Journal of Statistical Planning and Inference 74 (1998) 103–134

journal of statistical planning and inference

Estimation of multiple-regime regressions with least absolutes deviation

Jushan Bai[1]

Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 19 June 1997; received in revised form 4 February 1998; accepted 4 February 1998

Abstract

This paper considers least absolute deviations estimation of a regression model with multiple change points occurring at unknown times. Some asymptotic results, including rates of convergence and asymptotic distributions, for the estimated change points and the estimated regression coefficient are derived. Results are obtained without assuming that each regime spans a positive fraction of the sample size. In addition, the number of change points is allowed to grow as the sample size increases. Estimation of the number of change points is also considered. A feasible computational algorithm is developed. An application is also given, along with some Monte Carlo simulations. © 1998 Elsevier Science B.V. All rights reserved.

AMS classification: Primary 62F12; secondary 60J05, 62F35

Keywords: Multiple change points; Multiple-regime regressions; Least absolute deviation; Asymptotic distribution

1. Introduction

This paper considers the estimation of a multiple-regime regression in which the regime switch points are unknown. A common method of estimation is Gaussian maximum likelihood or the least squares method (e.g., Quandt, 1958). In this paper we consider the method of least absolute deviations (LAD). As is well known, for heavy tailed distributions, LAD is more efficient than least squares (LS). In the change point context, efficiency gains can be realized not only for the estimated regression parameters, but also for the estimated change points. The purpose of this paper is to study the consistency, rate of convergence, and asymptotic distributions for the estimated change points. We also study estimating the number of change points based on a Bayesian

[1] The author is grateful to two anonymous referees and an Associate Editor (Lajos Horváth) for very useful comments. This research is supported in part by a grant from the NSF No. SBR-9414083.

PII: S0378-3758(98)00082-2

information type criterion (BIC). Results are obtained allowing the number of change points to increase with the sample size.

Estimating multiple change points typically require enormous computation. As a result, computational feasibility becomes an important concern in selecting estimation methods. Two additional factors reinforce this concern. First, multiple change points typically occur in large samples. Second, even if there is only one change point, multiple ones are allowed when BIC criterion is used in estimating the true number. While possessing robust properties, LAD is computationally feasible, since optimization can be carried out via linear programming. In our Monte Carlo simulations, the BIC criterion is calculated up to 10 potential change points, and optimal solution is achieved quickly. In this regard, LAD has certain advantages over other robust procedures.

The LAD method has not been analyzed in the literature for estimating multiple change-points models. A related work is Bai (1995), who studies the method for a single change point. A different framework is needed for more than one change. In the case of a single change point, each of the two regimes has one fixed and known boundary; the first regime has its lower boundary known (i.e., first observation) and the second regime has its upper boundary known (i.e., the last observation). For multiple changes, each middle regime has boundaries completely unknown. The analysis must take into account the possibility that a hypothesized regime may not have overlapping observations with the true regime. In general, the objective function (sum of the absolute deviations) is a stochastic process indexed by a vector of integers [see Eq. (2) below]. This vector of integers must be allowed to take all possible combinations. Consequently, the analysis of multiple change points requires a different framework from that of a single change point. The purpose of this paper is to establish the underlying theory for the LAD method in the context of multiple change points. Furthermore, unlike the existing literature, we abandon the assumption that each regime spans a positive fraction of the total sample. In addition, we allow the number of change points to be unbounded. This setting needs a different argument from that of a bounded number of change points, a further departure from the existing framework.

There is a large body of literature on the change point problem, see the survey papers of Shaban (1980), Zacks (1983), and Krishnaiah and Miao (1988). The inference on a single change point has received the most attention, e.g. Picard (1985), Bhattacharya (1987), Kim and Siegmund (1989), Dumbgen (1991), Brodsky and Darkhovsky (1993), Gombay and Horváth (1994), Horváth (1995), Horváth et al. (1997), and Hušková (1996a). A procedure based on M-esitmation is proposed by Hušková (1996b) for the case of no covariates. For multiple changes, Yin (1988) proposes a moving-window estimation of change points occurring in a nonparametric function of time. Yao (1988) proposes the Schwarz criterion for estimating the number of change points in a sequence of normal means. Yao and Au (1989), and Huang and Chang (1993) consider the least squares estimation of change points in a sequence of random variables without covariates. Bai and Perron (1998) study the problem of estimating and testing multiple change points in regression models.

All above studies impose the restriction that each regime occupies at least a positive fraction of the total sample. That is, the length of each regime is $O(n)$, where n is the sample size. In this paper, we relax this assumption. We also allow the number of change points to grow as the sample size increases. Meanwhile, we consider multiple regression models, as well as a different estimation technique, namely LAD.

The rest of this paper is organized as follows. Section 2 gives the assumptions and main results. Rates of convergence and asymptotic distributions are derived. In Section 3, the issue of determining the number of change points is considered. Section 4 provides some numerical results, including computational issues, Monte Carlo simulations and an application with real data. Section 5 derives some preliminary results used for the main theorems and Section 6 provides the proofs.

2. Notation, assumptions and main results

Consider the following $(m+1)$-regime regression model:

$$\begin{aligned} y_i &= x_i'\beta_1 + \varepsilon_i, \qquad & i &= 1,2,\ldots,n_1,\\ y_i &= x_i'\beta_2 + \varepsilon_i, & i &= n_1+1,\ldots,n_2,\\ &\vdots & &\vdots\\ y_i &= x_i'\beta_{m+1} + \varepsilon_i, & i &= n_m+1,\ldots,n, \end{aligned} \tag{1}$$

where y_i is the dependent variable, x_i $(p\times 1)$ is a vector of regressors, and ε_i is a disturbance. The β_j $(p\times 1)$ are unknown parameters. This $(m+1)$-regime regression has m change points, $n_1,\ldots,n_m$, which are also unknown.

Let $\beta^0 = (\beta_1^0,\ldots,\beta_{m+1}^0)$ denote the vector of true regression parameters and let $(n_1^0,\ldots,n_m^0)$ denote the vector of true change points. Let $\mathscr{P} = (n_1,\ldots,n_m)$ denote a partition of the integers $1,\ldots,n-1$, such that $n_1 < \cdots < n_m$. Let $\hat\beta(\mathscr{P}) = (\hat\beta_1(\mathscr{P}),\ldots,\hat\beta_{m+1}(\mathscr{P}))$ denote the LAD estimator of β^0 for a given partition $\mathscr{P}$. Namely,

$$\hat\beta(n_1,\ldots,n_m) = \arg\min_{\beta} \sum_{j=1}^{m+1} \sum_{i=n_{j-1}+1}^{n_j} |y_i - x_i'\beta_j|,$$

where $n_0 = 0$ and $n_{m+1} = n$. Or equivalently, $\hat\beta_j(\mathscr{P})$ minimizes $\sum_{i=n_{j-1}+1}^{n_j} |y_i - x_i'\beta_j|$ $(j = 1,\ldots,m+1)$. Denote by $S_n(n_1,\ldots,n_m)$, the resulting sum of absolute values of residuals,

$$S_n(n_1,\ldots,n_m) = \sum_{j=1}^{m+1} \sum_{i=n_{j-1}+1}^{n_j} |y_i - x_i'\hat\beta_j(\mathscr{P})| = \sum_{j=1}^{m+1} \left(\inf_{\phi} \sum_{i=n_{j-1}+1}^{n_j} |y_i - x_i'\phi| \right). \tag{2}$$

The estimated change points, $(\hat n_1,\ldots,\hat n_m)$, are defined as a set of integers $n_1,\ldots,n_m$, which minimizes $S_n(n_1,\ldots,n_m)$. Finally, the estimators of regression parameters are defined as

$$\hat\beta = (\hat\beta_1,\ldots,\hat\beta_{m+1}) = \hat\beta(\hat n_1,\ldots,\hat n_m).$$

We shall study the asymptotic behavior of $(\hat n_1,\ldots,\hat n_m)$ and $\hat\beta$.

In what follows, we shall use $|y|$ to denote the Euclidean norm of y, i.e., $|y| = (\sum_{i=1}^{p} y_i^2)^{1/2}$ for $y \in R^p$. All limits are taken as n converges to infinity unless stated otherwise. We now state the assumptions:

A1. For each j, the length of regime j satisfies $n_j^0 - n_{j-1}^0 \geqslant c_1 n^{3/4}$ for some $c_1 > 0$. The number of change points satisfies $m = m(n) < c_2 n^{(1/4)-b}$ for some $c_2, b > 0$.

A2. The parameter vector β^0 is an interior point of a bounded set of $\mathbb{R}^{p(m+1)}$. In addition, $\min_{1 \leqslant j \leqslant m(n)+1} |\beta_{j+1}^0 - \beta_j^0| \geqslant c > 0$, where c does not depend on n.

A3. The regressors x_i are uniformly bounded, i.e., there exists $K < \infty$ such that $|x_i| \leqslant K$ for all i.

A4. The matrices $\frac{1}{k}\sum_{n_s^0+1}^{n_s^0+k} x_i x_i'$ and $(1/k)\sum_{n_s^0-k}^{n_s^0} x_i x_i'$ $(s = 0, \ldots, m+1)$ converge in probability to some nonrandom positive-definite matrices (not necessarily the same) as k increases.

A5. The disturbances ε_i are i.i.d. random variables with a zero median and a positive continuous density, f, at the neighborhood of zero. Moreover, ε_i is independent of x_k for all i and all k.

The assumptions on the number of change points and the regime length are not the weakest possible. They can be improved upon. For bounded m, the requirement of $n_{j+1}^0 - n_j^0 \geqslant c_1 n^{3/4}$ can be weakened to $n_{j+1}^0 - n_j^0 > c_1 n^{(1/2)+\delta}$ for some $\delta \in (0, 1/2)$. The assumption of a bounded parameter set in A2 is restrictive, although it may not be of any practical significance. Under a slightly stronger condition on the disturbances, namely the existence of a $1+\delta$ moment, A2 can be dispensed with so that the parameter set can be $\mathbb{R}^{p(m+1)}$. The uniform boundedness of regressors in A3 can also be dispensed with. A3 can be replaced by the following less restrictive assumption used by Pollard (1990) (p. 58) adapted to our case: for each $\varepsilon > 0$, there exists $K > 0$ such that

$$\frac{1}{k}\sum_{i=n_s^0+1}^{n_s^0+k} |x_i|^2 I(|x_i| > K) < \varepsilon \quad \text{and} \quad \frac{1}{k}\sum_{n_s^0-k}^{n_s^0} |x_i|^2 I(|x_i| > K) < \varepsilon$$

for all large k $(s = 0, \ldots, m+1)$, where $I(\cdot)$ is the indicator function, see Bai (1995). However, using these less stringent assumptions rather than A2 and A3 makes the argument much more complex. We thus retain A2 and A3 in this paper. Assumption A4 is used for bounded m. For unbounded m, we will require a stronger assumption (A6 below), under which A4 is automatically satisfied.

Remark 1. Assumption A4 does not cover the case of trending regressors. For example, let $h(t) = (1, t, \ldots, t^p)'$ for $t \in [0, 1]$ and $x_i = h(i/n)$. Then, unless k grows linearly in n, the matrix $(1/k)\ \sum_{n_s^0+1}^{n_s^0+k} x_i x_i$ converges to $h(\tau_s^0)h(\tau_s^0)'$, as n and k converge to infinity with $k = o(n)$, where $\tau_s^0 = \lim(n_s^0/n)$. The matrix $h(\tau_s^0)h(\tau_s^0)'$ has a rank of 1. Therefore, A4 rules out trending regressors. However, the regressor $x_i = h(i/n)$ has the following property. For every $\varepsilon > 0$ and for $k = [n\varepsilon]$,

$$\frac{1}{[n\varepsilon]}\sum_{n_s^0+1}^{n_s^0+[n\varepsilon]} x_i x_i' \to \frac{1}{\varepsilon}\int_{\tau_s^0}^{\tau_s^0+\varepsilon} h(t)h(t)'\,\mathrm{d}t > 0$$

and

$$\frac{1}{[n\varepsilon]}\sum_{n_s^0-[n\varepsilon]}^{n_s^0} x_i x_i' \to \frac{1}{\varepsilon}\int_{\tau_s^0-\varepsilon}^{\tau_s^0} h(t)h(t)'\,\mathrm{d}t > 0, \tag{3}$$

where, for a matrix, we write $A>0$ if A is positive definite. If we further assume that each regime occupies a positive fraction of observations such that

$$n_s^0 = [n\tau_s^0], \quad 0<\tau_1^0<\cdots<\tau_m^0<1 \text{ and } \beta_s^0 \neq \beta_{s+1}^0, \tag{4}$$

then Eqs. (3) and (4) are sufficient to establish the following result. For every $\varepsilon>0$ and $\delta>0$, for all large n, we have

$$P(|\hat{n}_s - n_s^0| > \delta n) < \varepsilon \quad (s=1,\ldots,m). \tag{5}$$

Under an additional assumption,

$$h(\tau_s^0)'(\beta_{s+1}^0 - \beta_s^0) \neq 0, \quad s=1,\ldots,m \tag{6}$$

we can improve the rate in Eq. (5) to obtain $\hat{n}_s - n_s^0 = \mathrm{O_p}(1)$. That is, Theorem 1 (below) still holds for trending regressors under assumptions (4) and (6). Because our general framework does not require Eq. (6) (i.e., positive fraction of the sample size for each regime), we will not give a separate proof for the case of trending regressors. A proof for this case is available from the author. In the sequel, we shall focus on regressors satisfying A1–A5. □

Throughout, the notation on the number of change points m is used interchangeably with $m(n)$ and m_n.

