Econometric Theory, **11**, 1995, 403–436. Printed in the United States of America.

LEAST ABSOLUTE DEVIATION ESTIMATION OF A SHIFT

JUSHAN BAI
Massachusetts Institute of Technology

This paper develops the asymptotic theory for least absolute deviation estimation of a shift in linear regressions. Rates of convergence and asymptotic distributions for the estimated regression parameters and the estimated shift point are derived. The asymptotic theory is developed both for fixed magnitude of shift and for shift with magnitude converging to zero as the sample size increases. Asymptotic distributions are also obtained for trending regressors and for dependent disturbances. The analysis is carried out in the framework of partial structural change, allowing some parameters not to be influenced by the shift. Efficiency relative to least-squares estimation is also discussed. Monte Carlo analysis is performed to assess how informative the asymptotic distributions are.

1. INTRODUCTION

Many approaches have been suggested in the literature for estimating parameter changes occurring at unknown times. Some well-known approaches include maximum likelihood (e.g., Quandt, 1958; Hinkley, 1970; Picard, 1985; Bhattacharya, 1987; Yao, 1987), the least-squares method (e.g., D.L. Hawkins, 1986; Bai, 1994b), the Bayesian method (e.g., Zacks, 1983; Broemeling and Tsurumi, 1987; Zivot and Phillips, 1994), and the nonparametric method (e.g., Carlstein, 1988; Duembgen, 1991).[1] References on various estimation techniques can be found in the review papers by Krishnaiah and Miao (1988) and Zacks (1983). A survey of empirical applications of the structural change problem in economics is given by Perron (1993). This paper explores the estimation of a shift point using the least absolute deviation (LAD) technique. The consideration of LAD is motivated by the possible efficiency loss resulting from the use of least squares when the data are observed from a thick-tailed distribution. It is well documented that least-squares estimation in the usual context (no structural change) is not efficient

I thank three anonymous referees and the editor for very useful comments that led to a substantial revision. This paper is part of my dissertation written at the University of California at Berkeley. I am grateful to my thesis advisor, Professor Thomas Rothenberg, for his support and numerous discussions. I also thank Professors Deborah Nolan and James Stock for their support and comments. Chris Snyder made many useful suggestions. Financial support from the Alfred P. Sloan Foundation Dissertation Fellowship is gratefully acknowledged. Any errors are mine. Address correspondence to: Jushan Bai, Department of Economics, E52-274B, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

for heavy-tailed distributions (e.g., Huber, 1981); this conclusion remains true when estimating a shift point. It is therefore of interest to study estimation methods less sensitive to extreme observations.

Despite the large body of literature on estimating structural changes, robust parametric estimation has not been widely examined. An exception is Hsu (1982a, 1982b), who investigated the robust estimation (not including LAD) in the Bayesian context and applied his work to shifts in the variability of stock market returns. In this paper, we study the LAD method for estimating models with a shift. The LAD method is perhaps the best known and the simplest technique that is robust against thick-tailed distributions.[2] As pointed out by Bloomfield and Steiger (1983), LAD is one of the oldest methods for curve fitting but was largely abandoned until recently because of computational difficulties. Fueled by today's advanced computing technology, there is a growing interest in the method in both the statistics and econometrics literature, as evidenced by the work of Amemiya (1982), Bassett and Koenker (1978), Honore (1992), Knight (1989, 1991), Pollard (1991), Powell (1984), Phillips (1991), and Weiss (1991), among others. As is well known, LAD is a special case of Koenker and Bassett's (1978, 1982) general quantile regressions, which find many applications in economics (e.g., Buchinsky, 1994; Chamberlain, 1991). Although the analysis in this paper is carried out in terms of LAD, the argument extends to quantile regressions without essential difficulty.

The primary concern of this paper is the joint behavior of estimated regression parameters and the estimated shift point. Our objective is to derive their rates of convergence and asymptotic distributions. The analysis is conducted in a model of partial structural change, in which some regression parameters do not change, as opposed to pure structural change, in which all regression parameters have a shift. Partial structural change includes pure structural change as a special case. The assumption of partial structural change is equivalent to imposing a parameter constraint across regimes. A cross-regime constraint, if valid, will yield a more efficient estimation of both the regression parameters and the shift point. Both independent and dependent disturbances are considered. The asymptotic distribution for the shift point estimator is studied under both fixed magnitude of shift and shift with magnitude tending to zero when sample size increases. We also derive asymptotic distributions in the presence of trending regressors. Finally, we perform Monte Carlo analysis to compare small sample distributions with asymptotic distributions.

Under conditions similar to those described by Pollard (1991) in the absence of a shift, we obtain the consistency of the shift point estimator. Regardless of moment conditions, the rate of convergence obtained here is the same (up to a factor) as that obtained by Hinkley (1970) for an independent and identically distributed (i.i.d.) normal sequence with a shift estimated by the maximum likelihood estimator. Examination of the asymptotic distribution offers further insight into the robustness of the LAD estimator: the shift point estimator has an asymptotic distribution depending inversely on the value of the density function at zero, not on the second moment in contrast to the least-squares estimator. The asymptotic distribution also reveals the way in which serial correlation affects the precision of the shift point estimator.

2. MODEL AND ASSUMPTION

Consider the following linear model with a single shift:

$$\begin{aligned} y_i &= x_i'\beta_0 + z_i'\delta_{10} + \varepsilon_i \qquad i = 1, \ldots, k_0, \\ y_i &= x_i'\beta_0 + z_i'\delta_{20} + \varepsilon_i \qquad i = k_0 + 1, \ldots, n, \end{aligned} \tag{1}$$

where $x_i \in R^p$ and $z_i \in R^q$ are vectors of regressors, and β_0, δ_{10}, and δ_{20} are unknown parameters. The shift point k_0 is also unknown and has to be estimated. The disturbances ε_i are assumed to be i.i.d. Dependent errors will be considered in Section 5. Model (1) is that of a partial structural change in the sense that the parameter vector β is constant throughout the whole sample period. When β plays no role ($\beta = 0$), a pure structural change model pertains (all parameters shift at k_0).

Testing for the existence of such a shift has received considerable attention in the recent econometric literature (e.g., Andrews, 1993; Andrews and Ploberger, 1992; Banerjee, Lumsdaine, and Stock, 1992; Christiano, 1992; Chu and White, 1992; Hansen, 1992; H.J. Kim and Siegmund, 1989; Kramer, Ploberger, and Alt, 1988; Perron and Vogelsang, 1992). The econometric literature actually treats a more general class of models than (1), permitting dynamic regressors, integrated or cointegrated regressors, and integrated error processes. Earlier studies for testing a shift in means can also be found in James, James, and Siegmund (1987), D.M. Hawkins (1977), Sen and Srivastava (1975a, 1975b), and Worsley (1979, 1986), among others.

The focus of this paper is to estimate the model under the maintained hypothesis that a shift exists as opposed to testing for its existence. Let $\theta_0 = (\beta_0', \delta_{10}', \delta_{20}')' \in R^{p+2q}$ be the true parameter, and let $\lambda_0 = \delta_1 - \delta_2$ be the vector of magnitudes of shift. We assume $\lambda_0 \neq 0$; i.e., at least one of the coefficients of z_i has a shift. The goal is to estimate θ_0 and k_0 with LAD and study the statistical properties of the resulting estimators, especially their rates of convergence and asymptotic distributions. The shift point is a discrete parameter and can be estimated by a grid search over all possible integer values.

Denote $\theta = (\beta, \delta_1, \delta_2)$, and define

$$S(\theta, k) = \sum_{i=1}^{k} |y_i - x_i'\beta - z_i'\delta_1| + \sum_{i=k+1}^{n} |y_i - x_i'\beta - z_i'\delta_2|. \tag{2}$$

Thus, $S(\theta,k)$ is simply the sum of absolute deviations for each fixed k. An estimator of (θ_0,k_0) is defined as a point $(\hat{\theta},\hat{k})$ that minimizes the sum of absolute deviations. To obtain such an estimator, a sequence of LAD estimations is performed. The parameter θ is concentrated out of the objective function, resulting in an objective function with parameter k only. A grid search is then performed to obtain $\hat{k}$. Notationally,

$$\hat{\theta}(k) = \underset{\theta}{\operatorname{argmin}}\ S(\theta,k),$$

$$\hat{k} = \underset{k}{\operatorname{argmin}}\ S(\hat{\theta}(k),k),$$

$$\hat{\theta} = \hat{\theta}(\hat{k}).$$

We shall study the asymptotic properties of the resulting estimators. Other estimation methods such as least squares may also be used. Bai (1994a) estimated model (1) with least sauares and established rates of convergence and asymptotic distributions. In light of the nondifferentiability of the objective function, this paper uses an entirely different approach to study the joint asymptotic behavior of the estimated regression parameters and shift point.

In what follows, we use $o_p(1)$ to denote a sequence of random variables converging to zero in probability and $O_p(1)$ to denote a sequence that is stochastically bounded. For a sequence of matrices B_n, we write $B_n = o_p(1)$ if each of its elements is $o_p(1)$ and likewise for $O_p(1)$. The notation $\|\cdot\|$ is used to denote the euclidean norm; i.e., $\|x\| = (\sum_1^p x_i^2)^{1/2}$ for $x \in R^p$. For a matrix A, $\|A\|$ is the vector-reduced norm; i.e., $\|A\| = \sup_{x\neq 0}\|Ax\|/\|x\|$. Finally, $[a]$ represents the integer part of a.

We make the following assumptions.

A1. The errors ε_i are i.i.d., admitting a positive and continuous density function in a neighborhood of zero and having a zero median. The ε_i are independent of the regressors.

A2. Let $w_i = (x_i', z_i')'$ and $X_i = (x_i', z_i', 0)'$ for $i \le k_0$ and $X_i = (x_i', 0', z_i')'$ for $i > k_0$; $w_i \in R^{p+q}$ and $X_i \in R^{p+2q}$. Then, both plim $\frac{1}{n}\sum_1^n w_i w_i'$ and plim $\frac{1}{n}\sum_1^n X_i X_i'$ exist, and the limits are positive definite matrices.

A3. For large j, both $\frac{1}{j}\sum_\ell^{\ell+j} z_i z_i'$ and its inverse are bounded above in probability for each ℓ.

A4. $n^{-1/2}\max_{1\le i\le n}\|w_i\|(\log n) = o_p(1)$. Furthermore, for each $\epsilon > 0$, there exists $K > 0$ such that for all large n

$$\frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2 I(\|w_i\| > K) < \epsilon$$

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

A5. $k_0 = [\tau_0 n]$ for some $\tau_0 \in (0,1)$.

Assumptions A1 and A2 are typical for LAD estimation (e.g., Pollard, 1991). Assumption A3 requires that the sum become a positive definite matrix when many observations on z_i are used. In our proof, we actually only use $\ell = 1$, $\ell = k_0$, and $\ell = n$. This assumption is needed for n-consistency (defined later) of the estimated shift point. The first part of Assumption A4 is also typical for LAD except the extra term $\log n$. Babu (1989) used $(\log n)^{1/2}$ instead of $\log n$ to obtain strong representations for LAD estimators. The second part of Assumption A4 is assumed by Pollard (1990, p. 58) in a different context. If w_i are i.i.d. with a finite covariance matrix or w_i are such that $E\|w_i\|^2 \times I(\|w_i\| > K)$ is uniformly small for large K, then the second part is satisfied. For uniformly bounded regressors, Assumption A4 is obviously satisfied. For asymptotic purposes, the shift point is assumed to be bounded away from the two ends, as in Assumption A5. A slightly more general setting is that $k_0 = \tau_n n$, where τ_n is one of the values $\{2/n, \dots, (n-1)/n\}$ such that $\tau_n \to \tau_0$ for some τ_0 in (0,1). Our results still hold under this assumption. Under these assumptions, we can obtain the convergence rates for the estimated parameters.

3. RATE OF CONVERGENCE

Let $\hat{\tau} = \hat{k}/n$ be an estimator for τ_0. Both $\hat{k}$ and $\hat{\tau}$ are referred to as the shift point estimator.

THEOREM 1. *Under Assumptions* A1–A5, *we have*

$$n(\hat{\tau} - \tau_0) = O_p\left(\frac{1}{\lambda_0'\lambda_0}\right) \quad \textit{and} \quad \sqrt{n}(\hat{\theta} - \theta_0) = O_p(1). \tag{3}$$

The theorem states that $\hat{\tau}$ is n-consistent and $\hat{\theta}$ is root-n-constant. We leave λ_0 in the notation $O_p(\cdot)$ to show explicitly the dependence of the rate of convergence on the magnitude of a shift (the larger the shift, the easier to identify it).

It is also possible to incorporate settings in which λ_0 depends on n and $\|\lambda_0\|$ converges to zero as n increases. This case is also useful because the shift point estimator admits an asymptotic distribution not depending on the underlying distribution of ε_i, as in Picard (1985) and Yao (1987). In the rest of this paper, we shall treat the general case in which λ_n depends on n, using the notation λ_n. The case of fixed λ_0 will be treated as a special case. For a fixed n, $\|\lambda_n\|$ should not be too small. More specifically, we make the following assumption.

A6. There exists $b \in (0,1/2)$ such that $n^{(1/2)-b}\|\lambda_n\| \to \infty$.

This assumption, of course, is trivially satisfied when λ_n does not vary with n and is nonzero.

THEOREM 1′. *Under Assumptions* A1–A6,

$$n(\hat{\tau} - \tau_0) = O_p\left(\frac{1}{\lambda_n'\lambda_n}\right) \quad \textit{and} \quad \sqrt{n}(\hat{\theta} - \theta_0) = O_p(1). \tag{4}$$

The convergence rate of (4) implies that $\hat{\tau}$ will be consistent as long as $n\|\lambda_n\|^2$ grows without bound.

