Statistica Sinica **18**(2008), 1619-1640

RATEWISE EFFICIENT ESTIMATION OF REGRESSION COEFFICIENTS BASED ON L_p PROCEDURES

P. Y. Lai and Stephen M. S. Lee

The University of Hong Kong

Abstract: We consider the problem of estimation of regression coefficients under general classes of error densities without assuming classical regularity conditions. Optimal orders of convergence rates of regression-equivariant estimators are established and shown to be attained in general by L_p estimators based on judicious choices of p. We develop a procedure for choosing p adaptively to yield L_p estimators that converge at approximately optimal rates. The procedure consists of a special algorithm to automatically select the correct mode of L_p estimation and the m out of n bootstrap to consistently estimate the log mean squared error of the L_p estimator. Our proposed adaptive L_p estimator is compared with other adaptive and non-adaptive L_p estimators in a simulation study, that confirms superiority of our procedure.

Key words and phrases: Adaptive, L_p estimator, m out of n bootstrap, ratewise efficient, regression.

1. Introduction

Consider a random sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ under a general linear regression setup, such that

$$Y_i = Z_i^{\mathrm{T}} \beta_0 + U_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $(U_1, \ldots, U_n)$ and $(Z_1, \ldots, Z_n)$ denote two independent random samples drawn from the univariate distribution function F_U and the d-variate distribution function F_Z respectively, and β_0 is an unknown d-variate parameter in $\mathbb{R}^d$. Under classical regularity conditions, the Cramér-Rao lower bound provides a benchmark for asymptotic efficiency, and the maximum likelihood estimator of β_0 is asymptotically efficient with convergence rate $n^{1/2}$. The situation is far less conclusive if F_U is not parametrically specified and does not satisfy the regularity conditions. There may, for example, exist hyper-efficient estimators having convergence rates faster than $n^{1/2}$. It would therefore be of interest to derive results analogous to the Cramér-Rao lower bound under general F_U or to at least establish a notion of "ratewise efficiency" to identify the best achievable convergence rates.

In this paper we relax the classical regularity conditions and ask only that F_U have a symmetric density f_U regularly varying at 0 with index $\zeta - 1 \in (-1, \infty)$. This allows for a broad variety of density shapes near 0, at which f_U may, for example, possess a cusp or a point of singularity. In Section 2 we establish, under this general class of error densities, the best convergence rates which can be achieved by any regression-equivariant estimators of β_0. The special case of location estimation under a known f_U and similar regular variation conditions has been extensively studied in the literature. However, if f_U is unspecified, no general estimation strategy has yet been found to yield ratewise efficient estimators under either the location or regression setup. A related study by Jurečková (1983) shows that under a certain type of singularity the convergence rate of a class of M-estimators can exceed the conventional $n^{1/2}$ but still not be ratewise efficient.

Conventionally the L_p estimator $\hat{\beta}(p)$ of β_0, for a fixed $p \in (0, \infty)$, is defined as the value of β that minimizes the criterion function $C_p(\beta) \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} |Y_i - Z_i^{\mathrm{T}}\beta|^p$. It is generally conceived that the L_2 estimator is efficient under normal f_U but is very sensitive to departures from normality, whilst the L_1 estimator is robust against outliers in the observed data or heavy tails of f_U. The last few decades have seen considerable work done on adaptive selection of $p \in [1, 2]$, which is found to be advantageous under certain contaminated or skewed error densities. Earlier literature focused predominantly on methods adaptive to the tail behaviour of F_U; see, for example, Hogg (1972), Harter (1974–1975), Money, Affleck-Graves, Hart and Barr (1982), Sposito and Hand (1983) and Nyquist (1983) for procedures targeted at specific error densities. However, these procedures remain largely exploratory and lack formal justification for their adaptivity under general classes of error distributions. A more rigorous treatment is given by Arcones (2005) in the context of location estimation. His procedure selects $p \in (1, \infty)$ by minimizing an estimate of the asymptotic mean squared error of $\hat{\beta}(p)$ under regularity conditions. Choices of $p < 1$ have received only sporadic attention in the literature, where focus is on the design of computational algorithms for L_p minimization; see, for example, Barrodale and Roberts (1970) or Ekblom (1974). None of the above adaptive L_p procedures yield ratewise efficient estimators under the nonregularity conditions presently considered.

Lai and Lee (2005) discuss the asymptotic properties of L_p estimators under different types of regular variation of f_U at 0. We review their results in Section 3 and show further that, in almost all cases, there exist L_p estimators that are ratewise efficient. When ζ is unknown, selection of the best p, which depends in general on ζ, poses a practical problem. A second difficulty arises when ratewise efficiency demands that p be chosen from the interval $(0, 1)$ so that the correct mode of L_p estimation should be carefully determined, that is, $\hat{\beta}(p)$ should be defined as a minimizer or local maximizer of $C_p(\beta)$ in accordance with the shape of the f_U in question. In the case of local maximization, computation of $\hat{\beta}(p)$ is further complicated by the existence of multiple local maxima which obscure, to some extent, the major feature of $C_p(\beta)$. We propose in Section 4 an adaptive procedure for computing L_p estimators that are approximately ratewise efficient. The procedure selects the optimal p adaptively, and computes $\hat{\beta}(p)$ by automatically locating the global minimum or local maximum of a smoothed version of $C_p(\beta)$ without having to fix the actual mode of optimization in advance. Its implementation is illustrated with a data set. Section 5 discusses the generalisation of our theory and adaptive procedure under asymmetric f_U. Section 6 reports a simulation study in which our adaptive L_p estimator is compared with other adaptive and non-adaptive alternative estimators in terms of mean squared error. Section 7 concludes our findings. Technical proofs are given in the Appendix.

2. Ratewise Efficiency

For the linear regression model (1.1), consider a class of symmetric error density functions $f_U = F_U'$ regularly varying at 0 with index $\zeta - 1$:

$$f_U(u) = |u|^{\zeta-1}\mathcal{L}(|u|) \quad \text{for } |u| \le \Delta, \tag{2.1}$$

for some $\zeta, \Delta > 0$ and nonnegative function $\mathcal{L}$ on $(0,\infty)$ which is slowly varying at 0. The class comprises density functions that may be continuous or discontinuous, differentiable or nondifferentiable, upwardly or downwardly sharp-pointed, infinite or zero at the origin. Special examples include normal densities, Laplace densities, symmetric gamma and Weibull densities. Ibragimov and Has'minskii (1981) term the origin a singularity of order $\zeta - 1$, of the first type if $\zeta \in (1,2)$ and of the third type if $\zeta \in (0,1)$. The form (2.1) represents a natural generalization of the classical regularity conditions, which assume that $\zeta = 1$ and that $\mathcal{L}$ is sufficiently well-behaved near 0. Smirnov (1952) provides an early reference to (2.1) and identifies with it three of his four domains of attraction for sample quantiles. Knight (1998) and Rogers (2001) focus on the same class of distributions, among others, in their investigation of L_1 regression asymptotics. Polfeldt (1970) focuses on the case $\zeta = 1$ and studies the location problem under different forms of $\mathcal{L}$. The condition of symmetry is imposed to ensure Fisher-consistency of L_p estimators for a common estimand β_0: $\mathbb{E}[\,\mathrm{sgn}(Y_1 - Z_1^{\mathrm{T}}\beta_0)|Y_1 - Z_1^{\mathrm{T}}\beta_0|^{p-1}] = \mathbb{E}[\,\mathrm{sgn}(U_1)|U_1|^{p-1}] = 0$ for all $p > 0$, provided that the moment exists. The case of asymmetric f_U will be discussed in Section 5.

An estimator $\hat{\beta} = \hat{\beta}(\{(Y_i, Z_i) : i = 1,\ldots,n\})$ of β_0 is regression-equivariant if for any $t \in \mathbb{R}^d$, $\hat{\beta}(\{(Y_i + Z_i^{\mathrm{T}}t, Z_i) : i = 1,\ldots,n\}) = \hat{\beta}(\{(Y_i, Z_i) : i = 1,\ldots,n\}) + t$. Denote by $\|\cdot\|$ the Euclidean norm. Assume that

(A0) $\mathbb{E}\|Z_1\|^2\,|\log\|Z_1\|\,|<\infty$ and $\mathbb{P}(Z_1^{\mathrm{T}}\beta=0)<1$ for any nonzero $\beta\in\mathbb{R}^d$.

To establish the best possible convergence rates of regression-equivariant estimators of β_0, we assume the following additional conditions on f_U:

(A1) If $\zeta>2$, then f_U is continuously differentiable and $I_U \overset{\text{def}}{=} \mathbb{E}(f_U'(U_1)/f_U(U_1))^2 <\infty$.

(A2) If $\zeta=2$, then $\mathcal{L}$ is twice continuously differentiable on $[0,\Delta)$, $\mathcal{L}(0)>0$, f_U is differentiable and $\mathbb{E}[(f_U'(U_1)/f_U(U_1))^2;|U_1|>\epsilon]<\infty$ for some $\epsilon>0$.