Theorem 1. *If* A1–A5 *hold and m is bounded, then*

$$\hat{n}_j - n_j^0 = \mathrm{O_p}(1) \quad (j=1,2,\ldots,m).$$

Although the number of change points in this theorem is bounded, the length of each regime is not assumed to be a positive fraction of n. That is, the assumption that $n_j^0=[n\tau_j^0]$ with $0<\tau_j^0<1$ is not needed. A1 assumes that each regime length is at least $c_1 n^{3/4}(c_1>0)$. This assumption can be weakened to $c_1 n^{(1/2)+\delta}$ for some $\delta\in(0,1/2)$, as long as m is bounded.

We shall not deal with this case because it would require a separate proof from the case of $m(n)\to\infty$, which will be considered below. To allow the number of changes to grow with the sample size, we need an additional assumption, under which the proof will be much easier:

A6. The regressors x_i are i.i.d. such that $E(x_i x_i')$ is positive definite.

Theorem 2. *If* A1–A6 *hold and* $m(n) \to \infty$ *but* $m(n) < cn^{(1/4)-b} (c, b > 0)$, *then* Theorem 1 *still holds. That is, for each* $j \leqslant m(n)$, $\hat{n}_j - n_j^0 = O_p(1)$.

Given this rate of convergence, it is not difficult to prove the following result.

Theorem 3. *For bounded* m, *assume* A1–A5. For $m = m(n) \to \infty$, *assume* A1–A6. *Then, for each* j

$$2f(0)(\hat{n}_j - \hat{n}_{j-1})^{1/2}(\hat{\beta}_j - \beta_j^0) \xrightarrow{d} N(0, V_j),$$

where $f(0)$ *is the density function of* ε_1 *at zero, and*

$$V_j = \text{plim} \frac{1}{n_j^0 - n_{j-1}^0} \sum_{i=n_{j-1}^0+1}^{n_j^0} x_i x_i'.$$

The limiting distribution of the estimated regression parameter is the same as if the change points were known. This result is well known for a single change point.

The next result concerns the limiting distributions of the estimated change points. To characterize the limiting distribution, we first define a stochastic process $W^{(j)}(k)$ on the set of integers as follows: $W^{(j)}(0) = 0$, $W^{(j)}(k) = W_1^{(j)}(k)$ for $k < 0$, and $W^{(j)}(k) = W_2^{(j)}(k)$ for $k > 0$ where, for $j = 1, \ldots, m(n)$:

$$W_1^{(j)}(k) = \sum_{l=k+1}^{0} |\varepsilon_l^{(j)} - \Delta_j' x_l^{(j)}| - |\varepsilon_l^{(j)}|, \quad k = -1, -2, \ldots, \tag{7}$$

$$W_2^{(j)}(k) = \sum_{l=1}^{k} |\varepsilon_l^{(j)} + \Delta_j' x_l^{(j)}| - |\varepsilon_l^{(j)}|, \quad k = 1, 2, \ldots \tag{8}$$

with $\Delta_j = (\beta_{j+1}^0 - \beta_j^0)$ and where $\{x_l^{(j)}, \varepsilon_l^{(j)}\}$ is an independent copy of $\{x_l, \varepsilon_l\}$.

Theorem 4. *Under assumptions* A1–A6, *and assuming that* $|\varepsilon_i \pm \Delta_j' x_i| - |\varepsilon_i|$ *has a continuous distribution, then for each* $j \leqslant m(n)$ ($m(n)$ *can be bounded or unbounded*),

$$\hat{n}_j - n_j^0 \xrightarrow{d} \arg\min_k W^{(j)}(k).$$

Furthermore, the estimated change points are asymptotically independent of each other and of the estimated regression parameters.

The assumption that $|\varepsilon_i \pm \Delta_j' x_i| - |\varepsilon_i|$ has a continuous distribution ensures the uniqueness (a.s.) of the minimum of $W^{(j)}$.

Remark 2. When $m = m(n)$ is fixed, Theorem 4 can be proved by showing that $S_n(n_1^0 + k_1, \ldots, n_m^0 + k_m) - S_n(n_1^0, \ldots, n_m^0)$ converges in distribution to $\sum_{j=1}^m W^{(j)}(k_j)$, for $|k_j| \leqslant M$, where $M<\infty$. However, the limiting process is not defined when $m=m(n)\to\infty$. This difficulty can be bypassed using the following small trick. We note that

$$\begin{aligned}\hat{n}_j &= \arg\min_{n_j} S_n(\hat{n}_1,\ldots,\hat{n}_{j-1},n_j,\hat{n}_{j+1},\ldots,\hat{n}_m)\\ &= \arg\min_{n_j}\{S_n(\hat{n}_1,\ldots,\hat{n}_{j-1},n_j,\hat{n}_{j+1},\ldots,\hat{n}_m) - S_n(\hat{n}_1,\ldots,\hat{n}_{j-1},n_j^0,\hat{n}_{j+1},\ldots,\hat{n}_m)\}.\end{aligned}$$

The limiting process above is indexed by $n_j - n_j^0$, a scalar, rather than a process with multiple indices. Further details are given in the proof of Theorem 4. □

Remark 3. Here we discuss the efficiency of LAD relative to LS. For simplicity, consider a single mean shift: $y_i=\mu_{1,n}+\varepsilon_i$ for $i\leqslant n^0$, and $y_i=\mu_{2,n}+\varepsilon_i$ for $i>n^0$, where $n=[n\tau]$, with $\tau\in(0,1)$ and ε_i are i.i.d. Let $\delta_n=\mu_{1,n}-\mu_{2,n}\neq 0$. For fixed magnitude of shift, it is difficult to compare the efficiency, so we assume shrinking shifts. Let $\delta_n\to 0$ but $\sqrt{n}\delta_n/\log n\to\infty$. Let $\hat{\tau}_{\mathrm{LS}}$ and $\hat{\tau}_{\mathrm{LAD}}$ denote the LS and LAD estimators of τ, respectively. Then Bai (1994) shows that,

$$n\delta_n^2(\hat{\tau}_{\mathrm{LS}}-\tau)\xrightarrow{\mathrm{d}}\sigma_\varepsilon^2\arg\max_v\{W(v)-|v|/2\}$$

where $\sigma_\varepsilon^2=\mathrm{Var}(\varepsilon_i)$, and $W(v)$ is a two-sided Brownian motion on $\mathbb{R}$. For LAD estimation, Bai (1995) shows that

$$n\delta_n^2(\hat{\tau}_{\mathrm{LAD}}-\tau)\xrightarrow{\mathrm{d}}(2f(0))^{-2}\arg\max_v\{W(v)-|v|/2\},$$

where $f(x)$ is the density function of ε_i. Obviously, if ε_i does not have a finite variance, LS estimation is less efficient than LAD. The same limiting distributions would result even if $\mu_{1,n}$ and $\mu_{2,n}$ were known and not estimated. In this sense, there is a direct gain in efficiency by LAD when estimating the change point for heavy-tailed distributions. On the other hand, efficiency gain is realized through LAD's consistent estimation of the regression coefficients when they are unknown. □

3. Determining the number of change points

In this section, we consider estimating the number of change points. Yao (1988) proposes the Schwarz criterion to estimate this number. If the underlying distribution is double exponential, then LAD is the maximum likelihood procedure. By the Schwarz criterion, the number of change points is determined by minimizing the objective function

$$\mathrm{LADBIC}(m)=n\log\hat{e}(m)+(1/2)(m+1)(p+1)\log n \tag{9}$$

where $\hat{e}(m)=S_n(\hat{n}_1,\ldots,\hat{n}_m)/n$. Note that the total estimated number of parameters $(m+1)(p+1)$, includes $(m+1)p$ regression parameters, m change points, and a scale parameter. Criterion (9) differs from Yao's criterion by an extra factor 1/2, which is absent for least squares estimation under the normality assumption. Of course, $\hat{e}(m)$ is the sample average of absolute deviations rather than squared values of residuals. Whether this criterion leads to a consistent estimate of the number of change points remains an open question. In this section, we study a modified criterion under which the estimated number of change points can be shown to be consistent for the true number of changes.

Consider the criterion of the form

$$B(m) = n \log \hat{e}(m) + mg(n). \tag{10}$$

Although there is some flexibility in choosing the penalty term $g(n)$, we shall consider $g(n) = \sqrt{n}$ to be specific. This choice of g is also used in the reported simulations.

We allow the true number $m_n^0 \to \infty$. Let $\hat{m}$ be the integer at which the criterion function is minimized over the integer set $\{0, 1, 2, \ldots, Lm_n^0\}$, where $L > 1$ is arbitrarily given.

Theorem 5. *If* A1–A6 *hold and* $E|\varepsilon_1| < \infty$, *then* $P(\hat{m} = m_n^0) \to 1$.

The theorem asserts that even if $m_n \to \infty$, with probability tending to 1, the estimated break point coincides with the true number.

4. Numerical result

In addition to the theoretical properties, we are also interested in LAD's implementation in practice. We develop a computer program for estimating multiple-regime regressions. The program allows one to choose the number of regimes based on the information criteria discussed earlier. Our program exploits linear programming for LAD estimation (Barrodale and Roberts, 1974) and dynamic programming for optimal segmentation (Guthery, 1974). Our program only requires $O(n^2m)$ number of LAD computations to achieve the global minimization, a considerable computational savings relative to the brute-force enumeration for $m > 2$. The computation is fast even with 10 change points as in the simulations reported later.

4.1. Monte Carlo simulation

This simulation focuses on the relative performance of LAD and LS. We consider the following simple model with 4 regimes (3 change points):

$$y_i = \alpha_k + \beta_k x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{11}$$

where the x_i are i.i.d. standard normal random variables, the vector (α_k, β_k) $(k = 1, \ldots, 4)$ is the parameter for regime k, and the ε_i are i.i.d. standard normal or double exponential random variables. In the latter case, the density function is $f(x) = 2^{-1}e^{-|x|}$, which has a variance of 2. We choose $n = 100$ and $m = 3$. The true change points are 25,

Table 1
Means and standard deviations of the estimated change points (500 repetitions)

Normal	LS	24.98	(1.49)	49.95	(1.75)	75.00	(1.31)
	LAD	24.95	(2.08)	49.99	(1.86)	74.97	(1.70)
Double	LS	24.99	(6.04)	49.32	(8.43)	74.41	(5.87)
exponential	LAD	25.18	(2.53)	50.11	(3.54)	75.15	(2.96)
t-distribution	LS	26.57	(10.15)	50.07	(10.98)	75.04	(7.26)
df = 3	LAD	25.11	(3.64)	50.00	(3.38)	75.18	(3.24)
Contaminated normal	LS	28.65	(13.96)	49.61	(14.47)	74.59	(9.72)
$(\varepsilon=0.1, \tau=5)$	LAD	25.42	(4.82)	49.91	(4.54)	75.26	(3.38)

Model: $y_i=\alpha_k+\beta x_i+\varepsilon_i$, where ε_i are, respectively, normal N(0, 1), double exponential with density $f(x)=2^{-1}e^{-|x|}$, student t with df = 3, and contaminated normal with cdf $F(x)=(1-\varepsilon)\Phi(x)+\varepsilon\Phi(\frac{x}{\tau})$, here $\varepsilon=0.1$ and $\tau=5$. Sample size 100, true change points 25, 50, and 75. Standard deviations are reported in parentheses.

50, and 75, respectively. Only the case of intercept changes with $\alpha_1=1, \alpha_2=3, \alpha_3=1$, $\alpha_4=-1$ and $\beta_k=1$ $(\forall k)$ is reported. The estimated means and standard deviations from 500 repetitions are reported in Table 1. The number of regimes is assumed to be known. Under normal errors, the LAD yields estimates with a larger spread than LS. The converse is true under double exponential errors. The LS gives estimates with a much larger spread than LAD. Additional simulations are done for ε_i being t distribution with df = 3 and contaminated normal distribution with cdf $F(x)=(1-\varepsilon)\Phi(x)+\varepsilon\Phi(x/\tau)$, here $\varepsilon=0.1$ and $\tau=5$. For these latter distributions, efficiency gain by LAD is striking. Though not reported in Table 1, this observation is true for jumps of different sizes and changes in slope parameters as well.

Monte Carlo simulations for estimating the number of change points are also performed. We only report the summary here. The model considered is still Eq. (11). Both criteria (9) and (10) are used. Each criterion function is minimized over the range $\{0,1,2,\ldots,10\}$. Criterion (9) correctly identifies the number of regimes 71% of the time for normal errors and 73% of the time for double exponential errors. This criterion has a tendency to overestimate the true number, suggesting that the penalty term is not heavy enough. In contrast, criterion (10) with $g(n)=\sqrt{n}$ correctly identifies the number of regimes 93% of the time for normal errors and 76% for double exponential errors. Here there is a tendency to underestimate the number. These results suggest the possibility of further improvement by adjusting the penalty term.

4.2. An application

This application concerns the response of market interest rates to changes in the Federal Reserve (Fed) discount rate, which is the rate at which the Fed lends money and is set by the Fed. The yield of three-month treasury bills is used as the market interest rate. The data range spans 1973–1989. Over this period the Fed made 56 changes in the discount rate. The details are described by Dueker (1992).

Changes in the market interest rate are often a complicated function of many factors in addition to the Fed discount rate. The most important of these is perhaps the state of the economy. As in Dueker (1992), we use the unemployment rate as an indicator of the performance of the economy. Dueker uses a mixture model by mixing 'high' and 'lower' response with mixing probability depending on other exogenous variables. His results suggest that the response is different over time. Here we use the simple change point model and estimate the response pattern over time. The following model is used:

$$\Delta \mathrm{TB}_i = \beta_{0k} + \beta_{1k}\Delta \mathrm{DR}_i + \beta_{2k} U_i + \varepsilon_i,$$

where ΔTB is the change in the T-bill rate, ΔDR is the change in the discount rate, U is the unemployment rate, and $(\beta_{0k}, \beta_{1k}, \beta_{2k})$ are the regression parameters of regime k. Both criteria (9) and (10) suggest the existence of three regimes. The estimated numbers of observations for the three regimes are 27, 15, and 14, respectively. The estimated regression parameters are $(0.331, 0.051, -0.058)$, $(3.256, 0.163, -0.383)$, and $(0.268, 0.064, -0.040)$, respectively. The second regime is markedly different from the rest, with responses being most sensitive to changes in the discount rate and in unemployment. Finally, it is interesting to note that the first change point occurs in October 1979, and the second occurs in November 1982. These estimated change points coincide with changes in the operating procedure of the Federal Reserve (Roley and Wheatley, 1990). Thus, policy changes that may not be directly linked to the variables under consideration can have an effect on those variables. This example highlights the potential use of the change-point model in social sciences.

