working paper
department
of economics

ESTIMATING MULTIPLE BREAKS ONE AT A TIME

Jushan Bai

95-18 Aug. 1995

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

ESTIMATING MULTIPLE BREAKS ONE AT A TIME

Jushan Bai

95-18 Aug. 1995

95-18

Estimating Multiple Breaks One at a Time

by
Jushan Bai[1]
Massachusetts Institute of Technology[2]

August, 1995

[1]Partial financial support is acknowledged from the National Science Foundation under Grant SBR-9414083.

[2]Please send correspondence to Jushan Bai, Department of Economics, E52-274B, MIT, Cambridge, MA 02139. Phone: (617) 253-6217. Email: jbai@mit.edu.

Abstract

Sequential (one by one) rather than simultaneous estimation of multiple breaks is investigated in this paper. The advantage of this method lies in its significant computational savings. The number of least squares 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 under-specification of the number of breaks. More interestingly and somewhat surprisingly, the estimated break points are shown to be T consistent, the same as the simultaneous estimation. Limiting distributions are also derived. Unlike simultaneous estimation, however, 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 having the same limiting distributions as those obtained via simultaneous estimation. Finally, a procedure is proposed to consistently estimate the number of breaks.

Keywords and Phrases: Multiple breaks, sequential estimation, simultaneous estimation, T consistency, limiting distribution, repartition method.

Running Head: MULTIPLE BREAKS

1. Introduction

Multiple breaks may exist in the trend function of many economic time series, as suggested by the studies of Cooper (1995), Garcia and Perron (1994), Papell and Lumsdaine (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 at once simultaneously.

A number of issues arise with the existence 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 are 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. Incidentally, taking advantage of dynamic programming, the simultaneous method requires $O(T^2)$ number of least squares irrespective of the number of break points. 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 required to compute 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) (also, see Bai (1994c) for an earlier exposition of the method). They argued that the estimated break point is consistent for one of the true break points. However, neither of the studies give the convergence rate of the estimated break point. In fact, the approach used in the previous studies does not allow one to study the convergence rate of sequential estimators. A different framework and more detailed analysis are necessary. The framework used in this paper is adapted from Bai (1994a). A major finding of this study is that the sequentially obtained estimated break points are T consistent, the same as the 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. All 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. Results corresponding to more than two breaks are stated in Section 6. The issue of the number of breaks is also discussed in this section. Section 7 proposes the "repartition method" that gives rise to estimators having the same asymptotic distribution as simultaneous estimation. Section 8 deals with shrinking shifts. Convergence rates and limiting distributions are also derived. Section 9 states the results for general models. Simulation results are reported in Section 10. The last section concludes. Mathematical proofs are provided in the appendix.

2. The Model

To present the major idea we shall consider a simple model with mean shifts in a linear process. The whole theory and results can be elaborated to general regression models using a combination of the argument of Bai (1994b) and this paper. To make the matter even simpler, the presentation and proof will be stated in terms of two breaks. Because of the nature of sequential estimation, the 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:

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

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

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

with $a(1) = \sum_{j=0}^{\infty} a_j \neq 0$. We assume that $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$, so that there are two break points in the model. In addition, we assume $k_1^0 = [T\tau_1^0]$ and $k_2^0 = [T\tau_2^0]$ with $\tau_1^0 < \tau_2^0$ and $\tau_1^0, \tau_2^0 \in (0,1)$. The unknown parameters are (τ_1^0, τ_2^0) [or (k_1^0, k_2^0)] and (μ_1, μ_2, μ_3). The focus will be the break points (τ_1^0, τ_2^0), because once they are obtained, the regression parameters can be easily computed.

The main thrust of sequential estimation is "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 (1994a). A single break point can be obtained by minimizing the sum of squared residuals among all possible sample splits. As in Bai (1994a), 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

$$\hat{k} = \text{argmin}_{1 \le k \le T-1} S_T(k).$$

Using the formula linking total variance with within-group and between-group variances, we can write, for each k $(1 \le k \le T-1)$,

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

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

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

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 the estimated break points. The former is also referred to as 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, $\hat{k}$ itself is not consistent for any of the k_i^0 $(i = 1, 2)$. For ease of exposition, we shall frequently say that $\hat{k}$ is T consistent with an understanding that we are actually referred to $\hat{\tau}$.

Below are assumptions that guarantee T consistency.

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

Assumption A2

$$\sum_{j=0}^{\infty} j|a_j| < \infty$$

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 $\tau_1^0 < \tau_2^0$.

These assumptions are used in Bai (1994a), except A3 which is stated in terms of a single break. Assumptions A1 and A2 are standard for linear processes. A3 assumes that there are two breaks. The next section proves the T consistency of $\hat{\tau}$ for one of the break points. The identification for the other break point will also be considered.

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 lead to the consistency result naturally. 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, i.e. $S_T(0) = S_T(T) = \sum_{t=1}^{T}(Y_t - \bar{Y})^2$. 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 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,*

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

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 $U(\tau)$. Typically, the function $U(\tau)$ has two local minima. To ensure the smallest value of $U(\tau)$ is unique, we assume:

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

This condition guarantees the uniqueness of the global minimum of $U(\tau)$. The condition is equivalent to, by (4) and (5),

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

Evidently, the condition assumes that the first break is more dominating in terms of the relative span of regimes and the magnitude of shifts. In other words, when the first break is more pronounced (larger τ_1^0 and/or larger $\mu_1 - \mu_2$), A4 will be true. The inequality will be reversed when the second break is more pronounced. Under A4 together with A1-A3, the estimated fraction $\hat{\tau}$ converges in probability to τ_1^0. This is true because only if the more pronounced break is chosen can the sum of squared residuals be reduced the most. If the inequality in A4 is reversed, then by mere symmetry, $\hat{\tau}$ converges in probability to τ_2^0. In the next section, we examine the case in which $U(\tau_1^0) = U(\tau_2^0)$. Under this condition, we show that $\hat{\tau}$ converges in distribution to a random variable with equal mass at τ_1^0 and τ_2^0 only. Incidentally, the set of parameters $\{(\tau_1^0, \tau_2^0, \mu_1, \mu_2, \mu_3)\}$ which makes $U(\tau_1^0) = U(\tau_2^0)$, defines a subset of $\mathcal{R}^5$ having a Lebesgue measure zero.