(A3) If $\zeta\in(0,1)\cup(1,2)$, then $\mathcal{L}$ is continuous on $[0,\Delta)$, $\mathcal{L}(0)>0$ and there exists $\zeta^*>\zeta$ such that, as $\epsilon\downarrow 0$,

$$\int_0^{\frac{\Delta}{2}}\left|\mathcal{L}(u+\epsilon)^{\frac{1}{\zeta^*}}-\mathcal{L}(u)^{\frac{1}{\zeta^*}}\right|^{\zeta^*}u^{\zeta-1}du=O(\epsilon^{\zeta^*}).$$

(A4) If $\zeta=1$, then $\mathcal{L}(0)>0$, $\gamma_0\overset{\text{def}}{=}\lim_{u\downarrow 0}u^{-q}(\mathcal{L}(0)-\mathcal{L}(u))\neq 0$ for some $q>0$, $\int f_U(u-\epsilon)^2/f_U(u)\,du<\infty$ for sufficiently small $|\epsilon|$, and for some fixed $\eta^*>0$,

$$0<\lim_{\epsilon\to 0}\epsilon^{-2}\int_{|u|>\eta^*}\left\{\frac{f_U(u-\epsilon)}{f_U(u)}-1\right\}^2 f_U(u)\,du<\infty.$$

The condition (A1) requires that f_U be sufficiently smooth with finite Fisher information. An "almost" finite Fisher information is assumed under (A2). Both (A2) and (A3) contain smoothness conditions on $\mathcal{L}$. In particular, (A3) holds if $\mathcal{L}$ is differentiable and

$$\int_0^{\frac{\Delta}{2}}\frac{u^{\zeta-1}|\mathcal{L}'(u)|^{\zeta^*}}{\mathcal{L}(u)^{\zeta^*-1}}du<\infty.$$

The condition (A4) requires that f_U be sufficiently smooth and second-order $(\zeta-1,q)$ regularly varying at 0. It has a peak or trough at 0 according as $\gamma_0>0$ or $\gamma_0<0$ respectively, and is sharp-pointed there if $q<1$.

The following theorem gives the orders of the smallest possible mean absolute or squared error of regression-equivariant estimators $\hat\beta$ under scenarios described by (A1)−(A4). The proof is given in the Appendix.

Theorem 1. *Assume that F_Z satisfies* (A0) *and f_U has the form* (2.1). *Let $\hat\beta$ be a generic regression-equivariant estimator of β_0. Then,*

(i) *under* (A1), $\liminf_{n\to\infty}\inf_{\hat\beta} n^{1/2}\mathbb{E}\|\hat\beta-\beta_0\|>0$;

(ii) *under* (A2), $\liminf_{n\to\infty}\inf_{\hat\beta}(n\log n)^{1/2}\mathbb{E}\|\hat\beta-\beta_0\|>0$;

(iii) *under* (A3), $\liminf_{n\to\infty} \inf_{\hat{\beta}} n^{1/\zeta}\mathbb{E}\|\hat{\beta} - \beta_0\| > 0$;

(iv) *under* (A4), $\liminf_{n\to\infty} \inf_{\hat{\beta}} \varphi_q(n)^2\mathbb{E}\|\hat{\beta}-\beta_0\|^2 > 0$, *where* $\varphi_q(n) = n^{1/2}$, $(n\log n)^{1/2}$ *and* $n^{1/(2q+1)}$ *for* $q > 1/2$, $= 1/2$ *and* $< 1/2$ *respectively.*

It is clear from Theorem 1 that the fastest possible convergence rates of $\hat{\beta}$ are $n^{1/2}$, $(n\log n)^{1/2}$, $n^{1/\zeta}$ and $\varphi_q(n)$ under (A1), (A2), (A3), and (A4) respectively. Any $\hat{\beta}$ which achieves the above rate is said to be ratewise efficient. Note that the best convergence rate exceeds the conventional $n^{1/2}$ under (A2), (A3) and, for $q \le 1/2$, under (A4).

In the special case of location estimation with $Z_1 \equiv 1$, Ibragimov and Has'minskii (1981) obtain results similar to Theorem 1 (i)−(iii). They show that if f_U is completely specified, then the maximum likelihood and Bayes estimators are ratewise efficient under $\zeta > 1$ and $\zeta > 0$, respectively. Smith (1985) and Ghosal and Samanta (1995) generalize these results to cases where f_U depends on an unknown vector parameter. Polfeldt (1970) proves that the minimum variances of unbiased location estimators have the same orders as those stated in Theorem 1(iv). Daniels (1960) and Prakasa Rao (1968) show under (A4) that the maximum likelihood location estimator is ratewise efficient for $q > 1/2$ and $q < 1/2$, respectively. If f_U is unspecified, no general strategy has yet been developed for constructing ratewise efficient equivariant estimators under either the location or regression model. We address this issue in the rest of the paper.

3. L_p Estimation: Asymptotic Theory

Under the general class of error densities (2.1), both ζ and p are determinative factors in the convergence rate, and hence accuracy, of $\hat{\beta}(p)$. Lai and Lee (2005) provide a detailed account of the asymptotics of L_p regression for each $\zeta > 0$ under weaker conditions on F_U. Specializing Theorems 1 and 2 of Lai and Lee (2005) to our present context, but slightly weakening their definition of L_p estimator, we establish in Theorem 2 the convergence rates and correct modes of L_p estimation under various scenarios. Denote hereafter by $\mathcal{B}$ an open neighbourhood containing β_0, the unique minimizer or maximizer of $\mathbb{E}|Y_1 - Z_1^{\mathrm{T}}\beta|^p$ over $\beta \in \mathcal{B}$.

THeorem 2. *Assume that* F_Z *satisfies* (A0), f_U *satisfies* (2.1), *the conditions on* $\mathcal{L}$ *specified in one of* (A1)−(A4), *and that* $\mathbb{E}\,|U_1|^{\max\{2p-2,0\}} < \infty$. *Define* r_n *according to the following:*

(i) $r_n = n^{1/2}$ *if* $p + \zeta > 2$ *and* $p \ne 1$;

(ii) $r_n = n^{1/2}\log n$ *if* $p + \zeta = 2$ *and* $p \ne 1$;

(iii) $r_n = n^{1/2(p+\zeta-1)}$ *if* $p+\zeta<2$, $2p+\zeta>2$ *and* $\zeta\neq 1$*;*

(iv) $r_n = n^{1/(2\zeta)}$ *if* $p=1$, $\zeta\in[1,2]$ *and* $\mathbb{E}\|Z_1\|^{\zeta+1}<\infty$*;*

(v) $r_n \sim \left\{n^{1/2}\mathcal{L}(1/r_n)\right\}^{1/\zeta}$ *if* $p=1$, $\zeta>2$ *and* $\mathbb{E}\|Z_1\|^{\zeta+1}<\infty$*;*

(vi) $r_n = (n/\log n)^{1/\zeta}$ *if* $2p+\zeta=2$ *and* $\zeta\neq 1$*;*

(vii) $r_n = n^{1/\zeta}$ *if* $2p+\zeta<2$ *and* $\zeta\neq 1$*;*

(viii) $r_n = \min\{n^{1/2}, n^{1/(3-2p)}\}$ *if* $\zeta=1$, $p+q>1$ *and* $1>p\neq 1/2$*;*

(ix) $r_n = (n/\log n)^{1/2}$ *if* $\zeta=1$, $p=1/2$ *and* $q>1/2$*;*

(x) $r_n = \min\{n^{1/(2p+2q)}, n^{1/(2q+1)}\}$ *if* $\zeta=1$, $p+q<1$ *and* $p\neq 1/2$*;*

(xi) $r_n = \min\{n^{1/2}\log n, (n^{1/2}\log n)^{2/(3-2p)}\}$ *if* $\zeta=1$, $p+q=1$ *and* $p\neq 1/2$*;*

(xii) $r_n = (n\log n)^{1/2}$ *if* $\zeta=1$ *and* $p=q=1/2$*;*

(xiii) $r_n = (n/\log n)^{1/(2q+1)}$ *if* $\zeta=1$, $p=1/2$ *and* $q<1/2$.

Let $\hat\beta(p)\in\mathcal{B}$ *satisfy,*

— *under case* (i),

$$C_p(\hat\beta(p))\begin{cases}\le \inf\limits_{\beta\in\mathcal{B}} C_p(\beta)+o_p(r_n^{-2}),\, p>1,\\ \ge \sup\limits_{\beta\in\mathcal{B}} C_p(\beta)-o_p(r_n^{-2}),\, p<1;\end{cases}$$

— *under case* (ii),

$$C_p(\hat\beta(p))\begin{cases}\le \inf\limits_{\beta\in\mathcal{B}} C_p(\beta)+o_p(r_n^{-2}\log r_n),\, p>1,\\ \ge \sup\limits_{\beta\in\mathcal{B}} C_p(\beta)-o_p(r_n^{-2}\log r_n),\, p<1;\end{cases}$$

— *under cases* (iii), (vi) *and* (vii),

$$C_p(\hat\beta(p))\begin{cases}\le \inf\limits_{\beta\in\mathcal{B}} C_p(\beta)+o_p(r_n^{-p-\zeta}),\, \zeta<1,\\ \ge \sup\limits_{\beta\in\mathcal{B}} C_p(\beta)-o_p(r_n^{-p-\zeta}),\, \zeta>1;\end{cases}$$

— *under cases* (iv) *and* (v), $C_p(\hat\beta(p))\le\inf_{\beta\in\mathcal{B}}C_p(\beta)+o_p(r_n^{-1-\zeta}\mathcal{L}(1/r_n))$;

— *under cases* (viii) *and* (ix),

$$C_p(\hat\beta(p))\begin{cases}\le \inf\limits_{\beta\in\mathcal{B}} C_p(\beta)+o_p(r_n^{-2}),\, \int|u|^{p-2}\left[f_U(u)-f_U(0)\right]\,du<0,\\ \ge \sup\limits_{\beta\in\mathcal{B}} C_p(\beta)-o_p(r_n^{-2}),\, \int|u|^{p-2}\left[f_U(u)-f_U(0)\right]\,du>0;\end{cases}$$

— *under cases* (x), (xi), (xii) *and* (xiii),

$$C_p(\hat{\beta}(p)) \begin{cases} \leq \inf_{\beta \in \mathcal{B}} C_p(\beta) + o_p(r_n^{-p-q-1} + r_n^{-2} \log r_n), \gamma_0 > 0, \\ \geq \sup_{\beta \in \mathcal{B}} C_p(\beta) - o_p(r_n^{-p-q-1} + r_n^{-2} \log r_n), \gamma_0 < 0. \end{cases}$$

Then $r_n(\hat{\beta}(p) - \beta_0) = O_p(1)$.

