i\}$ whose distributions are boundary visiting for a finite number of points $\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_q^*$. Assume the convergence of the point processes $M_{n1}^*, \cdots, M_{nq}^*$ defined in (16) to independent Poisson processes with mean measures given in (17). If the measure $\bar{\nu}$ in (19) puts all its probability mass on a hyperplane $\{\boldsymbol{x} : \boldsymbol{c}^T\boldsymbol{x} = 0\}$ for some $\boldsymbol{c} \neq \boldsymbol{0}$ then the extension of Theorem 1 for the L_∞ estimator does not hold; in practice, this means that the estimator, while uniquely defined for any finite sample, is somewhat unstable and may even be inconsistent.

What happens to the limiting distribution of an M-Chebyshev estimator the measure $\bar{\nu}$ in (19) puts all its probability mass on a lower dimensional space? For simplicity, we will consider only the LS-Chebyshev estimator although a similar argument will work for M-Chebyshev estimators defined for wider class of strictly convex functions. The LS-Chebyshev estimator is the solution of a quadratic program that can be approximated by a linear program whose solution in this scenario is non-unique. Under fairly mild conditions, the asymptotics of the LS-Chebyshev estimator will be determined by the quadratic part of the objective function in (23). The idea here is similar to that of Mangasarian (1984a,b) who shows that a linear program can be solved by solving an appropriately perturbed quadratic program.

Define $\xi_n(\boldsymbol{u}, v)$ to be the approximating linear objective function:

$$\xi_n(\boldsymbol{u}, v) = 2\gamma_0 v - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \boldsymbol{x}_i^T \boldsymbol{u} \tag{24}$$

and suppose that the limiting linear objective function

$$\xi(\boldsymbol{u}, v) = 2\gamma_0 v - \left\{\lim_{n\to\infty} \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \boldsymbol{x}_i\right\}^T \boldsymbol{u} = 2\gamma_0 v - 2\boldsymbol{\theta}^T\boldsymbol{u} \tag{25}$$

has a finite (that is, greater than $-\infty$) minimum subject to the constraints

$$\Gamma_i' \geq \boldsymbol{u}^T\boldsymbol{X}_i' - v \quad \text{for } i = 1, 2, 3, \cdots \tag{26}$$

where $\{\Gamma_i'\}$ is defined as in (18) and $\{\boldsymbol{X}_i'\}$ are i.i.d. random vectors (independent of $\{\Gamma_i'\}$ with distribution $\bar{\nu}$ as defined in (19); the minimum of (25) subject to (26) will be finite if the vector $-\boldsymbol{\theta}/\gamma_0$ lies in the convex hull of some $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s$ in the support of the measure $\bar{\nu}$.

Define $\mathcal{T}$ to be the (random) convex set of all $(\boldsymbol{U}, V)$ minimizing the limiting linear objective function (25) subject to the constraints (26) and define S_n to be the minimum value of $\xi_n(\boldsymbol{u}, v)$ in (24) subject to the constraints (10); then $S_n \xrightarrow{d} S$ where $S = \xi(\boldsymbol{u}, v)$ for all $(\boldsymbol{u}, v) \in \mathcal{T}$. For the LS-Chebyshev estimator, the quadratic remainder term is simply

$$a_n^{-1}\left\{v^2 + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{u})^2\right\}$$

and multiplying by a_n, we get

$$v^2 + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{u})^2 \to v^2 + \boldsymbol{u}^T C\boldsymbol{u} = \mathcal{Z}(\boldsymbol{u}, v) \tag{27}$$

where

$$C = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \int \boldsymbol{x}\boldsymbol{x}^T \, \mu(d\boldsymbol{x}). \tag{28}$$