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

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

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 by 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

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.

This proposition not only establishes consistency but also gives a convergence rate.

Proof:

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

The above 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 above inequality by T on both sides and using Lemma 2, we obtain the proposition immediately. □

The above convergence rate is obtained by examining the global behavior of the objective function. We can use this initial rate of convergence to obtain a better rate. 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 1 and away from the second break point with 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 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 an $M < \infty$ such that*

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

That is, the break point estimator is T consistent.

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

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

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

by Lemma 4. This proves the proposition. □

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

If it is known that $\hat{k}/T$ is consistent for τ_1^0, then an estimate for τ_2^0 can be easily constructed. Just apply the same technique to the random sample $[\hat{k}, T]$. Let $\hat{k}_2$ denote the resulting estimator. Then $\hat{\tau}_2 = \hat{k}_2/T$ must be T consistent for τ_2^0. This follows because $\hat{k} = k_1^0 + O_p(1)$, one is effectively using $[k_1^0, T]$ to estimate the second break. Alternatively, even if $\hat{k} < k_1^0$, the dominating break in the subsample $[\hat{k}, T]$ can only be k_2^0. Thus our previous analysis implies the T consistency of $\hat{\tau}_2$ for τ_2^0.

In summary, T consistent estimators for τ_1^0 and τ_2^0 can be obtained by sequential procedure. The total number of least squares required is no more than $2T$.

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 $U(\tau)$, as a function of τ, has two local minima at τ_1^0 and τ_2^0, respectively. 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. *Under A1-A3 together with $U(\tau_1^0) = U(\tau_2^0)$, the estimator $\hat{\tau}$ converges in distribution to a random variable with equal mass at τ_1^0 and τ_2^0, respectively. 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 an $M < \infty$ such that*

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

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

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

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

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

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

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

Let $\hat{k}_1$ be the location of the minimum of $S_T(k)$ for k such that $k \leq k_0^*$, that is, $\hat{k}_1 = \text{argmin}_{k \leq k_0^*} S_T(k)$. Let $\hat{k}_2 = \text{argmin}_{k_0^* < k} S_T(k)$. Note that this $\hat{k}_2$ is different from the one defined in the previous section. Also, $\hat{k}_1$ and $\hat{k}_2$ are not estimators as k_0^* is unknown. They are introduced here for theoretical purposes. It is clear that the global minimizer $\hat{k}$ satisfies:

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

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

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

$$\hat{\tau}_1 - \tau_1^0 = O_p(T^{-1/2}),$$

$$\hat{\tau}_2 - \tau_2^0 = O_p(T^{-1/2}).$$

The root T consistency is strengthened to T consistency using the following:

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

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

where

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

$$D_{T,M}^{(2)} = \{k : k_0^* + 1 \leq k \leq T(1 - \eta), |k - k_1^0| > M\}.$$

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

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

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

Proof of Proposition 3. By lemma 7, $P(\hat{\tau} = \hat{\tau}_i) \to 1/2$ ($i = 1, 2$). But $\hat{\tau}_i \xrightarrow{p} \tau_i^0$, 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 the T consistency of $\hat{\tau}_i$. □

It is clear that $\hat{k}$ is a good estimator for one of the breaks. If it is known that $\hat{k}$, for a given sample, is estimating k_1^0, then we can use the subsample $[\hat{k}, T]$ to estimate k_2^0. Note that this second stage estimator is not necessarily equal to $\hat{k}_2$, as the latter is based on the sum of squared residuals using the entire sample. Similarly, if $\hat{k}$ is estimating k_2^0, we can use the sample $[1, \hat{k}]$ to estimate k_1^0. Let $(\hat{k}^{(1)}, \hat{k}^{(2)})$ denote the ordered pair of the first stage and the second stage estimators such that $\hat{k}^{(1)} < \hat{k}^{(2)}$. It follows that the ordered pair forms a T consistent estimator for (k_1^0, k_2^0).

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.[1]

Let $\{X_t^{(1)}\}$ be an independent copy of the process of $\{X_t\}$. Define $W^{(1)}(\ell, \lambda) = W_1^{(1)}(\ell, \lambda)$ for $\ell < 0$ and $W^{(1)}(\ell, \lambda) = W_2^{(1)}(\ell, \lambda)$ for $\ell > 0$ and $W^{(1)}(0, \lambda) = 0$, where

$$W_1^{(1)}(\ell, \lambda) = -2(\mu_2 - \mu_1) \sum_{t=\ell+1}^{0} X_t^{(1)} + |\ell|(\mu_2 - \mu_1)^2(1 + \lambda), \quad \ell = -1, -2, \ldots$$

$$W_2^{(1)}(\ell, \lambda) = 2(\mu_2 - \mu_1) \sum_{t=1}^{\ell} X_t^{(1)} + \ell(\mu_2 - \mu_1)^2(1 - \lambda), \quad \ell = 1, 2, \ldots$$

Proposition 4. *Under assumptions A1-A5, together with the assumption of continuous distribution for X_t,*

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

where

$$\lambda_1 = \frac{1 - \tau_2^0}{1 - \tau_1^0}\Big(\frac{\mu_3 - \mu_2}{\mu_2 - \mu_1}\Big).$$

[1]This assumption allows one to express the limiting distribution free from the change point (k_1^0).The assumption can be dispensed with, see Bai (1994b).