Ratewise efficiency of L_p estimators now follows immediately from Theorems 1and 2. The results are summarized in the following corollary.

Corollary 1. *Assume the conditions of Theorem* 2. *Then*

(i) *under* (A1), L_p *estimators are ratewise efficient for* $p \neq 1$*;*

(ii) *under* (A2), L_p *estimators based on* $p \neq 1$ *have the fastest convergence rate* $n^{1/2}$*, which is slower than the optimal rate* $(n \log n)^{1/2}$*;*

(iii) *under* (A3), L_p *estimators are ratewise efficient for* $p < 1 - \zeta/2$*;*

(iv) *under* (A4), L_p *estimators are ratewise efficient for* $p > 1/2$ *if* $q > 1/2$*, for* $p = 1/2$ *if* $q = 1/2$*, and for* $p < 1/2$ *if* $q < 1/2$.

We see from Corollary 1 that ratewise efficient L_p estimators exist in all cases under our assumptions except when $\zeta = 2$. Even in the latter situation the most efficient L_p estimators have convergence rates only slightly slower than the optimal rate, by a factor slowly varying in n. Note that consideration of ratewise efficiency tends to favour use of a small p in the sense that $\hat{\beta}(p)$ can be made ratewise efficient for arbitrarily many ζ by choosing a sufficiently small $p > 0$. Indeed, if $p < 1/2$ is fixed, then $\hat{\beta}(p)$ is ratewise efficient for $\zeta \in (0,1) \cup (1, 2(1-p)) \cup (2, \infty)$, and for $q < 1/2$ under $\zeta = 1$.

Two other implications of Theorem 2 are of importance in practice. First, when f_U exhibits a trough at the origin and $p < 1$ is chosen, L_p estimation has to be undertaken by locally maximizing, rather than minimizing, $C_p(\beta)$ over β in an appropriate region. Secondly, the assertions of Theorem 2 hold for $\hat{\beta}(p)$ defined liberally as approximate minimizers or local maximizers of $C_p(\beta)$, enabling use of a smoothed version $\tilde{C}_p$ in place of C_p as criterion function for more convenient calculation of $\hat{\beta}(p)$.

Denote by $\hat{\beta}^*(p)$ the L_p estimator calculated, in a way analogous to the derivation of $\hat{\beta}(p)$, from a bootstrap sample $(Y_1^*, Z_1^*), \ldots, (Y_m^*, Z_m^*)$ drawn from $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. Theorem 3 of Lai and Lee (2005) establishes m out of n bootstrap (Bickel, Götze and van Zwet (1997)) consistency for L_p regression for a large majority of cases covered by Theorem 2. It has a trivial extension to the more liberally defined L_p estimators $\hat{\beta}(p)$ considered in Theorem 2.

Theorem 3. *Assume the conditions of Theorem* 2 *and that* $m = o(n)$ *and* $m \to \infty$. *Then the conditional distribution of* $r_m(\hat{\beta}^*(p) - \hat{\beta}(p))$, *given* (Y_1, Z_1), $\ldots$, (Y_n, Z_n), *converges weakly in probability to the same weak limit as that of* $r_n(\hat{\beta}(p) - \beta_0)$ *for cases* (i)−(vi), (viii), (ix), (xi) *and* (xii) *of Theorem* 2. *The result also holds for case* (x) *if* $\max\{p, q\} > 1/2$.

Note that we have not been able to assert m out of n bootstrap consistency for cases (vii), (xiii) or, for $p, q \leq 1/2$, under case (x) of Theorem 2. However, comparison of Corollary 1 and Theorem 3 shows that under the above situations m out of n bootstrappable L_p estimators can be found which are arbitrarily close to being ratewise efficient.

4. An Adaptive L_p Procedure

In practical situations where f_U is unknown and ζ is unspecified, the optimal value of p is in general unknown and it is uncertain whether $\hat{\beta}(p)$ should be obtained by minimizing or locally maximizing $C_p(\beta)$. Section 4.1 describes an algorithm for automatically computing $\hat{\beta}(p)$ without prior knowledge of f_U. Section 4.2 gives another algorithm for consistent estimation of the log mean squared error of $\hat{\beta}(p)$, $\log MSE(\hat{\beta}(p))$, using an m out of n bootstrap approach. An adaptive procedure is then given in Section 4.3 for calculating an approximately ratewise efficient $\hat{\beta}(p)$.

4.1. Algorithm (A): automatic search for L_p estimate

Recall that an open neighbourhood $\mathcal{B}$ exists around β_0 such that β_0 is the unique minimizer or maximizer of $\mathbb{E}|Y_1 - Z_1^{\mathrm{T}}\beta|^p$ over $\beta \in \mathcal{B}$. It is therefore natural to require that $\hat{\beta}(p)$ be searched within $\mathcal{B}$ so that it minimizes or maximizes $C_p(\beta)$ there. For $p \geq 1$, L_p estimation is necessarily done by minimization. Many efficient computational algorithms have been developed for this purpose: see, for example, Gonin and Money (1989, Chap. 1). For $p < 1$, $C_p(\beta)$ possesses multiple local minima at which the function is non-differentiable, and multiple local maxima at which it is differentiable. The number of local minima or maxima increases with the sample size n at a rate of order n^d. This poses difficulties in correct determination of $\hat{\beta}(p)$. We circumvent the problem by applying a specially-designed searching algorithm to automatically locate $\hat{\beta}(p)$. The algorithm first fixes a neighbourhood $\bar{\mathcal{B}}$ around an initial consistent estimate of β_0, smooths $C_p(\beta)$ by a moving weighted average method to get rid of the noisy local minima and maxima while retaining its major feature, and then proceeds to search for a global minimum or local maximum of the smoothed $C_p(\beta)$ over $\beta \in \bar{\mathcal{B}}$. Whether the result is a minimum or a maximum is determined adaptively by the feature of $C_p(\beta)$ without the user's prior interference. We now describe in detail Algorithm (A) for calculating $\hat{\beta}(p)$:

Step 1. Calculate an initial consistent estimate $\tilde{\beta}$ of β.

Step 2. Define $\bar{\mathcal{B}} = \{z \in \mathbb{R}^d : \|z - \tilde{\beta}\| \le r\}$ for some fixed $r > 0$. Let $z_1, \ldots, z_T$ denote points in the lattice $\{\tilde{\beta} + \delta[j_1, \ldots, j_d]^{\mathrm{T}} : j_1, \ldots, j_d = 0, \pm 1, \pm 2, \ldots\}$ which are contained in $\bar{\mathcal{B}}$, for some fixed $\delta > 0$.

Step 3. Construct a smooth approximation of $C_p(\beta)$ by the moving weighted average method given by

$$\tilde{C}_p(\beta) = \sum_{t=1}^{T} C_p(z_t) K\Big(\frac{\beta - z_t}{h}\Big) \Big/ \sum_{t=1}^{T} K\Big(\frac{\beta - z_t}{h}\Big),$$

where K is a d-variate kernel function and $h > 0$ denotes the bandwidth.

Step 4. Set $\bar{\mathcal{B}}^{[1]} = \bar{\mathcal{B}}$, $r^{[1]} = r$ and $l = 1$.

Step 5. Find the global maximizer, $\hat{\beta}^{[l]}$ say, of $\tilde{C}_p(\beta)$ over $\beta \in \bar{\mathcal{B}}^{[l]}$.

Step 6. If $\hat{\beta}^{[l]}$ lies in the interior of $\bar{\mathcal{B}}^{[l]}$, then terminate and return $\hat{\beta}(p) = \hat{\beta}^{[l]}$ as a local maximizer of $\tilde{C}_p(\beta)$. If $\hat{\beta}^{[l]}$ lies on the boundary of $\bar{\mathcal{B}}^{[l]}$, then shrink the ball $\bar{\mathcal{B}}^{[l]}$ to a new ball $\bar{\mathcal{B}}^{[l+1]} \subset \bar{\mathcal{B}}^{[l]}$, such that the two balls have boundaries touching at the point diametrically opposite to $\hat{\beta}^{[l]}$, share a common axis, and so that $\bar{\mathcal{B}}^{[l+1]}$ has radius $r^{[l+1]} = (1 - \rho) r^{[l]}$ for a fixed pre-determined shrinking factor $\rho \in (0, 1)$. If $r^{[l+1]}$ falls below a pre-determined lower limit R_L, then terminate and return $\hat{\beta}(p) = \hat{\beta}^{[l]}$ as a minimizer of $\tilde{C}_p(\beta)$; otherwise increment l and return to *Step* 5.

Algorithm (A) permits a local maximizer or minimizer to be adaptively identified according to the shape of $\tilde{C}_p(\beta)$. This proves to be crucial to our present regression context where the mode of L_p estimation cannot be a priori determined due to unavailability of f_U. For better and more stable performance of the algorithm, we can standardize the (Y_i, Z_i) beforehand to have zero sample means and unit sample variances.