The quadratic remainder term goes to 0 as $n \to \infty$; by subtracting S_n (the minimum value of ξ_n) and multiplying by a_n, we obtain an objective function that, in the limit, is quadratic and finite on the set $\mathcal{T}$ and infinite outside of $\mathcal{T}$. More precisely, $\big(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n(\widehat{\gamma}_n - \gamma_0)\big)$ minimizes

$$\mathcal{Z}_n(\boldsymbol{u}, v) = v^2 + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T \boldsymbol{u})^2 + a_n(\xi_n(\boldsymbol{u}, v) - S_n)$$

subject to the constraints (10); it follows that

$$a_n(\xi_n(\boldsymbol{u}, v) - S_n) \xrightarrow{e-d} \varphi(\boldsymbol{u}, v) = \begin{cases} 0 & \text{if } (\boldsymbol{u}, v) \in \mathcal{T} \\ +\infty & \text{if } (\boldsymbol{u}, v) \notin \mathcal{T}. \end{cases}$$

This suggests that (under appropriate regularity conditions)

$$\big(a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}), a_n(\widehat{\gamma}_n - \gamma_0)\big) \xrightarrow{d} \arg \min_{(\mathbf{u},v)\in\mathcal{T}} \mathcal{Z}(\boldsymbol{u}, v)$$

for $\mathcal{Z}(\boldsymbol{u}, v)$ defined in (27).

What these asymptotics suggest is that in the case of instability (where many estimates have nearly equal minimum maximum absolute residuals), the quadratic (or, more generally, convex) regularization preserves the convergence rate; the regularization allows us to pick a possibly sub-optimal (with respect to the L_∞ objective function) but stable solution. Note that the matrix C in (28) is defined in terms of the limiting measure μ of the design $\{\boldsymbol{x}_i\}$ (rather than $\bar{\nu}$) and hence is typically better conditioned.

EXAMPLE 4. Consider the simple linear regression model

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

where $x_i = i/n$ and $\{\varepsilon_i\}$ are independent with ε_i having the distribution (4) on the interval $[-1, 1]$ with $\alpha = x_i^{-\eta}$ for some $\eta \geq 0$. When $\eta > 0$, it can be verified that the L_∞ estimator satisfies

$$n^\gamma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{p} 0$$

for any $\gamma < 1$. Setting $a_n = n/\ln(n)$, it follows that the point process M_n defined as in (9) with $C_n = I$ converges in distribution to M^* where

$$M^*(A \times B) = \sum_{k=1}^{\infty} I(\eta\,\Gamma_k \in A, \boldsymbol{X}_k \in B)$$

where $\Gamma_k = E_1 + \cdots + E_k$ for unit mean exponential random variables $\{E_i\}$ and $\{\boldsymbol{X}_k\}$ are i.i.d. taking the values $\pm(1, 1)^T$ each with probability 1/2. The limiting linear program

