Econometric Theory, **13**, 1997, 315–352. Printed in the United States of America.

ESTIMATING MULTIPLE BREAKS ONE AT A TIME

JUSHAN BAI
Massachusetts Institute of Technology

Sequential (one-by-one) rather than simultaneous estimation of multiple breaks is investigated in this paper. The advantage of this method lies in its computational savings and its robustness to misspecification in the number of breaks. The number of least-squares regressions required to compute all of the break points is of order T, the sample size. Each estimated break point is shown to be consistent for one of the true ones despite underspecification of the number of breaks. More interestingly and somewhat surprisingly, the estimated break points are shown to be T-consistent, the same rate as the simultaneous estimation. Limiting distributions are also derived. Unlike simultaneous estimation, the limiting distributions are generally not symmetric and are influenced by regression parameters of all regimes. A simple method is introduced to obtain break point estimators that have the same limiting distributions as those obtained via simultaneous estimation. Finally, a procedure is proposed to consistently estimate the number of breaks.

1. INTRODUCTION

Multiple breaks may exist in the trend function of many economic time series, as suggested by the studies of Burdekin and Siklos (1995), Cooper (1995), Garcia and Perron (1996), Lumsdaine and Papell (1995), and others. This paper presents some theory and methods for making inferences in the presence of multiple breaks with unknown break dates. The focus is the sequential method, which identifies break points one by one as opposed to all simultaneously.

A number of issues arise in the presence of multiple breaks. These include the determination of the number of breaks, estimation of the break points given the number, and statistical analysis of the resulting estimators. These issues were examined by Bai and Perron (1994) when a different approach of estimation is used. The major results of Bai and Perron (1994) assume simultaneous estimation, which estimates all of the breaks at the same time. In this paper, we study an alternative method, which sequentially identifies the break points. The procedure estimates one break point even if multiple breaks exist. The number of least-squares regressions required to compute

I thank two referees and Bruce Hansen for helpful comments. Partial financial support is acknowledged from the National Science Foundation under grant SBR-9414083. Address correspondence to: Jushan Bai, Department of Economics, E52-274B, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; e-mail: jbai@mit.edu.

 0266-4666/97 $9.00 + .10

all of the breaks is proportional to the sample size. Obviously, simultaneous and sequential methods are not merely two different computing techniques; they are fundamentally different methodologies that yield different estimators. Not much is known about sequentially obtained estimators. This paper develops the underlying theory about them.

The method of sequential estimation was proposed independently by Bai and Perron (1994) and Chong (1994). They argued that the estimated break point is consistent for one of the true break points. However, neither of the studies gives the convergence rate of the estimated break point. In fact, the approach used in previous studies does not allow one to study the convergence rate of sequential estimators. A different framework and more detailed analysis are necessary. A major finding of this study is that the sequentially obtained estimated break points are T-consistent, the same rate as in the case of simultaneous estimation. This result is somewhat surprising in that, on first inspection, one might even doubt its consistency, let alone T consistency, in view of the incorrect specification of the number of breaks.

Furthermore, we obtain the asymptotic distribution of the estimated break points. The asymptotic distributions of sequentially estimated break points are found to be different from those of simultaneous estimation. We suggest a procedure for obtaining estimators having the same asymptotic distribution as the simultaneous estimators. We also propose a procedure to consistently estimate the number of breaks. These latter results are made possible by the T consistency. For example, one can construct consistent (but not T-consistent) break-point estimators for which the procedure will overestimate the number of breaks. In this view, the T-consistent result for a sequential estimator is particularly significant.

This paper is organized as follows. Section 2 states the model, the assumptions needed, and the estimation method. The T consistency for the estimated break points is established in Section 3. Section 4 studies a special configuration for the model's parameters that leads to some interesting asymptotic results. Limiting distributions are derived in Section 5. Section 6 proposes a "repartition method" that gives rise to estimators having the same asymptotic distribution as in the case of simultaneous estimation. Section 7 deals with shrinking shifts. Convergence rates and limiting distributions are also derived. Results corresponding to more than two breaks are stated in Section 8. The issue of the number of breaks is also discussed in this section. Section 9 reports the simulation results. Section 10 concludes the paper. Mathematical proofs are relegated to the Appendix.

2. THE MODEL

To present the main idea, we shall consider a simple model with mean shifts in a linear process. The theory and results can be extended to general regression models using a combination of the argument of Bai (1994a) and this paper. To make things even simpler, the presentation and proof will be stated in terms of two breaks. Because of the nature of sequential estimation, analysis in terms of two breaks incurs no loss of generality. This can also be seen from the proof. The general results with more than two breaks will be stated later. The model considered is as follows:

$$\begin{aligned} Y_t &= \mu_1 + X_t, \quad \text{if } t \le k_1^0, \\ Y_t &= \mu_2 + X_t, \quad \text{if } k_1^0 + 1 \le t \le k_2^0, \\ Y_t &= \mu_3 + X_t, \quad \text{if } k_2^0 + 1 \le t \le T, \end{aligned} \tag{1}$$

where μ_i is the mean of regime i ($i = 1,2,3$) and X_t is a linear process of martingale differences such that $X_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j}$; k_1^0 and k_2^0 are unknown break points.

The idea of sequential estimation is to consider one break at a time. The model is treated as if there were only one break point. Estimating one break point for a mean shift in linear processes is studied by Bai (1994b). A single break point can be obtained by minimizing the sum of squared residuals among all possible sample splits. As in Bai (1994b), we denote the mean of the first k observations by $\bar{Y}_k$ and the mean of the last $T - k$ observations by $\bar{Y}_k^*$. The sum of squared residuals is

$$S_T(k) = \sum_{t=1}^{k} (Y_t - \bar{Y}_k)^2 + \sum_{t=k+1}^{T} (Y_t - \bar{Y}_k^*)^2.$$

A break point estimator is defined as $\hat{k} = \operatorname{argmin}_{1 \le k \le T-1} S_T(k)$. Using the formula linking total variance with within-group and between-group variances, we can write, for each k ($1 \le k \le T - 1$),

$$\sum_{t=1}^{T} (Y_t - \bar{Y})^2 = S_T(k) + T V_T(k)^2, \tag{2}$$

where $\bar{Y}$ is the overall mean and

$$V_T(k) = \left(\frac{k(T-k)}{T^2} \right)^{1/2} (\bar{Y}_k^* - \bar{Y}_k). \tag{3}$$

It follows that

$$\hat{k} = \operatorname{argmin}_k S_T(k) = \operatorname{argmax}_k V_T(k)^2 = \operatorname{argmax}_k |V_T(k)|.$$

Consequently, the properties of $\hat{k}$ can be analyzed equivalently by examining $S_T(k)$ or $V_T(k)$. We define $\hat{\tau} = \hat{k}/T$. Both $\hat{\tau}$ and $\hat{k}$ are referred to as "estimated break points." The former is also referred to as an "estimated break fraction."

One of our major results is that $\hat{\tau}$ is T-consistent for one of the true breaks τ_i^0. It should be pointed out, however, that $\hat{k}$ itself is not consistent for either of the k_i^0 ($i = 1,2$). For ease of exposition, we shall frequently say that "$\hat{k}$ is T-consistent" with the understanding that we are actually referring to $\hat{\tau}$.

We need the following assumptions to derive T consistency.

Assumption A1. The ϵ_t are martingale differences satisfying $E(\epsilon_t | \mathcal{F}_{t-1}) = 0$, $E\epsilon_t^2 = \sigma^2$, and there exists a $\delta > 0$ such that $\sup_t E|\epsilon_t|^{2+\delta} < \infty$, where $\mathcal{F}_t$ is the σ-field generated by ϵ_s for $s \le t$.

Assumption A2.

$$X_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j}, \quad \sum_{j=0}^{\infty} j|a_j| < \infty, \quad \text{and} \quad a(1) = \sum_{j=0}^{\infty} a_j \neq 0.$$

Assumption A3. $\mu_i \neq \mu_{i+1}$, $k_i^0 = [T\tau_i^0]$, and $\tau_i^0 \in (0,1)$ $(i = 1,2)$ with $\tau_1^0 < \tau_2^0$.

Assumptions A1 and A2 are standard for linear processes. As the theory for a single-break model is worked out by Bai (1994b), we assume the existence of two breaks in Assumption A3.

3. CONSISTENCY AND RATE OF CONVERGENCE

In this section, a number of useful properties for the sum of squared residuals $S_T(k)$ will be presented. These properties naturally lead to the consistency result. Write $U_T(\tau) = T^{-1}S_T([T\tau])$ for $\tau \in [0,1]$. We define both $S_T(0)$ and $S_T(T)$ as the total sum of squared residuals with the full sample; that is, $S_T(0) = S_T(T) = \sum_{t=1}^{T} (Y_t - \bar{Y})^2$. This definition is also consistent with (2), as $V_T(0) = V_T(T) = 0$. In this way, $U_T(\tau)$ is well defined for all $\tau \in [0,1]$.

LEMMA 1. *Under Assumptions* A1–A3, $U_T(\tau)$ *converges uniformly in probability to a nonstochastic function* $U(\tau)$ *on* $[0,1]$.

The limit $U(\tau)$ is a continuous function and has different expressions over three different regimes. In particular,

$$U(\tau_1^0) = \sigma_X^2 + \frac{(1-\tau_2^0)(\tau_2^0 - \tau_1^0)}{1-\tau_1^0}(\mu_2 - \mu_3)^2 \tag{4}$$

and

$$U(\tau_2^0) = \sigma_X^2 + \frac{\tau_1^0}{\tau_2^0}(\tau_2^0 - \tau_1^0)(\mu_1 - \mu_2)^2, \tag{5}$$

where $\sigma_X^2 = EX_t^2$.

LEMMA 2. *Under Assumptions* A1–A3,

$$\sup_{1 \le k \le T} |U_T(k/T) - EU_T(k/T)| = O_p(T^{-1/2}).$$

This lemma says that the objective function (as a function of k) is uniformly close to its expected function. As a result, if the expected function is minimized at a certain point, then the stochastic function will be minimized at a neighborhood of that point with large probability. To study the extreme value of the expected function, we need an additional assumption, which is stated in terms of the limiting function $U(\tau)$. Note that $U(\tau)$ is also the limit of $T^{-1}ES_T(k)$ for $k = [T\tau]$. Typically, the function $U(\tau)$ has two local minima. To ensure the smallest value of $U(\tau)$ is unique, we assume the following.

Assumption A4. $U(\tau_1^0) < U(\tau_2^0)$.

By (4) and (5), this condition is equivalent to

$$\frac{\tau_1^0}{\tau_2^0}(\mu_1 - \mu_2)^2 > \frac{1-\tau_2^0}{1-\tau_1^0}(\mu_2 - \mu_3)^2. \tag{6}$$

Thus, the condition requires that the first break dominates in terms of the relative span of regimes and the magnitude of shifts. In other words, when the first break is more pronounced (large enough τ_1^0 and/or $|\mu_1 - \mu_2|$), Assumption A4 will be true. The inequality will be reversed when the second break is more pronounced. Under Assumptions A1–A4, the estimated fraction $\hat{\tau}$ converges in probability to τ_1^0 because the sum of squared residuals can be minimized only if the more pronounced break is chosen. If the inequality in Assumption A4 is reversed, then $\hat{\tau}$ converges in probability to τ_2^0 by mere symmetry. In the next section, we examine the case $U(\tau_1^0) = U(\tau_2^0)$, in which $\hat{\tau}$ converges in distribution to a random variable with equal mass at τ_1^0 and τ_2^0.

LEMMA 3. *Under Assumptions* A1–A4, *there exists a* $C > 0$, *depending only on* τ_i^0, *and* μ_j ($i = 1,2, j = 1,2,3$) *such that*

$$ES_T(k) - ES_T(k_1^0) \geq C|k - k_1^0| \quad \textit{for all large } T.$$

The lemma implies that the expected value of the sum of squared residuals is minimized at k_1^0 only. As mentioned earlier, because of the uniform closeness of the objective function to its expected function (Lemma 2), it is reasonable to expect that the minimum point of the stochastic objective function is close to k_1^0 with large probability. Precisely, we have the following result.

PROPOSITION 1. *Under Assumptions* A1–A4,

$$\hat{\tau} - \tau_1^0 = O_p(T^{-1/2}).$$

That is, the estimated break point is consistent for τ_1^0.

Proof.

$$\begin{aligned} S_T(k) - S_T(k_1^0) &= S_T(k) - ES_T(k) - [S_T(k_1^0) - ES_T(k_1^0)] \\ &\quad + ES_T(k) - ES_T(k_1^0) \\ &\geq -2 \sup_{1\leq j\leq T} |S_T(j) - ES_T(j)| + ES_T(k) - ES_T(k_1^0) \\ &\geq -2 \sup_{1\leq j\leq T} |S_T(j) - ES_T(j)| + C|k - k_1^0| \qquad \text{by Lemma 3.} \end{aligned}$$

The preceding holds for all $k \in [1,T]$. In particular, it holds for $\hat{k}$. From $S_T(\hat{k}) - S_T(k_1^0) \leq 0$, we obtain

$$|\hat{k} - k_1^0| \leq C^{-1}2 \sup_{1\leq j\leq T} |S_T(j) - ES_T(j)|.$$

Dividing the preceding inequality by T on both sides and using Lemma 2, we obtain the proposition immediately. ■

The preceding convergence rate is obtained by examining the global behavior of the objective function $S_T(k)$ $(k = 1,\ldots,T)$. This rate can be improved upon by examining $S_T(k)$ for k in a restricted range. Define $D_T = \{k: T\eta \leq k \leq T\tau_2^0(1-\eta)\}$, where η is a small positive number such that $\tau_1^0 \in (\eta, \tau_2^0(1-\eta))$ and $D_M = \{k: |k - k_1^0| \leq M\}$, where $M < \infty$ is a constant. Thus, for each $k \in D_T$, k is both away from 0 and away from the second break point for a positive fraction of observations. By Proposition 1, $\hat{k}$ will eventually fall into D_T. That is, for every $\epsilon > 0$, $P(\hat{k} \notin D_T) < \epsilon$ for all large T. We shall argue that $\hat{k}$ must eventually fall into D_M with large probability for large M, which is equivalent to T consistency.

Let $D_{T,M}$ be the intersection of D_T and the complement of D_M; that is, $D_{T,M} = \{k: T\eta \leq k \leq T\tau_2^0(1-\eta), |k - k_1^0| > M\}$.

LEMMA 4. *Under Assumptions* A1–A4, *for every* $\epsilon > 0$, *there exists an* $M < \infty$ *such that*

$$P\left(\min_{k\in D_{T,M}} S_T(k) - S_T(k_1^0) \leq 0\right) < \epsilon.$$

PROPOSITION 2. *Under Assumptions* A1–A4, *for every* $\epsilon > 0$, *there exists a finite* M *independent of* T *such that, for all large* T,

$$P(T|\hat{\tau} - \tau_1^0| > M) < \epsilon.$$

That is, the break point estimator is T-consistent.

Proof. Because $S_T(\hat{k}) \leq S_T(k_1^0)$, if $\hat{k} \in A$, it must be the case that $\min_{k\in A} S_T(k) \leq S_T(k_1^0)$, where A is an arbitrary subset of integers. Thus,

$$P(|\hat{k} - k_1^0| > M) \le P(\hat{k} \notin D_T) + P(\hat{k} \in D_T, |\hat{k} - k_1^0| > M)$$

$$\le \epsilon + P\left(\min_{k \in D_{T,M}} S_T(k) - S_T(k_1^0) \le 0 \right) \le 2\epsilon$$

by Lemma 4. This proves the proposition. ■

The rate of convergence is identical to that of simultaneous estimators (see Bai and Perron, 1994).