For a practical choice of r we suggest the following scheme. Solve the $N \equiv \binom{n}{d}$ subsets of d simultaneous equations of the form $Y_i = Z_i^{\mathrm{T}} \beta$, $i = 1, \ldots, n$, for β to obtain solutions $\beta_1^{\dagger}, \ldots, \beta_N^{\dagger}$ respectively; set r to be a constant multiple of $(\log n)^{-1}$ times the 75% quantile of the N distances $\|\beta_1^{\dagger} - \tilde{\beta}\|, \ldots, \|\beta_N^{\dagger} - \tilde{\beta}\|$. This choice of r prescribes a reasonably robust region $\bar{\mathcal{B}}$ which contains, and shrinks toward, the true parameter β_0 in probability, provided that $\tilde{\beta}$ converges to β_0 at a polynomial rate. Our empirical experience shows that Algorithm (A) performs stably and efficiently if the search is confined to the region $\bar{\mathcal{B}}$ for a start. Convenient choices of the initial consistent estimate $\tilde{\beta}$ are $\hat{\beta}(1)$ or $\hat{\beta}(2)$. The former is less sensitive to outliers in the dataset, whilst the latter maintains a more stable convergence rate over the entire class of error distributions under consideration.

Step 3 involves smoothing of C_p, which can be skipped so that the search is carried out directly on $C_p(\beta)$. In this case the resulting estimate corresponds typically to the value of β that gives rise to the highest local maximum or the minimum in $\bar{\mathcal{B}}$, and proves to be asymptotically valid. Finite-sample performance of our searching algorithm can nevertheless be improved by a small amount of smoothing as done in *Step* 3, which enables more convenient detection of the estimate by alleviating the obscuring effects of the multiple local maxima present in $C_p(\beta)$. The following lemma describes sufficient conditions for $\tilde{C}_p$ and C_p to be at most $o_p(\varepsilon_n)$ apart, for $\varepsilon_n \downarrow 0$. The proof is given in the Appendix.

Lemma 1. *Assume the conditions of Theorem* 2. *Let $\varepsilon_n \downarrow 0$ be fixed. Suppose that K is a d-variate differentiable density function such that, for some $\lambda_0 > d$ and $\lambda_1 > 0$,*

$$K(x) = O(\|x\|^{-\lambda_0}) \quad \textit{and} \quad \|\nabla K(x)\| = O(\|x\|^{-\lambda_1}), \tag{4.1}$$

and set, with $\bar{p} \overset{\text{def}}{=} \min\{p, 1\}$,

$$\delta = o\left(h^{1+\frac{d^2}{d+\lambda_1}}\right) \quad \textit{and} \quad h = o\left(\varepsilon_n^{\bar{p}^{-1}\frac{\lambda_0+\bar{p}}{\lambda_0-d}}\right). \tag{4.2}$$

Then $\sup_{\beta\in\mathcal{B}} |\tilde{C}_p(\beta) - C_p(\beta)| = o_p(\varepsilon_n)$.

Theorem 2 asserts that the asymptotics of $\hat{\beta}(p)$ remains unaltered provided $\hat{\beta}(p)$ minimizes or locally maximizes $C_p(\beta)$ up to order $o_p(r_n^{-\varrho} l(n))$, for some $\varrho > 0$ and function l slowly varying at infinity. It thus follows from Lemma 1, with ε_n set to $r_n^{-\varrho} l(n)$, that the above condition is guaranteed by choosing a kernel function K with properties (4.1), selecting a sufficiently small bandwidth h, and fixing a sufficiently fine lattice $\{z_1, \ldots, z_T\}$ such that (4.2) holds. We see from (4.2) that choices of both δ and h are subject to more stringent conditions in higher-dimensional problems as d increases. In practice h should not be chosen to be too small, or the smoothing effect will not be adequate to fend off the undesirable perturbations caused by the multiple local maxima. In view of the relations between (λ_0, λ_1) and (4.2), a kernel function with exponentially decaying tails, such as the standard d-variate normal density, is recommended to allow for more flexibility in the selections of h and δ.

4.2. Algorithm (B): m out of n bootstrap estimation of log mean squared error

We have established in Theorem 3 consistency of m out of n bootstrap estimation of the distribution of $r_n(\hat{\beta}(p) - \beta_0)$ for a majority of cases under

(A1)−(A4). However, that r_n is in general unknown renders m out of n bootstrap estimation of $MSE(\hat{\beta}(p))$ not immediately possible. We develop below Algorithm (B) for consistently estimating $\log MSE(\hat{\beta}(p))$ based on repeated applications of the m out of n bootstrap, without explicit specifications of r_m, r_n.

Step 1. For some fixed $0 < \alpha_1 < \cdots < \alpha_S < 1$, set m_s to be the integer part of n^{α_s}, $s = 1, \ldots, S$.

Step 2. For each $s = 1, \ldots, S$, draw B bootstrap samples, each of size m_s, from (Y_1, Z_1), …, (Y_n, Z_n); calculate the L_p estimate $\hat{\beta}_s^{*(b)}(p)$ using Algorithm (A) from the bth bootstrap sample, $b = 1, \ldots, B$.

Step 3. For each $s = 1, \ldots, S$, calculate $T_s^*(p) = \log\{B^{-1}\sum_{b=1}^{B}(\hat{\beta}_s^{*(b)}(p) - \hat{\beta}(p))^2\}$ and $U_s = \log(n/m_s)$. Then calculate $\bar{T}^*(p) = S^{-1}\sum_{s=1}^{S} T_s^*(p)$, $\bar{U} = S^{-1}\sum_{s=1}^{S} U_s$, $S_{UU} = \sum_{s=1}^{S}(U_s - \bar{U})^2$, $\hat{\gamma}(p) = (2S_{UU})^{-1}\sum_{s=1}^{S}(U_s - \bar{U})T_s^*(p)$ and $\hat{G}(p) = \bar{T}^*(p) - 2\bar{U}\hat{\gamma}(p)$.

We prove the following lemma in the Appendix. It asserts that $\hat{G}(p)$ consistently estimates $\log MSE(\hat{\beta}(p))$ under conditions sufficient for m out of n bootstrap consistency.

Lemma 2. *Suppose that $\mathbb{E}|U_1|^\nu < \infty$ for some $\nu > 0$ and that $p \in (0, 1 + \nu/2]$ is fixed. Assume the conditions of Theorem 3 and that $r_n^2(\hat{\beta}(p) - \beta_0)^2$ and $r_m^2(\hat{\beta}^*(p) - \hat{\beta}(p))^2$ are uniformly integrable, the latter being conditional on $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. Then $\hat{G}(p)/\log MSE(\hat{\beta}(p)) = 1 + o_p(1)$.*

To exclude cases where m out of n bootstrappability of L_p regression has not been confirmed by Theorem 3, we may restrict our choice of p to $[1, \infty) \cup \{p \in (1/2, 1) : 2(1-p)\hat{\gamma}(p) \le 1\} \cup \{p \in (0, 1/2] : \hat{\gamma}(p) \le 1/2\}$ in our empirical determination of p. Our remark following Theorem 3 implies that such restriction incurs no essential loss of ratewise efficiency.

The convergence rate of the m out of n bootstrap estimator $\hat{G}(p)$ clearly depends on both p and the choice of sizes m_s. In general, a smoother criterion function such as that corresponding to $p > 1$ results in a faster rate, whereas the bootstrap may work less satisfactorily for $p \le 1$ even if $m_s = o(n)$ is chosen optimally to achieve the best rate. Related discussions can be found in Hall and Martin (1988), DeAngelis, Hall and Young (1993) and Cheung and Lee (2005) for the case $p = 1$.

4.3. Adaptive procedure for determining p

With the aid of Algorithms (A) and (B), we propose an adaptive procedure for selecting p in L_p estimation of β_0 under conditions (A0)−(A4). The procedure consists of the following steps.

Step 1. Select a grid of candidate values $p_1, \ldots, p_W \in (0, \infty)$ for p.
Step 2. For each $w = 1, \ldots, W$, apply Algorithm (B) to estimate $\log MSE(\hat{\beta}(p_w))$.
Step 3. Set $\hat{p}$ to be that p_w which minimizes the $\hat{G}(p_w)$ values.

To fully exploit the range of achievable convergence rates, we should take a large W and select the p_w's from both sides of 1. An upper bound of at least 2 may be imposed on our choice of p without loss of generality. To see the optimality property of $\hat{p}$, consider two L_p estimators $\hat{\beta}(q_1)$ and $\hat{\beta}(q_2)$ with convergence rates $n^{\gamma_1}\ell_1(n)$ and $n^{\gamma_2}\ell_2(n)$, respectively, for some $\gamma_1, \gamma_2 > 0$ and some slowly-varying functions ℓ_1 and ℓ_2. Then

$$\begin{aligned}(\log n)^{-1} &\Big\{\log MSE(\hat{\beta}(q_1)) - \log MSE(\hat{\beta}(q_2))\Big\} \\ &= (\log n)^{-1}(\hat{G}(q_1) - \hat{G}(q_2)) + o_p(1) = -2(\gamma_1 - \gamma_2) + o_p(1).\end{aligned}$$

It follows that $\hat{\beta}(\hat{p})$ has the fastest convergence rate, up to a slowly varying factor, among the candidate values $p_1, \ldots, p_W$.