5. Auxiliary results

In this section, we derive a number of results in the absence of change points. In the next section, we show how these results can be used to establish Theorems 1–5. This framework of proof is useful for other estimation methods such as M-estimation. All needed is to prove the corresponding lemmas for a given estimation method. Consider the standard regression model:

$$y_i = w_i'\phi^0 + \varepsilon_i \quad (i = 1, \ldots, n),$$

where w_i is a $p \times 1$ vector of regressors, ϕ^0 is the true vector of parameters, and ε_i is a disturbance. We assume:

B1: The errors ε_i satisfy A5 with x_i interpreted as w_i.

B2: The regressors w_i are uniformly bounded as in A3. That is, there exists $K > 0$ such that $|w_i| < K$ for all i.

B3: The matrix $(1/k)\sum_{i=1}^{k} w_i w_i'$ converges in probability to a nonrandom positive-definite matrix.

Throughout this section, we assume B1–B3 are satisfied. We do not assume a bounded parameter set. The parameter space is $\mathbb{R}^p$. All the lemmas are true even if the regressors are not uniformly bounded, but satisfy the condition: for each $\varepsilon > 0$, there exists a $K > 0$ such that $(1/k)\sum_{i=1}^{k} |w_i|^2 I(|w_i| > K) < \varepsilon$ for all large k with large probability. We shall treat w_i as deterministic. Otherwise, conditional argument can be used because of the independence of disturbances and regressors. However, in Lemma 7 below, we do analyze the case of i.i.d. regressors, which allow us to strengthen some of the results. The case of i.i.d. regressors corresponds to Assumption A6.

We are interested in the behavior of the optimal objective function

$$\inf_{\phi} \sum_{i=1}^{n} |y_i - w_i'\phi| = \inf_{\phi} \sum_{i=1}^{n} |\varepsilon_i - w_i'(\phi - \phi^0)| = \inf_{\phi} \sum_{i=1}^{n} |\varepsilon_i - w_i'\phi|,$$

(redefining ϕ as $\phi - \phi^0$, or simply assuming $\phi^0 = 0$). Define the centered objective function as

$$\sum_{i=1}^{n} (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|).$$

To begin, we state a lemma due to Babu (1989), which is closely related to the Bernstein inequality.

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

$$P\left(\left| \sum_{i=1}^{k} Z_i \right| > a \right) \leqslant 2 \exp\{-a^2 s(1-s)/V\}. \qquad (12)$$

The following simple inequality will be used frequently:

$$||x - y| - |x - z|| \leqslant |y - z|. \qquad (13)$$

We define throughout

$$\eta_i(\phi) = |\varepsilon_i - w_i'\phi| - |\varepsilon_i| \quad \text{and} \quad \xi_i(\phi) = \eta_i(\phi) - E\eta_i(\phi). \qquad (14)$$

Lemma 2. (i) *For each $\delta \in (0, 1)$,*

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

(ii)

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

Proof. See Lemma 1 of Bai (1995). □

Lemma 3. *For every $\alpha > 1/2$ and every $M < \infty$, we have*

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

Proof. Let $\eta_i(\phi) = |\varepsilon_i - w_i'\phi| - |\varepsilon_i|$. Because $E\eta_i(\phi) \geqslant 0$ and $\eta_i(0) = 0$, we have

$$0 \geqslant \inf_{|\phi| \leqslant M} \sum_{i=l}^{k} \eta_i(\phi) \geqslant \inf_{|\phi| \leqslant M} \sum_{i=l}^{k} [\eta_i(\phi) - E\eta_i(\phi)].$$

Thus

$$\left| \inf_{|\phi| \leqslant M} \sum_{i=l}^{k} \eta_i(\phi) \right| \leqslant \sup_{|\phi| \leqslant M} \left| \sum_{i=l}^{k} [\eta_i(\phi) - E\eta_i(\phi)] \right|.$$

It follows that

$$\sup_{1 \leqslant l < k \leqslant n} \left| \inf_{|\phi| \leqslant M} \sum_{i=l}^{k} \eta_i(\phi) \right| \leqslant 2 \sup_{1 \leqslant k \leqslant n} \sup_{|\phi| \leqslant M} \left| \sum_{i=1}^{k} [\eta_i(\phi) - E\eta_i(\phi)] \right|.$$

Thus it suffices to prove the right-hand side above is bounded by $O_p(n^\alpha)$.

Let $\zeta_k = \sup_{|\phi| \leqslant M} |\sum_{i=1}^{k} [\eta_i(\phi) - E\eta_i(\phi)]|$. Then $\{\zeta_k, \mathscr{F}_k\}$ $(k = 1, \ldots, n)$ forms a sub-martingale, where $\mathscr{F}_k = \sigma - \text{field}\{\varepsilon_1, \ldots, \varepsilon_k\}$. By Doob's inequality,

$$P\left(\sup_{k \leqslant n} \zeta_k > n^\alpha \right) \leqslant n^{-\alpha m} C_m E(\zeta_n^m), \tag{15}$$

for some $C_m > 0$ (where $m > 1$ will be determined later). Next, divide the parameter set $|\phi| \leqslant M$ into $c_p n^{p/2}$ $(c_p > 0)$ cells such that the diameter of each cell is no larger than $Mn^{-1/2}$. For arbitrary s, t in a common cell,

$$\begin{aligned} \left| \sum_{i=1}^{n} \eta_i(s) - E\eta_i(s) - \eta_i(t) + E\eta(t) \right| &\leqslant 2 \sum_{i=1}^{n} |w_i||s - t| \\ &\leqslant 2Mn^{-1/2} \sum_{i=1}^{n} |w_i| \leqslant 2KMn^{1/2}. \end{aligned}$$

Let ϕ_r be a point in the rth cell $(r = 1, 2, \ldots, c_p n^{p/2})$. From $|h(\phi)| \leqslant |h(\phi_r)| + |h(\phi) - h(\phi_r)|$ for an arbitrary function $h(\phi)$, and $|a + b|^m \leqslant L_m |a|^m + L_m |b|^m$ for some constant L_m only depending on m, we have

$$\zeta_n^m \leqslant L_m \sup_r \left| \sum_{i=1}^{n} \eta_i(\phi_r) - E\eta_i(\phi_r) \right|^m + L_m (2KM)^m n^{m/2}. \tag{16}$$

Because $\eta_i(\phi_r) - E\eta_i(\phi_r)$ forms a sequence of bounded martingale differences for each fixed r, we have, for some $A > 0$,

$$E \left| \sum_{i=1}^{n} \eta_i(\phi_r) - E\eta_i(\phi_r) \right|^m \leqslant A n^{m/2}, \quad \forall r.$$

Taking expectations on both sides of Eq. (16) and using $E(\sup_r |\cdot|^m) \leqslant \sum_r E|\cdot|^m$, we obtain

$$E\zeta_n^m \leqslant AL_m c_p n^{p/2} n^{m/2} + O(n^{m/2}) = O(n^{(p+m)/2}).$$

Thus the right-hand side of Eq. (15) is bounded by, for some $C>0$, $Cn^{-\alpha m} n^{(p+m)/2}$, which converges to zero as $n\to\infty$ for $m>p/(2\alpha-1)$ and for $\alpha>1/2$. This proves the lemma. □

Lemma 4. *Let c_n be a positive sequence such that either $c_n \equiv c>0$ or $c_n\to 0$ and $nc_n^2/\log n\to\infty$. Then there exists a $C>0$ such that for each $\varepsilon>0$ and all large n,*

$$P\left(\sup_{|\phi|\leqslant c_n}\left|\frac{1}{nc_n^2}\sum_{i=1}^n \xi_i(\phi)\right| \geqslant \varepsilon\right) \leqslant \exp(-\varepsilon^2 nc_n^2 C),$$

where $\xi_i(\phi)$ is defined in Eq. (14).

Proof. Divide the region $|\phi|\leqslant c_n$ into $c_p n^{p/2}$ (for some $c_p<\infty$) cells such that the diameter of each cell is no larger than $n^{-1/2}c_n$. For ϕ_1,ϕ_2 belonging to a common cell, the incremental value

$$\frac{1}{nc_n^2}\left|\sum_{i=1}^n [\xi_i(\phi_1)-\xi_i(\phi_2)]\right| \leqslant 2K|\phi_1-\phi_2|/c_n^2 \leqslant 2K/(\sqrt{n}c_n)\to 0.$$

Let ϕ_r be a point in the rth cell ($r=1,\dots,c_p n^{p/2}$), we have

$$P\left(\sup_r\left|\frac{1}{nc_n^2}\sum_{i=1}^n \xi_i(\phi_r)\right| >\varepsilon\right) \leqslant \sum_r P\left(\left|\sum_{i=1}^n \xi_i(\phi_r)\right| > nc_n^2\varepsilon\right). \tag{17}$$

From $|\xi_i(\phi)|\leqslant 2|w_i'\phi|\leqslant 2Kc_n$, it follows that $\mathrm{Var}(\xi_i(\phi))\leqslant 4K^2c_n^2$ uniformly in $|\phi|\leqslant c_n$. Apply Lemma 1 with $d=2Kc_n$, $V=4K^2nc_n^2$, $s=1/2$, and $a=nc_n^2\varepsilon$, we have for large n

$$P\left(\left|\sum_{i=1}^n \xi_i(\phi_r)\right| > nc_n^2\varepsilon\right) \leqslant 2\exp(-\varepsilon^2 nc_n^2 C),$$

with $C=1/(16K^2)$. Thus the r.h.s. of Eq. (17) is bounded by $2c_p n^{p/2}\exp(-\varepsilon^2 nc_n^2 C)$, which is further bounded by $\exp(-\varepsilon^2 nc_n^2 C/2)$ for all large n, because $nc_n^2/\log n\to\infty$. □

Lemma 4 implies that, by the Borel–Cantelli lemma, for every $\varepsilon>0$,

$$\limsup_{n\to\infty}\sup_{|\phi|\leqslant c_n}\left|\frac{1}{nc_n^2}\sum_{i=1}^n \xi_i(\phi)\right| \leqslant \varepsilon, \quad \text{a.s.} \tag{18}$$

Remark 4. The following result will be used in subsequent proofs. Let $h(x)$ $(x\in R^p)$ be a convex function with $h(0)=0$. If $\inf_{|x|=c} h(x)=a>h(0)=0$, then $\inf_{|x|\geqslant c} h(x)=\inf_{|x|=c} h(x)$. That is, the extreme value of a convex function is attained on the boundary. To see this, suppose $|x'|>c$. Choose $\lambda\in(0,1)$ such that $x''=\lambda x'$ and $|x''|=\lambda|x'|$

$=c$. Then $a \leqslant h(x'') = h(\lambda x' + (1-\lambda)0) \leqslant \lambda h(x') + (1-\lambda)h(0) = \lambda h(x')$. Thus $h(x') \geqslant a/\lambda > a$. □

Lemma 5. *If $c_n \to 0$ and $nc_n^2/\log n \to \infty$, then there exists an $\eta > 0$, such that with probability* 1,

$$\liminf_n \inf_{|\phi| \geqslant c_n} \frac{1}{nc_n^2} \sum_{i=1}^n (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) \geqslant \eta > 0.$$

Proof. We prove the sum has a large expected value, and its deviation from its expected value is small. Because $|\varepsilon_i - w_i'\phi| - |\varepsilon_i|$ is convex in ϕ, and the sum of convex functions is still convex, it suffices to prove the lemma for $|\phi| = c_n$ (see Remark 4). Because $c_n \to 0$, we have (e.g., Pollard, 1991)

$$E\left(\sum_{i=1}^n (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|)\right) = \phi'\left(\sum_{i=1}^n w_i w_i'\right)\phi f(0)(1+o(1)) \geqslant n|\phi|^2 \lambda f(0)/2 = nc_n^2 \lambda f(0)/2, \tag{19}$$

where λ is a positive number which is no larger than the smallest eigenvalue of $(1/n)\sum_{i=1}^n w_i w_i'$. The existence of such a λ is guaranteed by assumption B3 for all large n. The lemma is proved with $\eta = \lambda f(0)/4$ if we take $\varepsilon = \lambda f(0)/4$ in Eq. (18). □

Lemma 6. *Let $\hat{\phi}_n$ be the LAD estimator of ϕ, i.e., $\hat{\phi}_n = \arg\min_\phi \sum_{i=1}^n (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|)$. Then for c_n in Lemma* 5, *there exists a $C > 0$ such that for all large n,*

$$P(|\hat{\phi}_n| > c_n) \leqslant \exp(-nc_n^2 C).$$

Proof. The lemma is implied by the following:

$$P\left(\inf_{|\phi| \geqslant c_n} \sum_{i=1}^n (|\varepsilon_i - w_i'\phi| - |\varepsilon_i|) < 0\right) \leqslant \exp(-nc_n^2 C). \tag{20}$$

We shall prove this inequality. By convexity, it is sufficient to consider $|\phi| = c_n$. Let $\eta_i(\phi) = |\varepsilon_i - w_i'\phi| - |\varepsilon_i|$ and $\xi_i(\phi) = \eta_i(\phi) - E\eta_i(\phi)$. Now

$$\begin{aligned}\inf_{|\phi|=c_n} \sum_{i=1}^n \eta_i(\phi) &\geqslant \inf_{|\phi|=c_n} \sum_{i=1}^n [\eta_i(\phi) - E\eta_i(\phi)] + \inf_{|\phi|=c_n} \sum_{i=1}^n E\eta_i(\phi) \\ &\geqslant -\sup_{|\phi|=c_n} \left|\sum_{i=1}^n \xi_i(\phi)\right| + \inf_{|\phi|=c_n} \sum_{i=1}^n E\eta_i(\phi).\end{aligned}$$

Thus

$$P\left(\inf_{|\phi|=c_n} \sum_{i=1}^n \eta_i(\phi) < 0\right) \leqslant P\left(\sup_{|\phi|=c_n} \left|\sum_{i=1}^n \xi_i(\phi)\right| \geqslant \inf_{|\phi|=c_n} \sum_{i=1}^n E\eta_i(\phi)\right)$$

$$\leqslant P\left(\sup_{|\phi|=c_n}\left|\sum_{i=1}^n \xi_i(\phi)\right| \geqslant nc_n^2\lambda f(0)/2\right)$$
$$\leqslant \exp(-nc_n^2 C\lambda^2 f(0)^2/4),$$

where the second inequality follows from $\inf_{|\phi|=c_n}\sum_{i=1}^n E\eta_i(\phi)\geqslant nc_n^2\lambda f(0)/2$ by Eq. (19); the last inequality follows from Lemma 4 with $\varepsilon=\lambda f(0)/2$. The lemma is proved by redefining C. □

The following lemma is an improved version of Lemma 3 under the i.i.d. assumption. The latter assumption is made in A6.