The rate of convergence of $\hat{\tau}$ can be obtained by evaluating the global behavior of the objective function $S(\theta, k)$ over the whole parameter space for θ and k. It is convenient to work with a reparameterized objective function, which will also be useful for obtaining the limiting distributions. Define

$$V_n(\theta, v) = S(\theta_0 + n^{-1/2}\theta, k(v)) - S(\theta_0, k_0), \tag{5}$$

where $k(v) = [k_0 + vc_n]$ with $c_n = O(\|\lambda_n\|^{-2})$ and v is a real scalar. When v varies, $k(v)$ visits all integers between 1 and n, assuming $k(v) = 1$ if $k(v) \leq 1$ and $k(v) = n$ if $k(v) \geq n$. This reparameterization conforms with the anticipated rate of convergence but does not preimpose the rate of convergence, because θ takes values in R^{p+2q} and $v \in R$ without any restriction. The minimization problem is not changed. Let $\tilde{\theta}$ and $\tilde{v}$ minimize $V_n(\theta, v)$; then $\tilde{\theta} = \sqrt{n}(\hat{\theta} - \theta_0)$ and $[\tilde{v}c_n] = n(\hat{\tau} - \tau_0) = \hat{k} - k_0$. Theorem 1′ is equivalent to $\tilde{\theta} = O_p(1)$ and $\tilde{v} = O_p(1)$. Because $V_n(0,0) = 0$, to prove the theorem it suffices to show that when θ or v is large, $V_n(0,0)$ must be large, thus, less likely to achieve its minimum. More specifically, Theorem 1′ is a consequence of the following result.

THEOREM 2. *Under the assumptions of Theorem* 1′,

(i) *for each* $\epsilon > 0$ *and each* $C > 0$, *there exists* $v_1 > 0$ *such that for large* n

$$P\left(\inf_{|v| \geq v_1} \inf_{\theta} V_n(\theta, v) < C\right) < \epsilon, \tag{6}$$

(ii) *for each* $\epsilon > 0$, $C > 0$, *and* $v_1 > 0$, *there exists* $M > 0$ *such that for large* n

$$P\left(\inf_{|v| \leq v_1, \|\theta\| \geq M} \inf V_n(\theta, v) < C\right) < \epsilon. \tag{7}$$

This theorem describes the global behavior of $V_n(\theta, v)$ or, equivalently, the global behavior of $S(\psi, k)$, the sum of absolute deviations. Expression (6) says that, when k is far from k_0, $S(\psi, k)$ simply cannot achieve its global minimum. A consequence of (7) is that, even when one knows the shift point, $S(\psi, k)$ cannot be minimized when ψ is not near the true parameter θ_0.

4. ASYMPTOTIC DISTRIBUTION

The rate of convergence is inferred from the global behavior of the objective function. To obtain the asymptotic distribution, we need to study the local behavior of the objective function. When explicit expressions for the estimated parameters are not available, examination of the objective function is generally the only way to deduce the limiting distribution.

The rate of convergence given in the previous section does not depend on a particular design of the regressors, provided relevant assumptions are fulfilled. The asymptotic distributions, however, depend on the behavior of the regressors. Two cases are to be considered. The first case concerns i.i.d. stochastic regressors, and the second pertains to trending regressors.

A7. The regressors z_i are i.i.d. (may include a constant component), and $Ez_iz_i' = Q_{zz}$ is nonsingular.

Let $\{(\varepsilon_i, z_i);\ -\infty < i < \infty\}$ be a sequence of i.i.d. random vectors, where ε_i is independent of z_i for all i. We define a two-sided random walk $W^\#$ on the integer set with a drift as follows: $W^\#(m) = W_1^\#(m)$ for $m < 0$ and $W^\#(m) = W_2^\#(m)$ for $m > 0$ and $W^\#(0) = 0$, where

$$W_1^\#(m) = \sum_{i=m+1}^{0} |\varepsilon_i - z_i'\lambda_0| - |\varepsilon_i|, \qquad m = -1, -2, \ldots,$$

$$W_2^\#(m) = \sum_{i=1}^{m} |\varepsilon_i + z_i'\lambda_0| - |\varepsilon_i|, \qquad m = 1, 2, \ldots.$$

Thus, $W_1^\#$ and $W_2^\#$ are two independent random walks with each having a positive linear drift. The drift is positive because the expected value of each summand is positive under Assumption A1. Consequently, with high probability, $W^\#(m)$ achieves its minimum near zero. The limiting distribution of $\hat{k}$ is closely related to $W^\#$.

Another case of interest is trending regressors, as follows.

A8. The regressors are functions of time trends: $z_i = g(i/n)$, where g is a vector-valued function defined on [0,1] and is continuously differentiable.

Let

$$Q = \text{plim}\ \frac{1}{n}\sum_{i=1}^{n} X_i X_i',$$

where $X_i = (x_i', z_i', 0)$ for $i \le k_0$ and $X_i = (x_i', 0', z_i')'$ for $i > k_0$.

THEOREM 3. *Under Assumptions* A1–A6, *we have the following*:

(i) *The estimated regression coefficient is asymptotically normal*:

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{4f(0)^2} Q^{-1}\right).$$

(ii) *Assume Assumption* A7 *together with the assumption of a continuous distribution for* $|\varepsilon_i \pm z_i\lambda_0| - |\varepsilon_i|$. *Then, for* $\lambda_n \equiv \lambda_0$,

$$\hat{k} - k_0 \xrightarrow{d} \underset{m}{\text{argmin}}\ W^\#(m).$$

(iii) *Assuming Assumption* A7 *and* $\lambda_n \to 0$, *then*

$$n\lambda_n' Q_{zz}\lambda_n(\hat{\tau} - \tau_0) \xrightarrow{d} \frac{1}{4f(0)^2} \operatorname*{argmax}_v \{W(v) - |v|/2\}.$$

(iv) *Assuming Assumption* A8 *and* $\lambda_n \to 0$, *then*

$$n\lambda_n' g(\tau_0) g(\tau_0)'\lambda_n(\hat{\tau} - \tau_0) \xrightarrow{d} \frac{1}{4f(0)^2} \operatorname*{argmax}_v \{W(v) - |v|/2\},$$

where $f(\cdot)$ *is the density function of* ε_i *and* $W(v)$ *is a two-sided Brownian motion process on* R. *Finally, the distribution of* $n^{1/2}(\hat{\theta} - \theta_0)$ *and that of* $\hat{k} - k_0$ *are asymptotically independent for all cases.*

A two-sided Brownian motion process $W(v)$ is defined as $W(v) = W_1(v)$ for $v \geq 0$ and $W(v) = W_2(-v)$ for $v < 0$, where $W_1(v)$ and $W_2(v)$ are two independent Brownian motion processes on the nonnegative half line with $W_i(0) = 0$ $(i = 1,2)$.

Part (i) of Theorem 3 asserts that estimated regression parameters have a limiting distribution as if the shift point were known provided that the magnitude of the shift is not too small. Part (ii) gives the limiting distribution of the shift point estimator under a fixed magnitude of shift. The assumption that $e_i = |\varepsilon_i \pm z_i'\lambda_0| - |\varepsilon_i|$ has a continuous distribution guarantees the uniqueness (almost surely) of the minimum for $W^{\#}(\cdot)$ because $P(W_i^{\#}(m') = W_i^{\#}(m'')) = 0$ $(i = 1,2)$ for $m' \neq m''$ and $P(W_1^{\#}(m') = W_2^{\#}(m'')) = 0$ for all m' and m''. This uniqueness enables us to invoke the continuous mapping theorem for the argmax functional. A continuous distribution for ε_i is, of course, not sufficient to have a continuous distribution for e_i. For example, for $z_i \equiv 1$ and $\lambda_0 > 0$, we have $|e_i| = \lambda_0$ when $|\varepsilon_i| \geq \lambda_0$. Thus, when ε_i has positive mass outside the interval $[-\lambda_0, \lambda_0]$, e_i cannot have a continuous distribution. In this case, $\operatorname{argmin}\{W^{\#}(m)\}$ is generally a set with more than one element, and the continuous mapping theorem no longer holds. However, one can modify the result as follows. Redefine $\hat{k}$ as the smallest value of the set $\{\ell;\ S(\hat{\theta}(\ell), \ell) = \min_h S(\hat{\theta}(h), h)\}$. Then, $\hat{k} - k_0$ converges in distribution to $\min\{\ell;\ W^{\#}(\ell) = \min_m W^{\#}(m)\}$, which is uniquely defined. It is not difficult to show that e_i will have a continuous distribution if $z_i'\lambda_0$ does. This is because, conditional on ε_i, e_i has a continuous conditional distribution. The distribution of e_i is the average of these conditional distributions.

The location of the minimum value of $W^{\#}(m)$ is stochastically bounded, because $W^{\#}$ has a positive drift. In other words, $W^{\#}(m)$ converges to infinity very quickly as $|m|$ grows unbounded. More precisely, $\operatorname{argmin}_m W^{\#}(m) = O_p(1)$, which, of course, is the result of Theorem 1. The distribution of $W^{\#}(m)$ (and, consequently, that of $\hat{k} - k_0$) is symmetric about zero if and only if $|\varepsilon_i - z_i\lambda_0| - |\varepsilon_i|$ and $|\varepsilon_i + z_i\lambda_0| - |\varepsilon_i|$ have the same distribution. This will be the case if either ε_i has a symmetric distribution about zero or

z_i has a symmetric distribution about zero. When z_i includes a constant regressor, symmetry in $W^{\#}$ requires the symmetry of ε_i.

It is interesting to note that the asymptotic distribution of $\hat{k} - k_0$ depends on the magnitude of shift λ_0 and on the distribution of ε_i and z_i but not on k_0 nor on other parameters of the model. The distributions of ε_i and z_i heavily influence the distribution of $\operatorname{argmin}_m W^{\#}(m)$, as the latter is essentially determined by a finite number of ε_i and z_i. This is in contrast to the case of vanishingly small shifts (part (iii)), where an infinite number of ε_i and z_i are involved as n grows, eventually bringing the Central Limit Theorem to relevance.

A remaining problem is to determine the distribution of $\operatorname{argmin}_m W^{\#}(m)$. Although in principle the problem can be addressed using the approach of Feller (1971, Ch. 18), this approach does not seem to permit analytically tractable results. Hinkley (1970) tried to solve a similar problem with different forms of summands under normality assumption. The solution seems to be too complicated to be of practical use. We shall not attempt any analytical solution because a solution must be solved case by case in view of the dependence on ε_i, z_i and on λ_0. If the distributions of ε_i and z_i together with λ_0 are known, however, the distribution of $\hat{k} - k_0$ can be easily simulated using Monte Carlo methods by constructing $W^{\#}(m)$ directly, with no LAD estimation needed. Details are discussed in Section 6.

Parts (iii) and (iv) of Theorem 3 concern the limiting distribution of the shift point estimator under small shifts. The asymptotic distribution does not depend on the distributions of ε_i and z_i, in contrast to the case in which the shift has a fixed magnitude. The density function of $\operatorname{argmax}\{W(v) - |v|/2\}$ is given by

$$3/2e^{|x|}\Phi(-3\sqrt{|x|}/2) - 1/2\Phi(-\sqrt{|x|}/2),$$

where Φ is the cumulative distribution function of a standard normal random variable (see, e.g., Picard, 1985). When sample size increases, because λ_n converges to zero, more observations in a neighborhood of the true shift point are needed to discern the shift point so that the Central Limit Theorem eventually applies. That gives the precise reason why a Brownian motion is embedded in the limiting process. The size of the neighborhood, however, increases at a much slower rate than the sample size n (more precisely, at the rate $\|\lambda_n\|^{-2}$). A remark here is that part (iii) holds for more general regressors. The i.i.d. assumption for z_i can be replaced by second order stationarity. As can be seen from the proof, all that is needed is

$$\operatorname*{plim}_{\ell\to\infty} \frac{1}{\ell} \sum_{i=k_0-\ell}^{k_0} z_i z_i' = \operatorname*{plim}_{\ell\to\infty} \frac{1}{\ell} \sum_{i=k_0+1}^{k_0+\ell} z_i z_i' = Q_{zz} \tag{8}$$

for some positive definite matrix Q_{zz}.

The asymptotic independence of $\sqrt{n}(\hat{\theta} - \theta_0)$ and $n(\hat{\tau} - \tau_0)$ is due to the fast rate of convergence for $\hat{\tau}$. The estimator $\hat{\tau}$ is determined by a small number of observations near τ_0, whereas $\hat{\theta}$ is determined by the entire set of observations. Whatever values are taken by a small number of observations contribute little (none asymptotically) to statistics comprised of the entire set of observations.

Limiting distributions are obtained by studying the local behavior of the objective function, i.e., the weak convergence of $V_n(\theta, v)$ on compact sets. When θ is constrained to be in a compact set, we essentially deal with parameters in an $n^{-1/2}$ neighborhood of θ_0 in terms of original parameters. This analysis is legitimate only if root-n-consistency for the estimated regression parameters is established. A similar comment applies to the estimated shift point. Obtaining the rate of convergence is necessary because the argmax functional, which is used to deduce the asymptotic distribution, is not a continuous functional for functions defined on an unbounded set. It is only a continuous functional on a set of functions with a compact domain and with a unique maximum. J. Kim and Pollard (1990) offered a rigorous analysis for the argmax functional.