We illustrate application of our adaptive procedure with a dataset drawn from Montgomery and Peck (1992) that contains twenty observations on the tool life (Y_i) and the lathe speed in revolutions per minute (S_i). A simple linear regression model $Y_i = \beta_1 + \beta_2 S_i + U_i$ was fitted to the data under an unspecified f_U. To improve stability of Algorithm (A) we first standardized the data (Y_i, S_i) to have zero sample means and unit sample variances. For $p < 1$, there are altogether $\binom{20}{2}$ local minima at which $C_p(\beta)$ is non-differentiable. The radius of the circular region $\bar{\mathcal{B}}$ was calculated to be $r = 2.3564$. We set $\delta = 0.04713$ and $h = 0.09426$ in Algorithm (A). For Algorithm (B), we selected six bootstrap sample sizes, $m_1 = 5$, $m_2 = 7$, $m_3 = 9$, $m_4 = 11$, $m_5 = 13$ and $m_6 = 15$, and drew 500 bootstrap samples for each size. A total of 121 distinct points from the interval $[0.03, 2]$ were selected as candidate values for p. Figure 1 displays the estimates $\hat{G}(p)$, which attain a minimum at about $p = 0.87$. The corresponding $L_{0.87}$ estimate is $(26.58, -0.002303)$. The plot of $\hat{G}(p)$ exhibits a sharp drop as p passes from 0.8 to 0.9, where the mode of L_p estimation switches from local maximization to global minimization. Figure 2 compares the fitted $L_{0.87}$ regression line with those obtained from the more conventional L_2 and L_1 fits.

5. L_p Estimation under Asymmetric f_U

Extension of our theory and adaptive L_p procedure to asymmetric f_U is possible if $d > 1$ and the regression model (1.1) contains an intercept term such that $Z_1 = [1, W_1]^{\mathrm{T}}$, with W_1 having a nondegenerate distribution in $\mathbb{R}^{d-1}$. In

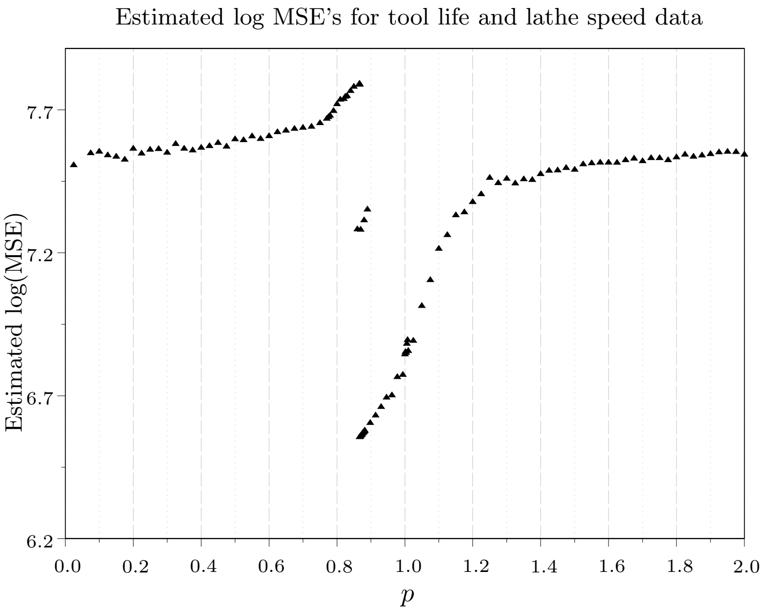
Estimated log MSE's for tool life and lathe speed data

Figure 1. Tool life and lathe speed example — estimates of log MSE's of L_p estimates under selected values of p.

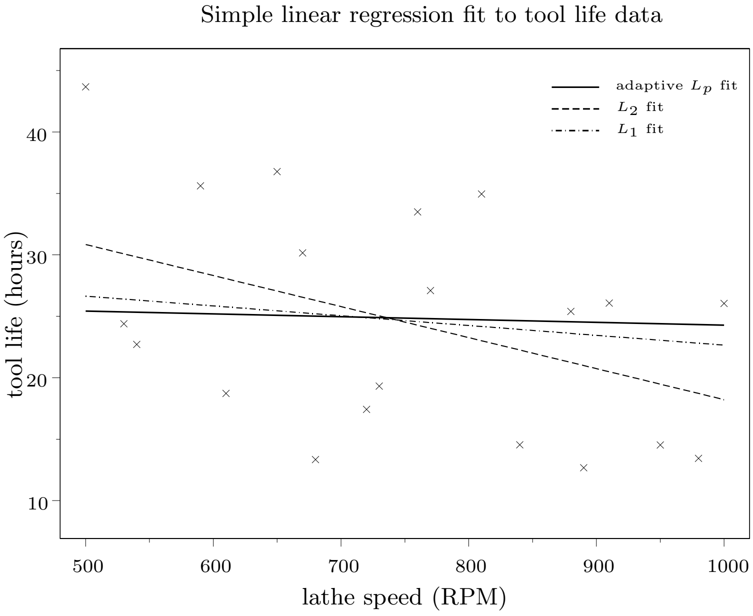
Simple linear regression fit to tool life data

Figure 2. Tool life and lathe speed example — fitted L_p regression lines for $p = 0.87$, 1 and 2.

this case $\hat{\beta}(p)$ converges in probability to $[\beta_{01}(p), \beta_{02}, \ldots, \beta_{0d}]^{\mathrm{T}}$, where $\beta_{01}(p)$ depends on p and satisfies $\mathbb{E}[\,\mathrm{sgn}(U_1 - \beta_{01}(p))|U_1 - \beta_{01}(p)|^{p-1}] = 0$, and the regression parameters $\beta_{02}, \ldots, \beta_{0d}$ remain independent of p. Thus adaptive L_p estimation still makes sense if we are interested in estimating only the regression parameters but not the intercept term.

To fix ideas, we consider a fixed $p > 0$ and suppose that $f_U(u) = (u - \beta_{01}(p))^{\zeta_p^+ - 1}\mathcal{L}_p^+(u - \beta_{01}(p))$ if $u > \beta_{01}(p)$ and $f_U(u) = (\beta_{01}(p) - u)^{\zeta_p^- - 1}\mathcal{L}_p^-(\beta_{01}(p) - u)$ otherwise, for some $\zeta_p^+, \zeta_p^- > 0$ and nonnegative functions $\mathcal{L}_p^+, \mathcal{L}_p^-$ slowly varying near 0. Assume without loss of generality that $\zeta_p^+ \geq \zeta_p^-$. We can show that the asymptotic properties of $\hat{\beta}(p)$ are determined by the shape of $f_U(u)$ for $u \leq \beta_{01}(p)$. Results of our Theorems 2 and 3 then carry over if we replace $(\zeta, \mathcal{L})$ by $(\zeta_p^-, \mathcal{L}_p^-)$ and slightly modify the boundary conditions of the various cases in Theorem 2. Technical details are given in Lai and Lee (2003). We remark also that the best possible convergence rates, as derived from Theorem 1, are in general ζ- and q-specific under (A3) and (A4), respectively. Ratewise efficiency should therefore be attributed to any estimator of $[\beta_{02}, \ldots, \beta_{0d}]^{\mathrm{T}}$ which achieves the fastest rate maximized over all location shifts of f_U that give rise to possibly different values of ζ or q. Our adaptive L_p procedure can be applied without change to this general situation. Whether the resulting adaptive L_p estimator is ratewise efficient remains, however, an open question.

6. Simulation Study

We conducted a simulation study to compare our adaptive L_p procedure with other methods for location estimation. Consider a location model $Y_i = \beta_0 + U_i$, where U_i has the density function $f_U(u) \propto |u|^{\zeta-1}L(|u|)1\{|u| \leq 1\}$ with probability 0.75 and $f_U(u) \propto |u|^{-3.01}1\{|u| > 1\}$ with probability 0.25, where $1\{\cdot\}$ is the indicator function. The true value β_0 was fixed at 0. Note that the specification of f_U above allows for a heavy-tailed component which accounts for 25% of the complete distribution. We considered in our study the following density shapes: (a) $\zeta = 0.3$, $L(|u|) \equiv 1$; (b) $\zeta = 0.8$, $L(|u|) \equiv 1$; (c) $\zeta = 1$, $L(|u|) = 2 - |u|^{0.25}$; (d) $\zeta = 1$, $L(|u|) = 2 - |u|^2$; (e) $\zeta = 1$, $L(|u|) = 1 + |u|^{0.25}$; (f) $\zeta = 1$, $L(|u|) = 1 + |u|^2$; (g) $\zeta = 1.3$, $L(|u|) \equiv 1$; and (h) $\zeta = 1.8$, $L(|u|) \equiv 1$. All cases except (d) and (f) favour use of $p < 1$ asymptotically. We approximated the mean squared error of each estimator by averaging over 1,000 random samples of size $n = 50$ drawn for each case. The range of p was restricted in $(0, 2]$.

Included in the study were the non-adaptive L_1, $L_{1.5}$ and L_2 estimates, as well as the following three adaptive approaches.

(i) Our adaptive L_p estimate, $\hat{\beta}(\hat{p})$ — for which the bootstrap sample sizes m_s were set to be 15, 20, 25, 30 and 35 in Algorithm (B).

(ii) Adaptive L_p estimate based on sample kurtosis — for which we combined the results of Harter (1974–1975) and Sposito and Hand (1983) to establish a rule for choosing p: set $p = 2$ if $\hat{\kappa}_4 \leq 2.2$, and set $p = 9/\hat{\kappa}_4^2 + 1$ if $\hat{\kappa}_4 > 2.2$, where $\hat{\kappa}_4 = n\sum_{i=1}^n (Y_i - \bar{Y})^4/\{\sum_{i=1}^n (Y_i - \bar{Y})^2\}^2$ is the sample kurtosis and $\bar{Y}$ denotes the sample mean.