Lemma 7. *Assuming that $\{w_i,\varepsilon_i\}_{i=1}^n$ are i.i.d., then for every $a>0$, $t>0$, and $M<\infty$,*

$$P\left(\sup_{1\leqslant l<k\leqslant n}\left|\inf_{|\phi|\leqslant M}\sum_{i=l}^k(|\varepsilon_i-w_i'\phi|-|\varepsilon_i|)\right|>n^a\right)=\mathrm{O}(n^{-t}).$$

Proof. Let $\eta_i(\phi)$ be defined as before. Then

$$P\left(\sup_{1\leqslant l<k\leqslant n}\left|\inf_{|\phi|\leqslant M}\sum_{i=l}^k\eta_i(\phi)\right|>n^a\right)\leqslant\sum_{l<k}P\left(\left|\inf_{|\phi|\leqslant M}\sum_{i=l}^k\eta_i(\phi)\right|>n^a\right)$$
$$=\sum_{l<k}P\left(\left|\inf_{|\phi|\leqslant M}\sum_{i=1}^{k-l}\eta_i(\phi)\right|>n^a\right)\quad\text{by i.i.d.}$$
$$\leqslant n^2\max_{1\leqslant k\leqslant n}P\left(\left|\inf_{|\phi|\leqslant M}\sum_{i=1}^k\eta_i(\phi)\right|>n^a\right).$$

From $|\eta_i(\phi)|\leqslant|w_i'\phi|\leqslant KM$, we have $|\sum_{i=1}^k\eta_i(\phi)|\leqslant kKM<n^a$ for $k<n^b$ and n large, where $0<b<a$. Thus, it is enough to consider $k\geqslant n^b$ for some $b\in(0,a)$. Let $\hat{\phi}_k=\arg\min_\phi\sum_{i=1}^k\eta_i(\phi)$, and let c_k be a sequence of positive numbers. Then,

$$n^2\max_{n^b\leqslant k\leqslant n}P\left(\left|\inf_{|\phi|\leqslant M}\sum_{i=1}^k\eta_i(\phi)\right|>n^a\right)$$
$$\leqslant n^2\max_{n^b\leqslant k\leqslant n}P(|\hat{\phi}_k|>c_k)+n^2\max_{n^b\leqslant k\leqslant n}P\left(\left|\inf_{|\phi|\leqslant c_k}\sum_{i=1}^k\eta_i(\phi)\right|>n^a\right).\tag{21}$$

Choose $c_k=k^{-1/2}\log k$. By Lemma 6,

$$P(|\hat{\phi}_k|>c_k)\leqslant\exp(-kc_k^2C)\leqslant\exp(-(b\log n)^2C)\quad\text{for } k\geqslant n^b.$$

It follows that, for every $t>0$

$$n^2\max_{n^b\leqslant k\leqslant n}P(|\hat{\phi}_k|>c_k)=\mathrm{O}(n^{-t})\tag{22}$$

for all large n. Next,

$$\left|\inf_{|\phi|\leqslant c_k}\sum_{i=1}^{k}\eta_i(\phi)\right| \leqslant \sup_{|\phi|\leqslant c_k}\left|\sum_{i=1}^{k}\xi_i(\phi)\right| + \sup_{|\phi|\leqslant c_k}\sum_{i=1}^{k}E\eta_i(\phi).$$

Because $c_k \to 0$,

$$\sum_{i=1}^{k}E\eta_i(\phi)=\phi'\sum_{i=1}^{k}w_iw_i'\phi f(0)(1+\mathrm{o}(1))\leqslant 2kK^2c_k^2\leqslant 2K^2(\log k)^2\leqslant n^a/2$$

for all large n. Moreover, by Lemma 4 (applied with $n=k$ and $\varepsilon=2^{-1}n^a/(kc_k^2)=2^{-1}n^a/(\log k)^2$),

$$P\left(\sup_{|\phi|\leqslant c_k}\left|\sum_{i=1}^{k}\xi_i(\phi)\right|>\frac{1}{2}n^a\right)\leqslant \exp(-4^{-1}Cn^{2a}/(\log k)^4)$$
$$\leqslant \exp(-4^{-1}Cn^{2a}/(b\log n)^4),$$

for $k>n^b$. Since $n^2\exp(-4^{-1}Cn^{2a}/(b\log n)^4)=\mathrm{O}(n^{-t})$ for every t, the last term of Eq. (21) is bounded by $\mathrm{O}(n^{-t})$. Combining with Eq. (22), we obtain the lemma. □

Lemma 8. *If $|\phi|=M>0$, then there exists a $\delta>0$ such that a positive fraction of observation satisfy $|w_i'\phi|>\delta$. More specifically, let $N_n(\phi)=\mathrm{card}\{i;|w_i'\phi|>\delta,\ 1\leqslant i\leqslant n\}$, then for some $\varepsilon_0>0$, uniformly in $|\phi|=M$, $N_n(\phi)\geqslant n\varepsilon_0$ for all large n.*

Proof. Note that

$$\sum_{i=1}^{n}(w_i'\phi)^2=\phi'\sum_{i=1}^{n}w_iw_i'\phi\geqslant\lambda n|\phi|^2=\lambda nM^2,$$

where λ is defined in the proof of Lemma 5. On the other hand,

$$\sum_{i=1}^{n}(w_i'\phi)^2 = \sum_{i=1}^{n}(w_i'\phi)^2I(|w_i'\phi|\leqslant\delta)+\sum_{i=1}^{n}(w_i'\phi)^2I(|w_i'\phi|>\delta)$$
$$\leqslant n\delta^2+(KM)^2\sum_{i=1}^{n}I(|w_i'\phi|>\delta).$$

Thus

$$N_n(\phi)=\sum_{i=1}^{n}I(|w_i'\phi|>\delta)\geqslant(KM)^{-2}n(\lambda M^2-\delta^2)\geqslant n\varepsilon_0$$

for $\varepsilon_0=(KM)^{-2}(\lambda M^2-\delta^2)$, which is positive for a small δ. □

Lemma 9. *For each $M>0$, there exists an $\eta>0$ and $C>0$ such that*

$$P\left(\inf_{|\phi|\geqslant M}\sum_{i=1}^{n}(|\varepsilon_i-w_i'\phi|-|\varepsilon_i|)\geqslant\eta n\right)\geqslant 1-\exp(-nC).$$

Proof. Again by convexity, we assume without loss of generality, $|\phi|=M$. Let $H(\mu)=E(|\varepsilon_i-\mu|-|\varepsilon_i|)$. Then $H(\mu)$ is nonnegative and $H(\mu)$ is an increasing function in

$|\mu|$ with a unique minimum at zero. By Lemma 8, there exist no less than $n\varepsilon_0$ (for some $\varepsilon_0>0$) observations such that $|w_i'\phi|>\delta$ for some $\delta>0$. Thus

$$E\sum_{i=1}^{n}(|\varepsilon_i-w_i'\phi|-|\varepsilon_i|)\geqslant n\varepsilon_0 H(\delta) \tag{23}$$

uniformly over $\{\phi;|\phi|=M\}$. Furthermore, by Lemma 4 (applied with $c_n\equiv M$), for each $\varepsilon>0$,

$$P\left(\sup_{|\phi|=M}\left|\sum_{i=1}^{n}(|\varepsilon_i-w_i'\phi|-|\varepsilon_i|-E[|\varepsilon_i-w_i'\phi|-|\varepsilon_i|])\right|>\varepsilon n\right)\leqslant\exp(-n\varepsilon^2C). \tag{24}$$

That is, the deviation from the mean is small. Take $\varepsilon=\varepsilon_0H(\delta)/2$, the lemma follows from Eqs. (23) and (24) with $\eta=\varepsilon_0H(\delta)/2$. □

Lemma 10. *Let n_1 and n_2 be two integers such that $n_1\geqslant n^\rho$ with $1\geqslant\rho\geqslant 3/4$ and $n_2\leqslant n^v$ with $v<1/4$. Consider*

$$y_i=w_i'\phi_1+\varepsilon_i,\quad i=1,\dots,n_1,$$
$$y_i=w_i'\phi_2+\varepsilon_i,\quad i=n_1+1,\dots,n_1+n_2.$$

Let $N=n_1+n_2$ and let $\hat\phi_N=\arg\min_{|\phi|\leqslant M}\sum_{i=1}^{N}|y_i-w_i'\phi|$, where M is large enough such that $|\phi_1|<M$ and $|\phi_2|<M$. Then

(i) *For every $\delta\in(0,\rho-v)$, with probability tending to 1,*

$$|\hat\phi_N-\phi_1|\leqslant n_1^{-1/2}n_1^{(v+\delta)/(2\rho)}\leqslant n^{-(\rho-v-\delta)/2}.$$

(ii) $\sum_{i=1}^{n_1}(|\varepsilon_i-w_i'(\hat\phi_N-\phi_1)|-|\varepsilon_i|)=O_p(1)$.

This lemma says that when the data are from two different models (two regimes in our application), the estimated regression parameter using the pooled data is close to the parameter of the model from where most of the data came. This is, of course, obvious, but (i) quantifies this intuition. Furthermore, similar to Lemma 2(i), the centered objective function of the 'dominating' model evaluated at the pooled estimator $\hat\phi_N$ is stochastically bounded, as asserted by (ii).

Proof. (i) Note that $\hat\phi_N$ minimizes

$$g_n(\phi)=\sum_{i=1}^{n_1}(|\varepsilon_i-w_i'(\phi-\phi_1)|-|\varepsilon_i|)+\sum_{i=n_1+1}^{n_1+n_2}(|\varepsilon_i-w_i'(\phi-\phi_2)|-|\varepsilon_i|). \tag{25}$$

The second term on the right-hand side of Eq. (25) is bounded by $\sum_{i=n_1+1}^{n_1+n_2}|w_i||\phi-\phi_2|\leqslant 2KMn_2=O(n^v)$ by the assumption of bounded regressors and $|\phi-\phi_2|\leqslant 2M$.

If $|\hat{\phi}_N-\phi_1|\geqslant n_1^{-1/2}n_1^{(v+\delta)/(2\rho)}$ for some $\delta>0$ with some positive probability η_0, then by Lemma 5, with $c(n_1)=n_1^{-1/2}n_1^{(v+\delta)/(2\rho)}$,

$$\sum_{i=1}^{n_1}(|\varepsilon_i-w_i'(\hat{\phi}_N-\phi_1)|-|\varepsilon_i|)\geqslant\eta n_1c(n_1)^2\geqslant\eta n_1^{(v+\delta)/\rho}\geqslant\eta n^{v+\delta}$$

with probability at least $\eta_0/2$ for large n. This implies that $g_n(\hat{\phi}_N)\geqslant\eta n^{(v+\delta)}-\mathrm{O}(n^v)\geqslant 2^{-1}\eta n^{v+\delta}$ with probability at least $\eta_0/2$ for large n. However, $\inf_\phi g_n(\phi)\leqslant g_n(\phi_1)=\mathrm{O}(n^v)$ with probability 1. Thus we arrive at a contradiction.

(ii) Rewrite $g_n(\phi)$ as

$$g_n(\phi)=f_n(\phi)+h_n(\phi)+\sum_{i=n_1+1}^{n_1+n_2}(|\varepsilon_i-w_i'(\phi_1-\phi_2)|-|\varepsilon_i|), \tag{26}$$

where

$$f_n(\phi)=\sum_{i=1}^{n_1}(|\varepsilon_i-w_i'(\phi-\phi_1)|-|\varepsilon_i|) \tag{27}$$

and

$$h_n(\phi)=\sum_{i=n_1+1}^{n_1+n_2}(|\varepsilon_i-w_i'(\phi-\phi_2)|-|\varepsilon_i-w_i'(\phi_1-\phi_2)|).$$

From (i), $|h_n(\hat{\phi}_N)|\leqslant\sum_{i=n_1+1}^{n_1+n_2}|w_i||\hat{\phi}_N-\phi_1|\leqslant Kn_2n^{-(\rho-v-\delta)/2}\leqslant Kn^{-(\rho-3v-\delta)/2}=\mathrm{o}(1)$, for $0<\delta<(\rho-3v)$. Because $f_n(\phi)+h_n(\phi)$ evaluated at $\phi=\phi_1$ is zero and $\hat{\phi}_N$ minimizes $f_n(\phi)+h_n(\phi)$, it follows that

$$0\geqslant f_n(\hat{\phi}_N)+h_n(\hat{\phi}_N)\geqslant f_n(\hat{\phi}_N)-|\mathrm{o_p}(1)|\geqslant\inf_\phi f_n(\phi)-|\mathrm{o_p}(1)|. \tag{28}$$

Thus,

$$|f_n(\hat{\phi}_N)|\leqslant\left|\inf_\phi f_n(\phi)\right|+\mathrm{o_p}(1).$$

By Lemma 2(i), $\inf_\phi f_n(\phi)=\mathrm{O_p}(1)$. This implies that $f_n(\hat{\phi}_N)=\mathrm{O_p}(1)$. □

The following result is an extension of Lemma 10.

