M-Estimators with Non-Standard Rates of Convergence and Weakly Dependent Data

Mehmet Caner[1]

Department of Economics

WW Posvar Hall 4509

University of Pittsburgh

Pittsburgh, PA 15260.

email:caner@pitt.edu.

Abstract

This paper analyzes M-estimators over general objective functions. We do not assume convexity and differentiability of the functions. A new result regarding M-estimators is derived. Unlike most of the former econometric literature, the rate of convergence is not square root n. The rate of convergence is non-standard and depends on the moment bounds of the objective function analyzed. We can actually connect the rate of convergence to the smoothness of the objective function in certain class of functions as decribed in van der Vaart and Wellner (1996). We also simplify this rate of convergence idea and extend to weakly dependent data from iid case. This rate is simple and usable in econometrics literature. We illustrate the techniques by deriving the rate of convergence for LAD estimator for censored regression and maximum score estimator with weakly dependent data.

Keywords: Empirical process methods, Censored Regression, Maximum Score Estimator.

Short Title: M-Estimators.

MSC: primary 62F12, secondary 62F10.

[1]The author thanks the editor, two referees, and Donald Andrews for their helpful comments.

1 Introduction

By maximizing certain objective functions statistical estimators are derived. These are called M-estimators. Several authors analyzed the limit laws of these estimators in various respects in the recent years. Van der Vaart and Wellner (1996) derived the limit law for M-estimators for the iid case. They largely benefited from the empirical process methods. Arcones (1998) considered the limit law for M-estimators over a convex kernel. By using the convexity assumption no tightness condition is needed for the objective functions. This makes the proof simple but less general.

There are also studies conducted in this literature when the rate of convergence is $n^{1/2}$. For example Bai, Romano, Wu (1992), Hjort and Pollard (1993) , Koltchinskii (1997). Huber (1967), Newey and McFadden (1994) dealt with non differentiable criterion function. However Newey and McFadden (1994) used the quadratic approximation method to derive the limit laws. This method results in root-n rate of convergence. In contrast, van der Vaart and Wellner (1996) , Kim and Pollard (1990) used the empirical process methods and developed results for cube-root rate of convergence with independent data. Andrews (1993,1994) also provided the limit theory for a similar type of estimators. He uses finite dimensional score equations to have the limit law for estimators. Thus his estimators are different from the ones considered in this study. This results in square root n convergence for m-dependent data.

This study uses the approach in van der Vaart and Wellner (1996) and extends the rate of convergence results to weakly dependent data . We should note that the techniques in van der Vaart and Wellner (1996) do not simply extend to the β mixing case, since the stochastic equicontinuity conditions for the iid and weakly dependent cases are different. A coupling lemma by Berbee (1979) is basically needed for this extension.

The main contribution of the article is to provide rate of convergence for M-estimators with weakly dependent data and non-standard rates of convergence. Under minimal conditions compared with the existing literature, the rate of convergence is obtained using empirical process methods. It is shown that this rate crucially depends on " the continuity modulus"of the empirical process. The empirical process is centered and scaled version

of the objective function. In the former econometrics literature the rate of convergence of certain classes of M-estimators is square root n. However in this article by benefiting from empirical process methods we are able to examine more general class of M-estimators and derive non-standard rates of convergence. The main example of the theory developed here is cube-root asymptotics in the econometrics literature. A typical application of these results is the maximum-score estimators in weakly dependent data.

In related literature, Yu (1994) considers the same problem but with uniformly bounded functions . This approach uses mainly the so-called Vapnik-Cveronenkis (V-C) classes of functions. Theorems in Yu (1994) are difficult to use in econometric setups. Here , in this study we benefit from entropy with bracketing approach which can be connected in a relatively simple fashion to the functions used in econometrics literature. We also do not restrict the function classes to be uniformly bounded. Our approach is simple and can be used in econometric problems as we demonstrate with our examples.

We illustrate our proposed techniques through an application to LAD estimator in censored regression and maximum score estimator . Other potential interesting applications may include the threshold models by Hansen (2000), Caner (2002) , Caner and Hansen (2001) and structural change with LAD as in Bai (1995).