4. THE CASE OF $U(\tau_1^0) = U(\tau_2^0)$

When $U(\tau_1^0) = U(\tau_2^0)$, it is easy to show that the function $U(\tau)$ has two local minima at τ_1^0 and τ_2^0. This leads to the conjecture that the estimated break point $\hat{\tau}$ may converge in distribution to a random variable with mass at τ_1^0 and τ_2^0 only. Indeed, we have the following result.

PROPOSITION 3. *If Assumptions* A1–A3 *hold and* $U(\tau_1^0) = U(\tau_2^0)$, *the estimator* $\hat{\tau}$ *converges in distribution to a random variable with equal mass at* τ_1^0 *and* τ_2^0. *Furthermore,* $\hat{\tau}$ *converges either to* τ_1^0 *or to* τ_2^0 *at rate T in the sense that for every* $\epsilon > 0$ *there exists a finite M, which is independent of T, such that, for all large T,*

$$P(|T(\hat{\tau} - \tau_1^0)| > M \quad \text{and} \quad |T(\hat{\tau} - \tau_2^0)| > M) < \epsilon.$$

To prove the proposition, we need a number of preliminary results. Analogous to Lemma 3, we have the following.

LEMMA 5. *Under the assumptions of Proposition* 3, *there exists* $C > 0$ *such that for all large T*

$$ES_T(k) - ES_T(k_1^0) \ge C|k - k_1^0| \quad \forall k \le k_0^*,$$

$$ES_T(k) - ES_T(k_2^0) \ge C|k - k_2^0| \quad \forall k \ge k_0^*,$$

where $k_0^* = (k_1^0 + k_2^0)/2$.

The choice of k_0^* in the preceding fashion is not essential. Any number between k_1^0 and k_2^0, but bounded away from k_1^0 and k_2^0 for a positive fraction of observations, is equally valid.

Let $\hat{k}_1^\dagger$ be the location of the minimum of $S_T(k)$ for k such that $k \le k_0^*$; that is, $\hat{k}_1^\dagger = \text{argmin}_{k \le k_0^*} S_T(k)$. Let $\hat{k}_2^\dagger = \text{argmin}_{k_0^* < k} S_T(k)$. Note that $\hat{k}_1^\dagger$ and $\hat{k}_2^\dagger$ are not estimators as k_0^* is unknown. It is clear that the global minimizer $\hat{k}$ satisfies

$$\hat{k} = \begin{cases} \hat{k}_1^\dagger & \text{if } S_T(\hat{k}_1^\dagger) < S_T(\hat{k}_2^\dagger), \\ \hat{k}_2^\dagger & \text{if } S_T(\hat{k}_1^\dagger) > S_T(\hat{k}_2^\dagger). \end{cases} \tag{7}$$

Note that $P(S_T(\hat{k}_1^\dagger) = S_T(\hat{k}_2^\dagger)) = 0$ if X_t has a continuous distribution. Even without the assumption of a continuous distribution for X_t, the event $\{S_T(\hat{k}_1^\dagger) = S_T(\hat{k}_2^\dagger)\}$ has probability approaching zero as the sample size increases because $T^{-1/2}\{S_T(\hat{k}_1^\dagger) - S_T(\hat{k}_2^\dagger)\}$ converges in distribution to a normal random variable (see the proof of Lemma 7, in the Appendix).

Let $\hat{\tau}_i^\dagger = \hat{k}_i^\dagger/T$ $(i = 1,2)$. Using Lemmas 2 and 5, we can easily obtain the following result analogous to Proposition 1:

$$\hat{\tau}_1^\dagger - \tau_1^0 = O_p(T^{-1/2}) \quad \text{and} \quad \hat{\tau}_2^\dagger - \tau_2^0 = O_p(T^{-1/2}).$$

The root T consistency is strengthened to T consistency using the following lemma.

LEMMA 6. *Under the assumptions of Proposition* 3, *for every* $\epsilon > 0$, *there exists an* $M > 0$ *such that*

$$P\left(\min_{k \in D_{T,M}^{(i)}} S_T(k) - S_T(k_i^0) \leq 0\right) < \epsilon, \quad \textit{for } i = 1,2,$$

where

$$D_{T,M}^{(1)} = \{k : T\eta \leq k \leq k_0^*, |k - k_1^0| > M\},$$

$$D_{T,M}^{(2)} = \{k : k_0^* + 1 \leq k \leq T(1 - \eta), |k - k_2^0| > M\}.$$

Lemma 6 together with consistency implies the T consistency of $\hat{k}_i^\dagger$ in the same way that Lemma 4 (together with consistency) implies the T consistency of $\hat{k}$ in Section 3. Using T consistency, we can prove the following.

LEMMA 7. *Under the assumptions of Proposition* 3,

$$\lim_{T \to \infty} P(\hat{k} = \hat{k}_i^\dagger) = \tfrac{1}{2}, \qquad i = 1,2.$$

Proof of Proposition 3. By Lemma 7, $P(\hat{\tau} = \hat{\tau}_i^\dagger) \to \frac{1}{2}$ $(i = 1,2)$, but $\hat{\tau}_i^\dagger \xrightarrow{p} \tau_i^0$, so it follows that $\hat{\tau}$ converges in distribution to random variable with equal mass at τ_1^0 and τ_2^0. The second part of the proposition follows from T consistency of $\hat{\tau}_i^\dagger$. ■

5. LIMITING DISTRIBUTION

Given the rate of convergence, it is relatively easy to derive the limiting distributions. We strengthen the assumption of second-order stationarity to strict stationarity.

Assumption A5. The process $\{X_t\}$ is strictly stationary.

This assumption allows one to express the limiting distribution free from the change point (k_1^0). The assumption can be eliminated (see Bai, 1994a).

Let $\{X_t^{(1)}\}$ and $\{X_t^{(2)}\}$ be two independent copies of the process of $\{X_t\}$. Define, for $i = 1,2$, $W^{(i)}(l,\lambda) = W_1^{(i)}(l,\lambda)$ for $l < 0$ and $W^{(i)}(l,\lambda) = W_2^{(i)}(l,\lambda)$ for $l > 0$ and $W^{(i)}(0,\lambda) = 0$, where

$$W_1^{(i)}(l,\lambda) = -2(\mu_{i+1} - \mu_i) \sum_{t=l+1}^{0} X_t^{(i)} + |l|(\mu_{i+1} - \mu_i)^2(1 + \lambda), \qquad l = -1,-2,\ldots,$$

$$W_2^{(i)}(l,\lambda) = 2(\mu_{i+1} - \mu_i) \sum_{t=1}^{l} X_t^{(i)} + l(\mu_{i+1} - \mu_i)^2(1 - \lambda), \qquad l = 1,2,\ldots.$$

PROPOSITION 4. *If Assumptions* A1–A5 *hold and* X_t *has a continuous distribution,*

$$\hat{k} - k_1^0 \xrightarrow{d} \operatorname{argmin}_l W^{(1)}(l,\lambda_1),$$

where

$$\lambda_1 = \frac{1 - \tau_2^0}{1 - \tau_1^0}\left(\frac{\mu_3 - \mu_2}{\mu_2 - \mu_1}\right). \tag{8}$$

Note that Assumption A4 (or, equivalently, (6)) guarantees that $|\lambda_1| < 1$. The assumption of a continuous distribution for X_t ensures the uniqueness of the global minimum for the process $W^{(1)}(l,\lambda_1)$, so that $\operatorname{argmin}_l W^{(1)}(l,\lambda_1)$ is well defined. The proof of this proposition is provided in the Appendix.

When λ is zero, the limiting distribution corresponds to that of a single break ($\tau_2^0 = 1$) or to that of the first break point estimator for simultaneous estimation of multiple breaks. If X_t has a symmetric distribution and λ is equal to zero, $W^{(1)}(l,\lambda)$ and $W^{(1)}(-l,\lambda)$ will have the same distribution and, consequently, $\hat{k} - k_1^0$ will have a symmetric distribution. Because $\lambda \neq 0$ in general, the limiting distribution from sequential estimation is not symmetric about zero. For positive λ (or, equivalently, for $\mu_2 - \mu_1$ and $\mu_3 - \mu_2$ having the same sign), the drift term of $W_2^{(1)}(l,\lambda)$ is smaller than that of $W_1^{(1)}(l,\lambda)$. This implies that the distribution of $\hat{k}$ will have a heavy right tail, reflecting a tendency to overestimate the break point relative to simultaneous estimation. For negative λ, there is a tendency to underestimate the break point. These theoretical implications are all borne out by Monte Carlo simulations.

When $\hat{k}/T$ is consistent for τ_1^0, an estimate for τ_2^0 can be obtained by applying the same technique to the subsample $[\hat{k}, T]$. Let $\hat{k}_2$ denote the resulting estimator. Then, $\hat{\tau}_2 = \hat{k}_2/T$ is T-consistent for τ_2^0, because in the subsample $[\hat{k}, T]$ k_2^0 is the dominating break. Moreover, we shall prove that the limiting distribution of $\hat{k}_2 - k_2^0$ is the same as that from a single break model. More precisely, we have the following proposition.

PROPOSITION 5. *Under Assumptions* A1–A5,

$$\hat{k}_2 - k_2^0 \xrightarrow{d} \operatorname{argmin}_l W^{(2)}(l,0)$$

and is independent of $\hat{k} - k_1^0$ *asymptotically.*

The proof is given in the Appendix. The asymptotic independence follows because $\hat{k}$ and $\hat{k}_2$ are determined by increasingly distant observations that are only weakly dependent. We call $\hat{k}$ the first-stage estimator (based on the full sample) and $\hat{k}_2$ the second-stage estimator (based on a subsample).

We now consider the case in which $U(\tau_1^0) = U(\tau_2^0)$. Let $(\hat{k}^{(1)}, \hat{k}^{(2)})$ denote the ordered pair of the first- and second-stage estimators. Let λ_1 be given in (8) and $\lambda_2 = (\tau_1^0/\tau_2^0)(\mu_2 - \mu_1)/(\mu_3 - \mu_2)$. We have, for $i = 1,2$,

$$\hat{k}^{(i)} - k_i^0 \xrightarrow{d} \begin{cases} \operatorname{argmin}_l W^{(i)}(l,\lambda_i) & \text{with probability } \frac{1}{2}, \\ \operatorname{argmin}_l W^{(i)}(l,0) & \text{with probability } \frac{1}{2}, \end{cases}$$

because, in the limit, $\hat{k}^{(1)}$ is the first-stage estimator with probability $\frac{1}{2}$ and the second-stage estimator with probability $\frac{1}{2}$. When $\hat{k}^{(1)}$ is the first-stage estimator, its limiting distribution is given by $\operatorname{argmin}_l W^{(1)}(l,\lambda_1)$. When $\hat{k}^{(1)}$ is the second-stage estimator, its limiting distribution is given by $\operatorname{argmin}_l W^{(1)}(l,0)$ because it is effectively estimated using the sample $[1, k_2^0]$, which contains only a single break. The argument for $k^{(2)}$ is similar.

6. FINE-TUNING: REPARTITION

The limiting distribution suggests that the estimation method has a tendency to over- or underestimate the true location of a break point depending on whether λ_i is positive or negative. We now discuss a procedure that yields an estimator having the same asymptotic distribution as the simultaneous estimators. We call the procedure repartition. The idea of repartition is simple and was first introduced by Bai (1994a) in an empirical application. Here we provide the theoretical basis for doing so. Suppose initial T-consistent estimators $\hat{k}_i$ for k_i^0 $(i = 1,2)$ are obtained. The repartition technique reestimates each of the break points based on the initial estimates. To estimate k_1^0 the subsample $[1, \hat{k}_2]$ is used, and to estimate k_2^0 the subsample $[\hat{k}_1, T]$ is used. We denote the resulting estimators by $\hat{k}_1^*$ and $\hat{k}_2^*$, respectively. Because of the proximity of $\hat{k}_i$ to k_i^0, we effectively use the sample $[k_{i-1}^0 + 1, k_{i+1}^0]$ to estimate k_i^0 $(i = 1,2$ with $k_0^0 = 1$, $k_3^0 = T)$. Consequently, $\hat{k}_i^*$ is also T-consistent for k_i^0, with a limiting distribution identical to what it would be for a single break point model (or for a model with multiple breaks estimated by the simultaneous method; see Bai and Perron [1994]). In summary, we have the next proposition.

PROPOSITION 6. *Under Assumptions* A1 *and* A2, *the repartition estimators satisfy the following*:

(i) *For each* $\epsilon > 0$, *there exists an* $M < \infty$ *independent of* T *such that, for all large* T,

$$P(|\hat{k}_i^* - k_i^0| > M) < \epsilon \qquad (i = 1,2).$$

(ii) *Under the additional Assumption* A5,

$$\hat{k}_i^* - k_i^0 \xrightarrow{d} \text{argmin}_l W^{(i)}(l,0) \qquad (i = 1,2).$$

Note that Assumption A4 is not required. The proposition only uses the fact that the initial estimators are T-consistent. As is shown in Section 4, T-consistent estimators can be obtained regardless of the validity of Assumption A4. We note that the repartitioned estimators have the same asymptotic distribution as those obtained via simultaneous estimation.

7. SMALL SHIFTS

The limiting distributions derived earlier, though of theoretical interest, are perhaps of limited practical use because the distribution of $\text{argmin}_l W^{(i)}(l,\lambda)$ depends on the distribution of X_t and is difficult to obtain. An alternative strategy is to consider small shifts in which the magnitude of shifts converges to zero as the sample size increases to infinity. The limiting distributions under this setup are invariant to the distribution of X_t and remain adequate even for moderate shifts. The result will be useful for constructing confidence intervals for the break points.

We assume that the mean $\mu_{i,T}$ for the ith regime can be written as $\mu_{i,T} = v_T \tilde{\mu}_i$ $(i = 1,2,3)$. We further assume the following.

Assumption B1. The sequence of numbers v_T satisfies

$$v_T \to 0, \qquad T^{(1/2)-\delta} v_T \to \infty \quad \text{for some } \delta \in (0, \tfrac{1}{2}). \tag{9}$$

Because v_T converges to zero, the function $U(\tau)$ defined in Section 2 will be a constant function for all τ. This can be seen from (4) and (5), with u_j interpreted as $v_T \tilde{\mu}_j$. Therefore, Assumption A4 is no longer appropriate. The correct condition for $\hat{\tau}$ to be consistent for τ_1^0 is given next.

Assumption B2.

$$p\lim v_T^{-2}[U_T(k_1^0/T) - U_T(k_2^0/T)] < 0.$$

This condition is identical to (6), with μ_j replaced by $\tilde{\mu}_j$.

Under Assumptions B1 and B2, we shall argue that $\hat{\tau}$ is consistent for τ_1^0. However, the convergence rate is slower than T, which is expected because it is more difficult to discern small shifts.

PROPOSITION 7. *Under Assumptions* A1–A3, B1, *and* B2, *we have* $Tv_T^2(\hat{\tau} - \tau_1^0) = O_p(1)$ *or, equivalently, for every* $\epsilon > 0$, *there exists a finite* M *independent of* T *such that*

$$P(T|(\hat{\tau} - \tau_1^0)| > Mv_T^{-2}) < \epsilon \quad \textit{for all large } T.$$

The proof of this proposition is again based on some preliminary results analogous to Lemmas 2 and 3. First, we modify the objective function to be

$$S_T(k) - \sum_{t=1}^{T} X_t^2,$$

which does not affect optimization because the second term is free of k.