We illustrate the basic step in deriving the asymptotic distributions. Let $D(M) = \{(\theta, v);\ \|\theta\| \le M, |v| \le M\}$ for an arbitrary $M > 0$. We use the uniform metric for functions defined on $D(M)$ (see Pollard, 1984, Ch. 5). The weak convergence of V_n on compact sets is sufficient for us to use the continuous mapping theorem for the argmax functional because of the rate of convergence established in Theorem 1′ (see Kim and Pollard, 1990). Central limit theorems and invariance principles enable us to deduce the limiting process for $V_n(\theta, v)$. For part (iii), if we let $c_n = (\lambda_n' Q_{zz} \lambda_n)^{-1}$, we find that $V_n(\theta, v)$ converges weakly on $D(M)$ for any $M < \infty$ to the process $V(\theta, v)$ given by

$$V(\theta, v) = \theta' Q^{1/2} Z + f(0)\theta' Q \theta + W(v) + f(0)|v|,$$

where Z is a vector of independent standard normal random variables, $f(\cdot)$ is the density function of ε_i, and $W(v)$ is a two-sided Brownian motion process on R independent of Z. The limiting process of $V(\cdot)$ has a unique minimum (almost surely for each sample path) and is minimized at

$$\theta^* = Q^{-1/2} Z$$

and

$$v^* = \underset{v}{\operatorname{argmin}}\{W(v) + f(0)|v|\}.$$

Using the facts that (a) $W(v) \overset{d}{=} W(-v)$, (b) $W(cv) \overset{d}{=} |c|^{1/2} W(v)$, and (c) for any function $h(x)$ and all $a > 0$, $\operatorname{argmin}_x ah(x) = \operatorname{argmin}_x h(x)$, we can show that, by a change in variable,

$$v^* \overset{d}{=} (2f(0))^{-2} \underset{v}{\operatorname{argmax}}\{W(v) - |v|/2\}.$$

The continuous mapping theorem for the argmax functional leads to $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \theta^*$ and $nc_n^{-1}(\hat{\tau} - \tau_0) \xrightarrow{d} v^*$, which gives parts (i) and (iii). Parts (ii) and (iv) are proved in an analogous way.

Confidence intervals. Confidence intervals for the regression parameters θ_0 can be constructed in the usual way in view of Theorem 3(i). Except for $f(0)$, all entities are available from LAD calculations. The density function at zero may be estimated by some nonparametric methods such as kernel based on residuals. One histogram estimator for $4f(0)^2$ suggested by Huber (1981) and reiterated by Buchinsky (1994) is given approximately by $n(\hat{\varepsilon}_{[s]} - \hat{\varepsilon}_{[t]})^2/16$, where $\hat{\varepsilon}_{[k]}$ represents the kth-order statistic, $s = [n/2 - \ell]$, and $t = [n/2 + \ell]$ with $\ell = \sqrt{n}$. Hahn (1992) suggested the bootstrap alternative for confidence intervals. Bootstrap avoids the estimation of the density at zero. Intervals for the shift point are computed analogously. For fixed shift, the asymptotic distribution in part (ii) is unknown, unless the distributions of ε_i and z_i are known so that the asymptotic distribution can be obtained by simulation (Section 6). One solution is to use parts (iii) and (iv) to approximate the distribution of part (ii). This approximation, however, gives a too narrow confidence interval, as illustrated by the Monte Carlo evidence in Section 6. Bootstrap Monte Carlo yields even narrower confidence intervals (a result not reported in Section 6).

5. DEPENDENT DISTURBANCES

In this section, we derive similar results for dependent errors under some mixing conditions. We assume that the sequence of random variables $\{\varepsilon_i\}_{i=1}^{\infty}$ is strongly mixing with an exponential mixing coefficient $\{\alpha_j\}_{i=1}^{\infty}$. For the definition of strong mixing and mixing coefficients, readers are referred to Rosenblatt (1956). Strong mixing is a weaker assumption than many other mixing conditions, as discussed in Bradley (1986). Linear processes, particularly ARMA processes, under some mild conditions, are strongly mixing with exponential mixing coefficients, as shown in Mokkadem (1988), Pham and Tran (1985), and Withers (1981). We make the following assumption.

A9. The errors $\{\varepsilon_i\}$ form a strictly stationary and strongly mixing sequence with exponential mixing coefficients. The existence of a nonzero density function $f(x)$ in a neighborhood zero together with a zero median is maintained.

Let $\phi^2 = 1 + 2\sum_2^{\infty} ED_1 D_j$, where $D_j = \text{sign}(\varepsilon_j)$. Thus, ϕ^2 is the spectral density of the sequence D_i at frequency zero. We assume $\phi^2 > 0$ to avoid degenerate limits. In addition, define

$$U = \lim \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E(X_i X_j' D_i D_j), \qquad (9)$$

where $X_i = (x_i', z_i', 0)$ for $i \le k_0$ and $X_i = (x_i', 0', z_i')'$ for $i > k_0$. The matrix U may be considered the spectral matrix at frequency zero for the vector $\{X_i D_i\}$. We assume U is positive definite. Similarly, let

$$U_{zz} = \lim \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E(z_i z_j' D_i D_j). \tag{10}$$

The estimators $\hat{\theta}$ and $\hat{k}$ are defined as in Section 2. Rates of convergence of the estimated parameters are not affected by dependence in the errors. The asymptotic distributions need to be modified to reflect serial correlation.

THEOREM 4. *Under Assumptions* A2–A6 *and* A9, *we have the following*:

(i) *The estimated regression parameters have a limiting distribution given by*

$$n^{1/2}(\hat{\theta} - \theta_0) \to N\left(0, \frac{1}{4f(0)^2} Q^{-1} U Q^{-1}\right).$$

(ii) *Assuming* $\{(\varepsilon_i, z_i);\ -\infty < i < \infty\}$ *is strictly stationary, Theorem* 3(ii) *holds, but* $W^{\#}$ *consists of two dependent drifted random walks.*
(iii) *Theorem* 3(iii) *holds with* $\lambda_n' Q_{zz} \lambda_n$ *replaced by* $(\lambda_n' Q_{zz} \lambda_n)^2 (\lambda_n' U_{zz} \lambda_n)^{-1}$.
(iv) *Theorem* 3(iv) *holds with* $[4f(0)]^{-2}$ *replaced by* $\phi^2 [4f(0)]^{-2}$, *where* $\phi^2 = 1 + 2\sum_{i=2}^{\infty} ED_1 D_j$ *and* $D_i = \text{sign}(\varepsilon_i)$.

The estimators of regression parameters and the shift point are asymptotically independent.

Again, the limiting distribution for the estimated regression parameters is standard and is the same as if the true change point were known. The scaling factor for the estimated shift point is similarly adjusted to reflect the dependence. To construct confidence intervals, one needs to estimate U and U_{zz}. Methods presented in Newey and West (1987) and Andrews (1991) can be used to estimate the matrices U and U_{zz}. The matrices Q and Q_{zz} are estimated by the corresponding sample moment. When estimating these matrices, one uses $\hat{k}$ in place of the unknown k_0.

Efficiency relative to least squares. In contrast to the least-squares estimation, the asymptotic distributions depend on the density function of ε_i evaluated at zero, not on the second moment, demonstrating the robustness of LAD to thick-tailed distributions. To evaluate the relative efficiency of LAD to least squares, we consider a simple case in which $\lambda_n \to 0$ and $z_t \equiv 1$, so that shift occurs only in the intercept. In this case, $(\lambda_n' Q_{zz} \lambda_n)^2 (\lambda_n' U_{zz} \lambda_n)^{-1} = \lambda_n^2 \phi^{-2}$. Theorem 4(iii) implies

$$n\lambda_n^2(\hat{\tau}_{LAD} - \tau_0) \to \phi^2 \operatorname{argmax}\{W(v) - |v|/2\},$$

where $\hat{\tau}_{LAD}$ represents the LAD shift point estimator. Least-squares estimation of a change point is similar to LAD, but absolute deviation is replaced by squared residuals. Bai (1994b) studied least-squares estimation and obtained some corresponding results. Assuming ε_i is an ARMA process such that

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \rho_2 \varepsilon_{i-2} + \cdots \rho_p \varepsilon_{i-p} + e_i + \phi_1 e_{i-1} + \cdots + \psi_q e_{i-q},$$

where e_i is white noise with variance σ^2, the asymptotic distribution of the change point estimator $\hat{\tau}_{LS}$ is given by

$$n\lambda_n^2(\hat{\tau}_{LS} - \tau_0) \rightarrow \bar{\sigma}^2 \operatorname*{argmax}_v \{W(v) - |v|/2\},$$

where $\bar{\sigma}^2 = \sigma^2(1 + \phi_1 + \cdots + \phi_q)(1 - \rho_1 - \cdots - \rho_p)^{-1}$. Clearly, the asymptotic distribution of $\hat{\tau}_{LS}$ has the same form as the LAD estimator but with a different scale. Also, the least-squares estimator depends on the second moment. So the given asymptotic distribution will not be valid if the second moment does not exist. The relative efficiency can be determined by comparing the scale coefficients. Define the rate of efficiency of LAD relative to least squares, e, as the ratio of their asymptotic variances so that

$$e = 4f(0)^2\bar{\sigma}^2/\phi^2.$$

The larger is e, the more efficient is LAD relative to least squares. When ε_i are i.i.d., the efficiency rate becomes $e = 4f(0)^2\sigma^2$, because $\phi^2 = 1$ and $\bar{\sigma}^2 = \sigma^2$ in this case. For i.i.d. normal distributions, $e = 0.637$, and so least squares is more efficient than LAD. For the double exponential distribution, $e = 2$; LAD is more efficient, indicating the robustness of LAD against thick-tailed distributions. When ε_i is contaminated normal having a distribution $(1 - \epsilon)N(0,1) + \epsilon N(0,\gamma)$, e tends to infinity as γ grows unbounded. For a Cauchy distribution, the variance does not exist, so the limiting distribution of the least-squares estimator is not well defined. If one considers a truncated Cauchy distribution and allows the truncation value to increase to infinity, e also increases to infinity.

More interesting is that under normality the relative efficiency of LAD increases as the correlation coefficient becomes larger. For simplicity, assume $\varepsilon_i = \rho\varepsilon_{i-1} + e_i$, where the e_i are i.i.d. normal $N(0,\sigma^2)$. The relative efficiency measure e becomes

$$e = \frac{\sigma^2/(1-\rho)^2}{\phi^2/(4f(0)^2)} = \frac{2(1-\rho)}{\pi\phi^2(1+\rho)}.$$

Gastwirth and Rubin (1975) showed in another context that e is an increasing function of ρ (note that ϕ^2 depends on ρ in a very complicated way). For $\rho = 0$ (i.e., i.i.d.), $e = 0.636$; for $\rho = 0.5$, $e = 0.828$; and for $\rho = 0.9$, $e =$ 0.911. Under normality, LAD is not as efficient as least squares, as expected. But the efficiency of LAD improves as (positive) correlation increases. For other heavy-tailed distributions such as double exponential, LAD is always more efficient than least squares.

6. MONTE CARLO SIMULATION

In this section, we report some Monte Carlo results for the behavior of the shift point estimator. Data are generated according to

$$\begin{aligned} y_i &= a_1 + b_1 x_i + \varepsilon_i \qquad i = 1, \ldots, k_0, \\ y_i &= a_2 + b_2 x_i + \varepsilon_i \qquad i = k_0 + 1, \ldots, n, \end{aligned} \tag{11}$$

where $n = 100$, $k_0 = 50$, x_i are i.i.d. $N(0,1)$, and ε_i are i.i.d. double exponential random variables. We consider three sets of parameters: (I) $a_2 - a_1 = 1$, $b_2 - b_1 = 1$; (II) $a_2 - a_1 = \sqrt{2}$, $b_1 = b_2$; (III) $a_2 - a_1 = 1$, $b_1 = b_2$. Actual values for a_i and b_i do not matter—only their differences matter. For each set of parameters, 10,000 repetitions are performed. Frequencies for the estimated shift point based on LAD are reported in Table 1 (columns 2, 5, and 8, respectively). These results serve as benchmarks for comparison with asymptotic frequencies. We point out that for the last two cases the restriction $b_1 = b_2$ is imposed, which leads to a smaller spread for the estimated shift points compared to the estimates when the restriction is not imposed. With set (II), the standard deviation of restricted estimates is 8.28 and that of unrestricted estimates is 10.58. With set (III), the restricted and unrestricted standard deviations are 4.30 and 4.67, respectively. (Unrestricted estimates are not reported here.)

Accompanying the LAD Monte Carlo benchmarks are frequencies obtained from two asymptotic distributions. The first asymptotic distribution corresponds to Theorem 3(ii) and is used to approximate the probability $P(\hat{k} - k_0 = \ell)$ by

$$\text{Asy(ii):} \quad P\left(\operatorname*{argmin}_{m} \; W^{\#}(m) = \ell\right).$$

The second asymptotic distribution is given in Theorem 3(iii) and is used to approximate the probability $P(\hat{k} - k_0 = \ell)$ by

$$\text{Asy(iii):} \quad \int_{\ell-0.5}^{\ell+0.5} \alpha^{-1} h(\alpha^{-1} x)\, dx,$$

where $h(x)$ is the density function of $\operatorname{argmin}\{W(v) - |v|/2\}$ and $\alpha = \lambda_n' Q_{zz} \lambda_n$ (note that $4f(0)^2 = 1$ for the double exponential distribution). Frequencies based on Asy(ii) are reported in columns 3, 6, and 9 of Table 1, and those based on Asy(iii) are reported in columns 4, 7, and 10, corresponding to sets (I), (II), and (III), respectively. Details of each approximation are explained next.