In this article we largely benefit from the empirical process methods in an excellent book by van der Vaart and Wellner (1996). The paper is organized as follows: Section 2 gives a brief explanation for empirical process methods. In section 3 we derive our main results, rate of convergence for M-estimators. Section 4 provides more specific results for "L^{2r}" continuous functions and monotone functions. Two specific examples end this section. Section 5 concludes. $\|.\|$ denotes the Euclidean norm, $\Longrightarrow$ shows weak convergence with respect to uniform metric. $\|f\|_{2r}$ denotes the usual $L_{2r}, r > 1$ norm, $\|f\|_{2r} = (E|f|^{2r})^{1/2r}$. The proofs are in the appendix.

2 Basic Empirical Process Theory

In this section we try to provide basic knowledge about empirical process theory. Following Andrews (1993, p.185), let (Ω, $\mathcal{A}$,P) be a probability space. Let $\{W_i : i \geq 1\}$ be a sequence

of $\mathcal{W}$ valued random variables defined on (Ω, $\mathcal{A}$,P) where $\mathcal{W}$ is a Borel measurable subset of R^m. Let Υ be a pseudometric space with pseudometric ρ. Furthermore, let

$$\mathcal{M} = \{f(.,\tau) : \tau \in \Upsilon\}$$

be a class of R^s valued function defined on $\mathcal{W}$ and indexed by $\tau \in \Upsilon$. In our case Υ is a subset of R^p. For any given $\tau \in \Upsilon$, $f(.,\tau)$ is an integrable function over W_i with respect to Borel measure.

An empirical process G_n is given by

$$G_n(\tau) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} (f(W_i,\tau) - Ef(W_i,\tau)) \tag{1}$$

for $\tau \in \Upsilon$.

We need two basic definitions in order to understand the developments in the following sections.

Definition 1. (L_{2r} Bracketing numbers) *Given the functions l and u , the bracket [l,u] is the set of all functions f with $l \leq f \leq u$. An ϵ bracket is a bracket [l,u] with $\|u - l\|_{2r} < \epsilon$ given the norm $\|.\|_{2r}$. The bracketing number $N(\epsilon, \mathcal{M}, \|.\|_{2r})$ is the minimum number of ϵ brackets neeeded to cover $\mathcal{M}$.*

Definition 2 .(Entropy). *The entropy with bracketing is the logarithm of the bracketing number.*

Definitions 1-2 are from van der Vaart and Wellner (1996,p.81-83).

Definition 3 . (β-mixing). *The absolute regular mixing coefficient $\beta(\mathcal{B},\mathcal{C})$ was defined by Volkonskii and Rozanov (1959) and is stronger than strong mixing and yet weaker than uniform mixing. The absolute regular mixing coefficient between σ fields $\mathcal{B},\mathcal{C}$ is:*

$$\beta(\mathcal{B},\mathcal{C}) = \frac{1}{2} \sup \sum_{(i,j)\in(I,J)} |P(B_i \cap C_j) - P(B_i)P(C_j)|,$$

where $B_i \subset B$, $C_j \subset C$ and the supremum is taken over all finite partitions $(B_i)_{i\in I}$ and $(C_j)_{j\in J}$ respectively $\mathcal{B}$ and $\mathcal{C}$ measurable.

Pham and Tran (1985) have shown that a wide class of linear processes with iid innovations (such as ARMA processes) are absolutely regular when the innovation has a bounded continuous density.

3 Rate of Convergence for Parametric Extremum Estimators

In this section we are interested in providing the rate of convergence theory for the estimators that maximize the following criterion function

$$S_n(\tau) = \frac{1}{n}\sum_{i=1}^{n} f(W_i, \tau)$$

In other words we try to find the limit for $\hat{\tau}_n$, being the near maximizer of $S_n(\tau)$ for all n , meaning

$$S_n(\hat{\tau}_n) \geq \sup_{\tau\in\Gamma} S_n(\tau) - o_p(1) \tag{2}$$

Instead of (2) we could have used the following

$$\hat{\tau}_n = argmax_{\tau\in\Upsilon}\frac{1}{n}\sum_{i=1}^{n} f(W_i, \tau)$$

However this could have raised the measurability and definitional problems regarding the argmax functional when multiple maxima exist (Kim and Pollard , 1990,p.194) .