Note that condition A4 [or equivalently (6)] guarantees that $|\lambda_1| < 1$. The assumption of continuous distribution ensures the uniqueness of the global minimum for the process $W^{(1)}(\ell, \lambda_1)$, so that $\text{argmin}_\ell W^{(1)}(\ell, \lambda_1)$ 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 ($\mu_3 = \mu_2$) or to that of the first break point estimator in the case of multiple breaks with simultaneous estimation. If X_t has a symmetric distribution and λ is equal to zero, $W^{(1)}(\ell, \lambda)$ and $W^{(1)}(-\ell, \lambda)$ will have the same distribution and consequently, $\hat{k} - k_1^0$ will have a symmetric distribution. Because $\lambda \neq 0$ generally, the limiting distribution from sequential estimation is not symmetric about zero. For positive λ (or equivalently, $\mu_2 - \mu_1$ and $\mu_3 - \mu_2$ have the same sign), the drift term of $W_2^{(1)}(\ell, \lambda)$ is smaller than that of $W_1^{(1)}(\ell, \lambda)$. This implies that the distribution of $\hat{k}$ will have a heavy right tail, a tendency to overestimate the break point relative to the 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.

Suppose the inequality in Assumption A4 is reversed, i.e. $U(\tau_1^0) > U(\tau_2^0)$. Then by mere symmetry,

$$\hat{k} - k_2^0 \xrightarrow{d} \text{argmin}_\ell W^{(2)}(\ell, \lambda_2)$$

where

$$W^{(2)}(\ell, \lambda) = \begin{cases} -2(\mu_3 - \mu_2) \sum_{t=\ell+1}^{0} X_t^{(2)} + |\ell|(\mu_3 - \mu_2)^2(1 + \lambda), & \ell = -1, -2, \dots \\ 2(\mu_3 - \mu_2) \sum_{t=1}^{\ell} X_t^{(2)} + \ell(\mu_3 - \mu_2)^2(1 - \lambda), & \ell = 1, 2, \dots \end{cases}$$

and

$$\lambda_2 = \frac{\tau_1^0}{\tau_2^0}\left(\frac{\mu_2 - \mu_1}{\mu_3 - \mu_2}\right),$$

with $\{X_t^{(2)}\}$ being an independent copy of the process $\{X_t\}$, and being also independent of $\{X_t^{(1)}\}$.

As discussed in Section 3, 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. We have argued that $\hat{\tau}_2 = \hat{k}_2/T$ is T consistent for τ_2^0. 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,

Proposition 5. *Under assumptions A1-A5,*

$$\hat{k}_2 - k_2^0 \xrightarrow{d} \text{argmin}_\ell W^{(2)}(\ell, 0)$$

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.

Similarly, if $\hat{k}/T$ is consistent for τ_2^0, then one can use the sample $[1, \hat{k}]$ to estimate τ_1^0. The resulting estimator must be T consistent. The limiting distribution is given by $\text{argmin}_\ell W^{(1)}(\ell, 0)$.

We now consider the case in which $U(\tau_1^0) = U(\tau_2^0)$. As in Section 4, let $(\hat{k}^{(1)}, \hat{k}^{(2)})$ denote the ordered pair of the first and second stage estimators. Then we have the following result: for $i = 1, 2$

$$\hat{k}^{(i)} - k_i^0 \xrightarrow{d} \begin{cases} \text{argmin}_\ell W^{(i)}(\ell, \lambda_i) & \text{with probability } 1/2 \\ \text{argmin}_\ell W^{(i)}(\ell, 0) & \text{with probability } 1/2. \end{cases}$$

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

6. More Than Two Breaks

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

$$\begin{aligned} Y_t &= \mu_1 + X_t, & &\text{if } t \le k_1^0 \\ Y_t &= \mu_2 + X_t, & &\text{if } k_1^0 + 1 \le t \le k_2^0 \\ &\vdots & &\vdots \\ Y_t &= \mu_{m+1} + X_t, & &\text{if } k_m^0 + 1 \le t \le T. \end{aligned} \tag{8}$$

where $\mu_i \ne \mu_{i+1}$ $k_i^0 = [T\tau_i^0]$, $\tau_i^0 \in (0,1)$ and $\tau_i^0 < \tau_{i+1}^0$ for $i = 1, ..., m$ with $\tau_{m+1}^0 = 1$. Assume the process X_t satisfies A1-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 $\hat{k} = \text{argmin} S_T(k)$ and $\hat{\tau} = \hat{k}/T$. From the proof for the earlier results in the appendix we can see that the assumption of two breaks is not essential. With more than two breaks, one just 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 $j \ne i$.

Proposition 6. *Under assumptions of A1-A3 and A6, the estimated break point $\hat{\tau}$ is T consistent for τ_i^0.*

Proposition 7. *Under the assumption of Proposition 6 and A5,*

$$\hat{k} - k_i^0 \xrightarrow{d} \text{argmin}_{\ell} W^{(i)}(\ell, \lambda_i)$$

where $W^{(i)}(\ell, \lambda)$ has the same form as $W^{(1)}(\ell, \lambda)$ with $(\mu_2 - \mu_1)$ replaced by $(\mu_{i+1} - \mu_i)$ 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 $|\lambda_i| < 1$.

A new terminology is appropriate here. A subsample $[k, \ell]$ is said to contain a nontrivial break point if both k and ℓ are bounded away from the break point with a positive fraction of observations. That is, $k^0 - k > T\epsilon_0$ and $\ell - k^0 > T\epsilon_0$ for some $\epsilon_0 > 0$ and for all large T, where k^0 is a break point inside $[k, \ell]$. 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 $S_{\hat{k}}(\ell)$ 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 4 subintervals are to be considered in the third round estimation. This procedure is repeated until each resulting subsample contains no nontrivial break point. Assuming the knowledge of the number of breaks as well as the knowledge of the existence of a nontrivial break in a given subsample, then all the breaks can be identified and all the estimated break fractions are T consistent. The total number of least squares required is no more than mT; here, m is the number of breaks.

A problem arises immediately in practice as to whether a subsample contains a nontrivial break, which is clearly tied up with 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 in the next section, such a decision rule leads to a consistent estimation of the number of breaks, and implicitly a correct judgment about the existence of a nontrivial break in a given subsample.