Because the theoretical distribution in Asy(ii) is unknown, we compute the probability by simulations as well. The simulation involves constructing a two-sided random walk $W^{\#}(m)$. Again, 10,000 repetitions are conducted. For each repetition, cumulative sums (random walks) on each side are com-

TABLE 1. Frequency of $\hat{k} - k_0$

	Set (I)			Set (II)			Set (III)		
$\hat{k} - k_0$	Monte Carlo[a]	Asy(ii)[b]	Asy(iii)[c]	Monte Carlo	Asy(ii)	Asy(iii)	Monte Carlo	Asy(ii)	Asy(iii)
<−20	72	30	1	26	8	1	257	98	44
−20	11	8	0	6	5	0	39	15	8
−19	12	5	0	10	5	0	20	18	10
−18	14	6	1	5	8	1	21	25	12
−17	19	18	1	12	5	1	23	19	14
−16	13	12	2	6	11	2	36	29	17
−15	28	15	2	18	13	2	34	33	21
−14	28	24	4	20	12	4	45	41	26
−13	27	31	5	17	19	5	54	75	32
−12	36	37	8	36	19	8	69	52	40
−11	53	39	11	40	20	11	77	64	50
−10	53	62	16	49	42	16	87	92	62
−9	60	62	24	60	42	24	110	125	79
−8	93	86	35	73	62	35	131	105	100
−7	123	122	53	90	95	53	146	158	130
−6	172	157	81	133	107	81	187	200	170
−5	235	218	126	171	200	126	232	271	228
−4	307	296	203	235	295	203	282	389	314
−3	402	412	345	376	391	345	420	469	449
−2	662	641	640	569	725	640	599	686	689
−1	1,103	1,180	1,443	1,141	1,122	1,443	935	953	1,224
0	2,833	3,094	3,977	3,618	3,579	3,977	2,251	2,467	2,542
1	1,174	1,193	1,443	1,172	1,185	1,443	989	1,012	1,224
2	666	658	640	695	805	640	633	575	689
3	434	439	345	389	379	345	426	438	449
4	282	268	203	258	242	203	318	307	314
5	226	188	126	187	157	126	280	216	228
6	177	169	81	122	112	81	179	188	170
7	120	108	53	96	91	53	170	146	130
8	106	81	35	77	50	35	139	103	100
9	90	67	24	68	44	24	114	97	79
10	59	59	16	48	35	16	82	80	62
11	34	38	11	31	26	11	70	70	50
12	42	33	8	29	19	8	59	67	40
13	35	33	5	22	18	5	62	50	32
14	28	18	4	15	8	4	61	27	26
15	24	19	2	13	10	2	47	30	21
16	16	14	2	10	9	2	28	31	17
17	13	16	1	8	4	1	36	21	14
18	11	9	1	5	3	1	32	20	12
19	10	8	0	8	1	0	20	17	10
20	21	3	0	4	4	0	20	10	8
>20	76	24	1	32	13	1	180	111	44

[a] From 10,000 repetitions.
[b] Based on Theorem 3(ii).
[c] Based on Theorem 3(iii).

puted with 500 observations ($W^{\#}(m)$, $m = 0, \pm 1, \ldots, \pm 500$). The location of the minimum value of $W^{\#}(m)$ is then found. This computation is extremely fast, because no LAD estimation is needed. The result is not sensitive to the number of observations used when constructing the random walk. Almost the same distribution is obtained as long as each side has more than 50 observations. For set (I), because $|\varepsilon_i \pm (1 + x_i)| - |\varepsilon_i|$ has a continuous distribution, $\text{argmin}_m W^{\#}(m)$ is uniquely defined. This is also consistent with the results of Monte Carlo simulations. For sets (II) and (III), the two-sided random walk has multiple minima. Columns 6 and 8 report the minimum value[3] of the set $\{\text{argmin}_m W^{\#}(m)\}$.

The frequencies in columns 4, 7, and 10 are based on the theoretical probabilities in Asy(iii) multiplied by 10,000. For set (I), $Q_{zz} = \text{diag}(1,1)$ and $\lambda_n = (1,1)'$; for set (II), $Q_{zz} = 1$ and $\lambda_n = \sqrt{2}$. So for both sets (I) and (II), $\lambda_n' Q_{zz} \lambda_n = 2$. Therefore, asymptotic theory based on shrinking shifts predicts the same frequency distribution for sets (I) and (II); yet Monte Carlo benchmarks for sets (I) and (II) (columns 2 and 5, respectively) show significant differences for their underlying distributions. For set (II), the asymptotic distribution gives a better approximation to the finite sample counterpart. These results are not surprising. With set (I), the asymptotic distribution treats essentially a finite average of $z_i z_i'$ as Q_{zz} (see equation (A.42) in Appendix), which is a poor approximation particularly for large shifts, whereas for set (II), $Q_{zz} = 1$ is an exact result. Thus, we expect that asymptotic distribution gives a better approximation for set (II) than for set (I).

Inspecting Table 1, we find that the asymptotic distribution based on random walks gives a very good approximation to the Monte Carlo LAD benchmark for all three cases. The asymptotic distribution based on shrinking shifts offers an unsatisfactory approximation. Although the general picture agrees with the benchmark, the predicted spread is too narrow compared with the benchmark.

Although not reported here, the simulated result is not sensitive to the choice of k_0, provided k_0 is reasonably bounded away from the two ends. When k_0 is set to 30, a quite similar frequency pattern is observed. We point out that when k_0 is too small, however, the distribution of $\hat{k}$ will be skewed to the right because $\hat{k} - k_0 \geq -k_0$, restricting the values of $\hat{k}$ from below. Also not reported here are some bootstrap results for estimating the underlying distribution. Preliminary bootstrap simulations exhibit variation that is much smaller than the Monte Carlo benchmark, even smaller than that predicted by the asymptotic distributions. We plan to further investigate bootstrap methodology for estimating a shift further.

7. DISCUSSION

This section considers potential avenues leading to improvements and generalizations of the results of this paper.[4]

Heteroskedasticity. The i.i.d. and zero median assumptions for the errors can be relaxed to a zero conditional median, allowing for an unspecified form of heteroskedasticity. Newey and Powell (1990) proposed an efficient estimation technique using weighted absolute sums, with weight equal to the conditional density at zero of disturbances. They also show how to estimate the conditional density at zero to make the approach feasible. Their approach can be adapted to estimate models with a shift to generate efficient estimation for both the regression coefficients and the shift point. Another direction of generalization is to consider more efficient estimation under serial correlation. For this purpose, a more concrete specification of the correlation structure seems necessary. Although we have not considered how to estimate the shift point efficiently, an efficient estimator for the regression parameter can be constructed by a one-step adaptation. Let sign(ε) be the vector of sign(ε_i) $(i = 1,\dots,n)$ and denote $E\,\text{sign}(\varepsilon)\text{sign}(\varepsilon)'$ by $\Omega_{n\times n}$. Let

$$\tilde{\theta} = \hat{\theta} + (\hat{X}\hat{\Omega}^{-1}\hat{X})^{-1}\hat{X}'\hat{\Omega}^{-1}\,\text{sign}(\hat{\varepsilon}), \tag{12}$$

where $\hat{\Omega}$ is an estimate of Ω based on sign$(\hat{\varepsilon})$ and $\hat{X}$ is an $n \times (p+q)$ matrix consisting of X_i' as its rows (with k_0 replaced by $\hat{k}$). All the entities with a circumflex are constructed based on preliminary estimators $\hat{k}$ and $\hat{\theta}$. When Ω only depends on a fixed number of parameters irrespective of the sample size (e.g., a moving average process for ε_i) the right-hand side of (12) will be a good approximation for the corresponding quantity of a known Ω. The matrix $\hat{X}$ always behaves like X because of the fast rate of convergence for the shift point. The estimator $\tilde{\theta}$ is more efficient than $\hat{\theta}$ and has an asymptotic variance U^{-1}.

Quantile regression. The analysis presented in this paper can be extended to quantile regressions proposed by Koenker and Bassett (1978). Let $q_\psi(x) = x[\psi I(x \geq 0) - (1-\psi)I(x<0)]$ for some $\psi \in (0,1)$. The quantile regression estimator of (θ_0, k_0) is obtained by minimizing

$$S(\theta,k) = \sum_{i=1}^{k} q_\psi(y_i - x_i'\beta - z_i'\delta_1) + \sum_{i=k+1}^{n} q_\psi(y_i - x_i'\beta - z_i'\delta_2).$$

Zero median of Assumption A1 is now changed to zero ψ-quantile. All remaining assumptions are maintained. All the proofs of this paper can be extended to quantile regressions without essential difficulty. Theorem 3 holds with $4f(0)^2$ replaced by $(\psi(1-\psi))^{-1}f(0)^2$. The random walk summands in part (ii) become $q_\psi(\varepsilon_i - z_i'\lambda_0) - q_\psi(\varepsilon_i)$. Theorem 4 also holds with similar amendment.

Multiple shifts. The model considered in the paper is restrictive for applications in economics because of the assumption of a single shift. An important generalization is to allow for more than one shift. A number of questions arise in this context. How should the hypothesis of s shifts versus $s + \ell$ shifts be tested? How can the computation problem be solved? What are the statistical properties of the resulting estimators including the shift point estimators? These issues may be addressed under the general quantile regression framework allowing for heteroskedasticity and serial correlation. The present study serves as a starting point for further research in this area.

8. SUMMARY

In this paper, we developed the asymptotic theory for the LAD estimation of a shift in linear regressions. We examined the joint statistical behavior of the estimated regression parameters and the estimated shift point. We showed that the asymptotic distribution of estimated regression parameters is the same as if the shift point were known, owing to the fast rate of convergence for the shift point estimator. We also derived the asymptotic distribution of the shift point estimator both in the case in which the magnitude of shift is fixed and in the case in which the magnitude becomes vanishingly small. Under the former, the asymptotic distribution is related to a drifted two-sided random walk defined on the integer set. For the latter, the asymptotic distribution is related to a drifted two-sided Brownian motion on the real line. Trending regressors and dependent disturbances are also considered. Monte Carlo analysis reveals that the asymptotic distribution based on random walks gives a very good approximation for the underlying distribution, but the asymptotic distribution based on Brownian motions is less satisfactory.

We conducted our analysis within the partial structural change framework, allowing some of the parameters to stay constant throughout the sample period. This generates a more efficient estimator for the regression parameters if the constraint is valid. For the shift point estimator, its asymptotic distribution is the same whether or not one imposes such a constraint. But in small samples, there is some gain in efficiency, as suggested by the Monte Carlo evidence.

NOTES

1. More extensive references can be found in the annotated bibliography of Hackl and Westlund (1989).
2. It should be pointed out that the notion of robustness here is not in the strict sense of Hampel (1974) or Huber (1981) because of the nonsmoothness of the objective function at zero. For further discussion, see Bloomfield and Steiger (1983).
3. One can also use the maximum value of this set as well as the median of this set, provided the LAD estimator $\hat{k}$ is defined correspondingly.
4. A referee suggested several of the ideas presented here.
5. Interested readers are referred to an earlier version of this paper (Bai, 1993).

REFERENCES

Amemiya, T. (1982) Two stage least absolute deviations estimators. *Econometrica* 50, 689-711.

Andrews, D.W.K. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817-858.

Andrews, D.W.K. (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821-856.

Andrews, D.W.K. & W. Ploberger (1992) Optimal Tests of Parameter Constancy. Manuscript, Cowles Foundation for Research in Economics, Yale University.

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

Bai, J. (1993) LAD Estimation of a Shift. Manuscript, Department of Economics, Massachusetts Institute of Technology.

Bai, J. (1994a) Estimation of Structural Change Based on Wald Type Statistics. Working paper 94-6, Department of Economics, Massachusetts Institute of Technology.

Bai, J. (1994b) Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis* 15, 453-472.

Banerjee, A., R.L. Lumsdaine, & J.H. Stock (1992) Recursive and sequential tests of the unit root and trend break hypothesis, theory and international evidence. *Journal of Business & Economic Statistics* 10, 271-287.

Bassett, G. & K. Koenker (1978) Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* 73, 618-622.

Bhattacharya, P.K. (1987) Maximum likelihood estimation of a change-point in the distribution of independent random variables, general multiparameter case. *Journal of Multivariate Analysis* 23, 183-208.

Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.

Birnbaum, Z.W. & A.W. Marshall (1961) Some multivariate Chebyshev inequalities with extensions to continuous parameter processes. *Annals of Mathematical Statistics* 32, 687-703.

Bloomfield, P. & W. Steiger (1983) *Least Absolute Deviations, Theory, Applications, and Algorithms*. Boston: Birkhauser.

Bradley, R.C. (1986) Basic properties of strong mixing conditions. In E. Eberlein and M.S. Taqque (eds.), *Dependence in Probability and Statistics*, pp. 165-192. Boston: Birkhauser.

Broemeling, L.D. & H. Tsurumi (1987) *Econometrics and Structural Change*. New York: Marcel Dekker.

Buchinsky, M. (1994) Changes in the U.S. wage structure 1963-1987, application of quantile regression. *Econometrica* 62, 405-458.

Carlstein, E. (1988) Nonparametric change point estimation. *Annals of Statistics* 16, 188-197.

Chamberlain, G. (1991) Quantile Regression, Censoring, and Structure of Wage. Discussion paper 1558, Harvard Institute of Economic Research.

Christiano, L.J. (1992) Searching for a break in GNP. *Journal of Business & Economic Statistics* 10, 237-250.

Chu, C.-S.J. & H. White (1992) A direct test for changing trend. *Journal of Business & Economic Statistics* 10, 289-300.

Duembgen, L. (1991) The asymptotic behavior of some nonparametric change point estimators. *Annals of Statistics* 19, 1471-1495.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*. New York: Wiley.

Gastwirth, J.L. & H. Rubin (1975) The behavior of robust estimators on dependent data. *Annals of Statistics* 3, 1070-1100.

Gyorfi, L., W. Hardle, P. Sarda, & P. Vieu (1990) *Nonparametric Curve Estimation from Time Series*. New York: Springer-Verlag.

Hackl, P. & A.H. Westlund (1989) Statistical analysis of "structural change", an annotated bibliography. In W. Kramer (ed.), *Econometrics of Structural Change*, pp. 103–128. Heidelberg: Physica-Verlag.

Hahn, J. (1992) Bootstrapping Quantile Regression Models. Manuscript, Department of Economics, Harvard University.