First we define some notation. This is already defined in Theorem 3 of Doukhan , Massart, Rio (1995) , from now on DMR (1995). Let L_{2r} represents functions f with the following characteristic $\|f(W_i,\tau)\|_{2r} \leq \sigma < \infty$ for all $\tau \in \Upsilon$ and σ be a positive number. Then let $f(W_i,\tau) \in \mathcal{M}_\delta$ where $\mathcal{M}_\delta \subset L_{2r}$ be a class of functions satisfying the condition, $\delta > 0$,

$$\mathcal{M}_\delta = \{f(W_i,\tau) - f(W_i,\tau_0) : \rho(\tau,\tau_0) < \delta\} \tag{3}$$

where ρ is the following pseudometric, $\rho(\tau_1,\tau_2) = [E|f(W_i,\tau_1) - f(W_i,\tau_2)|^{2r}]^{1/2r}$, $r > 1$.

Now we provide the assumptions that will be useful for the rate of convergence proof.

Assumption A.1. *Let W_i be strictly stationary, absolutely regular sequence of random variables with β mixing numbers $\beta(s)$ satisfying , $C > 0$,*

$$\beta(s) \leq Cs^{-L}, for some\, L > \frac{r}{r-1}, r > 1$$

.

Assumption A.2. *Let $\mathcal{M}_\delta$ be a class of functions $f(W_i, \tau)$ with envelopes $\bar{M}_\delta$ satisfying*

$$E\bar{M}_\delta^{2r} < \infty$$

.

Assumption A.3. *Let $N(\epsilon, \mathcal{M}_\delta, \|.\|_{2r})$ be the L_{2r} bracketing numbers and satisfy for each $\delta > 0$*

$$\int_0^1 [logN(\epsilon, \mathcal{M}_\delta, \|.\|_{2r})]^{1/2} \, d\epsilon < \infty$$

Assumption A.4. *Suppose there exists a point τ_0 such that*

$$S(\tau_0) > \sup_{\tau \notin \Delta} S(\tau)$$

for every open set Δ that contains τ_0 , where $S(\tau) = Ef(W_i, \tau)$.

Assumption A.5. *Assume for every τ in a neighborhood of τ_0, $S(\tau)$ is twice continuously differentiable at the point of maximum τ_0 with non-singular second derivative matrix V.*

Assumption A.6. *Assume for sufficiently large n and sufficiently small $\delta > 0$, $\psi(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$(not depending on n),where*

$$\psi(\delta) = \int_0^\sigma [logN(\epsilon, \mathcal{M}_\delta, \|.\|_{2r})^{1/2} d\epsilon \tag{4}$$

These are fairly standard assumptions and used for obtaining the rate of convergence , see van der Vaart and Wellner (1996,Chapter 3.2) and Kim and Pollard (1990, Theorem 1). A.1-A.4 are used in the consistency of the estimator. Our extension is the usage of A.1 rather than the iid assumption used in this literature. A.5-A.6 are used in van der Vaart and Wellner (1996), however we are able to specify the "modulus of continuity" in terms of the entropy of the class of functions rather than leaving as it is.

Since consistency arguments are pretty standard we show the proof of consistency in the rate of convergence proof. In Lemma 1 we prove the existence of maximal inequality in Theorem 3.2.5 of van der Vaart and Wellner (1996) with the weakly dependent data.

This Lemma provides the sufficient conditions for the existence of the maximal inequality benefiting from Theorem 3 in DMR (1992) .

This is shown in van der Vaart and Wellner (1996) only in iid case. So the modulus of continuity can be used in rate of convergence calculations as shown in the proofs. The proof is in the Appendix.

Lemma 1. *Under Assumptions A.1-A.3 there exists some positive constant A such that for sufficiently large n*

$$E \sup_{\rho(\tau,\tau_0)<\delta} |G_n(\tau) - G_n(\tau_0)| \leq 2A\psi(\delta) \tag{5}$$

where $G_n(\tau), G_n(\tau_0)$ are the empirical processes defined in (1), at τ and τ_0 respectively.

The following theorem extends the Corollary 3.2.6 of van der Vaart and Wellner (1996) from iid to weakly dependent data and provides primitive conditions compared with Corollary 3.2.6 of van der Vaart and Wellner (1996). Note that Lemma 1 can be shown for all n, under some additional conditions for the envelope function . This proof for all n, is tedious involving a lot of notation but simple. This can be obtained from the author on demand.