LEMMA 8. *Under the assumptions of Proposition* 7, *we have the following*:

(a) $$\sup_{1 \le k \le T} \left| U_T(k/T) - EU_T(k/T) - T^{-1} \sum_{t=1}^{T} (X_t^2 - EX_t^2) \right| = O_p(T^{-1/2} v_T).$$

(b) *There exists* $C_1 > 0$, *only depending on* τ_i^0 *and* $\tilde{\mu}_j$ $(i = 1,2, j = 1,2,3)$ *such that*
$$ES_T(k) - ES_T(k_1^0) \ge C_1 v_T^2 |k - k_1^0| \quad \textit{for all large } T.$$

COROLLARY 1. *Under the assumptions of Proposition* 7,

$$\hat{\tau} - \tau_1^0 = O_p\left(\frac{1}{\sqrt{T} v_T}\right).$$

Proof. Add and subtract $\sum_{t=1}^{T} (X_t^2 - EX_t^2)$ to the following identity

$$\begin{aligned} S_T(k) - S_T(k_1^0) &= S_T(k) - ES_T(k) - [S_T(k_1^0) - ES_T(k_1^0)] \\ &\quad + ES_T(k) - ES_T(k_1^0) \end{aligned}$$

to obtain

$$\begin{aligned} &S_T(k) - S_T(k_1^0) \\ &\quad \ge -2 \sup_{1 \le j \le T} \left| S_T(j) - ES_T(j) - \sum_{t=1}^{T} (X_t^2 - EX_t^2) \right| + ES_T(k) - ES_T(k_1^0) \\ &\quad \ge -2 \sup_{1 \le j \le T} \left| S_T(j) - ES_T(j) - \sum_{t=1}^{T} (X_t^2 - EX_t^2) \right| + C_1 v_T^2 |k - k_1^0|, \end{aligned}$$

where the second inequality follows from Lemma 8(b). From $S_T(\hat{k}) - S_T(k_1^0) \le 0$, we have

$$|\hat{k} - k_1^0| \le C_1^{-1} 2 v_T^{-2} \sup_{1 \le j \le T} \left| S_T(j) - ES_T(j) - \sum_{t=1}^{T} (X_t^2 - EX_t^2) \right|.$$

The corollary is obtained by dividing the preceding inequality by T on both sides and using Lemma 8(a). ■

Because $\sqrt{T}v_T \to \infty$, $\hat{\tau}$ is consistent for τ_1^0. Using this initial consistency, the rate of convergence stated in Proposition 7 can be proved. In view of the anticipated rate of convergence, we define D_{TM}^* the same as D_{TM} but replace M with Mv_T^{-2}. Thus, for $k \in D_{TM}^*$, it is possible for $k - k_1^0$ to diverge to infinity because v_T^{-2} diverges to infinity, although at a much slower rate than T.

LEMMA 9. *Under the assumptions of Proposition* 7, *for every* $\epsilon > 0$, *there exists an* $M > 0$ *such that, for all large* T,

$$P\left(\min_{k \in D_{T,M}^*} S_T(k) - S_T(k_1^0) \leq 0\right) < \epsilon.$$

Proof of Proposition 7. The proof is virtually identical to that of Proposition 2, but one uses Lemma 9 instead of Lemma 4. ■

Having obtained the rate of convergence, we examine the local behavior of the objective function in appropriate neighborhoods of k_1^0 to obtain the limiting distribution. Let $B_i(s)$ $(i = 1,2)$ be two independent and standard Brownian motions on $[0,\infty)$ with $B_i(0) = 0$ and define a two-sided drifted Brownian motion on $\mathcal{R}$ as

$$\Lambda(s,\lambda) = \begin{cases} 2B_1(-s) + |s|(1+\lambda) & \text{if } s < 0, \\ 2B_2(s) + |s|(1-\lambda) & \text{if } s > 0, \end{cases}$$

with $\Lambda(0,\lambda) = 0$.

PROPOSITION 8. *Under the assumptions of Proposition* 7,

$$T(\mu_{2T} - \mu_{1T})^2(\hat{\tau} - \tau_1^0) \xrightarrow{d} a(1)^2\sigma_\epsilon^2 \operatorname{argmin}_s \Lambda(s,\lambda_1),$$

where λ_1 *is defined in Proposition* 4 *with* μ_j *replaced by* $\tilde{\mu}_j$.

While the density function of $\operatorname{argmin}_s \Lambda(s,\lambda_1)$ is derived in Bai (1994a) so that confidence intervals can be constructed, it is suggested that the repartitioned estimators be used. For the repartitioned estimator, the limiting distribution corresponds to $\lambda_1 = 0$.

8. MORE THAN TWO BREAKS

In this section, we extend the procedure and the theoretical results to general multiple break points:

$$
\begin{aligned}
Y_t &= \mu_1 + X_t, && \text{if } t \le k_1^0, \\
Y_t &= \mu_2 + X_t, && \text{if } k_1^0 + 1 \le t \le k_2^0, \\
&\vdots && \vdots \\
Y_t &= \mu_{m+1} + X_t, && \text{if } k_m^0 + 1 \le t \le T.
\end{aligned} \tag{10}
$$

where $\mu_i \neq \mu_{i+1}$, $k_i^0 = [T\tau_i^0]$, $\tau_i^0 \in (0,1)$, and $\tau_i^0 < \tau_{i+1}^0$ for $i = 1, \ldots, m$ with $\tau_{m+1}^0 = 1$. Assume the process X_t satisfies Assumptions A1 and A2.

Define the quantities $S_T(k)$, $V_T(k)$, and $U_T(\tau)$ as before, and denote by $U(\tau)$ the limit of $U_T(\tau)$. Again, let $\hat{k} = \operatorname{argmin} S_T(k)$ and $\hat{\tau} = \hat{k}/T$. From the proof in the Appendix for the earlier results, we can see that the assumption of two breaks is not essential. With more than two breaks, one simply needs to deal with extra terms. The argument is virtually identical. Therefore, we state the major results without proof. First we impose the following.

Assumption A6. There exists an i such that $U(\tau_i^0) < U(\tau_j^0)$ for all $j \neq i$.

PROPOSITION 9. *If Assumptions* A1–A3 *and* A6 *hold, the estimated break point* $\hat{\tau}$ *is T-consistent for* τ_i^0.

PROPOSITION 10. *If Assumptions* A1–A3, A5, *and* A6 *hold,*

$\hat{k} - k_i^0 \xrightarrow{d} \operatorname{argmin}_l W^{(i)}(l, \lambda_i)$,

where $W^{(i)}(l,\lambda)$ *is defined earlier but uses another independent copy of* $\{X_t\}$ *and*

$$
\lambda_i = \frac{1}{\mu_{i+1} - \mu_i}\left[\frac{1}{1-\tau_i^0}\sum_{j=i+1}^{m}(1-\tau_j^0)(\mu_{j+1}-\mu_j) + \frac{1}{\tau_i^0}\sum_{h=1}^{i-1}\tau_h^0(\mu_{h+1}-\mu_h)\right].
$$

Again, Assumption A6 ensures that $|\lambda_i| < 1$.

We point out that Propositions 7 and 8 can also be extended to models with general multiple breaks.

A subsample $[k, l]$ is said to contain a nontrivial break point if both k and l are bounded away from a break point for a positive fraction of observations. That is, $k^0 - k > T\epsilon_0$ and $l - k^0 > T\epsilon_0$ for some $\epsilon_0 > 0$ and for all large T, where k^0 is a break point inside $[k, l]$. This definition rules out subsamples such as $[1, \hat{k}]$, where $\hat{k} = k_1^0 + O_p(1)$.

When it is known that the subsample $[1, \hat{k}]$ contains at least one nontrivial break point, the same procedure can be used to estimate a break point based on the sample $[1, \hat{k}]$. That is, the second break point is defined as the location where $S_{\hat{k}}(l)$ is minimized over the range $[1, \hat{k}]$. The resulting estimator must be T-consistent for one of the break points, assuming again Assumption A6 holds for this subsample. Furthermore, the resulting estimator has a limiting distribution as if the sample $[1, k_i^0]$ were used and thus has no connection with parameters in the sample $[k_i^0 + 1, T]$. This is because the first-stage estimator $\hat{k}/T$ is T-consistent for τ_i^0. A similar conclusion applies to the interval $[\hat{k}, T]$. Therefore, second-round estimation may yield an additional two breaks and, consequently, up to four subintervals are to be considered in the third-round estimation. This procedure is repeated until each resulting subsample contains no nontrivial break point. Assuming knowledge of the number of breaks as well as the existence of a nontrivial break in a given subsample, all the breaks can be identified and all the estimated break fractions are T-consistent.

In practice, a problem arises immediately as to whether a subsample contains a nontrivial break, which is clearly related to the determination of the number of breaks. We suggest that the decision be made based on testing the hypothesis of parameter constancy for the subsample. We prove below that such a decision rule leads to a consistent estimate of the number of breaks and, implicitly, a correct judgment about the existence of a nontrivial break in a given subsample.

Determining the number of breaks. The number of breaks, m, in practice is unknown. We show how the sequential procedure coupled with hypothesis testing can yield a consistent estimate for the true number of breaks. The procedure works as follows. When the first break point is identified, the whole sample is divided into two subsamples with the first subsample consisting of the first $\hat{k}$ observations and the second subsample consisting of the rest of the observations. We then perform hypothesis testing of parameter constancy for each subsample, estimating a break point for the subsample where the constancy test fails. Divide the corresponding subsample further into subsamples at the newly estimated break point, and perform parameter constancy tests for the hierarchically obtained subsamples. This procedure is repeated until the parameter constancy test is not rejected for all subsamples. The number of break points is equal to the number of subsamples minus 1.

Let $\hat{m}$ be the number of breaks determined in the preceding procedure, and let m_0 be the true number of breaks. We argue that $P(\hat{m} = m_0)$ converges to 1 as the sample size grows unbounded, provided the size of the tests slowly converges to zero. To prove this assertion, we need the following general result. Let

$$
\begin{aligned}
Y_t &= \mu_1 + X_t, \quad \text{if } -n_1 + 1 \le t \le 0, \\
Y_t &= \mu + X_t, \quad \text{if } 1 \le t \le n, \qquad\qquad (11) \\
Y_t &= \mu_2 + X_t, \quad \text{if } n + 1 \le t \le n + n_2,
\end{aligned}
$$

where n is a nonrandom integer and n_1 and n_2 are integer-valued random variables such that $n_i = O_p(1)$ as $n \to \infty$. The first and the third regimes are dominated by the second in the sense that $n_i/n = O_p(n^{-1})$. Let $N = n + n_1 + n_2$. The sup F–test is based on the difference between restricted and unrestricted sums of squared residuals. More specifically, let $\bar{S}_N =$

$\sum_{t=-n_1+1}^{n+n_2}(Y_t - \bar{Y})^2$ and $S_N(k) = \sum_{t=-n_1+1}^{k}(Y_t - \bar{Y}_k)^2 + \sum_{t=k+1}^{n+n_2}(Y_t - \bar{Y}_k^*)^2$, where $\bar{Y}_k$ represents the sample mean for the first $k + n_1$ observations and $\bar{Y}_k^*$ represents the sample mean for the last $n + n_2 - k$ observations. The sup F–test is then defined as, for some $\eta \in (0, \frac{1}{2})$,

$$\sup F_N = \sup_{N\eta \le k \le N(1-\eta)} \frac{\bar{S}_N - S_N(k)}{\hat{\sigma}^2},$$

where $\hat{\sigma}^2$ is a consistent estimator of $a(1)^2\sigma_\epsilon^2$. Note that $a(1)^2\sigma_\epsilon^2$ is proportional to the spectral density of X_t at zero, which can be consistently estimated in a number of ways.

LEMMA 10. *Under model* (11) *and Assumptions* A1 *and* A2, *as* $n \to \infty$,

$$\sup F_N \xrightarrow{d} \sup_{\eta \le \tau \le 1-\eta} \frac{|B(\tau) - \tau B(1)|^2}{\tau(1-\tau)}, \tag{12}$$

where $B(\cdot)$ *is standard Brownian motion on* $[0,1]$.

The limiting distribution is identical to what it would be in the absence of the first and last regimes in model (11). This is simply due to the stochastic boundedness of n_1 and n_2. We assume that the sup F–test is used in the sequential procedure and that the critical value and size of the test are based on the asymptotic distribution. Let ξ denote the random variable given on the right-hand side of (12). Then, for large z (see, e.g., DeLong, 1981), $P(\xi > z) \le K_\eta z^{1/2}\exp(-z/2) \le K_\eta \exp(-z/3)$, where $K_\eta = (2/\pi)^{1/2}\log[(1-\eta)^2/\eta^2]$. It follows that if $P(\xi > c) = \alpha$, then $c \le -3\log(\alpha/K_\eta)$. In particular, for $\alpha = K_\eta/T$, we have $c \le 3\log T$. Using these results and Lemma 10, we can prove the following proposition.

PROPOSITION 11. *Suppose that the size of the test* α_T *converges to zero slowly* ($\alpha_T \to 0$ *yet* $\liminf_{T\to\infty} T\alpha_T > 0$), *then under model* (10) *and Assumptions* A1 *and* A2,

$P(\hat{m} = m_0) \to 1$, *as* $T \to \infty$.

Proof. Consider the event $\{\hat{m} < m_0\}$. When the estimated number of breaks is less than the true number, there must exist a segment $[\hat{k}, \hat{l}]$ containing at least one true break point that is nontrivial. That is, $k_j^0 - \hat{k} > T\epsilon_0$ and $\hat{l} - k_j^0 > T\epsilon_0$ for some $\epsilon_0 > 0$, where $k_j^0 \in (\hat{k}, \hat{l})$ is a true break point. Then, the sup F–test statistic based on this subsample is of order T.[1] That is, there exists $\pi > 0$ such that, for every $\epsilon > 0$, $P(\sup F \ge \pi T) > 1 - \epsilon$ for all large T. For $\alpha_T = K_\eta T^{-1}$, then $c_T \le 3\log T$ (see the discussion following Lemma 10). Under this choice of α_T and hence c_T, we have $P(\sup F \ge c_T) > 1 - \epsilon$ for all large T. Or, equivalently, $P(\sup F \ge c_T) \to 1$ as $T \to \infty$. Thus, one will reject the null hypothesis of parameter constancy with probability tending to 1. This implies that $P(\hat{m} < m_0)$ converges to zero as the sample size increases. (Note that the argument holds for every $\alpha_T \ge K_\eta T^{-1}$.)