Hampel, F.R. (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 62, 1179–1186.

Hansen, B.E. (1992) Tests for parameter instability in regressions with I(1) processes. *Journal of Business & Economic Statistics* 10, 321–335.

Hawkins, D.L. (1986) A simple least square method for estimating a change in mean. *Communications in Statistics-Simulations* 15, 655–679.

Hawkins, D.M. (1977) Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association* 72, 180–186.

Hinkley, D. (1970) Inference about the change point in a sequence of random variables. *Biometrika* 57, 1–17.

Honore, B.E. (1992) Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 533–565.

Hsu, D.A. (1982a) A Bayesian robust detection of shift in the risk structure of stock market returns. *Journal of the American Statistical Association* 77, 29–39.

Hsu, D.A. (1982b) Robust inferences for structural shift in regression models. *Journal of Econometrics* 19, 89–107.

Huber, P.J. (1981) *Robust Statistics*. New York: Wiley.

Ibragimov, I.A. & Y.A. Linnik (1971) *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noordhoff.

James, B., K. James, & D. Siegmund (1987) Tests for a change point. *Biometrika* 74, 71–83.

Kim, J. & D. Pollard (1990) Cube root asymptotics. *Annals of Statistics* 18, 191–219.

Kim, H.J. & D. Siegmund (1989) The likelihood ratio test for a change point in simple linear regression. *Biometrika* 76, 409–423.

Knight, K. (1989) Limit theory for autoregressive parameter estimates in an infinite variance random walk. *Canadian Journal of Statistics* 17, 261–278.

Knight, K. (1991) Limit theory for M-estimates in an integrated infinite variance process. *Econometric Theory* 7, 200–212.

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

Koenker, R. & G. Bassett (1982) Robust tests for heteroskedasticity based on regression quantiles. *Econometrica* 56, 43–61.

Kramer, W., W. Ploberger, & R. Alt (1988) Testing for structural changes in dynamic models. *Econometrica* 56, 1355–1370.

Krishnaiah, P.R. & B.Q. Miao (1988) Review about estimation of change points. In P.R. Krishnaiah & C.R. Rao (eds.), *Handbook of Statistics*, vol. 7, pp. 375–402. New York: Elsevier.

Mokkadem, A. (1988) Mixing properties of ARMA processes. *Stochastic Processes and Their Applications* 29, 309–315.

Newey, W.K. & J.L. Powell (1990) Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* 6, 295–317.

Newey, W.K. & K. West (1987) A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Perron, P. (1993) Trend, Unit Root and Structural Change in Macroeconomic Time Series. Manuscript, C.R.E.D., University of Montreal.

Perron, P. & T. Vogelsang (1992) Nonstationarity and level shifts with an application to purchasing power parity. *Journal of Business & Economic Statistics* 10, 321–335.

Pham, T.D. & L.T. Tran (1985) Some mixing properties of time series models. *Stochastic Processes and Their Applications* 19, 297–303.

Phillips, P.C.B. (1991) A shortcut to LAD estimator asymptotics. *Econometric Theory* 7, 450–463.

Picard, D. (1985) Testing and estimating change-points in time series. *Advances in Applied Probability* 17, 841–867.

Pollard, D. (1984) *Convergence of Stochastic Processes.* New York: Springer-Verlag.

Pollard, D. (1990) *Empirical Processes: Theory and Applications*, vol. 2. CBMS Conference Series in Probability and Statistics. Hayward, CA: Institute of Mathematical Statistics.

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

Powell, J.L. (1984) Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303–325.

Quandt, R.E. (1958) The estimation of parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* 53, 873–880.

Rosenblatt, M. (1956) A central limit theory and a strong mixing condition. *Proceedings of the National Academy of Sciences USA* 42, 43–47.

Sen, A. & M.S. Srivastava (1975a) On tests for detecting change in mean. *Annals of Statistics* 3, 96–103.

Sen, A. & M.S. Srivastava (1975b) Some one-sided tests for change in level. *Technometrics* 17, 61–64.

Serfling, R.J. (1970) Moment inequalities for the maximum cumulative sum. *Annals of Mathematical Statistics* 41, 1227–1234.

Weiss, A. (1991) Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory* 7, 46–68.

Withers, C.S. (1981) Conditions for linear processes to be strong-mixing. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57, 477–480.

Wooldridge, J.M. & H. White (1988) Some invariance principles and central limit theorems for dependent and heterogeneous processes. *Econometric Theory* 4, 210–230.

Worsley, K.J. (1979) On the likelihood ratio test for a shift in locations of normal populations. *Journal of the American Statistical Association* 72, 180–186.

Worsley, K.J. (1986) Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* 73, 91–104.

Yao, Yi-Ching (1987) Approximating the distribution of the ML estimate of the change-point in a sequence in independent r.v.'s. *Annals of Statistics* 15, 1321–1328.

Zacks, S. (1983) Survey of classical and Bayesian approaches to the change-point problem, fixed sample and sequential procedures of testing and estimation. In M.H. Rivzi, J.S. Rustagi, & D. Siegmund (eds.), *Recent Advances in Statistics*, pp. 245–269. New York: Academic Press.

Zivot, E. & P.C.B. Phillips (1994) A Bayesian analysis of trend determination in economics time series. *Econometric Reviews* 13, 291–336.

APPENDIX: PROOFS

Proof of Theorem 2. For simplicity, we assume the regressors x_i and z_i are deterministic. All proofs go through for stochastic regressors by using conditional arguments because of the assumed independence of regressors and disturbances. We shall consider the case of $v \le 0$ without loss of generality because of symmetry. For $v \le 0$, implying $k(v) \le k_0$, $V_n(\theta, v)$ can be written as

$$V_n(\theta,v) = \sum_{i=1}^{k(v)} (|\varepsilon_i - x_i'\beta n^{-1/2} - z_i'\delta_1 n^{-1/2}| - |\varepsilon_i|)$$
$$+ \sum_{i=k(v)+1}^{k_0} (|\varepsilon_i - x_i'\beta n^{-1/2} - z_i'\delta_2 n^{-1/2} - z_i'\lambda_n| - |\varepsilon_i|)$$
$$+ \sum_{i=k_0+1}^{n} (|\varepsilon_i - x_i'\beta n^{-1/2} - z_i'\delta_2 n^{-1/2}| - |\varepsilon_i|). \qquad \textbf{(A.1)}$$

Letting $w_i = (x_i', z_i')'$ and

$$U_n(j,l,\phi) = \sum_{i=l}^{j} (|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|),$$

we have

$$V_n(\theta,v) = U_n(1,k(v),\phi_1) + U_n(k(v)+1,k_0,\sqrt{n}\lambda_n^* + \phi_2) + U_n(k_0,n,\phi_2), \qquad \textbf{(A.2)}$$

where $\phi_1 = (\beta,\delta_1)$, $\phi_2 = (\beta,\delta_2)$, and $\lambda_n^* = (0',\lambda_n')'$. (Note that $\|\lambda_n^*\| = \|\lambda_n\|$.) To study the behavior of V_n, it is sufficient to study the behavior of $U_n(1,k,\phi)$ for all k and all ϕ. We shall call $U_n(1,k,\phi)$ the sequential objective function (s.o.f.) and $\inf_s U_n(1,k,\phi)$ the optimized s.o.f. The following lemma gives various properties of the s.o.f. The results are general enough to allow us to deal with both fixed shifts and shrinking shifts.

LEMMA A.1. *Under the assumptions of Theorem* 2, *we have the following*:

(i) *For each* $\delta \in (0,1)$,

$$\sup_{n\ge k\ge n\delta} \left| \inf_{\phi} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \right| = O_p(1).$$

(ii)

$$\sup_{1\le k\le n} \left| \inf_{\phi} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \right| = O_p(\log n).$$

(iii) *For each* $\delta \in (0,1)$, $\epsilon > 0$, *and* $D > 0$, *we have for large* n

$$P\left(\inf_{n\ge k\ge n\delta,} \ \inf_{\|\phi\|\ge \log n} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|) < D\log n \right) < \epsilon.$$

(iv) *For each* $\epsilon > 0$ *and* $D > 0$, *there exists an* $M < \infty$ *such that for large* n

$$P\left(\inf_{n\ge k\ge n\delta} \ \inf_{\|\phi\|\ge M} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|) < D \right) < \epsilon.$$

(v) *Let* h_n *and* d_n *be positive sequences such that* h_n *is nondecreasing and* $d_n \to +\infty$ *and* $(h_n d_n^2)/n \to h > 0$, *where* $h < \infty$. *Then, for each* $\epsilon > 0$ *and* $D > 0$, *there exists an* $A > 0$ *such that*

$$Pr\left(\inf_{n\ge k\ge Ah_n} \ \inf_{\|\phi\|\ge d_n} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|) < D \right) < \epsilon.$$

(vi) *Under the same hypotheses as in part* (v), *we have for any given* $A > 0$

$$\sup_{k \le Ah_n} \left| \inf_{\|\phi\| \le d_n} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|) \right| = O_p(1).$$

The proof of the lemma is technical; we thus postpone its proof and take it as granted for the moment. Let us now look at some implications of the lemma. Part (i) simply says that when a positive fraction of data is used ($k \ge n\delta$), the optimized s.o.f. is then stochastically bounded (uniformly in $k \ge n\delta$). Part (ii) concerns the global behavior of the s.o.f. The optimized s.o.f. is uniformly bounded in k by $O_p(\log n)$. Parts (iii) and (iv) assert that when $\|\phi\|$ is large $U_n(1,k,\phi)$ will also be large, thus less likely achieving its minimum for large ϕ. Part (v) is similar to part (iii) but does not require that a positive fraction of data be used. The last part is similar to part (i), but again no positive fraction of data is required.

Proof of (6). This proof uses some ideas of Picard (1985). We divide the set $\{(\theta, v);\ \theta \in R,\ v \le -v_1\}$ into three regions:

$$B_1 = \{(\theta, v);\ \|\phi_2\| \le \tfrac{1}{2} n^{1/2}\|\lambda_n\|, n\delta \le k(v) \le k(-v_1)\},$$

$$B_2 = \{(\theta, v);\ \|\phi_2\| \le \tfrac{1}{2} n^{1/2}\|\lambda_n\|, 0 \le k(v) \le n\delta\},$$

$$B_3 = \{(\theta, v);\ \|\phi_2\| \ge \tfrac{1}{2} n^{1/2}\|\lambda_n\|, 0 \le k(v) \le k(-v_1)\},$$

where δ is a small positive number such that $\delta < \tau_o$. We examine the behavior of $V_n(\cdot)$ on each of the three sets.

On B_1, the behavior of $U_n(1, k(v), \phi_1)$ and $U_n(k_0, n, \phi_2)$ (the first and third terms of (A.2) on the right-hand side) is governed by Lemma A.1(i) because both terms involve a positive fraction of data. Because $\|n^{1/2}\lambda_n^* + \phi_2\| \ge n^{1/2}\|\lambda_n\| - \|\phi_2\| \ge n^{1/2}\|\lambda_n\|/2$, we apply Lemma A.1(v) with $d_n = n^{1/2}\|\lambda_n\|/2$, $h_n = \|\lambda_n\|^{-2}$, and $A = v_1$ to deduce that $U_n(k(v)+1, k_0, n^{1/2}\lambda_n^* + \phi_2) \ge D$ with high probability for any $D > 0$ as long as v_1 is large. (Note first that we have applied Lemma A.1(v) with the data order reversed, treating k_0 as the first observation. Second, note that at least $k_0 - k(-v_1) \ge v_1\|\lambda_n\|^{-2}$ observations are involved.) Thus, $V_n \ge O_p(1) + D + O_p(1)$, with high probability. Because $D > 0$ can be arbitrarily large by choosing a large v_1, V_n is large if v_1 is large.

On B_2, applying Lemma A.1(ii), (iii), and (i), respectively, to the three terms on the right-hand side of (A.2), we find that $V_n \ge O_p(\log n) + D\log n + O_p(1)$. Because D can be arbitrarily large, so is V_n. Note that we have utilized the fact that $\|n^{1/2}\lambda_n + \phi_2\| \ge n^{1/2}\|\lambda_n\|/2$, which is larger than $\log n$ by Assumption A6.

On B_3, we have $V_n \ge O_p(\log n) + O_p(\log n) + D\log n$ by Lemma A.1(ii) for the first two terms and by Lemma A.1(iii) for the third term of (A.2). Again, V_n can be large because D is arbitrary.

Proof of (7). The set $\{(\theta, v);\ \|\theta\| \ge M,\ -v_1 \le v \le 0\}$ is contained in the union of D_1 and D_2, where

$$D_1 = \{(\theta, v);\ \|\theta\| \ge M, \|n^{1/2}\lambda_n^* + \phi_2\| \le 2n^{1/2}\|\lambda_n\|, -v_1 \le v \le 0\},$$

$$D_2 = \{(\theta, v);\ \|n^{1/2}\lambda_n^* + \phi_2\| \ge 2n^{1/2}\|\lambda_n\|, -v_1 \le v \le 0\}.$$

On D_1, by Lemma A.1(iv), either $U_n(1, k(v), \phi_1)$ or $U_n(k_0, n, \phi_2)$ is large because either $\|\phi_1\| > M$ or $\|\phi_2\| > M$. The term $U_n(k(v)+1, k_0, n^{1/2}\lambda_n^* + \phi_2)$ is stochastically bounded in view of Lemma A.1(vi) by choosing d_n, h_n, and A as on set B_1. Thus, V_n will be large if M is large.