Before the theorem we need the definition of the rate of convergence (van der Vaart and Wellner, 1996,p.290).

Definition 4 . (The rate of convergence, r_n). *Let*

$$r_n^2 \psi(\frac{1}{r_n}) \leq n^{1/2}$$

for every n. $\psi(.)$ is defined in Assumption A.6 .

Now we can have the main theorem of the paper.

Theorem 1 . *Under Assumptions A.1-A.6,*

$$r_n(\hat{\tau}_n - \tau_0) = O_p(1).$$

Remarks.

1.Theorem 1 shows that the rate of convergence is calculated from the following inequality by benefiting from Definition 4,

$$r_n^2 \psi(1/r_n) \leq n^{1/2}$$

We substitute

$$\psi(\delta) = \int_0^\sigma [logN(\epsilon, \mathcal{M}_\delta, \|.\|_{2r})]^{1/2} d\epsilon$$

in calculating the convergence rate. So the " modulus of continuity " of the empirical process gives an upper bound on the rate of convergence. For example, if the modulus is $\psi(\delta) = \delta$, then the rate of convergence , $r_n = n^{1/2}$. If $\psi(\delta) < \delta$ then r_n is converging at a rate less than root n.

2. Even though Assumption A.6 is a high level condition, for the classes of functions analysed in section 4 below (uniformly Lipschitz and monotone), for sufficiently small δ and for large n , $\psi(\delta)/\delta^{\alpha}$ is decreasing for some $\alpha < 2$.

4. Assumptions A.1-A.6 are rather standard in this literature see van der Vaart and Wellner (1996) and Kim and Pollard (1990). In van der Vaart and Wellner (1996) the modulus of continuity is specified and tied to the entropy in iid case. In Kim and Pollard (1990) they only analyze cube root rate , here we do not impose any rate of convergence.

5. Unfortunately to have the limit law we need to localise the objective function and then try to obtain triangular array version of Theorem 1 of DMR (1995) . This is not clear and not trivial to do as can be seen from the Theorem 2.11.9 of van der Vaart and Wellner (1996) for the case of independent data.

4 Various Classes of Functions

In this section we try to show different classes of functions satisfy the entropy condition (Assumption A.3) hence resulting in simpler proofs and considerably weaker assumptions. So as long as we can show our objective function belongs to certain classes of functions that satisfy entropy conditions , then stochastic equicontinuity follows trivially.

4.1 L^{2k} continuous functions

One particular class of functions that satisfy Assumptions A.3 is the so-called L^{2k} continuous functions, see Andrews (1993).

Definition 5. (L^{2k} continuous functions).*A class $\mathcal{M}$ of real functions on W is called "L^{2k} continuous" under P with index 2k $\in [2, \infty]$ if each function in $\mathcal{M}$ satisfies $f(.) = f(., \tau)$ for some $\tau \in \Upsilon$ where Υ is some bounded subset of Euclidean space*

$$\left[E \sup_{\tau_1:\|\tau_1-\tau\|<\delta} |f(W_i,\tau_1) - f(W_i,\tau)|^{2k}\right]^{1/2k} \leq C\delta^{\nu}$$

for all $\tau \in \Upsilon$ and for all $\delta > 0$ in a neighborhood of 0 for some positive constant C and $\nu \in (0,1], \{W_i\}$ has distribution determined by P. .

This definition is due to Andrews (1993, 1994). The condition in Definition 5 allows for discontinuous functions such as sign and indicator functions. We have used $\nu \in (0,1]$. We try to provide some intuition regarding these function classes. We analyze $2k = 2$ case. If $\nu = 1$, even though our functions may be non-differentiable, very roughly we can say that variance of the " partial derivative" is constant. But if $\nu < 1$, then we see that the variance of "partial derivative" increases with decreasing δ, so this shows us that this type of objective function is less smooth compared with $\nu = 1$ case. Another important point to note is : in Andrews (1994) $\nu \in (0,\infty)$, but if $\nu > 1$ than this corresponds to very smooth functions, with zero variance of " partial derivatives". Since we are not using these type of objective functions we ruled this case out. A similar point can be seen in the case of functions that satisfy Lipschitz condition (van der Vaart and Wellner, 1996, p.198,294).

Using this definition we are able to simplify the Theorem 1 a little.