6.1. Determination of 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 in a similar way as described in the previous section. Along the way, hypothesis testing is used as an auxiliary tool to determine the existence of a break point for a given subsample. We summarize the procedure here. 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 accepted for all sequentially obtained 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 above procedure and m_0 is 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 converges to zero slowly. To prove this assertion, we need the following general result. Let

$$\begin{aligned} Y_t &= \mu_1 + X_t, && \text{if } -n_1 + 1 \le t \le 0, \\ Y_t &= \mu \;+ X_t, && \text{if } 1 \le t \le n \\ Y_t &= \mu_2 + X_t, && \text{if } n + 1 \le t \le n + n_2 \end{aligned} \tag{9}$$

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 one in the sense that $n_i/n = O_p(n^{-1})$. Let $N = n + n_1 + n_2$. The supF 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 supF test is then defined as, for some $\eta \in (0, 1/2)$,

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

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 8. *Under model (9) and assumptions A1-A2, as $n \to \infty$,*

$$\sup F_N \xrightarrow{d} \sup_{\eta \le \tau \le 1-\eta} \frac{|B(\tau) - \tau B(1)|^2}{\tau(1-\tau)} \tag{10}$$

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 the last regime in Model (9). This is simply due to the stochastic boundedness of n_1 and n_2. We assume that the supF test is used in the sequential procedure, and the critical value and size of the test are based on the asymptotic distribution. Using this lemma, we can prove

Proposition 8. *Suppose that the size of the test α_T converges to zero slowly. Then under model (8) and assumptions A1-A2,*

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

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 $[\hat{k}, \hat{\ell}]$ containing at least one true break point which is nontrivial in the sense that both the distance from $\hat{k}$ to the break point and the distance from $\hat{\ell}$ to the break point contain a positive fraction of observations. That is, $k^0 - \hat{k} > T\epsilon_0$ and $\hat{\ell} - k^0 > T\epsilon_0$ for some $\epsilon_0 > 0$, where $k^0 \in (\hat{k}, \hat{\ell})$ is a break point. Then the test statistic based on this subsample must converge to infinity because the sup F test is consistent, see Andrews (1993). Thus, one will reject the null hypothesis of parameter constancy (as long as α_T does not decrease too fast). This implies that $P(\hat{m} < m)$ converges to zero as the sample size increases.

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

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

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 the supF statistic computed for the subsample $[\hat{k}_i, \hat{k}_{i+1}]$ converges in distribution, by Lemma 8, to the right hand side of (10). Denote the limiting distribution by ξ. Suppose the size α_T and the critical value c_T are chosen using the asymptotic distribution such that $P(\xi > c_T) \leq \alpha_T$, then for large T (and hence large n), $P(\sup F_N > c_T) \leq 2\alpha_T$. Thus $P(\hat{m} > m_0) \leq (m_0+1)2\alpha_T$, which converges to zero if α_T converges to zero as T increases. This completes the proof of the proposition. □

It remains unanswered as to what rate α_T should converge to zero. Of course, the rate should be low so that the critical value will not increase too quickly in order to guarantee a rejection under the alternative hypothesis. With the existence of a nontrivial break, the statistic $\sup F_N$ is of order T. Therefore, any vanishing sequence of α_T making c_T a lower order than T is sufficient. A more accurate statement can also be made about the quickest rate for α_T. Such a rate is clearly linked to the tail behavior of the random variable ξ. In practice, the issue is perhaps more empirical than theoretical. An appropriate choice requires an assessment of the adverse effect of overestimating or underestimating on the problem under consideration. If underestimating is more costly, larger size may be used and vice versa. For most economic data with moderate size, we recommend a 5% significant level. Once the size is chosen, all the rest can be automated.[2]

Bai and Perron (1994) propose 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 reduced the most. For example, at the beginning of the ith round, $i-1$ breaks are already determined, yielding i subsamples. The ith break point is chosen in the subsample for which the reduction in the sum of squared residuals is the most. The procedure is repeated until the specified number of break

[2] A computer program written in GAUSS for both sequential and simultaneous estimations is available upon request. The program is introduced in Bai and Perron (1994).

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 is proposed by Yao (1987). Yao suggests the BIC criterion. His method requires simultaneous estimation.

6.2. Some Comments

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 explains partially why the conventional supF test may possess less power than the test proposed by Bai and Perron (1994) in the presence of multiple breaks.

The problem may 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 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 either selected 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.

7. Fine Tuning: Repartition

Although each estimated break point is T consistent, there is a tendency to over or underestimate the location of the breaks 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 in Bai (1994b) in an empirical application. This paper provides the theoretical basis for doing so. Suppose there are m breaks and initial T consistent estimators $\hat{k}_h$ $(h = 1, ..., m)$ are obtained. The repartition technique reestimates each of the break points based on the initial estimates. To estimate k_i^0, the subsample $[\hat{k}_{i-1}, \hat{k}_{i+1}]$ is used. We denote the resulting estimator by $\hat{k}_i^*$. Because of the proximity of $\hat{k}_h$ to k_h^0, we effectively use the sample $[k_{i-1}^0 + 1, k_{i+1}^0]$ to estimate k_i^0. 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:

Proposition 9. *Under model (8) and assumptions A1-A2, the repartition estimators satisfy: for each $\epsilon > 0$, there exists an $M < \infty$, such that for all large T,*

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

and under the additional assumption A5,

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

Note that assumption A6 is not required. The proposition only uses the fact that the initial estimators are T consistent. As is known in Section 4, T consistent estimators can be obtained regardless of the validity of A6 (or A4).

It is evident that the repartition method is straightforward to implement. Repartition requires an additional T least squares computations. All together, no more than $(m+1)T$ least squares are necessary to obtain break point estimators that have the same limiting distributions as those obtained by simultaneous estimation.

8. Small Shifts

The limiting distributions derived earlier, though of theoretical interest, are perhaps of limited practical use because the distribution of $\text{argmin}_\ell W^{(i)}(\ell, \lambda)$ depends on that 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.

For concreteness and ease of exposition, we consider the two-break model of Section 2. This also enables us to deliver a full proof of our results without much additional effort. 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

Assumption B1. The sequence of numbers v_T satisfies

$$v_T \to 0, \quad T^{(1/2)-\delta} v_T \to \infty \qquad \text{for some } \delta \in (0, 1/2) \tag{11}$$

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

Assumption B2.