Next, consider the event $\{\hat{m} > m_0\}$. For $\hat{m} > m_0$ to be true, it must be the case that for some i, at a certain stage in the sequential estimation, one rejects the null hypothesis for the interval $[\hat{k}_i, \hat{k}_{i+1}]$, where $\hat{k}_i = k_i^0 + O_p(1)$ and $\hat{k}_{i+1} = k_{i+1}^0 + O_p(1)$. That is, the given interval contains no nontrivial break point, but the null hypothesis is rejected. Thus,

$$\begin{aligned} P(\hat{m} > m_0) &\le P(\exists i, \text{ reject parameter constancy for } [\hat{k}_i, \hat{k}_{i+1}]) \\ &\le \sum_{i=0}^{m_0} P(\text{reject parameter constancy for } [\hat{k}_i, \hat{k}_{i+1}]), \end{aligned}$$

where $\hat{k}_0 = 1$ and $\hat{k}_{m_0+1} = T$. Because $\hat{k}_i = k_i^0 + O_p(1)$ and $\hat{k}_{i+1} = k_{i+1}^0 + O_p(1)$, if one lets $n = k_{i+1}^0 - k_i^0$ and $N = \hat{k}_{i+1} - \hat{k}_i$, then by Lemma 8 the test statistic computed for the subsample $[\hat{k}_i, \hat{k}_{i+1}]$, denoted by $\sup F_N^i$, converges in distribution to ξ, the right-hand side of (12). From $\alpha_T \to 0$, we have $c_T \to \infty$. Thus, for large T (and hence large n), $P(\sup F_N^i > c_T) \to 0$. Thus, $P(\hat{m} > m_0) \le (m_0 + 1)\max_{0 \le i \le m_0} P(\sup F_N^i > c_T) \to 0$, provided that $\alpha_T \to 0$ (i.e., $c_T \to \infty$). ■

Bai and Perron (1994) proposed an alternative strategy for selecting the number of breaks. We first describe their procedure for estimating the break points when the number of breaks is known. In each round of estimation, their method selects only one additional break. The single additional break is chosen such that the sum of squared residuals for the total sample is minimized. For example, at the beginning of the ith round, $i - 1$ breaks are already determined, giving rise to i subsamples. The ith break point is chosen in the subsample yielding the largest reduction in the sum of squared residuals. The procedure is repeated until the specified number of break points is obtained. It is necessary to know when to terminate the procedure when the number of breaks is unspecified. The stopping rule is based on a test for the presence of an additional break given the number of breaks already obtained. The number of breaks is the number of subsamples upon terminating the procedure minus 1. Again, assuming the size of the test approaches zero at a slow rate as the sample size increases, the number of breaks determined in this way is also consistent. A further alternative was proposed by Yao (1988), who suggested the Bayesian Information Criterion (BIC). His method requires simultaneous estimation.

9. SOME SIMULATED RESULTS

This section reports results from some Monte Carlo simulations. The data are generated according to a model with three mean breaks. Let $(\mu_1, \ldots, \mu_4)$ denote the mean parameters and (k_1^0, k_2^0, k_3^0) denote the break points. We consider two sets of mean parameters. The first set is given by (1.0, 2.0, 1.0, 0.0), and the second by (1.0, 2.0, −1.0, 1.0). The sample size T is taken to be 160 with break points at (40, 80, 120) for both sets of mean parameters.

The disturbances $\{X_t\}$ are independent and identically distributed standard normal. All reported results are based on 5,000 repetitions.

Estimating the break points. We assume the number of break points is known and focus on their estimation. To verify the theory and for comparative purposes, three different methods are used: sequential, repartition, and simultaneous methods. The sequential procedure employs the method of Bai and Perron (1994), "one and only one additional break" in each round of estimation. A chosen break point must achieve greatest reduction in total sum of squared residuals for that round of estimation.

Figure 1 displays the estimated break points for the first set of parameters (called model (I)). Because the magnitude of shift for each break is the same in model (I), we expect the three estimated break points to have a similar distribution for the repartition and simultaneous methods. This is indeed so, as suggested by the histograms. For sequential estimation, the distribution of the estimated break points shows asymmetry, as suggested by the theory. This asymmetry is removed by the repartition procedure.

Figure 2 displays the corresponding results for the second set of parameters (model (II)). Because the middle break has the largest magnitude of shift, it is estimated with the highest precision, then followed by the third, and then by the first. Note that the sequential method picks up the middle break point in the first place. This has two implications. First, the first and third estimated break points will have the same limiting distribution as in the case of simultaneous estimation, even without repartition. This explains why the results look homogeneous for the three different methods. Second, only the middle break point will have an asymmetric distribution for the sequential method. This asymmetry is again removed by repartition.

These simulation results are entirely consistent with the theory. Also remarkable is the match rate for the repartition and simultaneous methods. They yield almost identical results in the simulation. The match rate for model (I) is more than 92%, whereas for model (II) the rate is more than 99.5%.

Determining the number of breaks. Although the asymptotic theory implies that the sequential procedure will not underestimate (in a probabilistic sense) the number of breaks, Monte Carlo simulations show that the procedure has a tendency to underestimate. The problem was caused in part by the inconsistent estimation of the error variance in the presence of multiple breaks. When multiple breaks exist and only one is allowed in estimation, the error variance cannot be consistently estimated (because of the inconsistency of the regression parameters) and is biased upward. This decreases the power of the test. It is thus less likely to reject parameter constancy. This also partially explains why the conventional sup F-test possesses

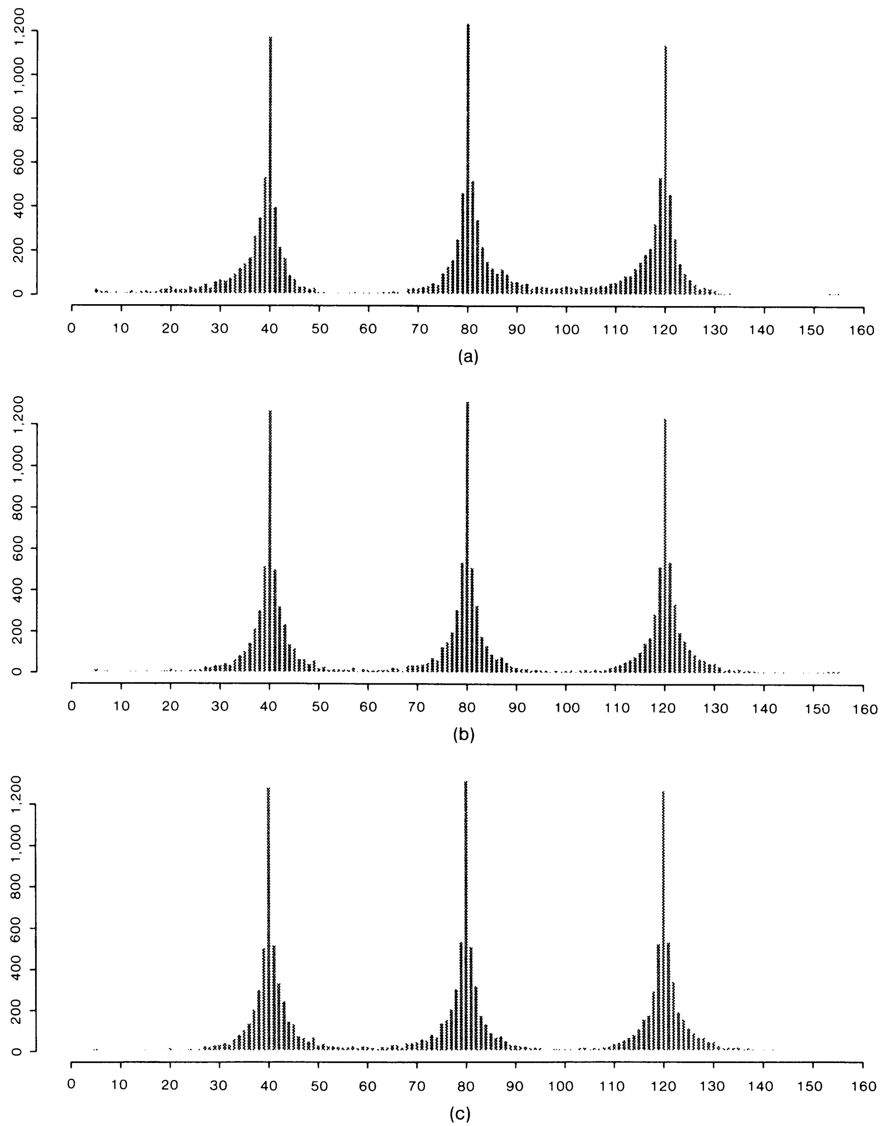

FIGURE 1. Histograms of the estimated break points for model (I): (a) sequential method, (b) repartition method, and (c) simultaneous method.

less power than the test proposed by Bai and Perron (1994) in the presence of multiple breaks. This observation is of practical importance.

The problem can be overcome by using a two-step procedure. In the first step, the goal is to obtain a consistent (or less biased) estimate for the error

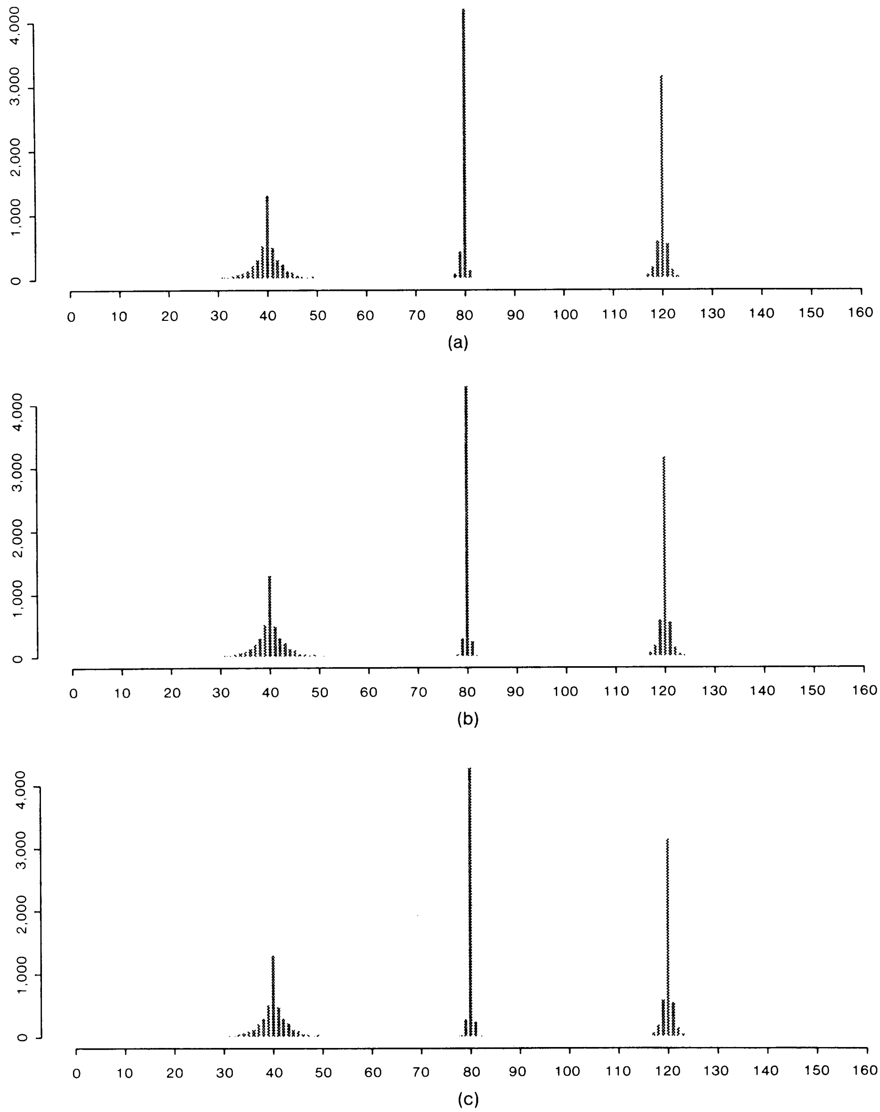

FIGURE 2. Histograms of the estimated break points for model (II): (a) sequential method, (b) repartition method, (c) simultaneous method.

variance. This can be achieved by allowing more breaks (solely for the purpose of constructing error variance). It is evident that as long as $m \geq m_0$ the error variance will be consistently estimated. Obviously, one does not know whether $m \geq m_0$, but the specification of m in this stage is not as important as in the final model estimation. When m is fixed, the m break points can be selected either by simultaneous estimation or by the "one additional break" sequential procedure described in Bai and Perron (1994) (no test is performed). In the second step, the number of breaks is determined by the sequential procedure coupled with hypothesis testing. The test statistics use the error variance estimator (as the denominator) obtained in the first step. This two-step procedure is used in our simulation.

In addition to the two sets of parameters considered earlier, we add a third set of parameters $(1.0, 2.0, 3.0, 4.0)$ (referred to as model (III)). For comparative purposes, estimates using the BIC method are also given. Figure 3 displays the estimates for both methods. The left three histograms, (a)–(c), are for the sequential method, and the right three (a′)–(c′), are for the BIC method. The sequential method uses a two-step procedure as already described. We assume the number of breaks is 4 in the first step. The size of the test is chosen to be 0.05 with corresponding critical value 9.63.

For the first set of parameters, the BIC does a better job than the sequential method; the latter underestimates the number of breaks. For a significant proportion of observations, the sequential method detects only a single break. For the second set of parameters, the two methods are comparable. Interestingly, the sequential method works better than the BIC for the third set of parameters.

The sequential method may be improved upon in at least two dimensions. First, the $\sup F$-test, which is designed for testing a single break, may be replaced by, or used in conjunction with, Bai and Perron's $\sup F(l)$-test for testing multiple breaks. The latter test is more powerful in the presence of multiple breaks. Other tests such as the exponential-type or average-type tests can also be used (see Andrews and Ploberger, 1994). Second, the critical values may be chosen using small sample distributions rather than limiting distributions. There are certain degrees of flexibility in the choice of sizes, as well. In any case, the sequential procedure seems promising. Further investigation is warranted.

10. SUMMARY

We have developed some underlying theory for estimating multiple breaks one at a time. We proved that the estimated break points are T-consistent, and we also derived their limiting distributions. A number of ideas have been presented to analyze multiple local minima, to obtain estimators having the same limiting distribution as those of simultaneous estimation, and to consistently determine the number of breaks in the data. The proposed repartition method is particularly useful because it allows confidence intervals to be constructed as if simultaneous estimation were used. Of course, the repartition estimators are not necessarily identical to simultaneous estimators.