Theorem 2. *Suppose f belongs to L^{2k} continuous class of functions with $\nu \in (0,1]$ and Assumptions A.1, A.4-A.5 hold then*

$$n^{\frac{1}{4-2\nu}}(\hat{\tau}_n - \tau_0) = O_p(1).$$

Remark. We do not use the high level entropy assumption that is used in Theorem 1 and we obtain also a simple convergence rate. This rate of convergence is the same obtained by van der Vaart and Wellner (1996) in the iid case for Lipschitz continuous functions, see example 3.2.12 van der Vaart and Wellner (1996). We establish the case for weakly dependent data for L^{2k} continuous functions for the first time in the literature . Assumption A.6 and Assumption A.4 is also satisfied by these classes of functions. The details are in the proof. Note that also with less smooth functions ($\nu < 1$) the rate of convergence is less than square root n. For maximum-score type estimators we have $\nu = 1/2$, which results in $n^{1/3}$ rate of convergence.

To derive the limit of cube-root consistent estimators the difficulty lies in the rate of convergence proof. So Theorems 1 and 2 contribute to the literature in facilitating and simplifying the rate of convergence proofs with weakly dependent data.

4.1.1 LAD Estimator for Censored Regression with Weakly Dependent Data

We now want to apply the ideas in the former section to a specific example. Powell (1984) proposed the so-called least absolute deviations estimator for the censored regression. He obtained the limit laws for the regression parameters in the case of independent random variables. Specifically he considered

$$\hat{\tau}_n = \arg\min_{\tau\in\Gamma} \frac{1}{n}\sum_{i=1}^{n} |y_i 1_{\{y_i\geq 0\}} - (x_i'\tau)1_{\{x_i'\tau\geq 0\}}|$$

So y_i is observed only if $y_i \geq 0$ and $x_i'\tau$ is observed only if $x_i'\tau \geq 0$. Set Υ as a subset of R^p. Pollard (1990) also analyzed the same problem using the empirical process methods and derived the limit law for the estimator in a simple way for iid random variables. For more information regarding these models see Powell (1994, p.2576). Note that the criterion function is non-differentiable and non-convex in τ. Here we extend the former literature to the weakly dependent errors " u_i".

Assumption B.1. *The u_i are strictly stationary β mixing random variables with exponential mixing coefficients.*

Assumption B.2. *The u_i have zero median and a continuous strictly positive density $f(.)$ near zero.*

Assumption B.3.

$$E|x_i|^2 < \infty$$

Assumption B.4. *For each $\epsilon > 0$, there is a $\kappa > 0$ such that for all large n*

$$1/n \sum |x_i|^2 1_{\{x_i'\tau_0|\leq\kappa\}} < \epsilon$$

with probability no less than $1-\epsilon$.

Assumption B.5.

$$E(\frac{1}{n}\sum_{i=1}^{n} |x_i|^2 1_{\{x_i'\tau_0>0\}}) < \infty$$

Assumptions B.2-B.5 are standard in the literature see Pollard (1990). Our main contribution comes from adding assumption B.1 and solving the problem for weakly dependent observations. We now briefly relate the assumptions in the former sections to B.1-B.5.

Proposition 1. *Under Assumptions B.1-B.5,*

$$n^{1/2}(\hat{\tau_n} - \tau_0) = O_p(1).$$

4.1.2 Maximum Score Estimator

We apply the ideas in sections 3 and 4.1 to a specific example . We analyze the following regression model

$$y_i = x_i'\tau_0 + u_i \quad i = 1, \cdots, n$$

where τ_0 is an unknown p dimensional vector of parameters, $\{x_i\}$ is a sequence of observed vectors and u_i are unobserved errors.

Manski (1985) proposed the maximum score estimator. He proved the consistency of the estimator and Kim and Pollard (1990) obtained the limit law for this estimator for the case of independent data. The maximum score estimator is:

$$\hat{\tau}_n = \arg\max_{\tau \in S} \frac{1}{n} \sum_{i=1}^{n} \left(1_{\{y_i \geq 0, x_i'\tau \geq 0\}} + 1_{\{y_i < 0, x_i'\tau < 0\}} \right)$$

This is the sample analog of the binary choice estimator see Newey and Mc Fadden (1994), Kim and Pollard (1990).