On D_2, notice that $\|\phi_2\| \geq \|n^{1/2}\lambda_n^* + \phi_2\| - n^{1/2}\|\lambda_n\| \geq 2n^{1/2}\|\lambda_n\| - n^{1/2}\|\lambda_n\| = n^{1/2}\|\lambda_n\| \geq \log n$. Therefore, by Lemma A.1(iii), $U_n(k_0, n, \phi_2) \geq D(\log n)$ for any $D > 0$ with high probability. The first two terms of (A.2) on the right are not less than $-|O_p(\log n)|$ by Lemma A.1(ii). Thus, V_n is large on D_2, completing the proof of Theorem 2. ■

LEMMA A.2 (Babu, 1989, Lemma 1). *Let Z_i be a sequence of independent random variables with mean zero and $|Z_i| \leq d$ for some $d > 0$. Let $V \geq \sum_{i=1}^k EZ_i^2$. Then, for all $0 < s < 1$ and $0 \leq a \leq V/(sd)$,*

$$P\left(\left|\sum_{i=1}^k Z_i\right| > a\right) \leq 2\exp\{-a^2 s(1-s)/V\}. \tag{A.3}$$

Proof of Lemma A.1. To prove this lemma, we use some results of Babu (1989) that are concerned with the strong representation of LAD estimators. The difficulty of the proof lies in the sequential nature of the problem. We need to bound the sequential objective function over all k ($k \geq 1$).

Proof of (i). Denote $\hat{\phi}_k = \operatorname{argmin}_\phi \sum_{i=1}^k |\varepsilon_i - w_i'\phi|$. Then, $\sup_{k\geq n\delta}|\hat{\phi}_k| = O_p(n^{-1/2})$ (Babu, 1989, Theorem 1). Thus, it is sufficient to prove that, for each $M > 0$,

$$\sup_{n\geq k\geq n\delta}\ \sup_{\|\phi\|\leq M}\left|\sum_{i=1}^k |\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|\right| = O_p(1).$$

However, a stronger result holds once ϕ is restricted to lie in a compact set. The requirement of $k \geq n\delta$ is no longer necessary. We have

$$\sup_{n\geq k\geq 1}\ \sup_{\|\phi\|\leq M}|G_{k,n}(\phi)| = O_p(1), \tag{A.4}$$

where $G_{k,n}(\phi) = \sum_{i=1}^k (|\varepsilon_i - n^{-1/2}w_i'\phi| - |\varepsilon_i|)$. This result can be proved easily using an argument of Pollard (1991). Denote $D_i = \operatorname{sign}(\varepsilon_i)$. Then, $ED_i = 0$ by the assumption of zero median. Letting

$$R_{i,n}(\phi) = |\varepsilon_i - n^{-1/2}w_i'\phi| - |\varepsilon_i| - n^{-1/2}\phi' w_i D_i, \tag{A.5}$$

we have

$$G_{k,n}(\phi) = n^{-1/2}\phi'\sum_{i=1}^k w_i D_i + \sum_{i=1}^k R_{i,n}(\phi). \tag{A.6}$$

Because the D_i are bounded, zero mean, and independent random variables, the invariance principle implies that $\|n^{-1/2}\sum_{i=1}^k w_i'D_i\|$ is $O_p(1)$ uniformly in k ($k \leq n$). It remains to show that

$$\sup_{1\leq k\leq n}\ \sup_{\|\phi\|\leq M}\left|\sum_{i=1}^k R_{i,n}(\phi)\right| = O_p(1). \tag{A.7}$$

By Pollard (1991),

$$|R_{i,n}(\phi)| \leq 2Mn^{-1/2}\|w_i\| I(|\varepsilon_i| \leq Mn^{-1/2}\|w_i\|). \tag{A.8}$$

Thus,

$$\left|\sum_{i=1}^{k} R_{i,n}(\phi)\right| \le 2Mn^{-1/2}\sum_{i=1}^{n}\|w_i\| I(|\varepsilon_i| \le Mn^{-1/2}\|w_i\|) \quad \forall k.$$

By the mean value theorem (ε_1 has a continuous density function at zero by assumption), (A.7) follows from

$$E\left\{n^{-1/2}\sum_{i=1}^{n}\|w_i\| I(|\varepsilon_i| \le Mn^{-1/2}\|w_i\|)\right\} \le 2M\left(\max_j |f(a_j)|\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2\right),$$

where $|a_i| \le M\|w_i\|n^{-1/2}$. By Assumption A4, $\max_j|a_j|$ converges to zero. Thus, $\max_j|f(a_j)| = O(1)$.

Proof of (ii). Lemma 6 in Babu (1989) implies that for some $C_0 > 0$, $\|\hat{\phi}_k\| \le C_0(k^{-1}\log k)^{1/2}$ with probability 1 for large k. Thus, for every $\epsilon > 0$, we can choose a $C > 0$ such that

$$P(\exists k > 1 \text{ such that } \|\hat{\phi}_k\| > C(k^{-1}\log k)^{1/2}) < \epsilon.$$

Denote $\eta_i(\phi) = |\varepsilon_i - w_i'\phi| - |\varepsilon_i|$ and $M_k = C(k^{-1}\log k)^{1/2}$. Then, for $A > 0$,

$$\begin{aligned} &P\left(\sup_{n\ge k>1}\left|\inf_{\phi}\sum_{i=1}^{k}\eta_i(\phi)\right| > 2A\log n\right) \\ &\quad\le P(\exists k \text{ such that } \|\hat{\phi}_k\| > M_k) + P\left(\sup_{n\ge k>1}\left|\inf_{\|\phi\|\le M_k}\sum_{i=1}^{k}\eta_i(\phi)\right| > 2A\log n\right) \\ &\quad\le \epsilon + P\left(\sup_{n\ge k>1}\sup_{\|\phi\|\le M_k}\left|\sum_{i=1}^{k}\eta_i(\phi)\right| > 2A\log n\right). \end{aligned} \tag{A.9}$$

The right-hand side of equation (A.9) is small if we can show that with probability 1

$$\limsup_{k\to\infty}\sup_{\|\phi\|\le M_k}\left|\sum_{i=1}^{k}\eta_i(\phi)\right|(\log k)^{-1} \le A \tag{A.10}$$

for some $A > 0$ (A to be determined later). This is because (A.10) implies that for any $\epsilon > 0$ there exists n_0 such that with probability not less that $1 - \epsilon$

$$\sup_{\|\phi|\le M_k}\left|\sum_{i=1}^{k}\eta_i(\phi)\right| < 2A\log k \le 2A\log n \tag{A.11}$$

for all k such that $n_0 \le k \le n$. For $k < n_0$, because $|\eta_i(\phi)| \le |w_i'\phi| \le \|w_i\|M_i$, we have $|\sum_{i=1}^{k}\eta_i(\phi)| \le \sum_{i=1}^{n_0}|\eta_i(\phi)| \le C(\sum_1^{n_0}\|w_i\|)n_0(\log n_0)^{1/2} < A\log n$ for large n. Thus, (A.11) holds for all $k \ge 1$, implying that (A.9) is small. To prove (A.10), we use a similar approach to that for the proof of Lemma 5 in Babu (1989). Divide the region $\|\phi\| \le M_k$ into $C_{p,q}k^{(p+q)/2}$ cells such that for g and h belonging to the same cell $\|g - h\| \le M_k k^{-1/2}$, where $C_{p,q}$ is a constant only depending on $p + q$. Notice that for a ϕ_r in the rth cell

$$\sup_{\|\phi\|\le M_k}\left|\sum_{i=1}^{k}\eta_i(\phi)\right| \le \sup_{r}\left|\sum_{i=1}^{k}\eta_i(\phi_r)\right| + \sup_{\|g-h\|\le M_k k^{-1/2}}\left|\sum_{i=1}^{k}\{\eta_i(g) - \eta_i(h)\}\right|. \tag{A.12}$$

Because

$$\left|\sum_{i=1}^{k}\{\eta_i(g)-\eta_i(h)\}/\log k\right| \le \sum_{i=1}^{k}\|w_i\|\,\|g-h\|/\log k \le C_1(\log k)^{-1/2}\to 0, \quad \textbf{(A.13)}$$

for g, h in the same cell ($C_1 \ge C\left(\frac{1}{k}\sum_{i=1}^{k}\|w_i\|\right)$ for all k), we need only consider the first term on the right-hand side of (A.12). Now $\sum_{i=1}^{k}\eta_i(\phi_r) = \sum_{i=1}^{k}\xi_i(\phi_r) + \sum_{i=1}^{k}E\eta_i(\phi_r)$, where $\xi_i(\phi) = \eta_i(\phi) - E\eta_i(\phi)$. But for $w_i'\phi$ near zero

$$E\eta_i(\phi) = \phi' w_i w_i' \phi f(0)(1+o(1)), \quad \textbf{(A.14)}$$

implying that, for $\|\phi_r\| \le M_k$,

$$\left|\sum_{i=1}^{k}E\eta_i(\phi_r)\right| \le 2f(0)\phi_r'\left(\sum_{i=1}^{k}w_iw_i'\right)\phi_r \le 2f(0)M\log k, \quad \textbf{(A.15)}$$

where M is a constant such that $\|\frac{1}{k}\sum_{i=1}^{k}w_iw_i'\| \le M$ for all k. Next, we shall bound the tail probability of $\sum_{i=1}^{k}\xi_i(\phi_r)$ by Lemma A.2. Notice that for each ϕ, $\xi_i(\phi)$ is a mean zero sequence with $|\xi_i(\phi)| \le 2\|w_i\|\,\|\phi\|$ and, thus, $\mathrm{Var}(\xi_i(\phi)) \le 4\|w_i\|^2\|\phi\|^2$. Because $\|\phi_r\| \le M_k = C(k^{-1}\log k)^{1/2}$, we have

$$|\xi_i(\phi_r)| \le 2C\max_{i\le k}\|w_i\|M_k \le 2C\max_{i\le k}\|w_i\|k^{-1/2}(\log k)^{1/2}, \quad \textbf{(A.16)}$$

$$\sum_{i=1}^{k}E\xi_i^2(\phi_r) \le 4M_k^2\sum_{i=1}^{k}\|w_i\|^2 \le 4C^2\left(k^{-1}\sum_{i=1}^{k}\|w_i\|^2\right)\log k \le M\log k, \quad \textbf{(A.17)}$$

for some $M<\infty$. Applying Lemma A.2 with Z_i equal to $\xi_i(\phi_r)$, d equal to the right-hand side of (A.16), V equal to $M\log k$, $s=1/2$, and a equal to $\lambda\log k$, we have

$$P\left(\left|\sum_{i=1}^{k}\xi_i(\phi_r)\right| > \lambda\log k\right) \le 2\exp\{-L\lambda^2\log k\} = 2k^{-L\lambda^2}, \quad \textbf{(A.18)}$$

where $L = 1/(4M)$, a constant not depending on k. (Note that $a \le 2V/(sd)$ is satisfied because $d\to 0$ by Assumption A4.) Thus,

$$P\left(\sup_{r\le C_{p,q}k^{(p+q)/2}}\left|\sum_{i=1}^{k}\xi_i(\phi_r)\right|\Big/\log k > \lambda\right) \le 2C_{p,q}k^{(p+q)/2}k^{-L\lambda^2}.$$

By choosing a large λ, the preceding is less than k^{-2}. The Borel-Cantelli Lemma leads to

$$\limsup_{k\to\infty}\sup_r\left|\sum_{i=1}^{k}\xi_i(\phi_r)\right|\Big/\log k \le \lambda \quad \textbf{(A.19)}$$

with probability one.

Proof of (iii). Because $\eta_i(\phi) = |\varepsilon_i - n^{-1/2}w_i'\phi| - |\varepsilon_i|$ is convex in ϕ, it is enough to consider $\|\phi\| = \log n$ (the notation $\eta_i(\phi)$ differs from the previous one in the extra factor $n^{-1/2}$). Let $\eta_i(\phi) = \xi_i(\phi) + E\eta_i(\phi)$. From (A.14) (replacing ϕ by $\phi/\sqrt{n}$),

$$E\sum_{i=1}^{k}\eta_i(\phi) = \phi'\left(\frac{1}{n}\sum_{i=1}^{k}w_iw_i'\right)\phi f(0)[1+o(1)] \ge \delta(\log n)^2L, \quad \forall k \ge n\delta, \quad \textbf{(A.20)}$$

where $L > 0$ by Assumptions A2 and A3. Next, we show

$$\sup_{n \ge k \ge n\delta} \sup_{\|\phi\| = \log n} \left| \sum_{i=1}^{k} \xi_i(\phi) \right| = (\log n)^{3/2} O_p(1). \tag{A.21}$$

The proof is similar to that of (ii). Only the outline is given here. Divide the region $\|\phi\| = \log n$ into $n^{(p+q)/2}$ cells with the diameter of each cell not exceeding $\log n/\sqrt{n}$. Then, incremental values within a cell are negligible (see the similar argument in (A.13)). For a point ϕ_r in the rth cell, we apply Lemma A.2 with $Z_i = \xi_i(\phi_r)$, $d = \max_{i \le n} \|w_i\| n^{-1/2} \log n$, $V = M(\log n)^2$, $s = 1/2$, and $a = \lambda(\log n)^{3/2}$ to deduce a similar inequality to (A.18). This leads easily to the inequality

$$\limsup_{n \to \infty} \sup_{1 \le k \le n} \sup_{r} \sum_{i=1}^{k} \xi_i(\phi_r)/(\log n)^{3/2} \le A,$$

which holds with probability one for some $A > 0$. Combining (A.20) and (A.21), we have

$$\inf_{n\delta \le k \le n} \inf_{\|\phi\| = \log n} \sum_{i=1}^{k} \eta_i(\phi) \ge (\log n)\{-|O_p(1)|(\log n)^{1/2} + \delta(\log n)L\}.$$

The right-hand side is larger than $D \log n$ for any $D > 0$ with high probability when n is large.

The proof of (iv) is quite similar and is thus omitted.