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

This condition turns out to be equivalent to (6), with μ_j replaced by $\tilde{\mu}_j$.

Under 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 10. *Under assumptions A1-A3 and B1-B2 we have* $T v_T^2(\hat{\tau} - \tau_1^0) = O_p(1)$ *or, equivalently, for every* $\epsilon > 0$, *there exists an* $M < \infty$ *such that*

$$P\Big(T|(\hat{\tau} - \tau_1^0)| > M v_T^{-2}\Big) < \epsilon.$$

The proof of this proposition is again based on some preliminary results analogous to Lemma 2 and 3. First we modified the objective function as

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

This does not change the problem, as the second term is free from k.

Lemma 9. *Under the assumptions of Proposition 10, we have*

(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

$$ES_T(k)-ES_T(k_1^0)\ge C_1v_T^2|k-k_1^0| \quad \textit{for all large } T.$$

Corollary 1. *Under the assumptions of Proposition 10,*

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

Proof: Adding and subtracting $\sum_{t=1}^{T}(X_t^2-EX_t^2)$ to the following identity

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

to obtain

$$\begin{aligned}
&S_T(k)-S_T(k_1^0)\\
&\ge\ -2\sup_{1\le j\le T}\left|S_T(j)-ES_T(j)-\sum_{t=1}^{T}(X_t^2-EX_t^2)\right|+ES_T(k)-ES_T(k_1^0)\\
&\ge\ -2\sup_{1\le j\le T}\left|S_T(j)-ES_T(j)-\sum_{t=1}^{T}(X_t^2-EX_t^2)\right|+C_1v_T^2|k-k_1^0|,
\end{aligned}$$

where the second inequality follows from Lemma 9(b). From $S_T(\hat{k})-S_T(k_1^0)\le 0$, we have

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

The corollary is obtained upon dividing the above inequality by T on both sides and using Lemma 9(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 10 can be proved. In view of the anticipated rate of convergence, we define D_{TM}^* the same as D_{TM} but replacing M by Mv_T^{-2}. Thus for $k\in D_{TM}^*$, it is possible for $k-k_1^0$ to converge to infinity because v_T^{-2} converges to infinity, although at a much slower rate than T.

Lemma 10. *Under assumptions of Proposition 10, for every $\epsilon>0$, there exists an $M>0$ such that*

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

Proof of Proposition 10. The proof is virtually identical to that of Proposition 2, but one uses Lemma 10 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 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 11. *Under the assumptions of Proposition 10,*

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

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

While the density function of $\text{argmin}_s\Lambda(s, \lambda_1)$ is derived in Bai (1994b) 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 $\lambda_1 = 0$.

9. Further Extensions

The preceding discussion has focused on multiple mean shifts in linear processes. The whole procedure can be elaborated to multiple regressions that are more useful in econometric applications. Here we give conditions that ensure T consistency. These conditions are similar to those in Bai (1994b) and Bai and Perron (1994).

Consider

$$\begin{array}{ll} Y_t = Z_t'\delta_1 + X_t, & t = 1, 2, \ldots, k_1^0, \\ Y_t = Z_t'\delta_2 + X_t, & t = k_1^0 + 1, \ldots, k_2^0, \\ \vdots & \vdots \\ Y_t = Z_t'\delta_{m+1} + X_t, & t = k_m^0 + 1, \ldots, T, \end{array} \tag{12}$$

where Y_t as before is the observed dependent variable at time t; Z_t $(q \times 1)$ is a vector of covariates; δ_j $(j = 1, \ldots, m + 1)$ are the corresponding vectors of coefficients with $\delta_i \neq \delta_{i+1}$ $(i = 1, \ldots, m)$; X_t is a linear process satisfying A1-A2.

Assumption C1: The regressors satisfy:

$$\text{plim} \frac{1}{T} \sum_{t=1}^{[Tv]} Z_t Z_t' = Q(v)$$

uniformly in $v \in [0,1]$, where $Q(v)$ is a positive definite matrix for each $v > 0$ and $Q(v) - Q(u) > 0$ for $v > u$.

Assumption C2: For large ℓ, the minimum eigenvalues of $\frac{1}{\ell}\sum_{k_i^0+1}^{k_i^0+\ell} Z_t Z_t'$ and of $\frac{1}{\ell}\sum_{k_i^0-\ell}^{k_i^0} Z_t Z_t'$ are bounded away from zero ($i = 1, ..., m+1$).

Assumption C3: The disturbances $\{X_t\}$ satisfy one of the following alternatives:

a) $\{X_t, \mathcal{F}_t\}$ forms a sequence of martingale differences where $\mathcal{F}_t = \sigma - field$ $\{Z_{s+1}, X_s; s \leq t\}$ with $\sup_t E|X_t|^{4+\delta} < \infty$.

b) X_t is independent of Z_s for all t and all s, but $\{X_t\}$ forms a sequence of mixingales satisfying conditions given in Bai and Perron (1994).

Assumption C4: $k_i^0 = [T\tau_i^0]$, $\tau_i^0 \in (0,1)$ with $\tau_i^0 < \tau_{i+1}$ ($i = 1, ..., m$).

Assumption C1 is satisfied by i.i.d. regressors having a finite variance. It is also satisfied by any second order stationarity process such that the strong law of large numbers holds for $Z_t Z_t'$. In these cases, $Q(v) = vQ$, where $Q = EZ_tZ_t'$. Trending regressors also satisfy C1. More interestingly, it is satisfied by autoregressive models with breaks. Suppose $Z_t = (1, Y_{t-1}, ..., Y_{t-q-1})$. Although Z_t is not globally stationary, its adjustment to its new stationary path is very quick after a break takes place. Thus for segment i, the limit $plim\frac{1}{\Delta k_i}\sum_{k_i^0+1}^{k_i^0+[v\Delta k_i^0]} Z_t Z_t'$ converges to vQ_i, where $\Delta k_i = k_{i+1}^0 - k_i^0$ and Q_i is the second moment matrix of a stationary autoregressive process with autoregressive parameters δ_i. Therefore, $Q(v) = \tau_1^0 Q_1 + \cdots + (\tau_\ell^0 - \tau_{\ell-1}^0)Q_\ell + (v - \tau_\ell^0)Q_{\ell+1}$ for $v \in [\tau_\ell^0, \tau_{\ell+1}^0]$.