Lemma 11. *Let n_1 and n_2 be the same as in the previous lemma. Consider*

$$y_i=w_i'\phi_1+\varepsilon_i,\quad i=1,\ldots,k,$$
$$y_i=w_i'\phi_2+\varepsilon_i,\quad i=k+1,\ldots,k+n_2,$$

where k is no smaller than a positive fraction of n_1 such that $k\in[n_1a,n_1]$ with $a\in(0,1]$. Let $\hat{\phi}_k=\arg\min_{|\phi|\leqslant M}\sum_{i=1}^{k+n_2}|y_i-w_i'\phi|$. We have

(i) *For every $a\in(0,1]$ and every $\delta\in(0,\rho-v)$, with probability tending to 1,*

$$\sup_{n_1a\leqslant k\leqslant n_1}|\hat{\phi}_k-\phi_1|\leqslant n_1^{-1/2}n_1^{(v+\delta)/(2\rho)}\leqslant n^{-(\rho-v-\delta)/2}.$$

(ii)

$$\sup_{n_1 a \leqslant k \leqslant n_1} \left| \sum_{i=1}^{k} |\varepsilon_i - w_i'(\hat{\phi}_k - \phi_1)| - |\varepsilon_i| \right| = \mathrm{O_p}(1).$$

Proof. (i) Let $c(k) = k^{-1/2} k^{(v+\delta)/(2\rho)}$. Then there exists a constant $A > 0$ such that, $c(n_1) \leqslant c(k) \leqslant Ac(n_1)$ for all $k \in [n_1 a, n_1]$. We prove (i) by reduction to absurdity. Now suppose $|\hat{\phi}_k - \phi_1| \geqslant c(n_1)$, then $|\hat{\phi}_k - \phi_1| \geqslant c(k)/A$. By Lemma 5,

$$\begin{aligned} \sum_{i=1}^{k} (|\varepsilon_i - w_i'(\hat{\phi}_k - \phi_1)| - |\varepsilon_i|) &\geqslant \eta k c(k)^2 A^{-2} \geqslant \eta k c(n_1)^2 A^{-2} \\ &\geqslant \eta a n_1^{(v+\delta)/\rho} A^{-2} \geqslant \eta a n^{(v+\delta)} A^{-2}. \end{aligned}$$

The above inequality implies that $g_k(\hat{\phi}_k) \geqslant \eta a n^{(v+\delta)}/A^2 - \mathrm{O}(n^v) \geqslant C n^{(v+\delta)}$. On the other hand, because $\hat{\phi}_k$ minimizes $g_k(\phi)$, we have $g_k(\hat{\phi}_k) \leqslant g_k(\phi_1)$. But $g_k(\phi_1) \leqslant \mathrm{O}(n^v)$. This gives rise to a contradiction.

(ii) Using part (i), it is easy to argue that $h_k(\hat{\phi}_k) = \mathrm{o_p}(1)$ uniformly in $k \in [n_1 a, n_1]$ as long as δ is small. Furthermore, Lemma 2(i) is equivalent to $\sup_{n_1 a \leqslant k \leqslant n_1} |\inf_\phi f_k(\phi)| = \mathrm{O_p}(1)$. The remaining argument is similar to the proof of the previous lemma. □

6. Proofs of Theorems 1–5

The proofs will use Lemmas 2, 3, 5, 7, 9, and 11. For the rest of the proofs, we assume that the infimum with respect to ϕ is taken over a bounded parameter set as stated in assumption A2. We need some preliminary results.

Proposition 1. *If m is bounded and assumptions* A1–A5 *hold, then for every* $\tau > 1/2$,

$$P(|\hat{n}_j - n_j^0| > n^\tau) \to 0 \quad (j = 1, \ldots, m).$$

Proof. Let $A_j = \{(n_1, \ldots, n_m)\colon\ n_1 < n_2 < \cdots < n_m, |n_l - n_j^0| \geqslant n^\tau, 1 \leqslant l \leqslant m\}$. It suffices to assume $\tau \leqslant 3/4$. Since

$$S_n(\hat{n}_1, \ldots, \hat{n}_m) \leqslant S_n(n_1^0, \ldots, n_m^0) \leqslant S_n(n_1^0, \ldots, n_m^0, \beta^0) = \sum_{i=1}^{n} |\varepsilon_i|$$

with probability 1, it suffices to show that

$$\min_{(n_1, \ldots, n_m) \in A_j} S_n(n_1, \ldots, n_m) > \sum_{i=1}^{n} |\varepsilon_i| \tag{29}$$

with probability tending to one as $n \to \infty$. Now, we extend the definition of S_n to every subset $\{n_1, \ldots, n_l\}$ of $\{1, \ldots, n-1\}$:

$$S_n(n_1, \ldots, n_l) = \sum_{r=1}^{l+1} \inf_\phi \sum_{i=n_{(r-1)}+1}^{n_{(r)}} |y_i - x_i'\phi|,$$

where $n_{(0)}=0, n_{(l+1)}=n$ and $0<n_{(1)}<\cdots<n_{(l)}<n$ is the ordered version of $n_1,\ldots,n_l$. For $(n_1,\ldots,n_m)\in A_j$

$$S_n(n_1,\ldots,n_m)\geqslant S_n(n_1,\ldots,n_m,n_1^0,\ldots,n_{j-1}^0,n_j^0-[n^\tau],n_j^0+[n^\tau],n_{j+1}^0,\ldots,n_m^0).$$

The right-hand side of the above can be expressed as $S_{n1}+S_{n2}$, where S_{n1} is the sum of at most $2(m+1)$ expressions of the form $\inf_\phi \sum_{i=l}^k |y_i-x_i'\phi|$, where l and k fall in a common true regime (i.e., $n_r^0\leqslant l<k\leqslant n_{r+1}^0$ for some r); and S_{n2} is given by

$$S_{n2}=\inf_\phi \sum_{n_j^0-[n^\tau]+1}^{n_j^0+[n^\tau]} |y_i-x_i'\phi|, \tag{30}$$

which can be rewritten as

$$S_{n2}=\inf_\phi\left\{\sum_{n_j^0-[n^\tau]+1}^{n_j^0}|\varepsilon_i-x_i'(\phi-\beta_j^0)|+\sum_{n_j^0+1}^{n_j^0+[n^\tau]}|\varepsilon_i-x_i'(\phi-\beta_{j+1}^0)|\right\}. \tag{31}$$

When l and k fall in a common true regime,

$$\inf_\phi\sum_{i=l}^k|y_i-x_i'\phi|=\inf_\phi\sum_{i=l}^k|\varepsilon_i-x_i'\phi|.$$

Thus

$$S_n(n_1,\ldots,n_m)-\sum_{i=1}^n|\varepsilon_i|\geqslant S_{n1}+S_{n2}-\sum_{i=1}^n|\varepsilon_i|$$

$$\geqslant -|2(m+1)\sup_{1\leqslant l<k\leqslant n}\left|\inf_\phi\sum_{i=l}^k(|\varepsilon_i-x_i'\phi|-|\varepsilon_i|)\right| \tag{32}$$

$$+\inf_\phi\left\{\sum_{n_j^0-[n^\tau]+1}^{n_j^0}(|\varepsilon_i-x_i'(\phi-\beta_j^0)|-|\varepsilon_i|)+\sum_{n_j^0+1}^{n_j^0+[n^\tau]}(|\varepsilon_i-x_i'(\phi-\beta_{j+1}^0)|-|\varepsilon_i|)\right\}. \tag{33}$$

From Lemma 3 and the boundedness of m, expression (32) is bounded by $O_p(n^\alpha)$ for every $\alpha>1/2$. Note that $\max\{|\phi-\beta_j^0|,|\phi-\beta_{j+1}^0|\}\geqslant(|\phi-\beta_j^0|+|\phi-\beta_{j+1}^0|)/2\geqslant|\beta_j^0-\beta_{j+1}^0|/2$. Thus if $|\phi-\beta_j^0|$ is bounded away from zero, then we can apply Lemma 9 to the first sum in Eq. (33), applied with the data order reversed (treating n_j^0 as the first observation, n^τ as n, and x_i as w_i). All conditions of the lemma are satisfied. If $|\phi-\beta_{j+1}^0|$ is bounded away from zero, then we can apply Lemma 9 to the second sum in Eq. (33), treating n_j^0+1 as the first observation. In each case, Lemma 9 implies that, for some $\eta>0$, Eq. (33) is larger than $[n^\tau]\eta$ with probability tending to one. Therefore $S_{n1}+S_{n2}>\sum_{i=1}^n|\varepsilon_i|+[n^\tau]\eta-O(n^\alpha)>\sum_{i=1}^n|\varepsilon_i|$ with probability tending to one for $\alpha\in(1/2,\tau)$. This proves Eq. (29) and hence the proposition. □

The rate of convergence given in the previous proposition can be improved upon under the additional assumption A6, even if the number of change points $m_n\to\infty$.

Proposition 2. *Under assumptions* A1–A6, *there exists a* $\delta>0$ *such that*

$$P\left(\sup_{1\leqslant j\leqslant m_n}|\hat{n}_j-n_j^0|>n^{1/(4+\delta)}\right)\to 0.$$

This proposition gives a uniform rate of convergence for bounded or unbounded m_n.

Proof. The argument is similar to that of Proposition 1, with Lemma 7 in place of Lemma 3 in the proof. For a $\delta>0$ (to be determined later), define $A_j=\{(n_1,\ldots,n_{m_n})$: $n_1<n_2<\cdots<n_{m_n},\ |n_l-n_j^0|\geqslant n^{1/(4+\delta)}, 1\leqslant l\leqslant m_n\}$. Then

$$P\left(\sup_{1\leqslant j\leqslant m_n}|\hat{n}_j-n_j^0|>n^{1/(4+\delta)}\right)\leqslant\sum_{j=1}^{m_n}P(|\hat{n}_j-n_j^0|>n^{1/(4+\delta)})$$
$$\leqslant\sum_{j=1}^{m_n}P\left(\inf_{A_j}S_n(n_1,\ldots,n_{m_n})-\sum_{i=1}^n|\varepsilon_i|\leqslant 0\right). \tag{34}$$

Using the previous arguments, we have [cf. Eqs. (32) and (33)]

$$\begin{aligned}
&S_n(n_1,\ldots,n_{m_n})-\sum_{i=1}^n|\varepsilon_i|\\
&\quad\geqslant -|2(m_n+1)\sup_{1\leqslant l<k\leqslant n}\left|\inf_\phi\sum_{i=l}^k(|\varepsilon_i-x_i'\phi|-|\varepsilon_i|)\right|\\
&\qquad+\inf_\phi\left\{\sum_{n_j^0-[n^{1/(4+\delta)}]+1}^{n_j^0}(|\varepsilon_i-x_i'(\phi-\beta_j^0)|-|\varepsilon_i|)\right.\\
&\qquad\left.+\sum_{n_j^0+1}^{n_j^0+[n^{1/(4+\delta)}]}(|\varepsilon_i-x_i'(\phi-\beta_{j+1}^0)|-|\varepsilon_i|)\right\}\\
&\quad\stackrel{\text{def}}{=}-2(m_n+1)U_n+V_{nj}.
\end{aligned}$$

Thus

$$\begin{aligned}
&P\left(\inf_{A_j}S_n(n_1,\ldots,n_{m_n})-\sum_{i=1}^n|\varepsilon_i|>0\right)\\
&\quad\geqslant P(V_{nj}>2(m_n+1)U_n)\\
&\quad\geqslant P(V_{nj}>2(m_n+1)U_n,\ U_n\leqslant n^\varepsilon)
\end{aligned}$$

$$\geqslant P(V_{nj} > 2(m_n+1)n^{\varepsilon},\ U_n \leqslant n^{\varepsilon})$$
$$\geqslant P(V_{nj} > 2(m_n+1)n^{\varepsilon}) + P(U_n \leqslant n^{\varepsilon}) - 1$$
$$\geqslant P(V_{nj} > 3n^{d+\varepsilon}) + P(U_n \leqslant n^{\varepsilon}) - 1$$

where $d < 1/4$ by the assumption on m_n. The fourth inequality follows from $P(A \cap B) \geqslant P(A) + P(B) - 1$. Lemma 7 implies that for every $\varepsilon > 0$ and $t > 0$, $P(U_n > n^{\varepsilon}) = O(n^{-t})$ for large n. Lemma 9 implies that [see the argument for Eq. (33)] there exists an $\eta > 0$ such that $P(V_{nj} \geqslant \eta n^{1/(4+\delta)}) \geqslant 1 - \exp(-n^{1/(4+\delta)}C)$, for some $C > 0$. Now, because $d < 1/4$, we can choose $\varepsilon > 0$ such that $d + \varepsilon < 1/4$. Furthermore, choose $\delta > 0$ such that $d + \varepsilon < 1/(4+\delta)$. Then, for every $\eta > 0$, $n^{d+\varepsilon} \leqslant \eta n^{1/(4+\delta)}$ for all large n. Thus

$$P(V_{nj} > 3n^{d+\varepsilon}) \geqslant P(V_{nj} \geqslant \eta n^{1/(4+\delta)}) \geqslant 1 - \exp(-n^{1/(4+\delta)}C).$$

Note that the constant C can be chosen independent of j because of the i.i.d. assumption and $\max\{|\phi - \beta_j^0|, |\phi - \beta_{j+1}|\} \geqslant |\beta_j^0 - \beta_{j+1}^0|/2 \geqslant c > 0$ for all j, by A2. This implies that, uniformly in j,

$$P\left(\inf_{A_j} S_n(n_1,\ldots,n_{m_n}) - \sum_{i=1}^{n}|\varepsilon_i| > 0\right) \geqslant (1 - \exp(-n^{1/(4+\delta)}C)) + (1 - O(n^{-t})) - 1,$$

which is $1 - O(n^{-t})$. Equivalently, uniformly in $j \leqslant m_n$,

$$P\left(\inf_{A_j} S_n(n_1,\ldots,n_{m_n}) - \sum_{i=1}^{n}|\varepsilon_i| \leqslant 0\right) \leqslant O(n^{-t})$$

for every $t > 0$ for large n. It follows from Eq. (34) that

$$P\left(\sup_j |\hat{n}_j - n_j^0| > n^{1/(4+\delta)}\right) \leqslant m_n O(n^{-t}) \to 0.$$

The proof of Proposition 2 is complete. □

The result of Proposition 1 can be further improved upon.