Since τ_n is determined only up to scalar multiples, it can be standardized to unit length. In the same manner τ_0 can be assumed to be a unit vector. The parameter space may be identified with the surface S on the unit sphere in R^p. Note that the methodology in Newey and McFadden (1994) does not apply here. The criterion function cannot be quadratically approximated in the way that is described in Newey and McFadden (1994). We also use a different method than Kim and Pollard (1990). Rather than using " uniform manageability" condition to prove stochastic equicontinuity we benefit from the L^{2k} continuity of the criterion function.

Now we present the assumptions that are needed to derive the rate of convergence.

Assumption C.1 . *x_i and u_i are β mixing with coefficients satisfying $\beta(s) \leq Cs^{-L}$, for $L > \frac{r}{r-1}$, $r > 1$, $C > 0$, and conditional median of u_i given x_i is zero and unique.*

Assumption C.2 . $\sup_i \|x_i\| < \infty$ *and x_i has a continuously differentiable density $p(x_i)$ and which the angular component of x_i has a bounded continuous density with respect to surface measure on S.*

Assumption C.3. *$x_i'\tau$ is continuously distributed for each τ.*

Assumption C.4. *If $\tau \neq \tau_0$ then*

$$1_{\{x'\tau > 0\}} \neq 1_{\{x'\tau_0 > 0\}}$$

Proposition 2.*Under assumptions C.1-C.4*

$$n^{1/3}(\hat{\tau}_n - \tau_0) = O_p(1)$$

5 Conclusion

This paper provides a rate of convergence proof for parametric M-estimators with weakly dependent observations. This is done without assuming convexity and differentiability of the objective functions. A natural extension may be to provide a limit law for M-estimators and with non-stationary data..

Appendix

Proof of Lemma 1. This is shown after Theorem 3 of DMR (1995) as a remark on p.410 of DMR (1995) . ■.

Proof of Theorem 1. Using Assumptions A.1-A.3 we obtain a specific $\psi(\delta)$ via Lemma 1 (equation 5). Assumptions A.1-A.3 supply a uniform version of the central limit theorem which is application 1 of Theorem 1 in DMR (1995). This is also given as application 3a on p.200 of Andrews (1993). Applying the Slutsky's theorem in van der Vaart and Wellner (1996, p.32) we have the uniform law of large numbers. Combining the uniform law of large numbers with Assumption A.4 and (2) , we obtain the consistency of the estimator $\hat{\tau}_n$ using Corollary 3.2.3i of Van der Vaart and Wellner (1996). Using Assumptions A.5-A.6 and combining Lemma 1 with Theorem 3.2.5 of Van der Vaart and Wellner gives the desired result. ■.

Proof of Theorem 2. The consistency can be shown easily given the assumptions A.1-A.4 via the proof of Theorem 1.

We try to find the rate of convergence for L^{2k} continuous functions . Assumptions A.2 and A.3 are satisfied by these classes of functions. Since by 3aiii in Andrews (1993, p.200) $logN(\epsilon, \mathcal{M}_\delta, \|.\|_{2r}) < C(\frac{1}{\epsilon})^B$ where $0 < B < 1/2$ we can show that

$$\begin{aligned} 2A\psi(\delta) &= 2A\int_0^{C\delta^{2\nu/(2-B)}} [logN(\epsilon, \mathcal{M}_\delta, \|.\|_{2r})]^{1/2} d\epsilon \\ &\leq 2A\int_0^{C\delta^{2\nu/(2-B)}} C^{1/2}(\frac{1}{\epsilon})^{B/2} d\epsilon \\ &\leq C_2\delta^\nu \end{aligned}$$

for large enough $C_2 > 0$. Note that via Holder continuity condition we replace σ by $C\delta^{2\nu/(2-B)}$. Since the modulus of the continuity is $C_2\delta^\nu$ from the above inequality using Definition 4

$$r_n^2 \frac{1}{r_n^\nu} \leq n^{1/2}$$

so

$$r_n = n^{1/(4-2\nu)}$$

The limit law does not simplify further and we obtain the result. Note that $\psi(\delta)/\delta$ is decreasing in δ. This can be seen since $\nu \in (0, 1]$ and $\psi(\delta) \leq C\delta^\nu$. Note that the proof is

for $r > 1$, however when $r = 1$ usage of the L_2 norm in DMR (1995) results in different mixing conditions. This can be seen in Andrews (1993) and application 5 of Theorem 1 in DMR (1995). ■.