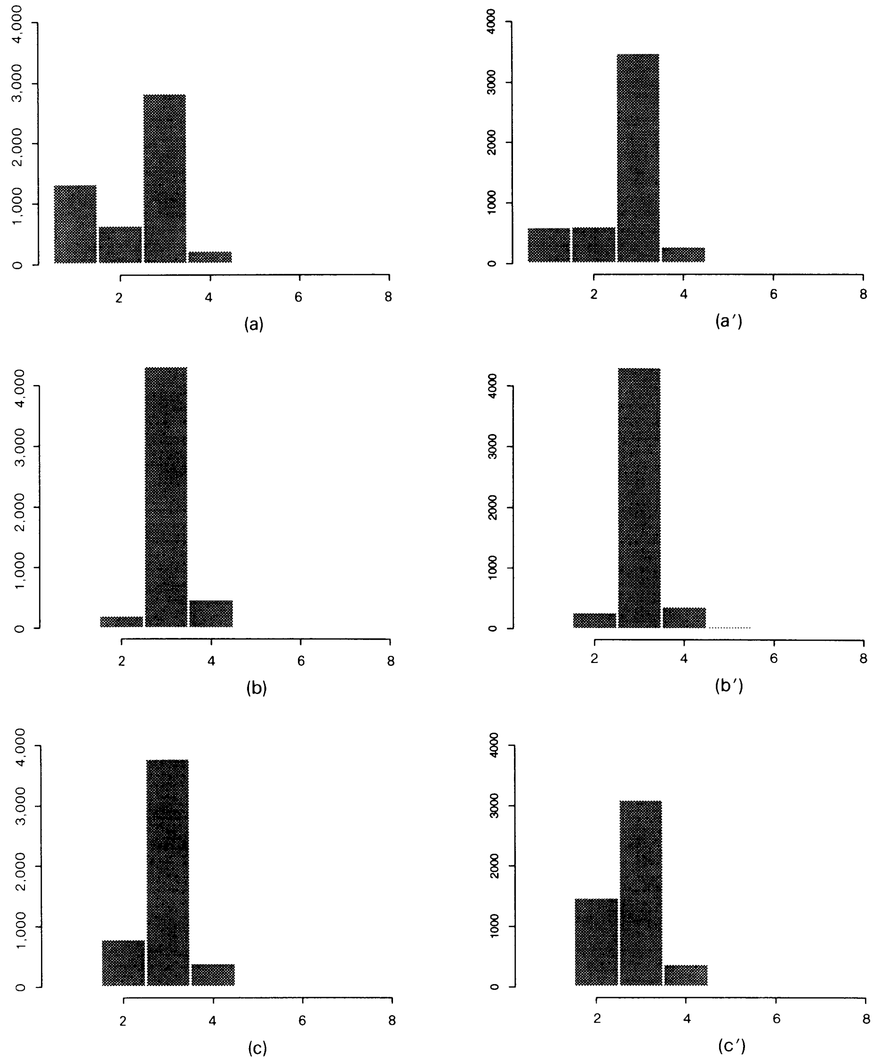

FIGURE 3. Histograms of the estimated numbers (of breaks) for models (I)–(III). (a)–(c) Sequential estimation; (a′)–(c′) BIC.

NOTES

1. To see this, suppose the sample $[1, T]$ has a break point at $k_0 = [T\tau_0]$ with prebreak parameter μ_1 and postbreak parameter μ_2. By (2) and (3), $\bar{S}_T - S_T(k_0) = TV_T(k_0)^2$, where $\bar{S}_T$ is the restricted sum of squared residuals (no break point is estimated). But $V_T(k_0)^2 \to \tau_0(1-\tau_0)(\mu_2-\mu_1)^2$. Thus, $\bar{S}_T - S_T(\hat{k}) \geq \bar{S}_T - S_T(k_0) = TV_T(k_0) = O_p(T)$. This implies that

the sup F-test based on the sample $[1, T]$ is $O_p(T)$. When the sample $[1, T]$ contains more than one break point (under the setup of Section 8), the sup F-test is still of order T because $V_T(k)$ for $k = [T\tau]$ has a limit, which is not identically zero in τ.

REFERENCES

Andrews, D.W.K. & W. Ploberger (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–1414.

Bai, J. (1994a) Estimation of Structural Change based on Wald Type Statistics. Working paper 94-6, Department of Economics, MIT, Cambridge, Massachusetts (forthcoming in *Review of Economics and Statistics*).

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

Bai, J. & P. Perron (1994) Testing for and Estimation of Multiple Structural Changes. Manuscript, Department of Economics, MIT, Cambridge, Massachusetts.

Burdekin R.C.K. & C.M. Siklos (1995) Exchange Rate Regimes and Inflation Persistence: Further Evidence. Manuscript, Department of Economics, University of California at San Diego.

Chong, T.T.-L. (1994) Consistency of Change-Point Estimators When the Number of Change-Points in Structural Change Models Is Underspecified. Manuscript, Department of Economics, University of Rochester, Rochester, New York.

Cooper, S.J. (1995) Multiple Regimes in U.S. Output Fluctuations. Manuscript, Kennedy School of Government, Harvard University, Cambridge, Massachusetts.

DeLong, D.M. (1981) Crossing probabilities for a square root boundary by a Bessel process. *Communications in Statistics—Theory and Methods, A* 10 (21), 2197–2213.

Garcia, R. & P. Perron (1996) An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics* 78, 111–125.

Lumsdaine, R.L. and D.H. Papell (1995) Multiple Trend Breaks and the Unit Root Hypothesis. Manuscript, Princeton University.

Phillips, P.C.B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.

Yao, Y.-C. (1988) Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters* 6, 181–189.

APPENDIX: PROOFS

The first two lemmas in the text are closely related. We first derive some results common to these two lemmas. We need to examine $U_T(k)$ for all $k \in [1, T]$.

For $k < k_1^0$,

$$\bar{Y}_k = \mu_1 + \frac{1}{k}\sum_{t=1}^{k} X_t,$$

$$\bar{Y}_k^* = \frac{1}{T-k}\sum_{t=k+1}^{T} Y_t = \frac{k_1^0 - k}{T-k}\mu_1 + \frac{k_2^0 - k_1^0}{T-k}\mu_2 + \frac{T-k_2^0}{T-k}\mu_3 + \frac{1}{T-k}\sum_{t=k+1}^{T} X_t.$$

Throughout, we define A_{Tk} and A_{Tk}^* as

$$A_{Tk} = \frac{1}{k}\sum_{t=1}^{k} X_t, \qquad A_{Tk}^* = \frac{1}{T-k}\sum_{t=k+1}^{T} X_t.$$

Thus,

$$\sum_{t=1}^{k} (Y_t - \bar{Y}_k)^2 = \sum_{t=1}^{k}\left(X_t - \frac{1}{k}\sum_{i=1}^{k} X_i\right)^2 = \sum_{t=1}^{k} (X_t - A_{Tk})^2 \tag{A.1}$$

and

$$\begin{aligned}
&\sum_{t=k+1}^{T} (Y_t - \bar{Y}_k^*)^2 \\
&\quad = \sum_{t=k+1}^{k_1^0} (\mu_1 + X_t - \bar{Y}_k^*)^2 + \sum_{k_1^0+1}^{k_2^0} (\mu_2 + X_t - \bar{Y}_k^*)^2 + \sum_{k_2^0+1}^{T} (\mu_3 + X_t - \bar{Y}_k^*)^2 \\
&\quad = \sum_{t=k+1}^{k_1^0}\left[\frac{1}{T-k}\{(T-k_1^0)(\mu_1-\mu_2) + (T-k_2^0)(\mu_2-\mu_3)\} + X_t - A_{Tk}^*\right]^2 \\
&\qquad + \sum_{k_1^0+1}^{k_2^0}\left[\frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1) + (T-k_2^0)(\mu_2-\mu_3)\} + X_t - A_{Th}^*\right]^2 \\
&\qquad + \sum_{k_2^0+1}^{T}\left[\frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1) + (k_2^0-k)(\mu_3-\mu_2)\} + X_t - A_{Tk}^*\right]^2.
\end{aligned}$$

The latter expression can be rewritten as

$$\begin{aligned}
\sum_{t=k+1}^{T} (Y_t - \bar{Y}_k^*)^2 &= (k_1^0 - k)a_{Tk}^2 + 2a_{Tk}\sum_{t=k+1}^{k_1^0}(X_t - A_{Tk}^*) \\
&\quad + (k_2^0 - k_1^0)b_{Tk}^2 + 2b_{Tk}\sum_{k_1^0+1}^{k_2^0}(X_t - A_{Tk}^*) \\
&\quad + (T - k_2^0)c_{Tk}^2 + 2c_{Tk}\sum_{k_2^0+1}^{T}(X_t - A_{Tk}^*) \\
&\quad + \sum_{t=k+1}^{T}(X_t - A_{Tk}^*)^2
\end{aligned} \tag{A.2}$$

where

$$a_{Tk} = \frac{1}{T-k}\{(T-k_1^0)(\mu_1-\mu_2) + (T-k_2^0)(\mu_2-\mu_3)\},$$

$$b_{Tk} = \frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1) + (T-k_2^0)(\mu_2-\mu_3)\},$$

$$c_{Tk} = \frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1) + (k_2^0-k)(\mu_3-\mu_2)\}.$$

Rewrite

$$\frac{1}{T}\sum_{t=k+1}^{T}(X_t - A_{Tk}^*)^2 = \frac{1}{T}\sum_{t=k+1}^{T}X_t^2 - \frac{T-k}{T}(A_{Tk}^*)^2 \tag{A.3}$$

and

$$\frac{1}{T}\sum_{t=1}^{k}(X_t - A_{Tk})^2 = \frac{1}{T}\sum_{t=1}^{k}X_t^2 - \frac{1}{T}\left(\frac{1}{\sqrt{k}}\sum_{t=1}^{k}X_t\right)^2. \tag{A.4}$$

Combining (A.1) and (A.2) and using (A.3) and (A.4), we have for $k \le k_1^0$

$$U_T(k/T) = \frac{1}{T}S_T(k)$$
$$= \frac{(k_1^0 - k)}{T}a_{Tk}^2 + \frac{(k_2^0 - k_1^0)}{T}b_{Tk}^2 + \frac{(T - k_2^0)}{T}c_{Tk}^2 + \frac{1}{T}\sum_{t=1}^{T}X_t^2 + R_{1T}(k), \tag{A.5}$$

where

$$R_{1T}(k) = \frac{1}{T}\left[2a_{Tk}\sum_{t=k+1}^{k_1^0}X_t + 2b_{Tk}\sum_{k_1^0+1}^{k_2^0}X_t + 2c_{Tk}\sum_{k_2^0+1}^{T}X_t\right]$$
$$- \frac{2}{T}[(k_1^0 - k)a_{Tk} + (k_2^0 - k_1^0)b_{Tk} + (T - k_2^0)c_{Tk}]A_{Tk}^*$$
$$- \frac{1}{T}\left(\frac{1}{\sqrt{k}}\sum_{t=1}^{k}X_t\right)^2 - \frac{T-k}{T}(A_{Tk}^*)^2. \tag{A.6}$$

We shall argue that

$$R_{1T}(k) = O_p(T^{-1/2}) \quad \text{uniformly in } k \in [1, k_1^0]. \tag{A.7}$$

Note that a_{Tk}, b_{Tk}, and c_{Tk} are uniformly bounded in T and in $k \le [T\tau_1^0]$ and $A_{Tk}^* = O_p(T^{-1/2})$, it is easy to see that the first two expressions on the right-hand side of (A.6) are $O_p(T^{-1/2})$ uniformly in $k \le [T\tau_1^0]$. The second to the last term is $T^{-1}O_p(\log^2 T)$ because $\sup_{1\le k\le T}|(1/\sqrt{k})\sum_{i=1}^{k}X_i| = O_p(\log T)$. Finally, the last term is $O_p(T^{-1})$ because $A_{Tk}^* = O_p(T^{-1/2})$ uniformly in $k \le k_1^0$. This gives $R_{1T}(k) = O_p(T^{-1/2})$ uniformly in $k \le k_1^0$.

Next, consider $k \in [k_1^0 + 1, k_2^0]$. We have

$$\bar{Y}_k = \frac{k_1^0}{k}\mu_1 + \frac{k - k_1^0}{k}\mu_2 + A_{Tk}, \qquad \bar{Y}_k^* = \frac{k_2^0 - k}{T - k}\mu_2 + \frac{T - k_2^0}{T - k}\mu_3 + A_{Tk}^*.$$

Thus,

$$Y_t - \bar{Y}_k = \begin{cases} \dfrac{k - k_1^0}{k}(\mu_1 - \mu_2) + X_t - A_{Tk} & \text{if } t \in [1, k_1^0], \\[2ex] \dfrac{k_1^0}{k}(\mu_2 - \mu_1) + X_t - A_{Tk} & \text{if } t \in [k_1^0 + 1, k], \end{cases}$$

$$Y_t - \bar{Y}_k^* = \begin{cases} \dfrac{T-k_2^0}{T-k}(\mu_2-\mu_3) + X_t - A_{Tk}^* & \text{if } t \in [k+1, k_2^0], \\ \dfrac{k_2^0-k}{T-k}(\mu_3-\mu_2) + X_t - A_{Tk}^* & \text{if } t \in [k_2^0+1, T]. \end{cases}$$

Hence, for $k \in [k_1^0+1, k_2^0]$,

$$\sum_{t=1}^{k}(Y_t-\bar{Y}_k)^2 = k_1^0 d_{Tk}^2 + 2d_{Tk}\sum_{t=1}^{k_1^0}(X_t - A_{Tk}) + \sum_{t=1}^{k_1^0}(X_t - A_{Tk})^2$$
$$+ (k-k_1^0)e_{Tk}^2 + 2e_{Tk}\sum_{t=k_1^0+1}^{k}(X_t - A_{Tk}) + \sum_{t=k_1^0+1}^{k}(X_t - A_{Tk})^2,$$

where $d_{Tk} = [(k-k_1^0)/k](\mu_1-\mu_2)$ and $e_{Tk} = (k_1^0/k)(\mu_2-\mu_1)$, and

$$\sum_{t=k+1}^{T}(Y_t-\bar{Y}_k^*)^2 = (k_2^0-k)f_{Tk}^2 + 2f_{Tk}\sum_{k+1}^{k_2^0}(X_t - A_{Tk}^*) + \sum_{k+1}^{k_2^0}(X_t - A_{Tk}^*)^2$$
$$+ (T-k_2^0)g_{Tk}^2 + 2g_{Tk}\sum_{k_2^0+1}^{T}(X_t - A_{Tk}^*) + \sum_{k_2^0+1}^{T}(X_t - A_{Tk}^*)^2, \tag{A.8}$$

where $f_{Tk} = [(T-k_2^0)/(T-k)](\mu_2-\mu_3)$ and $g_{Tk} = [(k_2^0-k)/(T-k)](\mu_3-\mu_2)$. Therefore,

$$U_T(k/T) = \frac{1}{T}S_T(k)$$
$$= \frac{k_1^0}{T}d_{Tk}^2 + \frac{k-k_1^0}{T}e_{Tk}^2 + \frac{k_2^0-k}{T}f_{Tk}^2 + \frac{T-k_2^0}{T}g_{Tk}^2$$
$$+ \frac{1}{T}\sum_{t=1}^{T}X_t^2 + R_{2T}(k)$$
$$= \frac{k_1^0(k-k_1^0)}{kT}(\mu_2-\mu_1)^2 + \frac{(k_2^0-k)(T-k_2^0)}{T(T-k)}(\mu_3-\mu_2)^2$$
$$+ \frac{1}{T}\sum_{t=1}^{T}X_t^2 + R_{2T}(k), \tag{A.9}$$

where

$$R_{2T}(k) = \frac{1}{T}\left[2d_{Tk}\sum_{t=1}^{k_1^0}X_t + 2e_{Tk}\sum_{t=k_1^0+1}^{k}X_t + 2f_{Tk}\sum_{k+1}^{k_2^0}X_t + 2g_{Tk}\sum_{k_2^0+1}^{T}X_t\right]$$
$$- \frac{2}{T}[k_1^0 d_{Tk} + (k-k_1^0)e_{Tk} + (k_2^0-k)f_{Tk} + (T-k_2^0)g_{Tk}]A_{Tk}^*$$
$$- \frac{k}{T}(A_{Tk})^2 - \frac{T-k}{T}(A_{Tk}^*)^2. \tag{A.10}$$

Using the uniform boundedness of d_{Tk}, e_{Tk}, f_{Tk}, and g_{Tk} as well as $A_{Tk} = O_p(T^{-1/2})$ and $A^*_{Tk} = O_p(T^{-1/2})$ uniformly in $k \in [k_1^0 + 1, k_2^0]$, we can easily show that $R_{2T}(k) = O_p(T^{-1/2})$ uniformly in $k \in [k_1^0 + 1, k_2^0]$. As for $k > k_2^0$, using the symmetry with the first regime, we have

$$U_T(k/T) = \frac{k_1^0}{T} h_{Tk}^2 + \frac{k_2^0 - k_1^0}{T} p_{Tk}^2 + \frac{k - k_2^0}{T} q_{Tk}^2 + \frac{1}{T}\sum_{t=1}^{T} X_t^2 + R_{3T}(k), \qquad \textbf{(A.11)}$$

where, similar to before, $R_{3T}(k) = O_p(T^{-1/2})$ uniformly for $k \in [k_2^0, T]$ and

$$h_{Tk} = \frac{1}{k}\,[(k - k_1^0)(\mu_1 - \mu_2) + (k - k_2)(\mu_2 - \mu_3)],$$

$$p_{Tk} = \frac{1}{k}\,[k_1^0(\mu_2 - \mu_1) + (k - k_2^0)(\mu_2 - \mu_3)],$$

$$q_{Tk} = \frac{1}{k}\,[k_1^0(\mu_2 - \mu_1) + k_2^0(\mu_3 - \mu_2)].$$

Proof of Lemma 1. Because $a_{Tk}, b_{Tk}, \ldots, q_{Tk}$ all have uniform limits for $k = [T\tau]$ and the stochastic terms in (A.5), (A.9), and (A.11) all have uniform limits in pertinent regions for $\tau \in [0,1]$, the uniform convergence of $U_T(\tau)$ follows easily. The uniform limit of $U_T(\tau)$ is also easy to obtain. Note that (4) and (5) are obtained, respectively, by taking $k = k_1^0$ and $k = k_2^0$ in (A.9) and letting $T \to \infty$. ■

Proof of Lemma 2. The only stochastic terms in (A.5), (A.9), and (A.11) are $R_{iT}(k)$ $(i = 1,2,3)$, each of which is $O_p(T^{-1/2})$ uniformly over pertinent regions for k. Furthermore, it is easy to see that $ER_{iT}(k) = O(T^{-1})$ uniformly in k $(i = 1,2,3)$. These results imply Lemma 2. ■

To prove Lemma 3, we need additional results.