Proof of (v). Because of convexity, we assume $\|\phi\| = d_n$. Define $\eta_i(\phi)$ and $\xi_i(\phi)$ as in the proof of (iii). In our application, $d_n = \sqrt{n}\|\lambda_n\|$ and $h_n = \|\lambda_n\|^{-2}$. Thus, $\|\phi\|/\sqrt{n} = d_n/\sqrt{n}$ is either a fixed constant or converging to zero depending on the magnitude of shift λ_n. We first consider the case that both $d_n/\sqrt{n}$ and h_n are constant (not depending on n). Then, (v) is equivalent to (absorbing $n^{-1/2}$ into ϕ),

$$\inf_{n \ge k \ge A} \inf_{\|\phi\| = C} \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \ge D \tag{A.22}$$

with high probability, where $C > 0$. To prove the preceding, we show that the expected value of the sum is large and its deviation from its expected value is small. Note that the summands do not depend on n. Also, $M(\mu) = E(|\varepsilon_i - \mu| - |\varepsilon_i|)$ has a unique minimum at zero and $M(\mu)$ increases in $|\mu|$. These facts together with Assumptions A3 and A4 assure

$$E \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \ge k\eta, \quad \text{for all large } k$$

uniformly over $\|\phi\| = C$, for some $\eta > 0$. For $k \ge A$, $k\eta \ge A\eta > D$ if A is large. Thus, to prove (A.22), it is sufficient to show that for large A

$$Pr\left(\sup_{n \ge k \ge A} \sup_{\|\phi\| = C} \frac{1}{k} \left| \sum_{i=1}^{k} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) - E(|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \right| > \eta/2 \right) < \epsilon.$$

However, the preceding is implied by the uniform strong law of large numbers of Pollard (1990, Theorem 8.3, p. 41). Pollard (1990, Ch. 8) proved that the summands are manageable for their envelops $\{2\|w_i\|C\}$. Also, we can show $\sum_{i=1}^{\infty} \|w_i\|^2/i^2 < \infty$

(see the identity (A.31) below and the proof of (A.30)). Thus, the conditions of Theorem 8.3 of Pollard are satisfied.

The proof for $d_n/\sqrt{n} \to 0$ is more demanding. First we show that for some $L > 0$ and for all large enough k and large n with $k \le n$

$$E\sum_{i=1}^{k}(|\varepsilon_i - w_i\phi n^{-1/2}| - |\varepsilon_i|) \ge Lkd_n^2/n, \quad \text{for all } \|\phi\| = d_n. \qquad \textbf{(A.23)}$$

Write $\eta_i(\phi) = |\varepsilon_i - w_i\phi n^{-1/2}| - |\varepsilon_i|$. Then,

$$E\sum_{i=1}^{k}\eta_i(\phi) = E\sum_{i=1}^{k}\eta_i(\phi)I(\|w_i\| \le K) + E\sum_{i=1}^{k}\eta_i(\phi)I(\|w_i\| > K). \qquad \textbf{(A.24)}$$

Let $M(\mu) = E(|\varepsilon_i - \mu| - |\varepsilon_i|)$. There exists a $C > 0$ such that $M(\mu) \le C\mu^2$ for all $\mu \in R$ (because $M(\mu) = f(0)\mu^2 + o(\mu^2)$ for small μ, $|(|\varepsilon_i - \mu| - |\varepsilon_i|)| \le |\mu|$, and $|\mu| < |\mu|^2$ for $|\mu| > 1$). Thus, for the ϵ and K in Assumption A4,

$$\left|\sum_{i=1}^{k}\eta_i(\phi)I(\|w_i\| > K)\right| \le C\frac{1}{n}\phi'\left(\sum_{i=1}^{k}w_iw_i'I(\|w_i\| > K)\right)\phi \le \epsilon Ckd_n^2/n. \qquad \textbf{(A.25)}$$

Next, for $\|w_i\| \le K$ ($\forall i$), we have $w_i'\phi n^{-1/2} = o(1)$ uniformly in $i \le n$ and uniformly over $\|\phi\| = d_n$. Thus, by Taylor expansion at zero,

$$\begin{aligned} E\sum_{i=1}^{k}\eta_i(\phi)I(\|w_i\| < K) &= \frac{1}{n}\phi'\left(\sum_{i=1}^{k}w_iw_i'I(\|w_i\| < K)\right)\phi f(0)(1 + o(1)) \\ &\ge k\frac{1}{n}\|\phi\|^2 f(0)A/2 = k(d_n^2/n)f(0)A/2 \qquad \textbf{(A.26)} \end{aligned}$$

for some $A > 0$ (taking A as one-half of the smallest eigenvalue of $Q = \text{plim}\,\frac{1}{k}\sum_{i=1}^{k}w_iw_i'$ is enough, because for large k and large K, Assumptions A1 and A4 assure $\frac{1}{k}\sum_{i=1}^{k} \times w_iw_i'I(\|w_i\| < K) > Q/2$). It follows from (A.24)-(A.26) that

$$E\left(\sum_{i=1}^{k}|\varepsilon_i - w_i\phi n^{-1/2}| - |\varepsilon_i|\right) \ge kd_n^2/nf(0)A/2 - \epsilon Ckd_n^2/n,$$

so (A.23) is obtained by letting $L = f(0)A/2 - \epsilon C$, which is positive for a small ϵ.

Now for $k \ge Ah_n$, we have $Lkd_n^2/n \ge LAh_nd_n^2/n \ge LAh/2 > D$ for large A, where h is the limit of $h_nd_n^2/n$. Thus, the expected value is large if A is large. To prove (v), it suffices to show that the quantity of interest is dominated by its expected value (or the deviation from the mean is small). More precisely, we shall show that for every $\epsilon > 0$, when A is large

$$P\left(\sup_{n\ge k\ge Ah_n}\ \sup_{\|\phi\|=d_n}\frac{1}{k}\left|\sum_{i=1}^{k}\xi_i(\phi)\right| > \frac{1}{2}(d_n^2/n)L\right) < \epsilon, \qquad \textbf{(A.27)}$$

where $\xi_i(\phi) = \eta_i(\phi) - E\eta_i(\phi)$. By (A.5) and (A.6), we have

$$\frac{1}{k}\sum_{i=1}^{k}\xi_i(\phi) = \frac{1}{k}n^{-1/2}\phi'\sum_{i=1}^{k}w_iD_i + \frac{1}{k}\sum_{i=1}^{k}[R_{i,n}(\phi) - ER_{i,n}(\phi)]. \qquad \textbf{(A.28)}$$

We first prove that for large A

$$P\left(\sup_{n\ge k\ge Ah_n}\frac{1}{k}\left\|\sum_{i=1}^{k} w_iD_i\right\| > \frac{L}{4}\,d_n/\sqrt{n}\right) < \epsilon. \tag{A.29}$$

Let n_1 be the smallest integer no less than Ah_n. By the Birnbaum and Marshall (1961) inequality, the left-hand side of (A.29) is bounded by

$$16(Ld_n/\sqrt{n})^{-2}E\left\|\frac{1}{n_1}\sum_{i=1}^{n_1} w_iD_i\right\|^2 + 16(Ld_n/\sqrt{n})^{-2}\sum_{k=n_1+1}^{n}\|w_i\|^2/k^2. \tag{A.30}$$

We now show that $\sum_{k=n_1+1}^{n}\|w_i\|^2/k^2 \le 5M/n_1$, where $M \ge \frac{1}{k}\sum_{i=1}^{k}\|w_i\|^2$ for all k. Apply the identity

$$\sum_{k=m+1}^{n}(a_k - a_{k-1})b_k = a_nb_n - a_mb_{m+1} + \sum_{k=m+1}^{n-1}a_k(b_k - b_{k+1}) \tag{A.31}$$

to $a_k = \sum_{i=1}^{k}\|w_i\|^2$ and $b_k = 1/k^2$ and notice that $b_k - b_{k+1} \le 3/k^3$; we have

$$\sum_{k=m+1}^{n}\|w_i\|^2/k^2 \le \frac{1}{n}M + \frac{1}{m}M + 3\sum_{k=m+1}^{n-1}\left(\frac{1}{k}\sum_{i=1}^{k}\|w_i\|^2\right)\Big/k^2$$

$$\le 2\,\frac{1}{m}M + 3M\sum_{k=m+1}^{\infty}1/k^2 \le 5M/m.$$

Thus, (A.30) is bounded by

$$16L^{-2}(d_n^2/n)^{-1}\,\frac{1}{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\|w_i\|^2\right) + 16L^{-2}(d_n^2/n)^{-1}(5M/n_1) \le C(n_1d_n^2/n)^{-1},$$

where $C = 96ML^{-1}$. By the definition of n_1, $(n_1d_n^2/n)^{-1} \le (Ah_nd_n^2/n)^{-1}$, which is small if A is large because $h_nd_n^2/n$ has a positive limit. This proves (A.29).

Next, consider the second term of the right-hand side of (A.28). Note that $R_{i,n}(\phi) = R_i(\phi/\sqrt{n})$, where $R_i(\psi) = |\varepsilon_i - w_i'\psi| - |\varepsilon_i| - \psi'w_iD_i$ not depending on n (see (A.5)). We shall prove

$$P\left(\sup_{n\ge k\ge Ah_n}\ \sup_{\|\psi\|=d_n/\sqrt{n}}\left|\frac{1}{k}\sum_{i=1}^{k}[R_i(\psi) - ER_i(\psi)]\right| > \frac{L}{4}\,d_n^2/n\right) < \epsilon. \tag{A.32}$$

By Pollard (1991, p. 63),

$$\frac{1}{k}\sum_{i=1}^{k}[R_i(\psi) - ER_i(\psi)] = o_p(\|\psi\|/\sqrt{k})$$

uniformly in k and uniformly over ψ in shrinking neighborhoods of the origin. Note that $k \to \infty$ implies $n \to \infty$. Thus, $A_n = \{\psi;\ \|\psi\| \le d_n/\sqrt{n}\}$ forms a sequence of shrinking neighborhoods of the origin as k grows. Also note that $\{\psi;\ \|\psi\| = d_n/\sqrt{n}\}$ is a subset (boundary) of A_n. Thus, Pollard's result implies

$$\sup_{\|\psi\|=d_n/\sqrt{n}}\frac{1}{k}\sum_{i=1}^{k}[R_i(\psi) - ER_i(\psi)] = o_p(\|\psi\|/\sqrt{k}) = o_p(d_n/\sqrt{nk}),$$

uniformly in $k \le n$.

Now for all $k \geq Ah_n$, $d_n/\sqrt{nk} \leq d_n/\sqrt{nAh_n}$, which is less than $Ld_n^2/(4n)$ if A is large because of the existence of a positive limit for $h_n d_n^2/n$. This proves (A.32). Combining (A.28), (A.29), and (A.32), we obtain (A.27).

Proof of (vi). For $\|\phi\| \leq d_n$, we have

$$\|n^{-1/2}\phi\| \leq (Ah_n)^{-1/2}\{(Ah_n)^{1/2}n^{-1/2}d_n\} \leq (Ah_n)^{-1/2}M' \leq n_1^{-1/2}M''$$

for some M' and $M'' > 0$ because of the existence of a limit for $h_n d_n^2/n$. Thus, the set $\{n^{-1/2}\phi;\ \|\phi\| \leq d_n\}$ is contained in the set $\{n_1^{-1/2}\psi;\ \|\psi\| \leq M''\}$. Consequently,

$$\sup_{1\leq k\leq Ah_n}\left|\inf_{\|\phi\|\leq d_n}\sum_{i=1}^{k}\eta_i(\phi)\right| \leq \sup_{1\leq k\leq n_1}\ \sup_{\|\psi\|\leq M''}\left|\sum_{i=1}^{k}(|\varepsilon_i - n_1^{-1/2}w_i'\psi| - |\varepsilon_i|)\right|.$$

It follows from (A.4) that the right-hand side is $O_p(1)$. ■

Proof of Theorem 3. We begin with the following lemma.

LEMMA A.3. *Let a_n be a sequence of integers such that $a_n/n^{1-\delta} \to 0$ for some $\delta > 0$. Then, under the assumptions of Theorem* 1,

$$\sum_{i=1}^{k}(|\varepsilon_i - w_i'\phi n^{-1/2}| - |\varepsilon_i|) = o_p(1) \qquad \textbf{(A.33)}$$

uniformly in $k \leq a_n$ and $\|\phi\| \leq M$.

Proof. Divide the region $\|\phi\| \leq M$ into $O(n^{(p+q)/2})$ cells such that the diameter of each cell is less than $Mn^{-1/2}$. As before, incremental values within a cell are negligible. Consider a point ϕ_r in the rth cell. Let $Z_i = \xi_i(\phi_r)$, $d = \max_{i\leq a_n}\|w_i\|n^{-1/2}$, $V = Ma_n/n$, $a = \epsilon$, and $s = \sqrt{a_n/n}$; then, Lemma A.2 implies

$$P\left(\left|\sum_{i=1}^{k}\xi_i(\phi_r)\right| > \epsilon\right) \leq 2\exp(-L\sqrt{n/a_n}) \leq 2\exp(-Ln^b)$$

for some $b > 0$ because of the assumption on a_n. The preceding bound does not depend on k and r. The lemma follows from $a_n n^{(p+q)/2}\exp(-Ln^b) \to 0$.