(iii) Arcones' (2005) adaptive L_p estimate based on asymptotic MSE — for which p was taken to minimize $(p-1)^{-2}n^{-1}\sum_{i=1}^n |Y_i - \hat{\beta}(1)|^{2p-2}\{n^{-1}\sum_{i=1}^n |Y_i - \hat{\beta}(1)|^{p-2})\}^{-2}$, which, as implied by Arcones (2005), is consistent for the MSE of $\hat{\beta}(p)$ for $p > 1$ under $n^{1/2}$-consistency conditions.

Note that the adaptive approaches (ii) and (iii) confine the choice of p to only values greater than 1.

We see from Figure 3 that the MSE of our adaptive estimate $\hat{\beta}(\hat{p})$ is generally the smallest under each density considered. Compared to our procedure, both adaptive approaches (ii) and (iii) are relatively less accurate, with approach (iii) outperforming (ii) slightly. The $L_{1.5}$ estimate lies somewhere between (ii) and (iii), but the L_1 and L_2 estimates are markedly less accurate. Particularly poor is the performance of $\hat{\beta}(2)$, which is adversely affected by the heavy tails present in our underlying densities. Accuracy of $\hat{\beta}(1)$ deteriorates as ζ increases, which

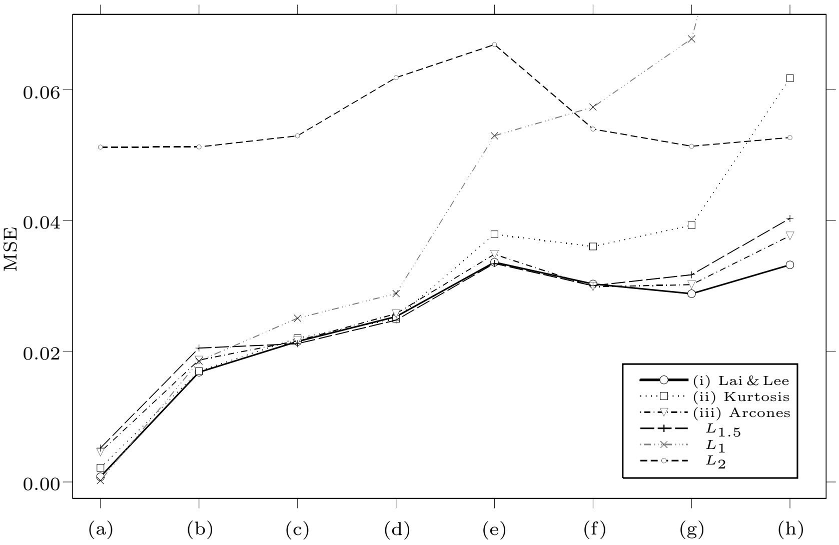

Figure 3. MSE's of L_p estimators under various error densities : (a) $\zeta = 0.3$; (b) $\zeta = 0.8$; (c) upwardly pointed, $\zeta = 1$, $q = 0.25$; (d) upwardly pointed, $\zeta = 1$, $q = 2$; (e) downwardly pointed, $\zeta = 1$, $q = 0.25$; (f) downwardly pointed, $\zeta = 1$, $q = 2$; (g) $\zeta = 1.3$; and (h) $\zeta = 1.8$.

Table 1. Means of adaptive choices of p over 1,000 samples under different procedures. Standard deviations are given in square parentheses.

Density	q	ζ	Lai & Lee's approach	Sample Kurtosis	Arcones's method
(a)		0.3	1.024 [0.067]	1.227 [0.225]	1.321 [0.127]
(b)		0.8	1.368 [0.380]	1.343 [0.328]	1.530 [0.232]
(c)	0.25	1	1.420 [0.472]	1.359 [0.332]	1.565 [0.252]
(d)	2	1	1.396 [0.481]	1.352 [0.325]	1.580 [0.249]
(e)	0.25	1	1.583 [0.574]	1.381 [0.346]	1.647 [0.271]
(f)	2	1	1.678 [0.507]	1.421 [0.363]	1.709 [0.276]
(g)		1.3	1.653 [0.502]	1.410 [0.369]	1.725 [0.270]
(h)		1.8	1.760 [0.455]	1.451 [0.385]	1.806 [0.252]

is not surprising in view of its convergence rate of $r_n = n^{1/(2\zeta)}$. The advantage of our adaptive procedure over $\hat{\beta}(1.5)$ is discernible in cases (a), (b), (g) and (h); whereas in cases (c)$-$(f), the two methods have very similar performance. A plausible explanation can be found by examining the ratio $MSE(\hat{\beta}(\hat{p}))/MSE(\hat{\beta}(1.5))$ which, as asserted by Theorems 1 and 2, has orders $n^{-65/12}$, $n^{-3/2}$, $n^{-1/3}$, n^0, $n^{-1/3}$, n^0, $n^{-7/13}$ and $n^{-1/9}$ under cases (a)$-$(h), respectively. This suggests that the discrepancy between the two methods should be most notable, at least asymptotically, in cases (a), (b) and, to a lesser extent, case (g).

Table 1 reports the means and standard deviations of p found under the adaptive approaches (i), (ii) and (iii), respectively. In general, the values of p selected by all three procedures decrease as ζ decreases. This agrees broadly with our theoretical assertion that a faster convergence rate is associated with a small p for small ζ. Note also that our adaptive procedure occasionally selected $\hat{p} < 1$ for calculating $\hat{\beta}(\hat{p})$ when ζ was small, and thus adapted more effectively to the shape of f_U around 0.

7 Conclusion and Discussion

We establish the best possible convergence rates for regression-equivariant estimators under a broad class of error densities, and thereby define a notion of ratewise efficiency. Similar to the parametric maximum likelihood approach, the L_p method is shown to provide a general strategy for constructing ratewise efficient estimators, at least up to a slowly-varying factor, provided p is selected from $(0, \infty)$ rather than the conventional, more restrictive, range $[1, 2]$. Indeed, ratewise efficiency can be achieved under an arbitrarily wide range of error density shapes by a sufficiently small p. We propose also an adaptive procedure for L_p regression which succeeds in adapting the choice of p to this general class of error densities to yield an approximately ratewise efficient estimator $\hat{\beta}(\hat{p})$. That the precise mode of L_p estimation and the convergence rate r_n of $\hat{\beta}(p)$ depend on p and f_U poses two practical difficulties that are successfully circumvented by Algorithms (A) and (B) respectively. Our simulation study shows that our adaptive procedure yields smaller MSE than other L_p estimates.

One major criticism against the classical L_2 method concerns its undesirable sensitivity to heavy tails of f_U. Such non-robustness is in fact characteristic of any L_p procedure with $p > 1$, for which the convergence rates never exceed and can be markedly slower than $n^{1/2}$. It is in this context that use of $p \leq 1$ offers another advantage, in addition to being ratewise efficient under the density class (2.1). We see from Theorem 2 that results for L_p estimators based on $p \leq 1$ require no moment conditions on f_U and are therefore robust against heavy tails. We also see that a convergence rate of at least $n^{1/2}$ is guaranteed by choosing $p \in (1/2, 1)$ irrespective of the central or tail behaviour of f_U. This suggests that our adaptive procedure may opt to restrict the candidate values $p_1, \ldots, p_W$ within the range $(1/2, 1)$ if we are willing to trade a little ratewise efficiency for robustness against heavy-tailed f_U. It should be remarked though that even such choices of p may not be robust against outliers in the covariate space created by, for example, points which are highly influential or have high leverages.

Acknowledgement

This research was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7104/01P). The second author for correspondence. Supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7128/02P).

Appendix

In what follows C stands for a universal positive constant which may vary from occurrence to occurrence.

A.1. Proof of Theorem 1

Proofs of parts (i)$-$(iii) are modelled after Ibragimov and Has'minskii (1981).

To prove (i), note first that the experiment $\mathcal{E}$ generated by (Y_1, Z_1) under model (1.1) is regular, which follows from continuity of f_U' and finiteness of the Fisher information $\mathbb{E}[Z_1 Z_1^{\mathrm{T}}]I_U$. Denote by $H(\beta_1, \beta_2)$ the Hellinger distance between the distributions of (Y_1, Z_1) under $\beta_0 = \beta_1$ and $\beta_0 = \beta_2$. Under regularity of $\mathcal{E}$, Theorem I.7.6 of Ibragimov and Has'minskii (1981) implies that $H(\beta_1, \beta_2)^2 \leq (1/4)\|\beta_1 - \beta_2\|^2 \mathbb{E}[Z_1^{\mathrm{T}} Z_1]I_U$. Part (i) then follows from the proof of Ibragimov and Has'minskii's (1981) Theorem I.6.1, and equivariance of $\hat{\beta}$.

Under (A2), the function $\psi(u) \overset{\text{def}}{=} f_U(u) - |u|\mathcal{L}(0)$ is clearly twice continuously differentiable with $\psi(0) = \psi'(0) = 0$. It follows that the arguments used for proving Ibragimov and Has'minskii's (1981) Theorem II.5.1 hold almost surely conditional on $(Z_1,\ldots,Z_n)$. Taking expectation with respect to $(Z_1,\ldots,Z_n)$ then verifies local asymptotic normality (LAN) of model (1.1) with convergence rate $(n\log n)^{1/2}$. Part (ii) now follows from the LAN property and Theorem II.12.1 of Ibragimov and Has'minskii (1981) with the loss function set to be the Euclidean norm.