Proof of Proposition 1. Note that

$$\left| |y_i 1_{\{y_i \geq 0\}} - (x_i'\tau_1)1_{\{x_i'\tau_1 \geq 0\}}| - |y_i 1_{\{y_i \geq 0\}} - (x_i'\tau_2)1_{\{x_i'\tau_2 \geq 0\}}| \right| \leq \|x_i\|\|\tau_1 - \tau_2\| \text{ for all} \tau_1, \tau_2 \in \Gamma \tag{A.4}$$

So

$$|y_i 1_{\{y_i \geq 0\}} - (x_i'\tau)1_{\{x_i'\tau \geq 0\}}|$$

are Lipschitz continuous and L^{2k} continuous under B.3. So the entropy assumption A.3 is satisfied.

Now we consider the consistency proof. By B.1 , B.3 and the Lipschitz continuity these classes of functions we see that this class of functions satisfy the entropy assumptions. So by B.1,B.3 and L^{2k} continuity , A.1-A.3 are satisfied so random variable version of Lemma A.1 is proven (This is application of Theorem 1 in DMR (1995)). Then use Slutsky's theorem in van der Vaart and Wellner (1996, p.32) to have the uniform law of large numbers. Then denoting the objective function to be minimized as

$$S_n(\tau) = 1/n \sum_{i=1}^{n} |y_i 1_{\{y_i \geq 0\}} - (x_i'\tau)1_{\{x_i'\tau \geq 0\}}| - |y_i 1_{\{y_i \geq 0\}} - (x_i'\tau_0)1_{\{x_i'\tau_0 \geq 0\}}|$$

We see that

$$ES_n(\tau) = 1/n \sum_{i=1}^{n} E|y_i 1_{\{y_i \geq 0\}} - (x_i'\tau)1_{\{x_i'\tau \geq 0\}}| - |y_i 1_{\{y_i \geq 0\}} - (x_i'\tau_0)1_{\{x_i'\tau_0 \geq 0\}}|$$

is uniquely minimized at $\tau = \tau_0$ by B.2 (see Pollard (1990), equation 11.2 or page 61).

Then use B.4 , B.5 to have

$$\lim_{n\to\infty} \inf_{|\tau - \tau_0| \geq v} ES_n(\tau) > 0 \tag{A.5}$$

for all $v > 0$ via Pollard (1990) equation (11.5). Then it is clear that Assumptions B.2,B.4,B.5 show that A.4 holds.

Combining Uniform law of large numbers with (A.5) we have the consistency.

Since classes of functions considered in this example are Lipschitz continuous with $\nu = 1$ in (A.4) , using Theorem 2 we have the rate of convergence under our assumptions as $r_n = n^{1/2}$. ■

Proof of Proposition 2 . Our objective function is

$$\frac{1}{n}\sum_{i=1}^{n}\left(1_{\{y_i\geq 0, x_i'\tau\geq 0\}} + 1_{\{y_i<0, x_i'\tau<0\}}\right) \tag{A.6}$$

Instead of maximizing (A.6), as in Kim and Pollard (1990, p.214) $\hat{\tau}_n$ equivalently maximizes the following criterion function

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i,u_i)(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}}) \tag{A.7}$$

where

$$h(x_i,u_i) = 1_{\{u_i + x_i'\tau_0\geq 0\}} - 1_{\{u_i+x_i'\tau_0<0\}} \tag{A.8}$$

Then

$$(E\sup_{\tau:\|\tau-\tau_0\|<\delta}|h(x_i,u_i)(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}})|^2) = (E\sup_{\tau:\|\tau-\tau_0\|<\delta}|h(x_i,u_i)^2(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}})^2|) \tag{A.9}$$

Note that given $\tau > \tau_0$, without losing any generality,by Assumption C.2

$$\begin{aligned} E\sup_{\tau:\|\tau-\tau_0\|<\delta}(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}})^2 &= E\sup_{\tau:\|\tau-\tau_0\|<\delta}(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}}) \\ &\leq C_1\delta \end{aligned} \tag{A.10}$$

where $C_1 = \sup_i \|x_i\|p(x_i)$.