Proposition 3. *If m is bounded and assumptions* A1–A5 *hold, then for every $\varepsilon > 0$ and for all large n*

$$P(|\hat{n}_j - n_j^0| > \log^2 n) < \varepsilon \quad (j = 1,\ldots,m).$$

Proof. Let $B = \{(n_1,\ldots,n_m)\colon |n_s - n_s^0| < n^{\tau},\ 1 \leqslant s \leqslant m\}$ for some $\tau \in (1/2, 3/4)$. Let B_j be a subset of B such that

$$B_j = \{(n_1,\ldots,n_m)\colon |n_j - n_j^0| > \log^2 n,\ |n_s - n_s^0| < n^{\tau},\ 1 \leqslant s \leqslant m\}.$$

By Proposition 1, $P(\{\hat{n}_1,\ldots,\hat{n}_m\} \in B) \to 1$. To prove Proposition 3, we show $P((\hat{n}_1,\ldots,\hat{n}_m) \in B_j) \to 0$, which is implied by the following:

$$\min_{(n_1,\ldots,n_m)\in B_j} S_n(n_1,\ldots,n_m) > \sum_{i=1}^{n}|\varepsilon_i| \tag{35}$$

with probability tending to 1. For $(n_1,\dots,n_m)\in B_j$,

$$S_n(n_1,\dots,n_m)\geqslant S_n(n_1,\dots,n_m,n_1^0,\dots,n_{j-1}^0,n_j^0-[\log^2 n],n_j^0+[\log^2 n],n_{j+1}^0,\dots,n_m^0)$$
$$\stackrel{\text{def}}{=} T_n(n_1,\dots,n_m)=T_n(\mathscr{P}).$$

Thus to prove Proposition 3, it is sufficient to show, with probability tending to 1,

$$\min_{(n_1,\dots,n_m)\in B_j} T_n(n_1,\dots,n_m)>\sum_{i=1}^{n}|\varepsilon_i|. \tag{36}$$

Let us introduce some terminology for ease of exposition. The diameter of $(l,k]$, denoted by $D(l,k)$, is defined as the sum of least absolute deviations for observations $i\in[l+1,k]$. That is, $D(l,k)=\inf_\phi\sum_{i=l+1}^{k}|y_i-x_i'\phi|$. The diameter of $(l,k]$ relative to a partition $\mathscr{P}=(n_1,\dots,n_m)$, denoted by $D(l,k,\mathscr{P})$, is defined as the sum of all the diameters of the form $(l,k]\cap(n_s,n_{s+1}]$ $(s=0,1,\dots,m)$. The diameter of an empty set is defined to be zero.

Because the length of each true regime is no smaller than $n^{3/4}$ and because $\tau<3/4$, it is clear that for each partition $\mathscr{P}=(n_1,\dots,n_m)\in B$, $(n_s^0,n_{s+1}^0]$ contains at most the two integers n_s and n_{s+1} of $\mathscr{P}$. Thus, $D(n_s^0,n_{s+1}^0,\mathscr{P})$ can be written as the sum of at most three diameters of subsets of $(n_s^0,n_{s+1}^0]$. Namely, if $n_s^0<n_s<n_{s+1}<n_{s+1}^0$, then $D(n_s^0,n_{s+1}^0,\mathscr{P})=D(n_s^0,n_s)+D(n_s,n_{s+1})+D(n_{s+1},n_{s+1}^0)$. Generally, let $r_s=\max\{n_s^0,n_s\}$, $r_{s+1}=\min\{n_{s+1}^0,n_{s+1}\}$, we have $D(n_s^0,n_{s+1}^0,\mathscr{P})=D(n_s^0,r_s)+D(r_s,r_{s+1})+D(r_{s+1},n_{s+1}^0)$ with the convention that $D(l,k)=0$ for $k\leqslant l$.

Given these preparations, we see that $T_n(\mathscr{P})$ can be written as:

$$\sum_{s\neq j-1,j} D(n_s^0,n_{s+1}^0,\mathscr{P})+D(n_{j-1}^0,n_j^0-[\log^2 n],\mathscr{P})+D(n_j^0+[\log^2 n],n_{j+1}^0,\mathscr{P}) \tag{37}$$

$$+D(n_j^0-[\log^2 n],n_j^0+[\log^2 n],\mathscr{P}). \tag{38}$$

Because the diameter of $(n_s^0,n_{s+1}^0]$ relative to $\mathscr{P}\in B_j$ involves observations from a common true regime, it can be written as

$$D(n_s^0,n_{s+1}^0,\mathscr{P})=\inf_\phi\sum_{n_s^0+1}^{r_s}|\varepsilon_i-x_i'\phi|+\inf_\phi\sum_{r_s+1}^{r_{s+1}}|\varepsilon_i-x_i'\phi|+\inf_\phi\sum_{r_{s+1}+1}^{n_{s+1}^0}|\varepsilon_i-x_i'\phi|. \tag{39}$$

That is, we can replace y_i by ε_i. All of the diameters in Eq. (37) have expressions similar to Eq. (39). The diameter in Eq. (38), however, involves observations from two different true regimes and hence it has an expression given by S_{n2} in Eq. (31) with $[n^\tau]$ replaced by $[\log^2 n]$. Now the difference between $T_n(\mathscr{P})$ and $\sum_{i=1}^{n}|\varepsilon_i|$ can be written as

$$T_n(\mathscr{P})-\sum_{i=1}^{n}|\varepsilon_i|=\sum_{s\neq j-1,j}\left(D(n_s^0,n_{s+1}^0,\mathscr{P})-\sum{}''|\varepsilon_i|\right) \tag{40}$$

$$+D(n_{j-1}^0,n_j^0-[\log^2 n],\mathscr{P})-\sum{}''|\varepsilon_i| \tag{41}$$

$$+ D(n_j^0 + [\log^2 n], n_{j+1}^0, \mathscr{P}) - \sum{}'' |\varepsilon_i| \tag{42}$$

$$+ D(n_j^0 - [\log^2 n], n_j^0 + [\log^2 n], \mathscr{P}) - \sum{}'' |\varepsilon_i|, \tag{43}$$

where $\sum''$ extends over the range over which the preceding diameter is defined. For example, the first sum $\sum''$ means $\sum_{n_s^0+1}^{n_{s+1}^0}$. Next we shall show that Eqs. (40)–(42) are all bounded by $O_p(\log n)$ uniformly in $\mathscr{P} \in B_j$, whereas Eq. (43) is larger than $\eta \log^2 n$, for some $\eta > 0$, with probability tending to 1. To this end, for $s \neq j-1, j$, by Eq. (39)

$$D(n_s^0, n_{s+1}^0, \mathscr{P}) - \sum_{n_s^0+1}^{n_{s+1}^0} |\varepsilon_i| = \left(\inf_{\phi} \sum_{n_s^0+1}^{r_s} |\varepsilon_i - x_i'\phi| - |\varepsilon_i| \right) \tag{44}$$

$$+ \left(\inf_{\phi} \sum_{r_s+1}^{r_{s+1}} |\varepsilon_i - x_i'\phi| - |\varepsilon_i| \right) + \left(\inf_{\phi} \sum_{r_{s+1}+1}^{n_{s+1}^0} |\varepsilon_i - x_i'\phi| - |\varepsilon_i| \right). \tag{45}$$

By Lemma 2(ii) (treating $n_s^0 + 1$ as the first observation), the term on the right of Eq. (44) is uniformly bounded in absolute value by $O_p(\log n)$ as r_s varies. Similarly, the second term of Eq. (45) is also uniformly bounded by $O_p(\log n)$ by Lemma 2(ii) (applied with the data order reversed and treating n_{s+1}^0 as the first observation). What is less obvious is that the first term of Eq. (45) is also bounded by $O_p(\log n)$. This is because r_s and r_{s+1} are not arbitrary, the interval $(r_s, r_{s+1}]$ must include $r_s^0 = [(n_s^0 + n_{s+1}^0)/2]$ by the definition of B_j, r_s and r_{s+1}. Thus we can break up the sum into two pieces with one piece summing over $(r_s, r_s^0]$ and the other summing over $(r_s^0, r_{s+1}]$. In this way Lemma 2(ii) can be applied to each piece (r_s^0 does not vary when r_s and r_{s+1} vary). Because m is bounded, the number of diameters in Eq. (40) is bounded. Thus Eq. (40) is bounded by $O_p(\log n)$.

Similarly, both Eqs. (41) and (42) are bounded uniformly on B_j by $O_p(\log n)$.

Next consider Eq. (43), which can be written as (see Eq. (31), replacing $[n^\tau]$ by $[\log^2 n]$):

$$\inf_{\phi} \left\{ \sum_{n_j^0-[\log^2 n]+1}^{n_j^0} (|\varepsilon_i - x_i'(\phi - \beta_j^0)| - |\varepsilon_i|) + \sum_{n_j^0+1}^{n_j^0+[\log^2 n]} (|\varepsilon_i - x_i'(\phi - \beta_{j+1}^0)| - |\varepsilon_i|) \right\}. \tag{46}$$

Because $\max\{|\phi - \beta_j^0|,\ |\phi - \beta_{j+1}^0|\} \geqslant |\beta_j^0 - \beta_{j+1}^0|/2$ for all ϕ, Lemma 9 implies that Eq. (46) is larger than $\eta \log^2 n$ for some $\eta > 0$, with probability tending to one [see the detailed argument concerning Eq. (33)]. Thus

$$\min_{(n_1,\ldots,n_m)\in B_j} T_n(n_1, \ldots, n_m) - \sum_{i=1}^{n} |\varepsilon_i| > -|O_p(\log n)| + \eta \log^2 n > 0$$

with probability tending to 1. Therefore Eq. (36) is proved and so is the proposition.

Proof of Theorems 1 and 2. Write $m = m(n)$. Define $G = \{(n_1, \ldots, n_m)\colon |n_k - n_k^0| \leqslant n^v,\ 1 \leqslant k \leqslant m\}$, where $v < 1/4$. For each fixed j and $C < \infty$ define $G_j(C)$ to be a subset

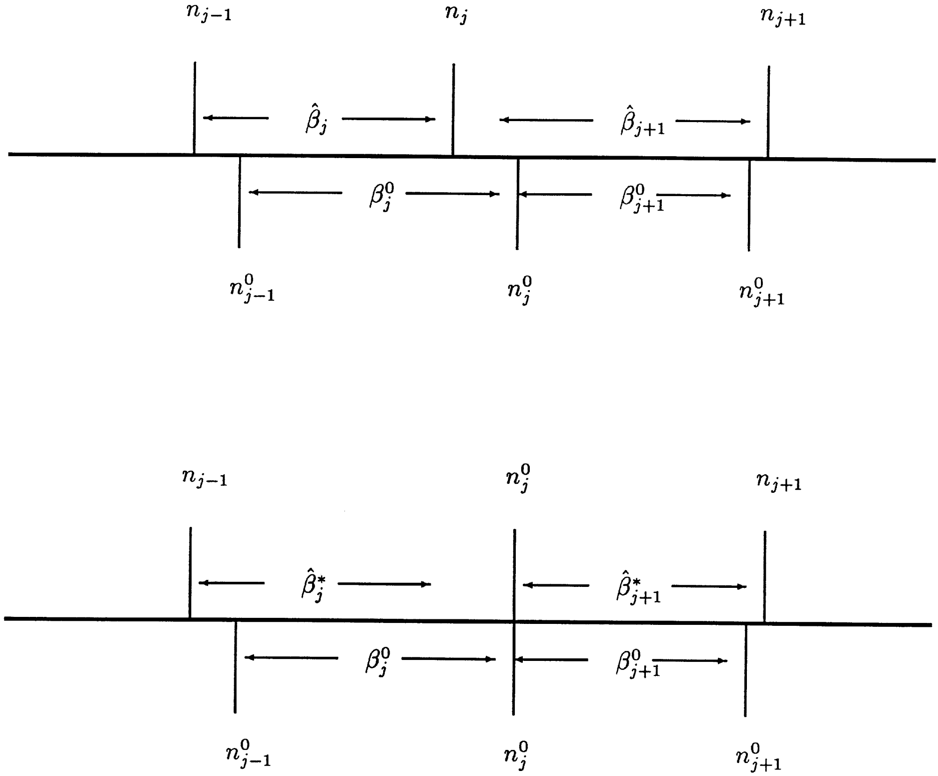

Fig. 1. A particular configuration of $(n_1, n_2, \ldots, n_m)$ in the set $G_j(C)$ defined in the proof of Theorems 1 and 2.

of G such that $G_j(C) = \{(n_1, \ldots, n_m) \in G;\ n_j < n_j^0 - C\}$. In $G_j(C)$, $n_j < n_j^0$; the case of $n_j > n_j^0$ is similar and is omitted. By Propositions 2 and 3, $P((\hat n_1, \ldots, \hat n_m) \in G) \to 1$. To prove the theorems, it suffices to show that for each $\varepsilon > 0$, $P((\hat n_1, \ldots, \hat n_m) \in G_j(C)) < \varepsilon$ for all large C and large n $(j = 1, \ldots, m)$. Because $(\hat n_1, \ldots, \hat n_m)$ must satisfy

$$S_n(\hat n_1, \ldots, \hat n_j, \ldots, \hat n_m) \leqslant S_n(\hat n_1, \ldots, \hat n_{j-1}, n_j^0, \hat n_{j+1}, \ldots, \hat n_m),$$

to show that $(\hat n_1, \ldots, \hat n_m)$ is not in $G_j(C)$, it suffices to show

$$\min_{(n_1, \ldots, n_m) \in G_j(C)} [S_n(n_1, \ldots, n_j, \ldots, n_m) - S_n(n_1, \ldots, n_{j-1}, n_j^0, n_{j+1}, \ldots, n_m)] > 0 \qquad (47)$$

with large probability for large C.