Under (A3), note by Theorem VI.1.1 of Ibragimov and Has'minskii (1981) that there exists $\epsilon^*>0$ such that

$$\Pi(\epsilon)\overset{\text{def}}{=}\int\left|f_U(u-\epsilon)^{\frac{1}{2}}-f_U(u)^{\frac{1}{2}}\right|^2du\le C|\epsilon|^{\zeta},\quad |\epsilon|<\epsilon^*. \tag{A.1}$$

Write $H(\beta_1,\beta_2)^2=H_1+H_2$, where $H_1=\mathbb{E}\left[\Pi(Z_1^{\mathrm{T}}(\beta_1-\beta_2));|Z_1^{\mathrm{T}}(\beta_1-\beta_2)|\ge\epsilon^*\right]$ and $H_2=\mathbb{E}\left[\Pi(Z_1^{\mathrm{T}}(\beta_1-\beta_2));|Z_1^{\mathrm{T}}(\beta_1-\beta_2)|<\epsilon^*\right]$. Note that

$$H_1\le 4\mathbb{P}(|Z_1^{\mathrm{T}}(\beta_1-\beta_2)|\ge\epsilon^*)\le 4\epsilon^{*-2}\|\beta_1-\beta_2\|^2\mathbb{E}\|Z_1\|^2\le C\|\beta_1-\beta_2\|^2,$$

and that, by (A.1), $H_2\le C\mathbb{E}\left[|Z_1^{\mathrm{T}}(\beta_1-\beta_2)|^{\zeta};|Z_1^{\mathrm{T}}(\beta_1-\beta_2)|<\epsilon^*\right]\le C\|\beta_1-\beta_2\|^{\zeta}$. Combining the above bounds on H_1,H_2 and noting that $\zeta<2$, we obtain that $H(\beta_1,\beta_2)^2\le C\|\beta_1-\beta_2\|^{\zeta}$ for sufficiently small $\|\beta_1-\beta_2\|$. Part (iii) now follows from the proof of Ibragimov and Has'minskii's (1981) Theorem I.6.1, and equivariance of $\hat\beta$.

To prove (iv), assume (A4) and note by equivariance of $\hat\beta$ and Schwarz's inequality that, for any $c,\gamma\in\mathbb{R}^d$ with $c^{\mathrm{T}}\gamma\ne 0$,

$$\begin{aligned}1&=\left\{\mathbb{E}\left[c^{\mathrm{T}}(\hat\beta-\beta_0)(c^{\mathrm{T}}\gamma)^{-1}\left\{\prod_{i=1}^n\left[\frac{f_U(U_i-Z_i^{\mathrm{T}}\gamma)}{f_U(U_i)}\right]-1\right\}\right]\right\}^2\\&\le\mathbb{E}\{c^{\mathrm{T}}(\hat\beta-\beta_0)\}^2(c^{\mathrm{T}}\gamma)^{-2}\mathbb{E}\left[\prod_{i=1}^n\frac{f_U(U_i-Z_i^{\mathrm{T}}\gamma)^2}{f_U(U_i)^2}-1\right]\\&=\mathbb{E}\{c^{\mathrm{T}}(\hat\beta-\beta_0)\}^2(c^{\mathrm{T}}\gamma)^{-2}\left\{(1+\mathcal{I}(\gamma))^n-1\right\},\end{aligned}$$

where $\mathcal{I}(\gamma)=\mathbb{E}\int f_U(u-Z_1^{\mathrm{T}}\gamma)^2/f_U(u)\,du-1$, so that

$$\mathbb{E}\{c^{\mathrm{T}}(\hat\beta-\beta_0)\}^2\ge(c^{\mathrm{T}}\gamma)^2\left\{(1+\mathcal{I}(\gamma))^n-1\right\}^{-1}. \tag{A.2}$$

Note that (A.2) holds trivially if $c^{\mathrm{T}}\gamma=0$; so we remove from now on the constraint $c^{\mathrm{T}}\gamma\ne0$. Putting c to be the d standard basis vectors in $\mathbb{R}^d$ successively in (A.2) and then summing, we get

$$\mathbb{E}\|\hat\beta-\beta_0\|^2\ge\|\gamma\|^2\left\{(1+\mathcal{I}(\gamma))^n-1\right\}^{-1}. \tag{A.3}$$

Define, for $x>0$, $\mathcal{I}_q(x)\overset{\text{def}}{=}x^{2q+1}$, $x^2|\log x|$ and x^2 for $q<1/2$, $=1/2$ and $>1/2$, respectively. It follows by Polfeldt's (1970) Theorem, under the moment condition on F_Z, that there exist constants $C_1,C_2>0$ such that $C_1<$

$|\mathcal{I}(\gamma)|/\mathcal{I}_q(\|\gamma\|) < C_2$ for sufficiently small $\|\gamma\|$. Thus we have, by putting $\|\gamma\| = \mathcal{I}_q^{-1}(n^{-1})$ in (A.3), that $\mathbb{E}\|\hat{\beta}-\beta_0\|^2 \geq C\mathcal{I}_q^{-1}(n^{-1})^2 \geq C\varphi_q(n)^{-2}$ for some positive constant C independent of n and the choice of $\hat{\beta}$, which implies (iv).

A.2. Proofs of Theorems 2 and 3

Details of the proof of Theorem 2 can be found in Appendices A.1 (for cases (i)−(vii)) and A.2 (for cases (viii)−(xiii)) of Lai and Lee (2005). Appendix A.3 of the same paper contains the proof of Theorem 3.

As an illustration we outline below the proof of Theorem 2 as adapted from Lai and Lee (2005). Define, for any fixed $s, t \in \mathbb{R}^d$ and $\lambda > 0$, $D(s/\lambda) = \mathbb{E}|U_1 - Z_1^{\mathrm{T}} s/\lambda|^p - \mathbb{E}|U_1|^p$ and $E(s/\lambda, t/\lambda) = \mathbb{E}\left[|U_1 - Z_1^{\mathrm{T}} s/\lambda|^p - |U_1 - Z_1^{\mathrm{T}} t/\lambda|^p\right]^2$. Then there exist $2\nu > \omega > 0$ and functions $L_1, L_2 : (0,\infty) \to [0,\infty)$, slowly varying at ∞, such that $\lambda^\omega L_1(\lambda) E(s/\lambda, t/\lambda)$ and $\lambda^\nu L_2(\lambda) D(s/\lambda)$ have finite, nontrivial, limits as $\lambda \to \infty$, for any $s, t \in \mathbb{R}^d$. The latter limit is either strictly positive or strictly negative for all $s \neq 0$. Define, for $\beta, z \in \mathbb{R}^d$ and $y \in \mathbb{R}$, $m_\beta(y,z)$ to be $-|y - z^{\mathrm{T}}\beta|^p$ times the sign of the above limit. For $s \in \mathbb{R}^d$, define $\tilde{m}_s = m_{\beta_0+s} - m_{\beta_0}$ and $\mathbb{M}_n(s) = n^{1/2} r_n^{\omega/2} L_1(r_n)^{1/2}\left(n^{-1}\sum_{i=1}^n \tilde{m}_{s/r_n}(Y_i, Z_i) - \mathbb{E}\,\tilde{m}_{s/r_n}(Y_1, Z_1)\right)$.

Consistency of $\hat{\beta}(p)$ for β_0 follows from a Glivenko-Cantelli theorem for the class of functions m_β and the fact that β_0 is a well-separated maximizer of $\mathbb{E}\, m_\beta(Y_1, Z_1)$. Note that r_n satisfies $r_n^{\nu-\omega/2} L_2(r_n) L_1(r_n)^{-1/2} \sim n^{1/2}$ in general. By assumption we have $n^{-1}\sum_{i=1}^n m_{\hat{\beta}(p)}(Y_i, Z_i) \geq n^{-1}\sup_{\beta\in\mathcal{B}}\sum_{i=1}^n m_\beta(Y_i, Z_i) - o_p(r_n^{-\nu} L_2(r_n)^{-1})$. Define $S_{j,n} = \{\beta : 2^{j-1} < r_n\|\beta - \beta_0\| \leq 2^j\}$. For fixed $\eta, M > 0$, we have, using properties of $\hat{\beta}(p)$ and maximal inequalities,

$$\begin{aligned}
&\mathbb{P}\left(r_n\|\hat{\beta}(p) - \beta_0\| > 2^M\right)\\
&\leq \sum_{M\leq j\leq \log_2 \eta r_n} \mathbb{P}\Big(n^{-1}\sup_{\beta\in S_{j,n}}\sum_{i=1}^n \tilde{m}_{\beta-\beta_0}(Y_i, Z_i) \geq -\, 2^{2M-1} r_n^{-\nu} L_2(r_n)^{-1}\Big) + o(1)\\
&= O\Big(\sum_{j\geq M} 2^{-j(\nu-\frac{\omega}{2})}\Big) + o(1) \to 0 \qquad \text{as } n, M \to \infty,
\end{aligned}$$

so that $r_n(\hat{\beta}(p) - \beta_0) = O_p(1)$.