We shall prove the weak convergence of $V_n(\theta, v)$ on the compact set $\{(\theta, v);\ \|\theta\| \leq M, |v| \leq M\}$. Again, we will only consider the case for $v \leq 0$, i.e., $k(v) \leq k_0$, because the case of $v > 0$ is similar. For $v \leq 0$, $V_n(\theta, v)$ is given by (A.2). Rewrite $U(1,k(v),\phi_1) = U(1,k_0,\phi_1) - U(k(v)+1,k_0,\phi_1)$. The second term is negligible. Because $k_0 - k(v) \leq M\|\lambda_n\|^{-2}$, applying Lemma A.3 with $a_n = O(\|\lambda_n\|^{-2})$, we have $U(k(v)+1,k_0,\phi_1) = o_p(1)$ uniformly in $0 \geq v \geq -M$ and $\|\phi_1\| \leq M$. Now $U(1,k_0,\phi_1) + U(k_0+1,n,\phi_2)$ can be written as

$$\sum_{i=1}^{n}(|\varepsilon_i - [x_i'\beta + z_i'I(i \leq k_0)\delta_1 + z_i'I(i > k_0)\delta_2]n^{-1/2}| - |\varepsilon_i|), \qquad \textbf{(A.34)}$$

where $I(\cdot)$ is the indicator function. Standard results of LAD estimation (e.g., Pollard, 1991) imply that the preceding is

$$n^{-1/2}\theta'\sum_{i=1}^{n}X_iD_i + f(0)\theta'\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)\theta + o_p(1), \qquad \textbf{(A.35)}$$

where $\theta = ((\beta', \delta_1', \delta_2')')$, $X_i = (x_i', z_i', 0)'$ for $i \le k_0$ and $X_i = (x_i', 0, z_i')'$ for $i > k_0$, and $D_i = \operatorname{sign}(\varepsilon_i)$. The $o_p(1)$ is uniform on any given compact set of θ. Thus, the preceding converges weakly on compact set to

$$\theta' Q^{1/2} Z + f(0)\theta' Q\theta, \tag{A.36}$$

where Z is a vector of independent standard normal random variables and Q is defined previously.

Next, adding and subtracting terms, $U(k(v) + 1, k_0, \sqrt{n}\lambda_n^* + \phi_2)$ can be written as

$$\sum_{i=k(v)+1}^{k_0} (|\varepsilon_i - z_i'\lambda_n| - |\varepsilon_i|) \tag{A.37}$$

$$+ \sum_{i=k(v)+1}^{k_0} (|\varepsilon_i - x_i'\beta n^{-1/2} - z_i'\delta_2 n^{-1/2} - z_i'\lambda_n| - |\varepsilon_i - z_i'\lambda_n|). \tag{A.38}$$

Expression (A.38) is $o_p(1)$, which follows from Lemma A.3 by simply renaming $\varepsilon_i - z_i'\lambda_n$ as ε_i. Thus, the limiting process of $V_n(\theta, v)$ is equal to (A.36) plus the limit of (A.37). The latter does not depend on θ. The limiting process is minimized with respect to θ at $Q^{-1/2}Z/(2f(0))$. By the continuous mapping theorem for the argmax functional (e.g., J. Kim and Pollard, 1990), we obtain part (i) of Theorem 3.

We point out that for $v > 0$ the corresponding term to (A.37) is given by

$$\sum_{i=k_0+1}^{k(v)} (|\varepsilon_i + z_i'\lambda_n| - |\varepsilon_i|). \tag{A.39}$$

Thus far, the assumption that λ_n converges to zero has not been referenced. The argument applies to a shift with a fixed magnitude and to shrinking shifts satisfying Assumption A6. To characterize the limiting distribution for the shift point estimator, we shall consider the two cases separately, as they possess different limiting distributions. The first case is that $\lambda_n \equiv \lambda_0$, not varying with sample size n. In this case, $k(v) = k_0 - vc_n$ does not change with n for a fixed v. Therefore, we do not need to use v as a parameterization; instead, we consider directly the convergence of $S(\theta_0 + n^{-1/2}\theta, k) - S(\theta_0, k_0)$. Let $W^*(k) = \sum_{i=k+1}^{k_0} |\varepsilon_i - z_i'\lambda_0| - |\varepsilon_i|$ for $k \le k_0$ (take $W^*(k_0)$ as 0) and $W^*(k) = \sum_{i=k_0+1}^{k} |\varepsilon_i + z_i'\lambda_0| - |\varepsilon_i|$ for $k > k_0$. Then, $W^*(k)$ has the same distribution as $W^{\#}(k - k_0)$, where $W^{\#}(\cdot)$ is defined previously. Let $k - k_0 = m$. Combining (A.35), (A.37), and (A.39), we have

$$S(\theta_0 + n^{-1/2}\theta, m + k_0) - S(\theta_0, k_0) \Rightarrow \theta' Q^{1/2} Z + \theta' Q\theta f(0) + W^{\#}(m) \tag{A.40}$$

on the set $\|\theta\| \le M$ and $|m| \le M$ for an arbitrary given M. Assuming $|\varepsilon_i \pm z_i\lambda_0| - |\varepsilon_i|$ has a continuous distribution, $\operatorname{argmin}_m W^{\#}(m)$ is uniquely defined. This implies $\hat{k} - k_0 \xrightarrow{d} \operatorname{argmin}_m W^{\#}(m)$ by the continuous mapping theorem.

Note that Z depends on (ε_i, z_i) appearing in (A.37) only through $U(k(v) + 1, k_0, \phi_2)$, which is $o_p(1)$, as shown earlier. The case of $v > 0$ is analogous. Thus, Z is independent of $W^*(k)$ for $|k - k_0| < M$ for any given $M > 0$. This implies that the limiting distributions of $\hat{\theta}$ and $\hat{k}$ are independent.

Proof of (iii). If $\lambda_n \to 0$, then (A.37) can be written as

$$\lambda_n' \sum_{i=k_0+vc_n}^{k_0} z_i D_i + \lambda_n' \left(\sum_{i=k_0+vc_n}^{k_0} z_i z_i' \right) \lambda_n f(0) + o_p(1), \tag{A.41}$$

which is a consequence of the standard result for the equivalence between (A.34) and (A.35) with a rescaling ($n^{-1/2}$ is replaced by λ_n because $k(v) = k_0 + vO(\|\lambda_n\|^{-2})$). Because $c_n = O(\|\lambda_n\|^{-2}) \to \infty$, we have

$$\frac{1}{c_n}\sum_{i=k_0+vc_n}^{k_0} z_i z_i' \to |v| Q_{zz}. \qquad \textbf{(A.42)}$$

If we choose c_n specifically such that $c_n = (\lambda_n' Q_{zz} \lambda_n)^{-1}$, then the second term of (A.41) converges to $|v|f(0)$. The first term of (A.41) converges weakly to a Brownian motion process, denoted by $W_1(-v)$, by the invariance principle for independent random variables (see, e.g., Billingsley, 1968). (Again, the scaling factor is λ_n instead of $n^{-1/2}$.) Thus, (A.41) converge weakly to $W_1(-v) + |v|f(0)$. The counterpart of (A.41) in the case of $v > 0$ (the sum is from $k_0 + 1$ to $k_0 + vc_n$) has a limit $W_2(v) + |v|f(0)$, where $W_2(v)$ is another Brownian motion process on the positive half line. The two processes are independent because they involve nonoverlapping disturbances. Thus, a two-sided Brownian motion process $W(v)$ can be defined based on the two processes so that $V_n(\theta, v)$ converges weakly to

$$\theta' Q^{1/2} Z + \theta' Q \theta f(0) + W(v) + f(0)|v|.$$

Part (iii) follows from the continuous mapping theorem.

Proof of (iv). Trending regressors $z_i = g(i/n)$ satisfy all assumptions required for consistency. It is thus enough to consider the limiting process of (A.41). By adding and subtracting terms, we can rewrite the second term of (A.41) (ignore $f(0)$) as

$$(k_0 - k)\lambda_n' g(\tau_0) g(\tau_0)' \lambda_n \qquad \textbf{(A.43)}$$

$$+ \lambda_n' \sum_{t=k+1}^{k_0} [g(i/n) - g(\tau_0)][g(i/n) - g(\tau_0)]' \lambda_n \qquad \textbf{(A.44)}$$

$$+ 2\lambda_n' \sum_{t=k+1}^{k_0} [g(i/n) - g(\tau_0)] g(\tau_0)' \lambda_n, \qquad \textbf{(A.45)}$$

where $k = k(v)$. Let $c_n = (\lambda_n' g(\tau_0) g(\tau_0)' \lambda_n)^{-1}$. Expression (A.43) converges to $-v = |v|$ uniformly in $v \in [-M, 0]$ because $k_0 - k = -[vc_n^{-1}]$. It is easy to show that (A.44) and (A.45) are both $o_p(1)$ uniformly in $v \in [-M, 0]$. For example, (A.44) is bounded by

$$\sup_x \left\| \frac{dg(x)}{dx} \right\|^2 \lambda_n' \lambda_n \sum_{i=k+1}^{k_0} (i - k_0)^2/n^2 \le B_1 \|\lambda_n\|^2 (k_0 - k)^3/n^2$$

$$\le B_2 (n^2 \|\lambda_n\|^4)^{-1} \to 0, \qquad \textbf{(A.46)}$$

for some constants B_1 and B_2. Next, the first term of (A.41) is

$$\lambda_n' \sum_{i=k+1}^{k_0} g(i/n) D_i = \lambda_n' g(k_0/n) \sum_{i=k+1}^{k_0} D_i + \lambda_n' \sum_{i=k+1}^{k_0} [g(i/n) - g(k_0/n)] D_i. \qquad \textbf{(A.47)}$$

The second term of (A.47) is uniformly negligible because its variance is equal to (A.44), which is $o_p(1)$ in view of (A.46). Thus, the limiting distribution of (A.47) is determined by $\lambda_n' g(\tau_0) \sum_{i=k+1}^{k_0} D_i$. Because $c_n = (\lambda_n' g(\tau_0) g(\tau_0)' \lambda_n)^{-1}$, by the invariance principle for independent variables

$$\lambda_n' g(\tau_0) \sum_{i=k(v)+1}^{k_0} D_i \Rightarrow W_1(-v). \tag{A.48}$$

Thus, (A.41) converges weakly to $W_1(-v) + f(0)|v|$. Part (iv) is obtained by considering the case of $v > 0$ and then using the continuous mapping theorem. ■

Proof of Theorem 4. The rate of convergence is not affected by the assumption of dependence under strong mixing with exponential mixing coefficients. This is because Lemma A.1 still holds. In proving Lemma A.1, we use some results of Babu (1989), which are still valid under the dependent assumption. The exponential inequality of Lemma A.2 is replaced by a similar inequality suitable for mixing process (see Babu, 1989, Lemma 7). We can also use an exponential inequality of Carbon (e.g., Gyorfi, Hardle, Sarda, and Vieu, 1990). The Birnbaum and Marshall inequality is replaced by an inequality due to Serfling (1970, Theorem 5.1). All arguments for consistency carry over to the dependent case. Details are omitted.[5] Thus, we focus on the asymptotic distribution.

Proof of (i). The equivalence of (A.34) and (A.35) is still a standard result under the mixing assumption (e.g., Babu, 1989). The limit of (A.35) is $\theta' U^{1/2} Z + f(0)\theta' Q \theta$. This quadratic form is minimized at $[2f(0)]^{-1} Q^{-1} U^{1/2} Z$, which is $N(0, [2f(0)]^{-2} Q^{-1} U Q^{-1})$.

Proof of (ii). Recall that (A.37) determines the limiting distribution of the shift point estimator. Combined with (A.39), we still have a two-sided random walk as a limit. It is noted that $W_1^{\#}$ and $W_2^{\#}$ are no longer independent. The correlation between $W_1^{\#}(k)$ and $W_2^{\#}(j)$ does not disappear even when $|k - j|$ becomes large.

Proof of (iii). The goal is to determine the limit of (A.41) under strong mixing. For $c_n = (\lambda_n' Q_{zz} \lambda_n)^{-1}$, the second term of (A.41) has the same limit. Suppose

$$\lim_{n \to \infty} \frac{\lambda_n' U_{zz} \lambda_n}{\lambda_n' Q_{zz} \lambda_n} = \pi^2 > 0.$$

Then, functional central limit theorems for dependent variables (e.g., Wooldridge and White, 1988) with a rescaling imply that

$$\lambda_n' \sum_{i=k_0+vc_n}^{k_0} z_i D_i \Rightarrow \pi W_1(-v).$$

For the case of $v > 0$, the corresponding limit is $\pi W_2(v)$, a separate Brownian motion process. The processes W_i $(i = 1,2)$ are independent. To see this, assume for simplicity that z_i are bounded such that $\|z_i\| \leq M$. Then, by an inequality of Ibragimov and Linnik (1971),

$$E\left[\lambda_n' \left(\sum_{i=k_0+vc_n}^{k_0} z_i D_i \right) \left(\sum_{i=k_0+1}^{k_0+vc_n} z_i D_i \right) \lambda_n \right] \leq 4M^2 \lambda_n' \lambda_n \left(\sum j\alpha_j \right),$$

which converges to zero as n increases to infinity, implying the independence of W_1 and W_2 (where the α_j are the mixing coefficients). Thus, we can define a two-sided Brownian motion process as before, so the limit of (A.41) under dependence is

$\pi W(v)+f(0)|v|$. This implies that $n\lambda_n' Q_{zz}\lambda_n \pi^{-2}(\hat{\tau}-\tau_0) \xrightarrow{d} [2f(0)]^{-2} \operatorname{argmax}_v\{W(v)-|v|/2\}$, proving part (iii) in view of the definition of π^2.

Proof of (iv). We only need to revise the limit of (A.48) under dependence. Because

$$\lim_h \frac{1}{h}\sum_1^h \sum_1^h E(D_i D_j) \to \phi^2,$$

where ϕ^2 is defined previously, by the invariance principle for dependent variables, we have

$$\lambda_n' g(\tau_0) \sum_{k(v)+1}^{k_0} D_i \Rightarrow \phi W_1(-v).$$

Similar to the argument of (iii), this implies $n\lambda_n' g(\tau_0)g(\tau_0)'\lambda_n \phi^{-2}$ is the scaling factor for $(\hat{\tau}-\tau_0)$, establishing (iv). ■