The inequality in (A.10) is derived through mean value expansion. Benefiting from Cauchy-Schwartz inequality in (A.9) and using (A.8), (A.10) we have

$$(E\sup_{\tau:\|\tau-\tau_0\|<\delta}|h(x_i,u_i)^2(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}})^2|) \leq C\delta \tag{A.11}$$

since $Eh(x_i,u_i)^2 < \infty$. By (A.11)

$$(E\sup_{\tau:\|\tau-\tau_0\|<\delta}|h(x_i,u_i)(1_{\{x_i'\tau\geq 0\}} - 1_{\{x_i'\tau_0\geq 0\}})|^2)^{1/2} \leq C\delta^{1/2} \tag{A.12}$$

. where C is a positive constant.

From (A.12) it is clear that these class of functions are L^{2k} continuous with $\nu = 1/2$ in Theorem 3.

Assumptions C.1,C.3,C.4 are sufficient conditions for the consistency of $\hat{\tau}_n$. This is already shown in Theorem 2.10 and Lemma 2.4 of Newey and McFadden (1994) . Since $\nu = 1/2$ by Theorem 2 the rate of convergence is $n^{1/3}$ which is the rate of convergence also found by Kim and Pollard (1990) for independent variables. ■.

References

Andrews,D.W.K., 1993. An introduction to econometric applications to empirical process theory for dependent random variables. Econometric Reviews 12,183-216.

Andrews,D.W.K., 1994. Empirical process methods in econometrics. Handbook of Econometrics IV.

Arcones , M., 1998. Asymptotic Theory for M-estimators over a convex kernel. Econometric Theory 14,387-422.

Babu,G.J., 1989. Strong representations for LAD estimators in linear models. Probability Theory and Related Fields 83, 547-558.

Bai, J., 1995. Least absolute deviation estimation of a shift. Econometric Theory , 11, 403-436.

Bai,Z.,D., C.R. Romano and Y.Wu, 1992. M-estimation of multivariate linear regression parameters under a convex discrepancy function. Statistica Sinica 2, 237-254.

Berbee, H.C.P., 1979, Random walks with stationary increments and renewal theory. Mathematics Centre Tracts, Amsterdam.

Caner, M. , 2002, A note on LAD estimation of threshold model. Econometric Theory 18, 800-814.

Caner , M. and B.E. Hansen , 2001, Threshold Autoregressions with a unit root, Econometrica, 69, 1555-1596.

Doukhan, P, P. Massart and E.Rio, 1995. Invariance Principles for absolutely regular empirical procesess. Annales Institut Henri Poincare Probability and Statistics,31 393-427.

Hansen, B.E. 2000. Sample Splitting and Threshold Estimation. Econometrica 68,575-605.

Hjort, N.L, D.Pollard, 1993. Asymptotics for minimisers of convex processes , Preprint, Department of Statistics, Yale University.

Huber, P., 1967. The behavior of maximum likelihood estimates under nonstandard conditions, in L.M. LeCam and J. Neyman , eds., Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Berkeley, University of California Press.

Kim J. and D.Pollard, 1990. Cube root asymptotics. Annals of Statistics 25, 435-477.

Koltchinskii, V.I, 1997. M-estimation, convexity and quantiles. Annals of Statistics 25, 435-477

Manski, C.F., 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics 27, 313-333.

Newey, W.K., and D. McFadden , 1994. Large sample estimation and hypothesis testing . Handbook of Econometrics vol. IV.

Pham T.D., and L.T.Tran , 1985. Some mixing properties of time series models. Stochastic Processes and Their Applications, 19, 297-303.

Pollard D., 1990. Empirical Processes: Theory and applications. NSF.CBMS.Regional Conference Series, vol.2.

Powell, J.L., 1984. Least absolute deviations estimation for the censored regression model. Journal of Econometrics 25, 3030-325.

Powell, J.L., 1994. Estimation of semiparametric models. Handbook of Econometrics vol.IV.

Van der Vaart A. and J.Wellner, 1996. Weak convergence and Empirical Procesess.

Springer VERLAG.

Volkonskii V.A. , Y.A. Rozanov, 1959, Some limit theorems for random functions , Theory and Probability Applications,4,178-197.

Yu, B., 1994. Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability, 22, 94-116.