It is easy to verify that any finite-dimensional covariance matrix of $\mathbb{M}_n$ converges to a nonsingular limit. Define $M_\delta(y,z)$ to be the minimum or maximum of $\left\{\delta\|z\|\,|y - z^{\mathrm{T}}\beta_0|^{p-1}, \delta^p\|z\|^p\right\}$ according as $p < 1$ or $p \geq 1$, respectively. The Lindeberg-Feller condition is implied by the condition that for any $c, \eta > 0$,

$$r_n^\omega L_1(r_n)\mathbb{E}\left[M_{\frac{c}{r_n}}(Y_1, Z_1)^2; M_{\frac{c}{r_n}}(Y_1, Z_1) > \eta n^{\frac{1}{2}} r_n^{-\frac{\omega}{2}} L_1(r_n)^{-\frac{1}{2}}\right] \to 0.$$

This in turn follows from the fact that $r_n^{p\vee 1}n^{1/2}r_n^{-\omega/2}L_1(r_n)^{-1/2}\to\infty$ for cases (i)−(vii) and that $r_n=o(n)$ and $r_n\to\infty$ for cases (viii)−(xiii), thus confirming finite-dimensional weak convergence of $\mathbb{M}_n$ to a zero-mean Gaussian process $\mathbb{G}$. This, together with its stochastic equicontinuity, which follows by applying Theorem 2.11.22 of Van der Vaart and Wellner (1996, p.220), proves that $\mathbb{M}_n$ converges weakly to $\mathbb{G}$ as a process.

Note that $r_n(\hat\beta(p)-\beta_0)$ maximizes the process $s\mapsto n^{-1}\sum_{i=1}^n\tilde m_{s/r_n}(Y_i,Z_i)$ up to $o_p(r_n^{-\nu}L_2(r_n)^{-1})$, so that, for sufficiently large n,

$$\begin{aligned}&\mathbb{M}_n(r_n(\hat\beta(p)-\beta_0))+r_n^{\nu}L_2(r_n)\mathbb{E}\,\tilde m_{\hat\beta(p)-\beta_0}(Y_1,Z_1)\\ &\geq\sup_{s\in\mathbb{R}^d}\Big\{\mathbb{M}_n(s)+r_n^{\nu}L_2(r_n)\mathbb{E}\,\tilde m_{\frac{s}{r_n}}(Y_1,Z_1)\Big\}-o_p(1).\end{aligned}$$

Theorem 2 then follows by existence of $\lim_{\lambda\to\infty}\lambda^{\nu}L_2(\lambda)\mathbb{E}\,\tilde m_{s/\lambda}(Y_1,Z_1)$ and weak convergence of $\mathbb{M}_n$ to $\mathbb{G}$.

A.3. Proof of Lemma 1

Note first that, for any $\beta\in\mathcal{B}$,

$$\Big|\tilde C_p(\beta)-C_p(\beta)\Big|\leq\frac{\sum_{t=1}^T|C_p(\beta)-C_p(z_t)|K(\frac{\beta-z_t}{h})}{\sum_{t=1}^T K(\frac{\beta-z_t}{h})}. \tag{A.4}$$

Approximation by integration gives that, for some constant $C>0$,

$$\inf_{\beta\in\mathcal{B}}\sum_{t=1}^T K(\frac{\beta-z_t}{h})\geq C\Big(\frac{h}{\delta}\Big)^d(1+o(1)), \tag{A.5}$$

provided δ satisfies (4.2). For $p<1$, we have

$$|C_p(\beta)-C_p(z_t)|=O_p\Big(n^{-1}\sum_{i=1}^n|(\beta-z_t)^{\mathrm{T}}Z_i|^p\Big)=O_p(\|\beta-z_t\|^p). \tag{A.6}$$

For any arbitrary $a>0$, observe, by considering $\|\beta-z_t\|\leq ah$ and $\|\beta-z_t\|>ah$ separately, that

$$\sum_{t=1}^T\|\beta-z_t\|^pK(\frac{\beta-z_t}{h})=O_p\left\{(ah)^p\Big(\frac{h}{\delta}\Big)^d+a^{-\lambda_0}\delta^{-d}\right\}. \tag{A.7}$$

It follows by (A.4), (A.5), (A.6), (A.7), and setting $a=h^{-(p+d)/(p+\lambda_0)}$ that

$$\sup_{\beta\in\mathcal{B}}\Big|\tilde C_p(\beta)-C_p(\beta)\Big|=O_p\Big(h^{\frac{p(\lambda_0-d)}{\lambda_0+p}}\Big). \tag{A.8}$$

For $p \geq 1$, we have $|C_p(\beta) - C_p(z_t)| = O_p\left(\|\beta - z_t\|^p + \|\beta - z_t\|\right)$. Applying similar arguments as for $p < 1$ gives that

$$\sup_{\beta \in \mathcal{B}} \left| \tilde{C}_p(\beta) - C_p(\beta) \right| = O_p\Big(h^{\frac{\lambda_0 - d}{\lambda_0 + 1}} \Big). \tag{A.9}$$

Lemma 1 then follows from (A.8) and (A.9), provided h satisfies (4.2).

A.4. Proof of Lemma 2

Let $p \leq 1 + \nu/2$ be fixed. Assume $B = \infty$. Denote by $\hat{\beta}_s^*(p)$ the L_p estimate calculated from a generic bootstrap sample of size m_s. Then

$$\exp(T_s^*(p)) = \mathbb{E}\left[(\hat{\beta}_s^*(p) - \hat{\beta}(p))^2 \mid (Y_1, Z_1), \ldots, (Y_n, Z_n) \right].$$

We see from Theorem 2 that the convergence rate r_n typically has the form $n^{\varrho}\ell(n)$ for some $\varrho > 0$ and some function ℓ slowly varying at infinity. Using the fact that $\log\{\ell(n)/\ell(m_s)\} = o(U_s)$, we have $T_s^*(p) = \log MSE(\hat{\beta}(p)) + 2\varrho U_s + o_p(U_s)$, $s = 1, \ldots, S$. We linearly regress the $T_s^*(p)$'s on the U_s's to estimate the intercept term $\log MSE(\hat{\beta}(p))$ by $\hat{G}(p)$: see *Step 3* of Algorithm (B). Note that the choices of $m_1, \ldots, m_S$ guarantee that $\bar{U}$ and S_{UU} have magnitudes of orders $\log n$ and $(\log n)^2$, respectively. This, together with the fact that $\log MSE(\hat{\beta}(p))$ has order $\log n$, implies Lemma 2.

References

Arcones, M. A. (2005). Convergence of the optimal M-estimator over a parametric family of M-estimators. *Test* **14**, 281-315.

Barrodale, I. and Roberts, F. D. K. (1970). Applications of mathematical programming to L_p approximation. *Nonlinear Programming* (Edited by J. B. Rosen, O. L. Mangasarian, K. Ritter), 447-463. Academic Press, New York.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica* **7**, 1-31.

Cheung, K. Y. and Lee, S. M. S. (2005). Variance estimation for sample quantiles using the m out of n bootstrap. *Ann. Inst. Statist. Math.* **57**, 279-290.

Daniels, H. (1960). The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, 151-163.

DeAngelis, D., Hall, P. and Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in L_1 regression. *J. Amer. Statist. Assoc.* **88**, 1310-1316.

Ekblom, H. (1974). L_p-methods for robust regression. *BIT* **14**, 22-32.

Ghosal, S. and Samanta, T. (1995). Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.* **4**, 361-388.

Gonin, R. and Money, A. H. (1989). *Nonlinear L_p-norm Estimation.* Marcel Dekker, Inc., New York.

Hall, P. and Martin, M. (1988). Exact convergence rate of bootstrap quantile variance estimator. *Probab. Theory Related Fields* **80**, 261-268.

Harter, H. L. (1974–1975). The method of least squares and some alternatives – Part I-VI. *Internat. Statist. Rev.* **42**, 147-174; **42**, 159-264, 282 (notes added in proof); **43**, 1-44; **43**, 125-190; **43**, 269-278.

Hogg, R. V. (1972). More light on the kurtosis and related statistics. *J. Amer. Statist. Assoc.* **67**, 422-424.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

Jurečková, J. (1983). Asymptotic behavior of M-estimators of location in nonregular cases. *Statist. Decisions* **1**, 323-340.

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

Lai, P. Y. and Lee, S. M. S. (2003). L_p Regression Under General Classes of Error Densities: an Asymptotic Overview. Research Report No. 370, The University of Hong Kong, Department of Statistics and Actuarial Science.

Lai, P. Y. and Lee, S. M. S. (2005). An overview of asymptotic properties of L_p regression under general classes of error distributions. *J. Amer. Statist. Assoc.* **100**, 446-458.

Money, A. H., Affleck-Graves, J. F., Hart, M. L. and Barr, G. D. I. (1982). The linear regression model: L_p norm estimation and the choice of p. *Comm. Statist. Simulation Comput.* **11**, 89-109.

Montgomery, D. C and Peck, E. A. (1992). *Introduction to Linear Regression Analysis.* Wiley, New York.

Nyquist, H. (1983). The optimal L_p norm estimator in linear regression models. *Comm. Statist. Simulation Comput.* **12**, 2511-2524.

Polfeldt, T. (1970). Minimum variance order when estimating the location of an irregularity in the density. *Ann. Math. Statist.* **41**, 673-679.

Prakasa Rao, B. L. S. (1968). Estimation of the location of the cusp of a continuous density. *Ann. Math. Statist.* **39**, 76-87.

Rogers, A. J. (2001). Least absolute deviations regression under nonstandard conditions. *Econometric Theory* **17**, 820-852.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67-90.

Smirnov, N. V. (1952). *Limit Distributions for the Terms of a Variational Series*. Amer. Math. Soc. Translation, 67.

Sposito, V. A. and Hand, M. L. (1983). On the efficiency of using the sample kurtosis in selecting optimal L_p estimators. *Comm. Statist. Simulation Comput.* **12**, 265-272.

Van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: pylaipy@graduate.hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: smslee@hkusua.hku.hk

(Received July 2006; accepted April 2007)