LEMMA 11. *There exists an $M < \infty$ such that for all i and all $j > i$*

$$\left| E\left\{ \left(\sum_{t=1}^{i} X_t\right)\left(\sum_{s=i+1}^{j} X_s\right)\right\} \right| \le M.$$

Proof. Let $\gamma(h) = E(X_t X_{t+h})$. Then, under Assumptions A1 and A2, it is easy to argue that $\sum_{h=1}^{\infty} h|\gamma(h)| < \infty$. Now,

$$\left| E\left(\sum_{t=1}^{i} X_t\right)\left(\sum_{s=i+1}^{j} X_s\right)\right| = \left| \sum_{t=1}^{i}\sum_{s=i+1}^{j} \gamma(s-t)\right| \le \sum_{h=1}^{j} h|\gamma(h)| \le \sum_{h=1}^{\infty} h|\gamma(h)| < \infty.$$

■

We will also use the following result: there exists an $M < \infty$, such that for arbitrary $i < j$,

$$E\,\frac{1}{j-i}\left(\sum_{t=i+1}^{j} X_t\right)^2 < M. \qquad \textbf{(A.12)}$$

In the sequel, we shall use a_{Tk} and $a_T(k)$ interchangeably. Similar notations are also adopted for A_{Tk} and A^*_{Tk} as well as for $b_{Tk}, c_{Tk}, \ldots$.

LEMMA 12. *Under Assumptions* A1–A3, *there exists an* $M < \infty$ *such that*

$$T|ER_{1T}(k) - ER_{1T}(k_1^0)| \le \frac{|k_1^0 - k|}{T} M.$$

Proof. The expected value of the first two terms on the right-hand side of (A.6) is zero. We thus need to consider the last two terms. For $k < k_1^0$,

$$\begin{aligned}
&\frac{1}{k}\left(\sum_{t=1}^{k} X_t\right)^2 - \frac{1}{k_1^0}\left(\sum_{t=1}^{k_1^0} X_t\right)^2 \\
&\quad = \left(\frac{1}{k} - \frac{1}{k_1^0}\right)\left(\sum_{t=1}^{k} X_t\right)^2 - 2\frac{1}{k_1^0}\left(\sum_{t=1}^{k} X_t\right)\left(\sum_{t=k+1}^{k_1^0} X_t\right) - \frac{1}{k_1^0}\left(\sum_{t=k+1}^{k_1^0} X_t\right)^2 \\
&\quad = \frac{k_1^0 - k}{k_1^0}\,\frac{1}{k}\left(\sum_{t=1}^{k} X_t\right)^2 - 2\frac{1}{k_1^0}\left(\sum_{t=1}^{k} X_t\right)\left(\sum_{t=k+1}^{k_1^0} X_t\right) - \frac{k_1^0 - k}{k_1^0}\,\frac{1}{k_1^0 - k}\left(\sum_{t=k+1}^{k_1^0} X_2\right)^2.
\end{aligned} \tag{A.13}$$

Apply Lemma 11 to the second term above and apply (A.12) to the first term and the third term above; we see that the absolute value of the expectation of (A.13) is bounded by $M|k_1^0 - k|/T$. This result holds for $k > k_1^0$ (we only need to use $\Sigma_{t=1}^{k} = \Sigma_{t=1}^{k_1^0} + \Sigma_{t=k_1^0+1}^{k}$ in the proof). By symmetry,

$$|(T-k)E(A_T^*(k))^2 - (T-k_1^0)E(A_T^*(k_1^0))^2| \le M|k_1^0 - k|/T.$$

Combining these results, we obtain Lemma 12. ∎

Note that the expected values of $R_{jT}(k)$ for $j = 1,2,3$ have an identical expression as functions of k. We thus have

$$T|ER_{iT}(k) - ER_{iT}(k_1^0)| \le \frac{|k_1^0 - k|}{T} M \qquad (i = 1,2,3). \tag{A.14}$$

LEMMA 13. *Under Assumptions* A1–A3, *for* $k \le k_1^0$,

$$ES_T(k) - ES_T(k_1^0) \ge T\{ER_{1T}(k) - ER_{1T}(k_1^0)\} \ge -M|k_1^0 - k|/T \tag{A.15}$$

and, for $k \ge k_2^0$,

$$ES_T(k) - ES_T(k_2^0) \ge T\{ER_{3T}(k) - ER_{3T}(k_2^0)\} \ge -M|k_2^0 - k|/T. \tag{A.16}$$

Proof. For $k \le k_1^0$, using (A.5) with some algebra we obtain

$$\begin{aligned}
&ES_T(k) - ES_T(k_1^0) \\
&\quad = (k_1^0 - k)a_T(k)^2 + (k_2^0 - k_1^0)[b_T(k)^2 - b_T(k_1^0)^2] \\
&\qquad + (T - k_2^0)[c_T(k)^2 - c_T(k_1^0)^2] + T\{ER_{1T}(k) - ER_{1T}(k_1^0)\} \\
&\quad = \frac{k_1^0 - k}{(1 - k/T)(1 - k_1^0/T)}[(1 - k_1^0/T)(\mu_1 - \mu_2) + (1 - k_2^0/T)(\mu_2 - \mu_3)]^2 \\
&\qquad + T\{ER_{1T}(k) - ER_{1T}(k_1^0)\}.
\end{aligned} \tag{A.17}$$

The first term on the right-hand side of the preceding is nonnegative. This together with Lemma 12 yields (A.15). Inequality (A.16) follows from symmetry (which can be thought of as reversing the data order). ■

LEMMA 14. *Under Assumptions* A1–A3 *and* $U(\tau_1^0) \le U(\tau_2^0)$ (*equality is allowed*), *for* $k \le k_1^0$, *there exists a* $C > 0$ *such that*

$ES_T(k) - ES_T(k_1^0) \ge C|k - k_1^0|$ *for all large T.*

Proof. Because $|(k_i^0/T) - \tau_i^0| \le T^{-1}$ $(i = 1,2)$, we can rewrite (A.17) as

$$ES_T(k) - ES_T(k_1^0)$$
$$= \frac{k_1^0 - k}{(1 - k/T)(1 - k_1^0/T)} [(1 - \tau_1^0)(\mu_1 - \mu_2) + (1 - \tau_2^0)(\mu_2 - \mu_3)]^2$$
$$+ O\left(\frac{k_1^0 - k}{T}\right) + T\{ER_{1T}(k) - ER_{1T}(k_1^0)\}.$$

We claim that when $U(\tau_1^0) \le U(\tau_2^0)$,

$$C = (1 - \tau_1^0)(\mu_1 - \mu_2) + (1 - \tau_2^0)(\mu_2 - \mu_3) \neq 0. \tag{A.18}$$

Condition $U(\tau_1^0) \le U(\tau_2^0)$ is equivalent to

$$\frac{1 - \tau_2^0}{1 - \tau_1^0} (\mu_2 - \mu_3)^2 \le \frac{\tau_1^0}{\tau_2^0} (\mu_1 - \mu_2)^2.$$

Multiplying $(1 - \tau_2^0)(1 - \tau_1^0)$ on both sides of the preceding and using $(1 - \tau_2^0)\tau_1^0/\tau_2^0 < 1 - \tau_1^0$ we obtain

$$(1 - \tau_2^0)^2(\mu_2 - \mu_3)^2 < (1 - \tau_1^0)^2(\mu_1 - \mu_2)^2.$$

This verifies (A.18). Together with Lemma 12, we have, for all large T,

$$ES_T(k) - ES_T(k_1^0) \ge (k_1^0 - k)C^2 - O\left(\frac{k_1^0 - k}{T}\right) \ge (k_1^0 - k)C^2/2. \tag{A.19}$$

■

Proof of Lemma 3. For $k \le k_1^0$, Lemma 3 is implied by Lemma 14. For $k \in [k_1^0 + 1, k_2^0]$, use the last equality of (A.9) with some algebra,

$$ES_T(k) - ES_T(k_1^0) = (k - k_1^0)\left[\frac{k_1^0}{k} (\mu_2 - \mu_1)^2 - \frac{(T - k_2^0)^2}{(T - k)(T - k_1^0)} (\mu_3 - \mu_2)^2\right]$$
$$+ T\{ER_{2T}(k) - ER_{2T}(k_1^0)\}. \tag{A.20}$$

Factor out k_2^0/k, and use $k(T - k_2^0)/\{k_2^0(T - k)\} \le 1$, for all $k \le k_2^0$,

$$ES_T(k) - ES_T(k_1^0) \ge (k - k_1^0)\frac{k_2^0}{k}\left[\frac{k_1^0}{k_2^0} (\mu_2 - \mu_1)^2 - \frac{(T - k_2^0)}{(T - k_1^0)} (\mu_3 - \mu_2)^2\right]$$
$$+ T\{ER_{2T}(k) - ER_{2T}(k_1^0)\}.$$

Denote $C^* = (\tau_1^0/\tau_2^0)(\mu_2 - \mu_1)^2 - [(1 - \tau_2^0)/(1 - \tau_1^0)] (\mu_3 - \mu_2)^2$. By (6), $C^* > 0$. From

$$\frac{k_1^0}{k_2^0} - \frac{\tau_1^0}{\tau_2^0} = O(T^{-1}), \qquad \frac{T-k_2^0}{T-k_1^0} - \frac{1-\tau_2^0}{1-\tau_1^0} = O(T^{-1}), \tag{A.21}$$

we have

$$\begin{aligned} &ES_T(k) - ES_T(k_1^0) \\ &\quad \geq (k-k_1^0)\frac{k_2^0}{k}C^* - (k-k_1^0)O(T^{-1}) + T\{ER_{2T}(k) - ER_{2T}(k_1^0)\} \\ &\quad \geq (k-k_1^0)C^* - M\frac{k-k_1^0}{T} \end{aligned}$$

for some $M < \infty$ by (A.14). Thus, for large T,

$$ES_T(k) - ES_T(k_1^0) \geq (k-k_1^0)C^*/2. \tag{A.22}$$

It remains to consider $k \in [k_2^0 + 1, T]$. From (A.16), $ES_T(k) - ES_T(k_2^0) \geq -(k-k_2^0)M/T \geq -(T-k_2^0)M/T$. Thus,

$$\begin{aligned} ES_T(k) - ES_T(k_1^0) &= ES_T(k) - ES_T(k_2^0) + ES_T(k_2^0) - ES_T(k_1^0) \\ &\geq ES_T(k_2^0) - ES_T(k_1^0) - (T-k_2^0)M/T \\ &\geq (k-k_1^0)\frac{T-k_2^0}{T-k_1^0}\left[\frac{ES_T(k_2^0) - ES_T(k_1^0)}{T-k_2^0} - \frac{M}{T}\right]; \end{aligned}$$

the last inequality follows from $(k-k_1^0)/(T-k_1^0) \leq 1$. Using (A.22) with $k = k_2^0$, we see that the term in the bracket is no smaller than $[(\tau_2^0 - \tau_1^0)/(1-\tau_2^0)]C^*/4$ for large T. Thus,

$$ES_T(k) - ES_T(k_1^0) \geq (k-k_1^0)\frac{\tau_2^0 - \tau_1^0}{1-\tau_1^0}C^*/8 \tag{A.23}$$

for all large T. Combining (A.19), (A.22), and (A.23), we obtain Lemma 3. ∎

Proof of Lemma 4. Rewrite

$$S_T(k) - S_T(k_1^0) = S_T(k) - ES_T(k) - [S_T(k_1^0) - ES_T(k_1^0)] + ES_T(k) - ES_T(k_1^0).$$

From Lemma 3, $S_T(k) - S_T(k_1^0) \leq 0$ implies that

$$\{S_T(k) - ES_T(k) - [S_T(k_1^0) - ES_T(k_1^0)]\}/|k-k_1^0| \leq -C.$$

This further implies that the absolute value of the left-hand side of the preceding is at least as large as C. We show this is unlikely for $k \in D_{T,M}$. More specifically, for every $\epsilon > 0$ and $\eta > 0$, there exists an $M > 0$ such that for all large T

$$P\left(\sup_{k \in D_{T,M}} |S_T(k) - ES_T(k) - [S_T(k_1^0) - ES_T(k_1^0)]|/|k-k_1^0| > \eta\right) < \epsilon.$$

First, note that

$$\begin{aligned} &|S_T(k) - ES_T(k) - \{S_T(k_1^0) - ES_T(k_1^0)\}| \\ &\quad = |T\{R_{1T}(k) - ER_{1T}(k)\} - T\{R_{1T}(k_1^0) - ER_{1T}(k_1^0)\}| \\ &\quad \leq |T\{R_{1T}(k) - R_{1T}(k_1^0)\}| + M'|k-k_1^0|T \end{aligned}$$

for some $M' < \infty$ by Lemma 12. Thus, it suffices to show

$$P\left(\sup_{k \in D_{T,M}} T|R_{1T}(k) - R_{1T}(k_1^0)|/|k - k_1^0| > \eta\right) < \epsilon. \tag{A.24}$$