For a fixed j, let $\hat\beta_j$ be the LAD estimator based on observations $(n_{j-1}, n_j]$, viz., $\hat\beta_j = \arg\min_\beta \sum_{n_{j-1}+1}^{n_j} |y_i - x_i'\beta|$. Similarly, let $\hat\beta_{j+1}$ be the LAD estimator based on observations $(n_j, n_{j+1}]$. For notational simplicity, we omit the dependence of $\hat\beta_j$ on the partition. Let $\hat\beta_j^*$ and $\hat\beta_{j+1}^*$ be the LAD estimators based on observations $(n_{j-1}, n_j^0]$ and $(n_j^0, n_{j+1}]$, respectively (see Fig. 1). By the definition of G, $\hat\beta_k$ and $\hat\beta_k^*$ $(k = j, j+1)$ are estimated with at least a positive fraction of $n_k^0 - n_{k-1}^0$ observations belonging to a common true regime (because n_k is close to n_k^0) and with at most $O(n^v)$ observations from another true regime. Note that $n_k^0 - n_{k-1}^0 \geqslant n^\rho$ with $\rho \geqslant 3/4$ by assumption. Thus by Lemma 11(i), we have, for each $\delta \in (0, \rho - v)$, with probability tending to 1,

$$|\hat{\beta}_k - \beta_k^0| \leqslant n^{-(\rho - v - \delta)/2}, \quad k = j, j+1, \tag{48}$$

and similarly,

$$|\hat{\beta}_k^* - \beta_k^0| \leqslant n^{-(\rho - v - \delta)/2}, \quad k = j, j+1. \tag{49}$$

These inequalities hold uniformly on G. We further assume, for the sake of concreteness, that $n_{j-1} \leqslant n_{j-1}^0$ and $n_{j+1} \geqslant n_{j+1}^0$ (other cases can be analyzed similarly and are actually simpler). For $n_{j-1} \leqslant n_{j-1}^0$ and $n_{j+1} \geqslant n_{j+1}^0$ (see Fig. 1),

$$\begin{aligned} &S_n(n_1, \ldots, n_j, \ldots, n_m) \\ &\quad = \sum_{k=1}^{j-1} D(n_{k-1}, n_k) + \sum_{n_{j-1}+1}^{n_{j-1}^0} |\varepsilon_i - x_i'(\hat{\beta}_j - \beta_{j-1}^0)| + \sum_{n_{j-1}^0+1}^{n_j} |\varepsilon_i - x_i'(\hat{\beta}_j - \beta_j^0)| \\ &\qquad + \sum_{n_j+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_{j+1} - \beta_j^0)| + \sum_{n_j^0+1}^{n_{j+1}^0} |\varepsilon_i - x_i'(\hat{\beta}_{j+1} - \beta_{j+1}^0)| \\ &\qquad + \sum_{n_{j+1}^0+1}^{n_{j+1}} |\varepsilon_i - x_i'(\hat{\beta}_{j+1} - \beta_{j+2}^0)| + \sum_{k=j+2}^{m} D(n_{k-1}, n_k) \\ &\quad \stackrel{\text{def}}{=} \sum_{k=1}^{j-1} D(n_{k-1}, n_k) + a + b + c + d + e + \sum_{k=j+2}^{m} D(n_{k-1}, n_k), \end{aligned} \tag{50}$$

where $D(l,k) = \inf_\phi \sum_{i=l+1}^{k} |y_i - x_i'\phi|$, as defined earlier. Similarly,

$$\begin{aligned} &S_n(n_1, \ldots, n_{j-1}, n_j^0, n_{j+1}, \ldots, n_m) \\ &\quad = \sum_{k=1}^{j-1} D(n_{k-1}, n_k) + \sum_{n_{j-1}+1}^{n_{j-1}^0} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_{j-1}^0)| + \sum_{n_{j-1}^0+1}^{n_j} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_j^0)| \\ &\qquad + \sum_{n_j+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_j^0)| + \sum_{n_j^0+1}^{n_{j+1}^0} |\varepsilon_i - x_i'(\hat{\beta}_{j+1}^* - \beta_{j+1}^0)| \\ &\qquad + \sum_{n_{j+1}^0+1}^{n_{j+1}} |\varepsilon_i - x_i'(\hat{\beta}_{j+1}^* - \beta_{j+2}^0)| + \sum_{k=j+2}^{m} D(n_{k-1}, n_k) \\ &\quad \stackrel{\text{def}}{=} \sum_{k=1}^{j-1} D(n_{k-1}, n_k) + a^* + b^* + c^* + d^* + e^* + \sum_{k=j+2}^{m} D(n_{k-1}, n_k). \end{aligned} \tag{51}$$

A major distinction between $S_n(n_1, \ldots, n_m)$ and $S_n(n_1, \ldots, n_j^0, \ldots, n_m)$ lies in the fourth expression on the right hand of each, c and c^*. Expression c involves $\hat{\beta}_{j+1}$ and c^* involves $\hat{\beta}_j^*$; with $\hat{\beta}_{j+1}$ and $\hat{\beta}_j^*$ being estimators of β_{j+1}^0 and β_j^0, respectively. We now consider the difference between Eqs. (50) and (51). First, by the simple inequality (13), Assumption A3, and Eqs. (48) and (49),

$$|a-a^*|\leqslant(n_{j-1}^0-n_{j-1})K|\hat{\beta}_j-\hat{\beta}_j^*|\leqslant 2Kn^{v}n^{-(\rho-v-\delta)/2}=2Kn^{-(\rho-3v-\delta)/2}=\mathrm{o_p}(1)$$

for $\delta\in(0,\rho-3v)$, where $\mathrm{o_p}(1)$ is uniform on G. Similarly, $|e-e^*|=\mathrm{o_p}(1)$ uniformly on G. Next, $b-b^*$ can be written as (by adding and subtracting $|\varepsilon_i|$),

$$b-b^*=\sum_{n_{j-1}^0+1}^{n_j}(|\varepsilon_i-x_i'(\hat{\beta}_j-\beta_j^0)|-|\varepsilon_i|)-\sum_{n_{j-1}^0+1}^{n_j}(|\varepsilon_i-x_i'(\hat{\beta}_j^*-\beta_j^0)|-|\varepsilon_i|). \quad (52)$$

By Lemma 11(ii), each term on the right-hand side is $\mathrm{O_p}(1)$ uniformly on G. To see this, consider the first term on the right. Note that $\hat{\beta}_j$ is estimated with observations $(n_{j-1},n_j]$. But $|n_{j-1}-n_{j-1}^0|\leqslant n^v$ and $|n_{j-1}^0-n_j|\geqslant a|n_{j-1}^0-n_j^0|$ for some $a\in(0,1)$ because n_j is close to n_j^0. Thus, the conditions of Lemma 11 are satisfied (treating $n_j^0-n_{j-1}^0$ as the n_1, and treating $n_j-n_{j-1}^0$ as the k of the lemma). By the same reasoning, $d-d^*=\mathrm{O_p}(1)$ uniformly on G. It remains to deal with c and c^*. Adding and subtracting terms,

$$c-c^*=\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\beta_{j+1}^0-\beta_j^0)|-|\varepsilon_i|) \quad (53)$$

$$+\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\hat{\beta}_{j+1}-\beta_j^0)|)-\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\beta_{j+1}^0-\beta_j^0)|) \quad (54)$$

$$-\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\hat{\beta}_j^*-\beta_j^0)|-|\varepsilon_i|). \quad (55)$$

Expression (54) is bounded by $(n_j^0-n_j)K|\hat{\beta}_{j+1}-\beta_{j+1}^0|\leqslant n^vKn^{-(\rho-v-\delta)/2}=\mathrm{o_p}(1)$. Similarly, Eq. (55) is bounded by $(n_j^0-n_j)K|\hat{\beta}_j^*-\beta_j^0|=\mathrm{o_p}(1)$ [see Eq. (49)]. Expression (53) will be treated later. Summarizing these results, we obtain, uniformly on G,

$$S_n(n_1,\dots,n_m)-S_n(n_1,\dots,n_j^0,\dots,n_m)=\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\beta_{j+1}^0-\beta_j^0)|-|\varepsilon_i|)+\mathrm{O_p}(1). \quad (56)$$

Next, for $(n_1,\dots,n_m)\in G_j(C)\subset G$, we shall show that the r.h.s. term above is large. Because $|\beta_{j+1}^0-\beta_j^0|>0$ is fixed and $n_j^0-n_j\geqslant C$, Lemma 9 implies that the first term on the r.h.s. of Eq. (56) is greater than $\eta(n_j^0-n_j)\geqslant\eta C$ for some $\eta>0$ with probability tending to 1 as C tends to infinity. Thus on $G_j(C)$,

$$\min_{(n_1,\dots,n_m)\in G_j(C)}[S_n(n_1,\dots,n_m)-S_n(n_1,\dots,n_j^0,\dots,n_m)]\geqslant\eta C+\mathrm{O_p}(1).$$

The r.h.s. above is positive with large probability if C is large. This proves Eq. (47) and thus the theorems. □

Proof of Theorem 3. We note that $\hat{\beta}_j(\hat{n}_1,\dots,\hat{n}_m)$ only depends on $\hat{n}_{j-1}$ and $\hat{n}_j$ so we can write it as $\hat{\beta}_j(\hat{n}_{j-1},\hat{n}_j)$. Further note that $\hat{\beta}_j(n_{j-1}^0,n_j^0)$ has the stated limiting distribution [see, e.g., Bassett and Koenker (1978)]. But $\hat{n}_i=n_i^0+\mathrm{O_p}(1)$, thus with large probability, $\hat{\beta}_j(\hat{n}_{j-1},\hat{n}_j)$ is estimated using the same set of observations as $\hat{\beta}_j(n_{j-1}^0,n_j^0)$ with at most a finite number of different observations. A finite number of different observations will not alter the limiting distribution. The proof of Theorem 3 is now complete. □

Proof of Theorem 4. The key to the proof lies in the following fact. Let $(\hat{n}_1,\dots,\hat{n}_m)$ be the jointly estimated change points, where $m=m(n)$, not necessarily bounded. Then for each j, it must be true that

$$\hat{n}_j=\operatorname*{arg\,min}_{1\leqslant n_j\leqslant n} S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j,\hat{n}_{j+1},\dots,\hat{n}_m).$$

This fact effectively transforms the problem into that of a single change point. The above is equivalent to $\hat{n}_j-n_j^0=\arg\min_k S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j^0+k,\hat{n}_{j+1},\dots,\hat{n}_m)$. In view of the rate of convergence of $\hat{n}_j$ given by Theorems 1 and 2, to prove Theorem 4 it suffices to show that, for $|k|\leqslant M$ ($M<\infty$ arbitrarily given)

$$S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j^0+k,\hat{n}_{j+1},\dots,\hat{n}_m)-S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j^0,\hat{n}_{j+1},\dots,\hat{n}_m)\xrightarrow{\mathrm{d}} W^{(j)}(k). \tag{57}$$

Let $n_j=n_j^0+k$. Then $(\hat{n}_1,\dots,\hat{n}_{j-1},n_j,\hat{n}_{j+1},\dots,\hat{n}_m)\in G$ with probability approaching to 1. Thus Eq. (56) implies that, for $k<0$ (the case of $k>0$ is similar and is omitted),

$$\begin{aligned}&S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j,\hat{n}_{j+1},\dots,\hat{n}_m)-S_n(\hat{n}_1,\dots,\hat{n}_{j-1},n_j^0,\hat{n}_{j+1},\dots,\hat{n}_m)\\&\quad=\sum_{n_j+1}^{n_j^0}(|\varepsilon_i-x_i'(\beta_{j+1}^0-\beta_j^0)|-|\varepsilon_i|)+\mathrm{O_p}(1).\end{aligned} \tag{58}$$

The first term on the r.h.s. above has the same distribution as $W^{(j)}(n_j^0-n_j)=W^{(j)}(k)$ under the i.i.d. assumption. Thus Eq. (57) and hence Theorem 4 will be proved if the $\mathrm{O_p}(1)$ term in Eq. (58) can be strengthened to be Eq. $\mathrm{o_p}(1)$, under $|n_j-n_j^0|\leqslant M$. Note that the $\mathrm{O_p}(1)$ term represents $b-b^*$ and $d-d^*$ defined in the previous proof. We next show that $b-b^*=\mathrm{o_p}(1)$. The proof for $d-d^*$ being $\mathrm{o_p}(1)$ is similar.