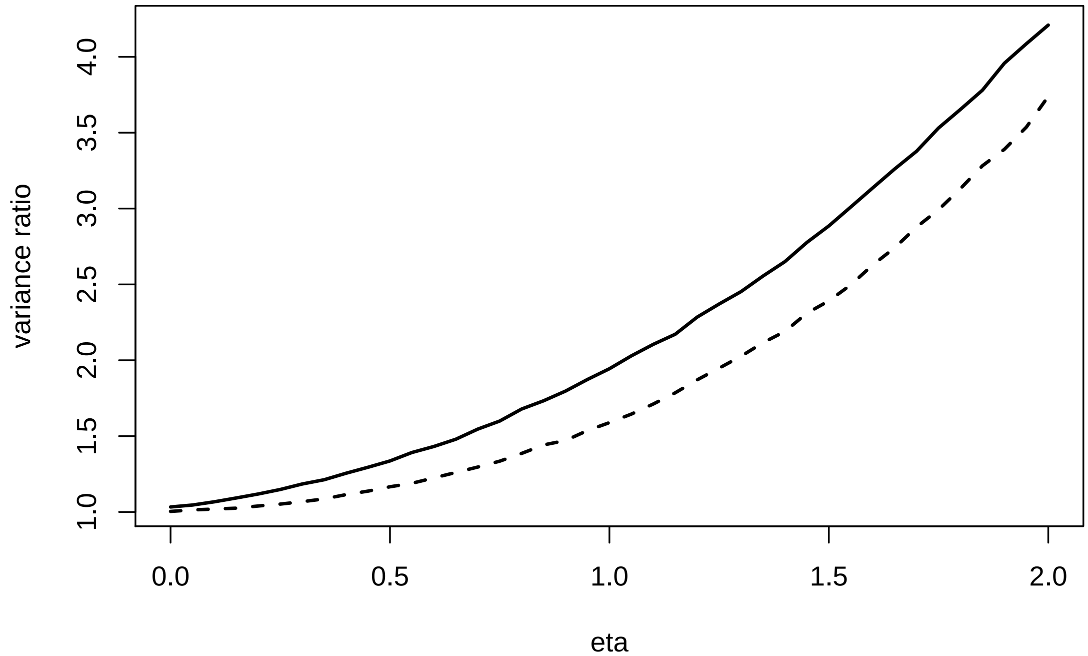

Figure 5: Ratio of the variance of the L_∞ estimator of β_1 to the variance of the LS-Chebyshev estimator in Example 4 for $n = 100$ (solid line) and $n = 1000$ (dashed line) for $0 \le \eta \le 2$.

based on the limiting Poisson process does not have a unique solution and hence $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is not necessarily bounded in probability. However, the discussion above suggests that the LS-Chebyshev estimator will be a_n-consistent. Define $\mathcal{T}$ as the set of minimizers v subject to the constraints from the limiting Poisson process and $\mathcal{Z}(\boldsymbol{u}, v)$ as in (27). We then obtain (after some straightforward calculations) for $\eta > 0$,

$$\mathcal{T} = \left\{ (u_0, u_1, v) : u_0 + u_1 = \frac{V_1 - V_2}{2}, v = -\frac{V_1 + V_2}{2} \right\}$$

where V_1 and V_2 are independent exponential random variables with common mean $\eta/2$ and

$$\mathcal{Z}(\boldsymbol{u}, v) = v^2 + \boldsymbol{u}^T C \boldsymbol{u}$$

where

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

Some simple calculations give

$$a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \begin{pmatrix} -\dfrac{V_1 - V_2}{4} \\ \dfrac{3(V_1 - V_2)}{4} \end{pmatrix}.$$

Figure 5 shows the ratio of the variance of the L_∞ estimator of β_1 to that of the LS-Chebyshev estimator (based on 100000 Monte Carlo samples for $n = 100$ and $n = 1000$). In both cases, this ratio increases as η increases although (perhaps strangely) the advantage of the LS-Chebyshev estimator is smaller for $n = 1000$.

Figure 6 shows the B-spline LS-Chebyshev estimate for the motorcycle data in Example 1 using the same model used to produce the estimates in Figures 1 and 2. Clearly, this estimate is much more stable than the L_∞ estimate and quite comparable to the least squares estimate; moreover, the estimate of γ is 63.0378, which is only slightly larger than the estimate of γ for the L_∞ estimate (62.8916). Perhaps as significant is the fact that the LS-Chebyshev estimate yields an estimate of the acceleration as a function of time that is as plausible as the least squares estimate.