We consider the case $k < k_1^0$. From (A.6),

$$\begin{aligned} &T\{R_{1T}(k) - R_{1T}(k_1^0)\} \\ &\quad = 2\left(a_{Tk}\sum_{t=k+1}^{k_1^0} X_t\right) + 2\left(\{b_{Tk} - b_T(k_1^0)\}\sum_{k_1^0+1}^{k_2^0} X_t\right) + 2\left(\{c_{Tk} - c_T(k_1^0)\}\sum_{k_2^0+1}^{T} X_t\right) \\ &\quad\quad - (k_1^0 - k)a_{Tk}A_{Tk}^* - (k_2^0 - k_1^0)\{b_{Tk}A_{Tk}^* - b_T(k_1^0)A_T(k_1^0)^*\} \\ &\quad\quad - (T - k_2^0)\{c_{Tk}A_{Tk}^* - c_T(k_1^0)A_T^*(k_1^0)\} \\ &\quad\quad + \left[\frac{1}{k_1^0}\left(\sum_{t=1}^{k_1^0} X_t\right)^2 - \frac{1}{k}\left(\sum_{t=1}^{k} X_t\right)^2\right] + [(T - k_1^0)(A_T^*(k_1^0))^2 - (T - k)(A_{Tk}^*)^2]. \end{aligned} \tag{A.25}$$

We shall show that each term on the right-hand side divided by $k_1^0 - k$ is arbitrarily small in probability as long as M is large and T is large. Because a_{Tk}, b_{Tk}, and c_{Tk} are all uniformly bounded, with an upper bound, say, L, the first term divided by $k_1^0 - k$ is bounded by $L|[1/(k_1^0 - k)]\sum_{k+1}^{k_1^0} X_t|$, which is uniformly small in $k < k_1^0 - M$ for large M by the strong law of large numbers. For the rest of the terms, we will use the following easily verifiable facts:

$$|b_{Tk} - b_T(k_1^0)| \le \left|\frac{k_1^0 - k}{T - k}\right| C, \qquad |c_{Tk} - c_T(k_1^0)| \le \left|\frac{k_1^0 - k}{T - k}\right| C, \tag{A.26}$$

for some $C < \infty$, and

$$A_T^*(k_1^0) - A_T^*(k) = \frac{k_1^0 - k}{(T - k)(T - k_1^0)}\sum_{t=k_0}^{T} X_t - \frac{1}{T - k}\sum_{t=k+1}^{k_1^0} X_t. \tag{A.27}$$

By (A.26), the second term on the right-hand side of (A.25) divided by $k_1^0 - k$ is bounded by

$$C\frac{1}{T - k}\left|\sum_{k_1^0+1}^{k_2^0} X_t\right| = C\frac{k_2^0 - k_1^0}{T - k}\left(\frac{1}{k_2^0 - k_1^0}\right)\left|\sum_{k_1^0+1}^{k_2^0} X_t\right| \le C'\frac{1}{k_2^0 - k_1^0}\left|\sum_{k_1^0+1}^{k_2^0} X_t\right|$$

for some $C' < \infty$, which converges to zero in probability by the law of large numbers (note that $T - k \ge T(1 - \tau_2^0)$ for all $k \in D_T$). The third term is treated similarly. The fourth term divided by $k_1^0 - k$ is bounded by $L|A_{Tk}^*| = O_p(T^{-1/2})$ uniformly in $k \in D_T$. The fifth term can be rewritten as

$$(k_2^0 - k_1^0)\{b_{Tk} - b_T(k_1^0)\}A_{Tk}^* + (k_2^0 - k_1^0)b_T(k_1^0)\{A_T^*(k_1^0) - A_T^*(k)\}. \tag{A.28}$$

Using (A.26), the first expression of (A.28) divided by $k_1^0 - k$ is readily seen to be $o_p(1)$. The second expression divided by $k_1^0 - k$ is equal to, by (A.27),

$$\frac{k_2^0 - k_1^0}{(T - k)(T - k_1^0)}\sum_{t=k_0}^{T} X_t - \frac{k_2^0 - k_1^0}{T - k}\left(\frac{1}{k_1^0 - k}\right)\sum_{t=k+1}^{k_1^0} X_t, \tag{A.29}$$

with the first term being $o_p(1)$ and the second term being small for large M. Thus, the fifth term of (A.25) is small if M is large. The sixth term is treated similarly to the fifth one. It is also elementary to show that the seventh term and the eighth term of (A.25) divided by $k_1^0 - k$ can be arbitrarily small in probability provided that M and T are large. This proves Lemma 4 for $k < k_1^0$. The case of $k > k_1^0$ is similar; the details are omitted. ∎

Proof of Lemma 5. We prove the first inequality; the second follows from symmetry. For $k \le k_1^0$, Lemma 5 is implied by Lemma 14, which holds for $U(\tau_1^0) = U(\tau_2^0)$. Next, consider $k \in [k_1^0 + 1, k_0^*]$. From (A.20), (A.21), and the condition $U(\tau_1^0) = U(\tau_2^0)$ (i.e., $(\tau_1^0/\tau_2^0)(\mu_2 - \mu_1)^2 = [(1-\tau_2^0)/(1-\tau_1^0)]\,(\mu_3 - \mu_2)^2$), we have

$$\begin{aligned} &ES_T(k) - ES_T(k_1^0) \\ &\quad = (k - k_1^0)\left(\frac{k_2^0}{k} - \frac{T - k_2^0}{T - k}\right)\frac{\tau_1^0}{\tau_2^0}(\mu_2 - \mu_1)^2 \\ &\qquad + (k - k_1^0)O(T^{-1}) + T\{ER_{2T}(k) - ER_{2T}(k_1^0)\}. \end{aligned} \tag{A.30}$$

Note that for all $k \le k_0^* = (k_1^0 + k_2^0)/2$

$$\frac{k_2^0}{k} - \frac{T - k_2^0}{T - k} = \frac{(k_2^0 - k)T}{k(T - k)} \ge 2^{-1}\frac{(k_2^0 - k_1^0)T}{k(T - k)} \ge 2\,\frac{k_2^0 - k_1^0}{T} \ge \tau_2^0 - \tau_1^0.$$

The last two terms of (A.30) on the right-hand side are dominated by the first term. The lemma is proved. ∎

Proof of Lemma 6. It is enough to prove the lemma for $i = 1$. The case of $i = 2$ follows from symmetry. The proof is virtually identical to that of Lemma 4. One uses Lemma 5 instead of Lemma 3. The rest can be copied here. ∎

Proof of Lemma 7. We shall prove $P(S_T(\hat{k}_1^\dagger) - S_T(\hat{k}_2^\dagger) < 0) \to \frac{1}{2}$ or, equivalently, $P(T^{-1/2}\{S_T(\hat{k}_1^\dagger) - S_T(\hat{k}_2^\dagger)\} < 0) \to \frac{1}{2}$. Because $\hat{k}_i^\dagger = k_i^0 + O_p(1)$, $S_T(\hat{k}_i) = S_T(k_i^0) + O_p(1)$ (see the proof of Proposition 4). Because $T^{-1/2}O_p(1) \xrightarrow{p} 0$, it suffices to prove

$$P(T^{-1/2}\{S_T(k_1^0) - S_T(k_2^0)\} < 0) \to \tfrac{1}{2}.$$

Using $|(k_i^0/T) - \tau_i^0| \le 1/T$, it is easy to show that $|ES_T(k_i^0) - TU(\tau_i^0)| < A$ for some $A < \infty$. This implies $|ES_T(k_1^0) - ES_T(k_2^0)| < 2A$ because $U(\tau_1^0) = U(\tau_2^0)$. Thus,

$$T^{-1/2}\{S_T(k_1^0) - S_T(k_2^0)\} = \sqrt{T}\{R_{2T}(k_1^0) - R_{2T}(k_2^0)\} + O(T^{-1/2}),$$

where $R_{2T}(k_i^0) = T^{-1}\{S_T(k_i^0) - ES_T(k_i^0)\}$ $(i = 1,2)$ (see (A.9)). Note that we have used the fact that $S_T(k)$, when $k = k_1^0$, can be represented by both (A.5) and (A.9) and we have used (A.9). Thus, it suffices to prove $P\big(\sqrt{T}\{R_{2T}(k_1^0) - R_{2T}(k_2^0)\} < 0\big) \to \frac{1}{2}$. From (A.10),

$$T^{1/2}R_{2T}(k_1^0) = 2f_T(k_1^0)\frac{1}{\sqrt{T}}\sum_{k_1^0+1}^{k_2^0} X_t + 2g_T(k_1^0)\frac{1}{\sqrt{T}}\sum_{k_2^0+1}^{T} X_t + O_p(T^{-1/2}).$$

The preceding follows from $(k_2^0 - k_1^0)f_T(k_1^0) + (T - k_2^0)g_T(k_1^0) = 0$ and $(T - k_1^0)T^{-1/2}(A_T^*(k_1^0)) = O_p(T^{-1/2})$. Similarly,

$$T^{1/2}R_{2T}(k_2^0) = 2d_T(k_2^0)\frac{1}{\sqrt{T}}\sum_{t=1}^{k_1^0} X_t + 2e_T(k_2^0)\frac{1}{\sqrt{T}}\sum_{k_1^0+1}^{k_2^0} X_t + O_p(T^{-1/2}).$$

Thus, $T^{1/2}\{R_{2T}(k_1^0) - R_{2T}(k_2^0)\}$ converges in distribution to a mean zero normal random variable by the central limit theorem. The lemma follows because a mean zero normal random variable is symmetric about zero. ■

Proof of Proposition 4. Consider the process $S_T(k_1^0 + l) - S_T(k_1^0)$ indexed by l, where l is an integer (positive or negative). Suppose that the minimum of this process is attained at $\hat{l}$. By definition, $\hat{l} = \hat{k} - k_1^0$. By Proposition 2, for each $\epsilon > 0$, there exists an $M < \infty$ such that $P(|\hat{k} - k_1^0| > M) = P(|\hat{l}| > M) < \epsilon$. Thus, to study the limiting distribution of $\hat{l} = \hat{k} - k_1^0$, it suffices to study the behavior of $S_T(k_1^0 + l) - S_T(k_1^0)$ for bounded l. We shall prove that $S_T(k_1^0 + l) - S_T(k_1^0)$ converges in distribution for each l to $(1 + \lambda_1)W^{(1)}(l,\lambda_1)$, where λ_1 and $W^{(1)}(l,\lambda_1)$ are defined in the text. This will imply that $\hat{k} - k_1^0 \xrightarrow{d} \text{argmin}_l(1 + \lambda_1)W^{(1)}(l,\lambda_1)$ (see Bai, 1994a). Because $(1 + \lambda_1) > 0$, $\text{argmin}_l(1 + \lambda_1)W^{(1)}(l,\lambda_1) = \text{argmin}_l W^{(1)}(l,\lambda_1)$, giving rise to the proposition. First consider the case of $l > 0$ and $l \le M$, where $M > 0$ is an arbitrary finite number. Let

$$\hat{\mu}_1^* = \frac{1}{k_1^0 + l}\sum_{t=1}^{k_1^0+l} Y_t \quad \text{and} \quad \hat{\mu}_2^* = \frac{1}{T - k_1^0 - l}\sum_{t=k_1^0+l+1}^{T} Y_t, \tag{A.31}$$

$$\hat{\mu}_1 = \frac{1}{k_1^0}\sum_{t=1}^{k_1^0} Y_t \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{T - k_1^0}\sum_{t=k_1^0+1}^{T} Y_t. \tag{A.32}$$

Thus, $\hat{\mu}_1^*$ is the least-squares estimator of μ_1 using the first $k_1^0 + l$ observations and $\hat{\mu}_2^*$ is the least-squares estimator of a weighted average of μ_2 and μ_3 using the last $T - k_1^0 - l$ observations. The interpretation of $\hat{\mu}_i$ $(i = 1,2)$ is similar. The estimators $\hat{\mu}_i^*$ $(i = 1,2)$ depend on l. This dependence will be suppressed for notational simplicity. It is straightforward to establish the following result:

$$\hat{\mu}_1^* - \mu_1 = O_p(T^{-1/2}) \quad \text{and} \quad \hat{\mu}_1 - \mu_1 = O_p(T^{-1/2}) \tag{A.33}$$

$$\hat{\mu}_2^* - \mu_2 - \frac{1 - \tau_2^0}{1 - \tau_1^0}(\mu_3 - \mu_2) = O_p(T^{-1/2}) \quad \text{and}$$

$$\hat{\mu}_2 - \mu_2 - \frac{1 - \tau_2^0}{1 - \tau_1^0}(\mu_3 - \mu_2) = O_p(T^{-1/2}) \tag{A.34}$$

$$\hat{\mu}_i^* - \hat{\mu}_i = O_p(T^{-1}) \qquad (i = 1,2) \tag{A.35}$$

where the $O_p(\cdot)$ terms are uniform in l such that $|l| \le M$. Now,

$$S_T(k_1^0 + l) = \sum_{t=1}^{k_1^0} (Y_t - \hat{\mu}_1^*)^2 + \sum_{t=k_1^0+1}^{k_1^0+l} (Y_t - \hat{\mu}_1^*)^2 + \sum_{k_1^0+l+1}^{T} (Y_t - \hat{\mu}_2^*)^2. \tag{A.36}$$

Similarly,

$$S_T(k_1^0) = \sum_{t=1}^{k_1^0} (Y_t - \hat{\mu}_1)^2 + \sum_{t=k_1^0+1}^{k_1^0+l} (Y_t - \hat{\mu}_2)^2 + \sum_{k_1^0+l+1}^{T} (Y_t - \hat{\mu}_2)^2. \tag{A.37}$$

The difference between the two first terms on the right-hand sides of (A.36) and (A.37) is

$$\sum_{t=1}^{k_1^0} (Y_t - \hat{\mu}_1^*)^2 - \sum_{t=1}^{k_1^0} (Y_t - \hat{\mu}_1)^2 = k_1^0(\hat{\mu}_1^* - \hat{\mu}_1)^2 = O_p(T^{-1}). \tag{A.38}$$

Similarly, the difference between the two third terms on the right-hand sides of (A.36) and (A.37) is also $O_p(T^{-1})$. Next, consider the difference between the two middle terms. For $t \in [k_1^0 + 1, l]$, $Y_t = \mu_2 + X_t$. Hence,

$$\sum_{t=k_1^0+1}^{k_1^0+l} (Y_t - \hat{\mu}_1^*)^2 - \sum_{t=k_1^0+1}^{k_1^0+l} (Y_t - \hat{\mu}_2)^2$$

$$= 2\{\mu_2 - \hat{\mu}_1^* - (\mu_2 - \hat{\mu}_2)\} \sum_{t=k_1^0+1}^{k_1^0+l} X_t + l\{(\mu_2 - \hat{\mu}_1^*)^2 - (\mu_2 - \hat{\mu}_2)^2\}. \tag{A.39}$$

From (A.33) and (A.34), we have

$$\mu_2 - \hat{\mu}_1^* - (\mu_2 - \hat{\mu}_2) = (\mu_2 - \mu_1)(1 + \lambda_1) + O_p(T^{-1/2})$$

and

$$(\mu_2 - \hat{\mu}_1^*)^2 - (\mu_2 - \hat{\mu}_2)^2 = (\mu_2 - \mu_1)^2(1 - \lambda_1^2) + O_p(T^{-1/2}).$$

Thus, (A.39) is equal to

$$2(\mu_2 - \mu_1)(1 + \lambda_1) \sum_{t=k_1^0+1}^{k_1^0+l} X_t + l(\mu_2 - \mu_1)^2(1 - \lambda_1^2) + O_p(T^{-1/2}). \tag{A.40}$$