Assumption C2 requires that there be sufficient data near the break point. It is used for T consistency. Part (a) of Assumption C3 allows for autoregressive models or models with lagged dependent variables. Part (b) allows for general serial correlated disturbances. A mixingale sequence includes many dependent processes as special cases. These assumptions are similar to those of Bai (1994b) and Bai and Perron (1994).

Under the assumptions C1-C4, using the argument presented earlier in this paper together with that of Bai (1994b), it can be shown that the sequential estimators are T consistent. The repartitioned estimators have limiting distributions identical to simultaneous estimators. Therefore, confidence intervals can be constructed in the way given in Bai (1994b) and Bai and Perron (1994).

10. 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 $(\mu_1, ..., \mu_4)$ 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 i.i.d. standard normal. All reported results are based on 5000 repetitions.

First we assume the number of break points is known and focus on their estimation. The break points are chosen using the suggestion of Bai and Perron (1994), "one and only one additional break" in each round. 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)]. To verify the theory and for comparison purposes, three different methods are used– sequential, repartition, and simultaneous methods. Because, for model (I), the magnitude of shift for each break is the same, we expect three estimated break points should 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. One, the first and third estimated break points will have the same limiting distribution as simultaneous estimation, even without repartition. This explains why the results look homogeneous for the three different methods. Two, 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 over 92%, while for model (II) the rate is over 99.5%.

We also perform some limited Monte Carlo simulations for estimating the number of breaks using the sequential method. In addition to the two sets of parameters considered earlier, we add a third set of parameters, which is $(1.0, 2.0, 3.0, 4.0)$ [referred to as model (III)]. For comparison purposes, estimates using the BIC method are also given. Figure 3 displays the estimated numbers for both methods. The left three histograms (a, b, c) are for the sequential method and the right three (a', b', c') are for the BIC method. The sequential method uses a two-step procedure described in Section 6.2. We assume the number of breaks is 4 in the first step and estimate the error variance based on repartitioned estimators. The size of the test is chosen to be 0.05 with the corresponding critical value 9.63.

For the first set of parameters, the BIC criterion does a better job than the sequential method. The latter underestimates the number of breaks. For a significant proportion of observations, the sequential method only detects a single break. We find that the single break identified by the sequential method in most cases is the third break. Put another way, the sequential method has difficulties in finding breaks if the first 80 observations are used. Indeed, the supF test has lower power in detecting "hat" shaped mean changes (especially for small samples and less pronounced shifts). For the second set of parameters, the two methods are comparable. Interestingly, the sequential method works better than the BIC criterion for the third set of parameters.

The sequential method may be improved upon in at least two dimensions. First, the supF test which is designed for testing a single break may be replaced by, or used in conjunction with, Bai and Perron's supF(ℓ) test for testing multiple breaks. The latter test has better power 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.

11. 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.

Appendix: Mathematical Proofs

Throughout the proof, the notation $o_p(1)$ $[O_p(1)]$ is used to denote a sequence of random variables converging to zero in probability [stochastically bounded]. All limits are taken as the sample size T converges to infinity, unless stated otherwise. We may write $X \overset{d}{=} Y$ when X and Y having the same distribution.

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

$$\sum_{t=1}^{k}(Y_t - \bar{Y}_k)^2 = \sum_{t=1}^{k}(X_t - \frac{1}{k}\sum_{i=1}^{k} X_i)^2 = \sum_{t=1}^{k}(X_t - A_{Tk})^2 \tag{13}$$

and

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

The latter expression can be rewritten as

$$\sum_{t=k+1}^{T} (Y_t - \bar{Y}_k^*)^2 \;=\; (k_1^0 - k)a_{Tk}^2 + 2a_{Tk}\sum_{t=k+1}^{k_1^0} (X_t - A_{Tk}^*)$$

$$
\begin{aligned}
&+(k_2^0-k_1^0)b_{Tk}^2+2b_{Tk}\sum_{k_1^0+1}^{k_2^0}(X_t-A_{Tk}^*)\\
&+(T-k_2^0)c_{Tk}^2+2c_{Tk}\sum_{k_2^0+1}^{T}(X_t-A_{Tk}^*) \qquad (14)\\
&+\sum_{t=k+1}^{T}(X_t-A_{Tk}^*)^2
\end{aligned}
$$

where

$$
\begin{aligned}
a_{Tk} &= \frac{1}{T-k}\{(T-k_1^0)(\mu_1-\mu_2)+(T-k_2^0)(\mu_2-\mu_3)\}\\
b_{Tk} &= \frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1)+(T-k_2^0)(\mu_2-\mu_3)\}\\
c_{Tk} &= \frac{1}{T-k}\{(k_1^0-k)(\mu_2-\mu_1)+(k_2^0-k)(\mu_3-\mu_2)\}.
\end{aligned}
$$

Rewrite

$$
\frac{1}{T}\sum_{t=k+1}^{T}(X_t-A_{Tk}^*)^2=\frac{1}{T}\sum_{t=k+1}^{T}X_t^2-\frac{T-k}{T}(A_{Tk}^*)^2 \qquad (15)
$$

and

$$
\frac{1}{T}\sum_{t=1}^{k}(X_t-A_{Tk})^2=\frac{1}{T}\sum_{t=1}^{k}X_t^2-\frac{1}{T}\Big(\frac{1}{\sqrt{k}}\sum_{t=1}^{k}X_t\Big)^2. \qquad (16)
$$