Let $\hat{\beta}_j$ and $\hat{\beta}_j^*$ be the LAD estimators of β_j^0 based on observations $[\hat{n}_{j-1},n_j]$ and $[\hat{n}_{j-1},n_j^0]$, respectively. Because $|n_j,n_j^0|\leqslant M$ by assumption and $\hat{n}_{j-1}-n_{j-1}^0=\mathrm{O_p}(1)$ by Theorem 2, it follows from the classical result that

$$\hat{\beta}_j-\beta_j^0=n_a^{-1/2}\mathrm{O_p}(1)\quad\text{and}\quad\hat{\beta}_j^*-\beta_j^0=n_a^{-1/2}\mathrm{O_p}(1), \tag{59}$$

where $n_a = n_j^0 - n_{j-1}^0$. Using the Bahadur type of representation (Babu, 1989), we have, by cancelling the common term of representations,

$$|\hat{\beta}_j^* - \hat{\beta}_j| \leqslant n_a^{-3/4} (\log n_a) \mathrm{O_p}(1). \tag{60}$$

We now prove $b - b^* = \mathrm{o_p}(1)$ (cf. Eq. (52)). Replacing n_j by n_j^0 in Eq. (52), which is equivalent to adding an $\mathrm{o_p}(1)$ term $\sum_{n_j}^{n_j^0}(\cdot)$. [The term being $\mathrm{o_p}(1)$ follows because $|\sum_{n_j}^{n_j^0}| \leqslant \sum_{n_j}^{n_j^0} |x_i||\hat{\beta}_j - \hat{\beta}_j^*| \leqslant MK|\hat{\beta}_j - \hat{\beta}_j^*| = \mathrm{o_p}(1)$ by Eq. (60)], we can rewrite $b - b^*$ as

$$\begin{aligned} b - b^* &= \sum_{n_{j-1}^0+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_j - \beta_j^0)| - \sum_{n_{j-1}^0+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_j^0)| + \mathrm{o_p}(1) \\ &= \sum_{n_{j-1}^0+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_j^0) - x_i'(\hat{\beta}_j - \hat{\beta}_j^*)| - \sum_{n_{j-1}^0+1}^{n_j^0} |\varepsilon_i - x_i'(\hat{\beta}_j^* - \beta_j^0)| + \mathrm{o_p}(1). \end{aligned}$$

The following lemma together with Eqs. (59) and (60) implies that $b - b^*$ is $\mathrm{o_p}(1)$. □

Lemma 12. *Under* A3 *and* A5, *for every* $L < \infty$,

$$\sup_{|\phi_1| < L, |\phi_2| < L} \left| \sum_{i=1}^n (|\varepsilon_i - x_i'\phi_1 n^{-1/2} - x_i'\phi_2 n^{-3/4} \log n| - |\varepsilon_i - x_i'\phi_1 n^{-1/2}|) \right| = \mathrm{o_p}(1).$$

Proof. Denote the ith summand by $\psi_{in}(\theta)$, where $\theta = (\phi_1, \phi_2)$. From $E(|\varepsilon_i - t| - |\varepsilon_i|) = t^2 f(0) + \mathrm{o}(t^2)$, it is easy to verify that $\sum_{i=1}^n E\psi_{in}(\theta) = \mathrm{o}(1)$ uniformly in $|\theta| \leqslant M = 2L$. Thus the lemma will be true if we can prove it with $\psi_{in}(\theta)$ replaced by $\psi_{in}(\theta) - E\psi_{in}(\theta)$. From $|\psi_{in}(\theta)| \leqslant |x_i'\phi_2| n^{-3/4} \log n \leqslant KLn^{-3/4} \log n$, we obtain $E[\psi_{in}(\theta)]^2 \leqslant (KL)^2 n^{-3/2} (\log n)^2$. Apply Lemma 1 with $a = \varepsilon$, $s = 1/2$, $V = (KL)^2 n^{-1/2} (\log n)^2$, we obtain, for each fixed θ,

$$P\left(\left| \sum_{i=1}^n [\psi_{in}(\theta) - E\psi_{in}(\theta)] \right| > \varepsilon \right) \leqslant 2 \exp(-\varepsilon^2 C n^{1/2} / (\log n)^2)$$

for some $C > 0$. Next divide the region $|\theta| \leqslant M$ into $\mathrm{O}(n^p)$ cells such that for θ', θ'' belonging to a common cell, $|\theta' - \theta''| \leqslant Mn^{-1/2}$. In this way, the incremental value,

$$\begin{aligned} \left| \sum_{i=1}^n (\psi_{in}(\theta') - E\psi_{in}(\theta') - \psi_{in}(\theta'') + E\psi_{in}(\theta'')) \right| &\leqslant \sum_{i=1}^n |x_i'(\phi_2' - \phi_2'')| n^{-3/4} \log n \\ &\leqslant 2KMn^{-1/4} \log n = \mathrm{o}(1). \end{aligned}$$

Thus

$$P\left(\sup_{|\theta|\leqslant M}\left|\sum_{i=1}^{n}\psi_{in}(\theta)-E\psi_{in}(\theta)\right|>\varepsilon\right)\leqslant \mathrm{O}(n^{p})\exp(-\varepsilon^{2}Cn^{1/2}/(\log n)^{2})\to 0,$$

proving the lemma. □

Proof of Theorem 5. We first prove $P(\hat{m}<m_n^0)\to 0$ by arguing that $P(\inf_{m<m_n^0}B(m)-B(m_n^0)\leqslant 0)\to 0$. Write $m_0=m_n^0$. First note that when $m<m_0$, there must exist at least one change point that cannot be estimated. Because each of the regime lengths is at least $n^{3/4}$, there exists a segment $[l,k]$ which contains no estimated change point and satisfies $n_j^0-l\geqslant n^{3/4}/2$ and $k-n_j^0\geqslant n^{3/4}/2$ for some n_j^0. Using a similar argument as in proving Eq. (33) and Lemma 9, we can show that, for some $C>0$,

$$S_n(\hat{n}_1,\ldots,\hat{n}_m)-\sum_{i=1}^{n}|\varepsilon_i|>Cn^{3/4} \tag{61}$$

with probability tending to 1. Next,

$$B(m)-B(m_0)=n\log\left(1+\frac{\hat{e}(m)-\hat{e}(m_0)}{\hat{e}(m_0)}\right)+(m-m_0)g(n).$$

Without loss of generality, we may assume that $|\hat{e}(m)-\hat{e}(m_0)|/\hat{e}(m_0)$ is small (if it is large, it is even less unlikely for $B(m)\leqslant B(m_0)$). Using $\log(1+x)\sim x$,

$$\begin{aligned}B(m)-B(m_0)&\sim n[\hat{e}(m)-\hat{e}(m_0)]/\hat{e}(m_0)+(m-m_0)g(n)\\&=\left[S_n(\hat{n}_1,\ldots,\hat{n}_m)-\sum_{i=1}^{n}|\varepsilon_i|-\left\{S_n(\hat{n}_1,\ldots,\hat{n}_{m_0})-\sum_{i=1}^{n}|\varepsilon_i|\right\}\right]\Big/\hat{e}(m_0)\\&\quad+(m-m_0)g(n).\end{aligned} \tag{62}$$

We need the following lemma:

Lemma 13. *Under* A1–A6,

$$0\geqslant S_n(\hat{n}_1,\ldots,\hat{n}_{m_0})-\sum_{i=1}^{n}|\varepsilon_i|=\mathrm{O_p}(n^{1/4}).$$

Proof. From $S_n(\hat{n}_1,\ldots,\hat{n}_m)\geqslant S_n(\hat{n}_1,\ldots,\hat{n}_m,n_1^0,\ldots,n_{m_0}^0)$ for all m, we have

$$\sum_{i=1}^{n}|\varepsilon_i|-S_n(\hat{n}_1,\ldots,\hat{n}_m)\leqslant(m+m_0+1)U_n,\quad\forall m \tag{63}$$

where $U_n=\sup_{1\leqslant l<k\leqslant n}|\inf_{\phi}\sum_{i=l}^{k}|\varepsilon_i-x_i'\phi|-|\varepsilon_i||$. Thus, for $m=m_0$,

$$P\left(\sum_{i=1}^{n}|\varepsilon_i|-S_n(\hat{n}_1,\ldots,\hat{n}_{m_0})>n^{1/4}\right)\leqslant P(3m_0U_n>n^{1/4})=P(3U_n>n^{1/4}/m_0).$$

The above probability converges to zero by Lemma 7 because $n^{1/4}/m_0 > n^a$ for some $a>0$. □

The lemma implies that $\hat{e}(m_0) = n^{-1}\sum_{i=1}^n |\varepsilon_i| + O_p(n^{-3/4}) \to E(|\varepsilon_1|)$. From Eqs. (61), (62) and Lemma 13, we have, for all large n

$$B(m) - B(m_0) > Cn^{3/4} - O_p(n^{1/4}) + (m - m_0)g(n) \geqslant 2^{-1}Cn^{3/4} - m_0 g(n) > 0$$

because $m_0 g(n) \leqslant c_2 n^{(1/4)-b} n^{1/2} = o(n^{3/4})$ by A1 and $g(n) = n^{1/2}$. This implies that $P(\hat{m} < m_0) \to 0$.

Next consider $m > m_0$. We assume $m < Lm_0$ for a large given L. For $m > m_0$,

$$\begin{aligned} S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) &\geqslant S_n(\hat{n}_1, \ldots, n_m) \\ &= S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) + \left\{ S_n(\hat{n}_1, \ldots, \hat{n}_m) - \sum_{i=1}^n |\varepsilon_i| \right\} - \left\{ S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) - \sum_{i=1}^n |\varepsilon_i| \right\} \\ &\geqslant S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) - (Lm_0 + m_0 + 1)U_n - (2m_0 + 1)U_n \end{aligned}$$

where the last inequality follows from Eq. (63). By Lemma 7, $U_n = O_p(n^{\varepsilon})$ for every $\varepsilon > 0$. Choose a small $\varepsilon > 0$ such that $m_0 U_n = O_p(n^{1/4})$, we have

$$S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) \geqslant S_n(\hat{n}_1, \ldots, \hat{n}_m) \geqslant S_n(\hat{n}_1, \ldots, \hat{n}_{m_0}) - O_p(n^{1/4}).$$

Divide by n on both sides above to obtain

$$0 \leqslant \hat{e}(m_0) - \hat{e}(m) = O_p(n^{-3/4}).$$

Thus

$$n \log \hat{e}(m_0) - n \log \hat{e}(m) = -n \log \left(1 + \frac{\hat{e}(m) - \hat{e}(m_0)}{\hat{e}(m_0)} \right) = O_p(n^{1/4}).$$

Because $g(n)/n^{1/4} \to \infty$,

$$n \log \hat{e}(m_0) - n \log \hat{e}(m) = O_p(n^{1/4}) < g(n) \leqslant (m - m_0)g(n)$$

for all $m > m_0$. That is, for $m > m_0$,

$$n \log \hat{e}(m) + mg(n) > n \log \hat{e}(m_0) + m_0 g(n)$$

for all large n. This implies that $P(\hat{m} > m_0) \to 0$. □

References

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

Bai, J., 1994. Least squares estimation of a shift in linear processes. J. Time Ser. Anal. 15, 453–472.

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

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66, 47–78.

Barrodale, I., Roberts, F.D.K., 1974. Algorithm 478: solution of an overdetermined system of equations in the L_1 norm. Comm. ACM 17, 319–320.

Bassett, G., Koenker, K., 1978. Asymptotic theory of least absolute error regression. J. Amer. Statist. Assoc. 73, 618–622.

Bhattacharya, P.K., 1987. Maximum likelihood estimation of a change-point in the distribution of independent random variables, General multiparameter case. J. Multi. Anal. 23, 183–208.

Brodsky, B.E., Darkhovsky, B.S., 1993. Nonparametric Methods in Change Point Problems. Kluwer, Dordrecht.

Dueker, M.J., 1992. The response of market interest rates to discount rate changes. Federal Reserve Bank of St. Louis Review 74 (4), 78–91.

Dümbgen, L., 1991. The asymptotic behavior of some nonparametric change point estimators. Ann. Statist. 19, 1471–1495.

Gombay, E., Horváth, L., 1994. Limit theorems for change in linear regression. J. Multivariate Anal. 48, 43–69.

Guthery, S.B., 1974. Partition regression. J. Ameri. Stat. Assoc. 69, 945–947.

Horváth, L., 1995. Detecting changes in linear regressions. Statistics 26, 189–208.

Horváth, L., Hušková, M., Serbinowska, M., 1997. Estimators for the time of change in linear models. Statistics 29, 109–130.

Huang, W.-T., Chang, Y.-P., 1993. Nonparametric estimation in change-point models. J. Statist. Plann. Inference 35, 335–347.

Hušková, M., 1996a. Tests and estimators for change point based on M-statistics. Statist. Decisions 14, 115–136.

Hušková, M., 1996b. Estimation of a change in linear models. Statist. Probab. Lett. 26, 13–24.

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

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

Picard, D., 1985. Testing and estimating change-points in time series. Adv. Appl. Probab. 17, 841–867.

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

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

Quandt, R.E., 1958. The estimation of parameters of a linear regression system obeying two separate regimes. J. Amer. Statist. Assoc. 53, 873–880.

Roley, V., Wheatley, S.M., 1990. Temporal variation in the interest rate response to money announcements. NBER No. 3471.

Shaban, S.A., 1980. Change point problem and two-phase regression: an annotated bibliography. Internat. Statist. Rev. 48, 83–93.

Yao, Y.-C., 1988. Estimating the number of change-points via Schwarz' criterion. Statist. Prob. Lett. 6, 181–189.

Yao, Y.-C., Au, S.T., 1989. Least squares estimation of a step function. Sankhyā 51 Ser. A 370–381.

Yin, Y.Q., 1988. Detection of the number, locations, and magnitudes of jumps. Comm. Stat. Stoch. Models 4, 445–455.

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