Under strict stationarity, $\sum_{t=k_1^0+1}^{k_1^0+l} X_t$ has the same distribution as $\sum_{t=1}^{l} X_t$. Thus, (A.40) or, equivalently, (A.39) converges in distribution to $(1 + \lambda_1)W_2^{(1)}(l,\lambda_1)$. This implies that $S_T(k_1^0 + l) - S_T(k_1^0)$ converges in distribution to $(1 + \lambda_1)W_2^{(1)}(l,\lambda_1)$ for $l > 0$. It remains to consider $l < 0$. We replace l by $-l$ and still assume l positive. In particular, $\hat{\mu}_1^*$ and $\hat{\mu}_2^*$ are defined with $-l$ in place of l. Then, (A.36) and (A.37) are replaced, respectively, by

$$S_T(k_1^0 - l) = \sum_{t=1}^{k_1^0-l} (Y_t - \hat{\mu}_1^*)^2 + \sum_{t=k_1^0-l+1}^{k_1^0} (Y_t - \hat{\mu}_2^*)^2 + \sum_{k_1^0+1}^{T} (Y_t - \hat{\mu}_2^*)^2 \tag{A.41}$$

and

$$S_T(k_1^0) = \sum_{t=1}^{k_1^0-l} (Y_t - \hat{\mu}_1)^2 + \sum_{t=k_1^0-l+1}^{k_1^0} (Y_t - \hat{\mu}_1)^2 + \sum_{k_1^0+1}^{T} (Y_t - \hat{\mu}_2)^2. \tag{A.42}$$

The major distinction between (A.36) and (A.41) lies in the change of $\hat{\mu}_1^*$ to $\hat{\mu}_2^*$ for the middle terms on the right-hand side. One can observe a similar change for (A.37) and (A.42). Similar to (A.38), the difference between the two first terms on the right-hand sides of (A.41) and (A.42) is $O_p(T^{-1})$. The same is true for the difference between the two third terms. Using $Y_t = \mu_1 + X_t$ for $t \leq k_1^0$, we have

$$\sum_{t=k_1^0-l+1}^{k_1^0} (Y_t - \hat{\mu}_2^*)^2 - \sum_{t=k_1^0-l+1}^{k_1^0} (Y_t - \hat{\mu}_1)^2$$
$$= 2(\mu_1 - \hat{\mu}_2^*) \sum_{t=k_1^0-l+1}^{k_1^0} X_t + (\mu_1 - \hat{\mu}_2^*)^2 l + O_p(T^{-1/2})$$
$$= -2(\mu_2 - \mu_1)(1 + \lambda_1) \sum_{t=k_1^0-l+1}^{k_1^0} X_t + (\mu_2 - \mu_1)^2(1 + \lambda_1)^2 l + O_p(T^{-1/2}).$$

Ignoring the $O_p(T^{-1/2})$ term and using strict stationarity, we see that the preceding has the same distribution as $(1 + \lambda_1)W_1^{(1)}(-l,\lambda_1)$. In summary, we have proved that $S_T(k_1^0 + l) - S_T(k_1^0)$ converges in distribution to $(1 + \lambda_1)W^{(1)}(l,\lambda_1)$. This convergence implies that $\hat{k} - k_1^0 \xrightarrow{d} \text{argmin}_l(1 + \lambda_1)W^{(1)}(l,\lambda_1) = \text{argmin}_l W^{(1)}(l,\lambda_1)$. The proof of Proposition 4 is complete. ■

Proof of Proposition 5. The argument is virtually the same as in the proof of Proposition 4. The reason for $\lambda = 0$ is that regression coefficients can be consistently estimated in this case, in contrast with the inconsistent estimation given in (A.34). The details will not be presented to avoid repetition. ■

Proof of Proposition 6. By the T consistency of $\hat{k}_{i-1}$ and $\hat{k}_{i+1}$, we see that k_i^0 is a nontrivial and dominating break point in the interval $[\hat{k}_{i-1}, \hat{k}_{i+1}]$. Thus, the T consistency of $\hat{k}_i^*$ for k_i^0 follows from the property of sequential estimator. The argument for the limiting distribution is the same as that of Proposition 5. ■

Proof of Lemma 8.

(a) First, consider $k \leq k_1^0$. From (A.5),

$$\left| U_T(k/T) - EU_T(k/T) - T^{-1} \sum_{t=1}^{T} (X_t^2 - EX_t^2) \right| = |R_{1T}(k) - ER_{1T}(k)|. \quad \textbf{(A.43)}$$

The first two terms of $R_{1T}(k)$ (see (A.6)) are linear in μ_{iT} and, thus, are $v_T O_p(T^{-1/2})$. The last two terms do not depend on μ_{iT} but are of higher order than $v_T O_p(T^{-1/2})$. Moreover, $ER_{1T}(k) = O(T^{-1})$ uniformly in k. Thus, $|R_{1T}(k) - ER_{1T}(k)| = O_p(T^{-1/2}v_T)$ uniformly in $k \leq k_1^0$. This proves the lemma for $k \leq k_1^0$. The proof for $k \geq k_1^0$ is the same and follows from $R_{iT}(k) - ER_{iT}(k) = O_p(T^{-1/2}v_T)$ $(i = 2,3)$.

(b) Consider first $k \leq k_1^0$. By the second equality of (A.17), the first term of $ES_T(k) - ES_T(k_1^0)$ on the right-hand side depends on the squared and the cross-product of $\mu_{iT} - \mu_{(i+1)T}$ $(i = 1,2)$ (hence, on v_T^2). Factor out v_T^2 and replace μ_j by $\tilde{\mu}_j$; the rest of proof will be the same as that of Lemma 3. This implies that

$$ES_T(k) - ES_T(k_1^0) \geq v_T^2(k_1^0 - k)C^2/2,$$

where C is given by (A.18) with $\tilde{\mu}_j$ in place of μ_j. The proof for $k > k_1^0$ is similar and the details are omitted. ■

Proof of Lemma 9. As in the proof of Lemma 4, it suffices to show that for every $\eta > 0$ there exists an $M > 0$ such that, for all large T,

$$P\left(\sup_{k\in D^*_{T,M}} T|R_{1T}(k)-R_{1T}(k_1^0)|/|k-k_1^0| > \eta v_T^2\right) < \epsilon. \tag{A.44}$$

The preceding is similar to (A.24) with η replaced by ηv_T^2 and $D_{T,M}$ replaced by $D^*_{T,M}$. Note for $k \in D^*_{TM}$ we either have $k < k_1^0 - Mv_T^{-2}$ or $k > k_1^0 + Mv_T^{-2}$. Consider $k < k_1^0 - Mv_T^{-2}$. We need to show that each term on the right-hand side of (A.25) divided by $k_1^0 - k$ is no larger than ηv_T^2 as long as M is large and T is large. The proof requires the Hajek and Renyi inequality, extended to linear processes by Bai (1994b): there exists a $C_1 < \infty$ such that for each $l > 0$

$$P\left(\sup_{k\ge l} \frac{1}{k}\left|\sum_{t=1}^{k} X_t\right| > \alpha\right) < \frac{C_1}{\alpha^2 l}.$$

Now consider the first term on the right-hand side of (A.25). Note that $|a_{Tk}| \le v_T L$ for some $L < \infty$. Thus, it is enough to show

$$P\left(\sup_{k<k_1^0-Mv_T^{-2}} \frac{1}{k_1^0-k}\left|\sum_{t=k+1}^{k_1^0} X_t\right| > \eta v_T L^{-1}\right) < \epsilon$$

for large M. By the Hajek and Renyi inequality (applied with the data order reversed by treating k_1^0 as 1),

$$P\left(\sup_{k<k_1^0-Mv_T^{-2}} \frac{1}{k_1^0-k}\left|\sum_{t=k+1}^{k_1^0} X_t\right| > \eta v_T L^{-1}\right) \le \frac{C_1L^2}{\eta^2 v_T^2 M v_T^{-2}} = \frac{C_1 L^2}{\eta^2 M}.$$

The preceding probability is small if M is large. The proof of Lemma 4 demonstrates that all other terms are of smaller or equal magnitude than the term just treated. This proves the lemma for k less than k_1^0. The proof for $k > k_1^0$ is analogous. ■

Proof of Proposition 8. The proof is similar to that of Proposition 4. In view of the rate of convergence, we consider the limiting process of $\Lambda_T(s) = S_T(k_1^0 + [sv_T^{-2}]) - S_T(k_1^0)$, for $s \in [-M, M]$ for an arbitrary given $M < \infty$. First, consider $s > 0$. Let $l = [sv_T^{-2}]$. Define $\hat{\mu}_i^*$ and $\hat{\mu}_i$ as in (A.31) and (A.32), respectively. Then, (A.33)–(A.35) still hold with μ_j interpreted as μ_{jT}. For example,

$$\sqrt{T}(\hat{\mu}_1 - \mu_{1T}) = \frac{\sqrt{T} l v_T(\tilde{\mu}_2 - \tilde{\mu}_1)}{k_1^0 + l} + \frac{T}{k_1^0+l}\frac{1}{\sqrt{T}}\sum_{t=1}^{k_1^0+l} X_t = O_p(1).$$

This follows because, from $|l| \le Mv_T^{-2}$, the first term on the right-hand side is of $O\big(1/(\sqrt{T}v_T)\big)$, which converges to zero, and the second term is $O_p(1)$. Equations (A.36) and (A.37) are simply identities and still hold here. Similar to the proof of Proposition 4, the difference between the two first terms and the difference between the two third terms of (A.36) and (A.37) converge to zero in probability. Equation (A.40) in the present case is reduced to

$$2(1+\lambda_1)(\tilde{\mu}_2 - \tilde{\mu}_1)v_T \sum_{t=k_1^0+1}^{k_1^0+l} X_t + l v_T^2(\tilde{\mu}_2-\tilde{\mu}_1)^2(1-\lambda_1)^2 + O_p(T^{-1/2}).$$

Note that λ_1 is free from v_T because it is canceled out due to its presence in the denominator and the numerator. From $l = [sv_T^{-2}]$, using the functional central limit theorem for linear processes (e.g., Phillips and Solo, 1992),

$$v_T \sum_{t=k_1^0+1}^{k_1^0+[sv_T^{-2}]} X_t = v_T \sum_{t=1}^{[sv_T^{-2}]} X_{t-k_1^0} \Rightarrow a(1)\sigma_\epsilon B_2(s),$$

where $B_2(s)$ is a Brownian motion process on $[0,\infty)$ and

$lv_T^{-2} = [sv_T^{-2}]v_T^2 \to s$ uniformly in $s \in [0,M]$.

In summary, for $s > 0$,

$$S_T(k_1^0 + [sv_T^{-2}]) - S_T(k_1^0) \Rightarrow 2(1+\lambda_1)(\tilde{\mu}_2 - \tilde{\mu}_1)a(1)\sigma_\epsilon B_2(s) + s(\tilde{\mu}_2 - \tilde{\mu}_1)^2(1-\lambda_1^2).$$

The same analysis shows that for $s < 0$

$$S_T(k_1^0 + [sv_T^{-2}]) - S_T(k_1^0)$$
$$\Rightarrow 2(1+\lambda_1)(\tilde{\mu}_2 - \tilde{\mu}_1)a(1)\sigma_\epsilon B_1(-s) + |s|(\tilde{\mu}_2 - \tilde{\mu}_1)^2(1+\lambda_1)^2,$$

where $B_1(\cdot)$ is another Brownian motion process on $[0,\infty)$ independent of $B_2(\cdot)$. Introduce

$$\Gamma(s,\lambda) = \begin{cases} 2a(1)\sigma_\epsilon B_1(-s) + |s|(1+\lambda) & \text{if } s < 0, \\ 2a(1)\sigma_\epsilon B_2(s) + |s|(1-\lambda) & \text{if } s > 0, \end{cases}$$

with $\Gamma(0,\lambda) = 0$. The process Γ differs from Λ in the extra term $a(1)\sigma_\epsilon$. By a change of variable, it can be show that $\operatorname{argmin}_s \Gamma(s,\lambda) \stackrel{d}{=} a(1)^2\sigma_\epsilon^2 \operatorname{argmin}_s \Lambda(s,\lambda)$. Now, because $cB_i(s)$ has the same distribution as $B_i(c^2 s)$, we have

$$S_T(k_1^0 + [sv_T^{-2}]) - S_T(k_1^0) \Rightarrow (1+\lambda_1)\Gamma((\tilde{\mu}_2 - \tilde{\mu}_1)^2 s, \lambda_1).$$

This implies that

$$\begin{aligned} Tv_T^2(\hat{\tau} - \tau_1^0) &\xrightarrow{d} \operatorname{argmin}_s (1+\lambda_1)\Gamma((\tilde{\mu}_2 - \tilde{\mu}_1)^2 s, \lambda_1) \\ &\stackrel{d}{=} (\tilde{\mu}_2 - \tilde{\mu}_1)^{-2} \operatorname{argmin}_v \Gamma(v,\lambda_1) \\ &\stackrel{d}{=} (\tilde{\mu}_2 - \tilde{\mu}_1)^{-2} a(1)^2 \sigma_\epsilon^2 \operatorname{argmin}_v \Lambda(v,\lambda_1). \end{aligned}$$

We have used the fact that $\operatorname{argmin}_x af(x) = \operatorname{argmin}_x f(x)$ for $a > 0$ and $\operatorname{argmin}_x f(a^2 x) = a^{-2} \operatorname{argmin}_x f(x)$ for an arbitrary function $f(x)$ on $\mathcal{R}$. ■

Proof of Lemma 10. From identity (2), $\bar{S}_N - S_N(k) = NV_N(k)^2$, where $V_N(k) = \{(k/N)(1-k/N)\}^{1/2}(\bar{Y}_k^* - \bar{Y}_k)$. It is enough to consider k such that $k \in [n\eta, n(1-\eta)]$ because N and n are of the same order. Now,

$$\begin{aligned} N^{1/2}V_N(k) &= N^{1/2}\{(k/N)(1-k/N)\}^{1/2}\Bigg(\frac{1}{n+n_2-k}\sum_{k+1}^{n+n_2} X_t - \frac{1}{k+n_1}\sum_{-n_1+1}^{k} X_t \\ &\qquad + \frac{n_2}{n+n_2-k}\mu_2 - \frac{n_2}{n+n_2-k}\mu \\ &\qquad - \frac{n_1}{k+n_1}\mu_1 + \frac{n_1}{k+n_1}\mu\Bigg) \end{aligned}$$

$$= N^{1/2}\{(k/N)(1-k/N)\}^{1/2}\left(\frac{1}{n+n_2-k}\sum_{k+1}^{n+n_2} X_t - \frac{1}{k+n_1}\sum_{-n_1+1}^{k} X_t + O_p\left(\frac{1}{n}\right)\right)$$

$$= n^{1/2}\{(k/n)(1-k/n)\}^{1/2}\left(\frac{1}{n-k}\sum_{k+1}^{n} X_t - \frac{1}{k}\sum_{1}^{k} X_t\right) + O_p(n^{-1/2})$$

$$= \{(k/n)(1-k/n)\}^{-1/2}\left\{\frac{k}{n}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t\right) - \frac{1}{\sqrt{n}}\sum_{t=1}^{k} X_t\right\} + O_p(n^{-1/2}),$$

where the second equality follows from $n_i = O_p(1)$ and $k^{-1} = O(n^{-1})$, the third follows from the asymptotic equivalence of N and n, and the fourth follows from some simple algebra. For $k = [n\tau]$, $N^{1/2}V_N(k)$ converges in distribution to $a(1)\sigma\{\tau(1-\tau)\}^{-1/2}[\tau B(1) - B(\tau)]$. This gives the finite-dimensional convergence. The rest follows from the functional central limit theorem and the continuous mapping theorem. ■