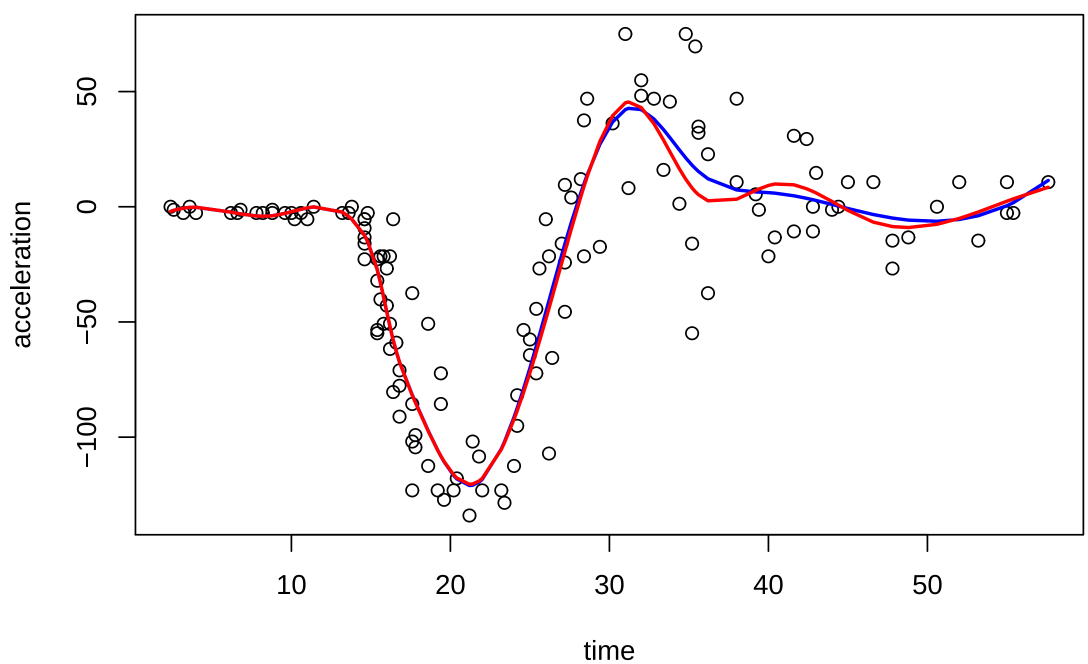

Figure 6: LS-Chebyshev estimate (red) for motorcycle data compared to the least squares estimate (blue).

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

References

Akçay, H., At, N., 2006. Convergence analysis of central and minimax algorithms in scalar regression models. Mathematics of Control, Systems, and Signals 18, 66-99.

Akçay, H., Hjalmarsson, H., Ljung, L., 1996. On the choice of norm in system identification. IEEE Transactions on Automatic Control 41, 1367-1372.

Alecu, A., Munteanu, A., Cornelis, J.P.H., Schelkens, P., 2006 Wavelet-based scalable L-infinity-oriented compression. IEEE Transactions on Image Processing 15, 2499-2512.

Aoki, M., 1965. Successive generation of Chebyshev approximation solution. Journal of Basic Engineering (Transactions of the ASME, Series D) 87, 17-22.

Beck, A., Eldar, Y.C., 2007. Regularization in regression with bounded noise: a Chebyshev center approach. SIAM Journal on Matrix Analysis and Applications 29, 606-625.

Berenguer-Rico, V., Nielsen, B., Johansen, S., 2019. Models where the Least Trimmed Squares and Least Median of Squares estimators are maximum likelihood (unpublished manuscript).

Bertsch, G.F., Sabbey, B., Uusnäkki, M., 2005. Fitting theories of nuclear binding energies. Physical Review C 71, 054311(7).

Brenner, M.J., 2002. Aeroservoelastic model uncertainty bound estimation from flight data. Journal of Guidance, Control, and Dynamics 25, 748-754.

Broffitt, J.D., 1974. An example of the large sample behaviour of the midrange. American Statistician 28, 69-70.

Castillo, E., Castillo, C., Hadi, A.S., Sarabia, J.M., 2009. Combined regression models. Computational Statistics 24, 27-66.

Chernozhukov, V., 2005. Extremal quantile regression. Annals of Statistics 33, 806-839.

Chernozhukov, V., Hong, H., 2004. Likelihood estimation and inference in a class of non-regular econometric models. Econometrica 77, 1445-1480.

Chuah, S., Dumitrescu, S., Wu, X., 2013. ℓ_2 optimized predictive image coding with ℓ_∞ bound. IEEE Transactions on Image Processing 22, 5271-5281.

Clason, C., 2012. L^∞ fitting for inverse problems with uniform noise. Inverse Problems 28, 104007.

Cline, A.K., 1972. Rate of convergence of Lawson's algorithm. Mathematics of Computation 26, 167-176.

Croux, C., Rousseeuw, P.J., Hössjer. O., 1994. Generalized S-estimators. Journal of the American Statistical Association 89, 1271-1281.

Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S., 2010. Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied

Mathematics 63, 1-38.

Dou, Y., Hu, H., Han, B. 2013. Solving a linear inverse problem by using an L^∞ fitting and a hybrid penalty with an application to an inverse heat conduction problem. Journal of Information and Computational Science 10, 5935-5943.

Du, J., Cao, S., Hunt, J.H., Huo, X. 2019. Optimal shape control via L_∞ loss for composite fuselage assembly. arXiv: 1911.03592

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.

Harter, H.L., 1975a. The method of least squares and some alternatives – part III. International Statistical Review 43, 1-44.

Harter, H.L., 1975b. The method of least squares and some alternatives – part IV. International Statistical Review 43, 125-190.

Hayashi, K., Yoshida, Y., 2020. Testing proximity to subspaces: approximate ℓ_∞ minimization in constant time. Algorithmica 82, 1277-1297.

James, F., 1983a, Fitting tracks in wire chambers using the Chebyshev norm instead of least squares. Nuclear Instruments and Methods 211, 145-152.

James, F., 1983b. Probability, statistics, and associated computing techniques. In: Techniques and concepts of high-energy physics II (pp. 189-231). Springer, New York.

Jaschke, S., 1998. Arbitrage bounds for the term structure of interest rates. Finance and Stochastics 2, 29-40.

Jaschke, S., Küchler, U., 2001. Coherent risk measures and good-deal bounds. Finance and Stochastics 5, 181-200.

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

Knight, K., 1999. Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript.

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

Koenker, R., 2005. Quantile regression. Cambridge University Press.

Lai, P.Y., Lee, S.M.S., 2005. An overview of asymptotic properties of L_p regression under general classes of error distributions. Journal of the American Statistical Association 100, 446-458.

Lawson, C.L., 1961. Contributions to the Theory of Linear Least Maximum Approximations. Ph.D. Thesis, UCLA.

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

Mäkilä, P.M., 1991. Robust identification and Galois sequences. International Journal of Control 54, 1189-1200.

Mangasarian, O.L., 1984a. Normal solutions of linear programs. Mathematical Programming Study 22, 206-216.

Mangasarian, O.L., 1984b. Sparsity-preserving SOR algorithms for separable quadratic and linear programming. Computers and Operations Research 11, 105-112.

Milanese, M., Belforte, G., 1982. Estimation theory and uncertainty intervals evaluation in the presence of unknown but bounded errors – linear families of models and estimators. IEEE Transactions on Automatic Control 27, 408-414.

Molchanov, I., 2005. Theory of Random Sets. Springer, London.

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

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

Qi, C., 2015. Theoretical uncertainties of the Duflo-Zuker shell-model mass formulae. Journal of Physics G: Nuclear and Particle Physics 42, 045104.

Rider, P.J., 1957. The midrange of a sample as an estimator of the population midrange. Journal of the American Statistical Association 52, 537-542.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871-880.

Schechtman, E., Schechtman, G., 1986. Estimating the parameters in regression with uniformly distributed errors. Journal of Statistical Computation and Simulation 26, 269-281.

Schmidt, G., Mattern, R., Schueler, F., 1981. Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without potective helmet under the effects of impact. EEC Research Program on Biomechanics of Impacts, Final Report, Phase III, Project G5, Institut für Rechtsmedizin, University of Heidelberg.

Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). Journal of the Royal Statistical Society, series B 47, 1-52.

Sposito, V.A., 1990. Some properties of L_p estimators. In: Robust Regression: Analysis and Applications, (eds K.D. Lawrence and J.L. Arthur) 23-58. CRC Press, Boca Raton, FL.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal

Statistical Society, series B 58, 267-288.

Tse, D.N.C., Daleh, M.A., Tsitsiklis, J.N., 1993. Optimal asymptotic identification under bounded disturbances. IEEE Transactions on Automatic Control 38, 1176-1190.

Zolghadri, A., Henry, D., 2004. Minimax statistical models for air pollution time series. Application to ozone time series data measured in Bordeaux. Environmental Monitoring and Assessment 98, 275-294.