Combining (13) and (14) and using (15) and (16), we have for $k \le k_1^0$,

$$
\begin{aligned}
U_T(k/T) &= \frac{1}{T}S_T(k)\\
&= \frac{(k_1^0-k)}{T}a_{Tk}^2+\frac{(k_2^0-k_1^0)}{T}b_{Tk}^2+\frac{(T-k_2^0)}{T}c_{Tk}^2+\frac{1}{T}\sum_{t=1}^{T}X_t^2+R_{1T}(k) \qquad (17)
\end{aligned}
$$

where

$$
\begin{aligned}
R_{1T}(k) \;&=\; \frac{1}{T}\Big[2a_{Tk}\sum_{t=k+1}^{k_1^0}X_t+2b_{Tk}\sum_{k_1^0+1}^{k_2^0}X_t+2c_{Tk}\sum_{k_2^0+1}^{T}X_t\Big]\\
&\quad -\frac{2}{T}\Big[(k_1^0-k)a_{Tk}+(k_2^0-k_1^0)b_{Tk}+(T-k_2^0)c_{Tk}\Big]A_{Tk}^* \qquad (18)\\
&\quad -\frac{1}{T}\Big(\frac{1}{\sqrt{k}}\sum_{t=1}^{k}X_t\Big)^2-\frac{T-k}{T}(A_{Tk}^*)^2
\end{aligned}
$$

We shall argue that

$$
R_{1T}(k)=O_p(T^{-1/2}) \qquad \text{uniformly in } k\in[1,k_1^0]. \qquad (19)
$$

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 first two expression on the right hand side of (18) 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 $\sup_{1\le k\le T} |\frac{1}{\sqrt{k}}\sum_{i=1}^k X_i| = O_p(\log T)$. 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

$$\bar{Y}_k = \frac{k_1^0}{k}\mu_1 + \frac{k - k_1^0}{k}\mu_2 + A_{Tk},$$

$$\bar{Y}_k^* = \frac{k_2^0 - k}{T - k}\mu_2 + \frac{T - k_2^0}{T - k}\mu_3 + A^*_{Tk}.$$

Thus

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

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

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

$$\begin{aligned} &\sum_{t=1}^{k}(Y_t - \bar{Y}_k)^2 \\ &= k_1^0 d_{Tk}^2 + 2d_{Tk}\sum_{t=1}^{k_1^0}(X_t - A_{Tk}) + \sum_{t=1}^{k_1^0}(X_t - A_{Tk})^2 \\ &+(k - k_1^0)e_{Tk}^2 + 2e_{Tk}\sum_{t=k_1^0+1}^{k}(X_t - A_{Tk}) + \sum_{t=k_1^0+1}^{k}(X_t - A_{Tk})^2 \end{aligned} \tag{20}$$

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

$$\begin{aligned} &\sum_{t=k+1}^{T}(Y_t - \bar{Y}_k^*)^2 \\ &= (k_2^0 - k)f_{Tk}^2 + 2f_{Tk}\sum_{k+1}^{k_2^0}(X_t - A^*_{Tk}) + \sum_{k+1}^{k_2^0}(X_t - A^*_{Tk})^2 \\ &+(T - k_2^0)g_{Tk}^2 + 2g_{Tk}\sum_{k_2^0+1}^{T}(X_t - A^*_{Tk}) + \sum_{k_2^0+1}^{T}(X_t - A^*_{Tk})^2 \end{aligned} \tag{21}$$

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

$$
\begin{aligned}
U_T(k/T) &= \frac{1}{T} S_T(k) \\
&= \frac{k_1^0}{T} d_{Tk}^2 + \frac{k - k_1^0}{T} e_{Tk}^2 + \frac{k_2^0 - k}{T} f_{Tk}^2 + \frac{T - k_2^0}{T} g_{Tk}^2 + \frac{1}{T}\sum_{t=1}^{T} X_t^2 + R_{2T}(k) \qquad (22) \\
&= \frac{k_1^0(k - k_1^0)}{kT}(\mu_2 - \mu_1)^2 + \frac{(k_2^0 - k)(T - k_2^0)}{T(T-k)}(\mu_3 - \mu_2)^2 + \frac{1}{T}\sum_{t=1}^{T} X_t^2 + R_{2T}(k)
\end{aligned}
$$

where

$$
\begin{aligned}
R_{2T}(k) \;=\; & \frac{1}{T}\Big[2d_{Tk}\sum_{t=1}^{k_1^0} X_t + 2e_{Tk}\sum_{t=k_1^0+1}^{k} X_t + 2f_{Tk}\sum_{k+1}^{k_2^0} X_t + 2g_{Tk}\sum_{k_2^0+1}^{T} X_t\Big] \\
& -\frac{2}{T}\Big[k_1^0 d_{Tk} + (k - k_1^0)e_{Tk} + (k_2^0 - k)f_{Tk} + (T - k_2^0)g_{Tk}\Big]A_{Tk}^* \qquad (23) \\
& -\frac{k}{T}(A_{Tk})^2 - \frac{T-k}{T}(A_{Tk}^*)^2.
\end{aligned}
$$

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

$$
U_T(k/T) = \frac{k_1^0}{T} h_{Tk}^2 + \frac{k_2^0 - k_1^0}{T} p_{Tk}^2 + \frac{k - k_2^0}{T} q_{Tk}^2 + \frac{1}{T}\sum_{t=1}^{T} X_t^2 + R_{3T}(k) \qquad (24)
$$

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

$$
\begin{aligned}
h_{Tk} &= \frac{1}{k}[(k - k_1^0)(\mu_1 - \mu_2) + (k - k_2)(\mu_2 - \mu_3)] \\
p_{Tk} &= \frac{1}{k}[k_1^0(\mu_2 - \mu_1) + (k - k_2^0)(\mu_2 - \mu_3)] \\
q_{Tk} &= \frac{1}{k}[k_1^0(\mu_2 - \mu_1) + k_2^0(\mu_3 - \mu_2)].
\end{aligned}
$$

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 (17), (22), and (24) 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 (22) and letting $T \to \infty$.

Proof of Lemma 2. The only stochastic terms in (17), (22), and (24) 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 $j > i$,*

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

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

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

□

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

$$E\frac{1}{j-i}(\sum_{t=i+1}^{j} X_t)^2 < M. \tag{25}$$

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

Lemma 12. *Under A1-A3, there exists an $M < \infty$ such that*

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

Proof: The expected value of the first two terms on the right hand side of (18) is zero. We thus need to consider the last two terms. For $k < k_1^0$,

$$\begin{aligned}
&\frac{1}{k}(\sum_{t=1}^{k} X_t)^2 - \frac{1}{k_1^0}(\sum_{t=1}^{k_1^0} X_t)^2 \\
&= (\frac{1}{k} - \frac{1}{k_1^0})(\sum_{t=1}^{k} X_t)^2 - 2\frac{1}{k_1^0}(\sum_{t=1}^{k} X_t)(\sum_{t=k+1}^{k_1^0} X_t) - \frac{1}{k_1^0}(\sum_{t=k+1}^{k_1^0} X_t)^2 \qquad (26)\\
&= \frac{k_1^0 - k}{k_1^0}\frac{1}{k}(\sum_{t=1}^{k} X_t)^2 - 2\frac{1}{k_1^0}(\sum_{t=1}^{k} X_t)(\sum_{t=k+1}^{k_1^0} X_t) - \frac{k_1^0 - k}{k_1^0}\frac{1}{k_1^0 - k}(\sum_{t=k+1}^{k_1^0} X_2)^2
\end{aligned}$$

Apply Lemma 11 to the second term above and apply (25) to the first term and the third term above, we see that the absolute value of the expectation of (26) is bounded by $M|k_1^0 - k|/T$. This result holds for $k > k_1^0$ (only need to use $\sum_{t=1}^{k} = \sum_{t=1}^{k_1^0} + \sum_{t=k+1}^{k_1^0}$ 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

$$T|ER_{iT}(k) - ER_{iT}(k_1^0)| \le \frac{|k_1^0 - k|}{T} M, \quad (i = 1,2,3) \tag{27}$$

Proof of Lemma 3. For $k < k_1^0$, using (17) with some algebra to obtain

$$\begin{aligned} & ES_T(k) - ES_T(k_1^0) \\ = \; & (k_1^0 - k)a_T(k)^2 + (k_2^0 - k_1^0)[b_T(k)^2 - b_T(k_1^0)^2] + (T - k_2^0)[c_T(k)^2 - c_T(k_1^0)^2] \\ & + T\{ER_{1T}(k) - ER_{1T}(k_1^0)\} \\ = \; & \frac{k_1^0 - k}{(1 - k/T)(1 - k_1^0/T)} \left[(1 - k_1^0/T)(\mu_1 - \mu_2) + (1 - k_2^0/T)(\mu_2 - \mu_3)\right]^2 \qquad (28) \\ & + T\{ER_{1T}(k) - ER_{1T}(k_1^0)\}. \end{aligned}$$

Because $|(k_i^0/T) - \tau_i^0| \le T^{-1}$ $(i = 1,2)$,

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

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

$$C = (1 - \tau_1^0)(\mu_1 - \mu_2) + (1 - \tau_2^0)(\mu_2 - \mu_3) \ne 0. \tag{29}$$

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 above and using $(1 - \tau_2^0)\tau_1^0/\tau_2^0 < 1 - \tau_1^0$ to obtain

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

This verifies (29). Together with Lemma 12, we have

$$ES_T(k) - ES_T(k_1^0) \geq (k_1^0 - k)C^2 - O(\frac{k_1^0 - k}{T}) \geq (k_1^0 - k)C^2/2 \tag{30}$$

for all large T.

Remark 1 (a) Regardless of the validity of Assumption A4, by (28) and Lemma 12, for $k \leq k_1^0$

$$ES_T(k) - ES_T(k_1^0) \geq T\{ER_{1T}(k) - ER_{1T}(k_1^0)\} \geq -M|k_1^0 - k|/T.$$

By symmetry (which can be thought of as reversing the data order), for $k \geq k_2^0$,

$$ES_T(k) - ES_T(k_2^0) \geq T\{ER_{3T}(k) - ER_{3T}(k_2^0)\} \geq -M|k_2^0 - k|/T.$$

This property will be used later.

(b) Even if the strict inequality in Assumption A4 is replaced by an equality, i.e. $U(\tau_1^0) = U(\tau_2^0)$, the previous proof shows that Lemma 3 still holds for $k \in [1, k_1^0]$. The strict inequality is only needed for $k \in [k_1^0 + 1, k_2^0)]$, which is considered below. □

For $k \in [k_1^0 + 1, k_2^0)]$, use the last equality of (22) with some algebra,

$$\begin{aligned} & ES_T(k) - ES_T(k_1^0) \\ = \; & (k - k_1^0)\left[\frac{k_1^0}{k}(\mu_2 - \mu_1)^2 - \frac{(T - k_2^0)^2}{(T - k)(T - k_1^0)}(\mu_3 - \mu_2)^2\right] \\ & + T\{ER_{2T}(k) - ER_{2T}(k_1^0)\}. \end{aligned} \tag{31}$$

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

$$\begin{aligned} & ES_T(k) - ES_T(k_1^0) \\ \geq \; & (k - k_1^0)\frac{k_2^0}{k}\left[\frac{k_1^0}{k_2^0}(\mu_2 - \mu_1)^2 - \frac{(T - k_2^0)}{(T - k_1^0)}(\mu_3 - \mu_2)^2\right] \\ & + T\{ER_{2T}(k) - ER_{2T}(k_1^0)\}. \end{aligned}$$

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

$$\frac{k_1^0}{k_2^0} - \frac{\tau_1^0}{\tau_2^0} = O(T^{-1}), \quad \frac{T - k_2^0}{T - k_1^0} - \frac{1 - \tau_2^0}{1 - \tau_1^0} = O(T^{-1}) \tag{32}$$

we have

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

for some $M < \infty$ by Lemma 12. Thus for large T,

$$ES_T(k) - ES_T(k_1^0) \geq (k - k_1^0)C^*/2. \tag{33}$$

It remains to consider $k \in [k_2^0 + 1, T]$. From Remark 1(a), $ES_T(k) - ES_T(k_2^0) \geq -T|ER_{3T}(k) - ER_{3T}(k_2^0)| \geq -(k - k_2^0)M/T \geq -(T - k_2^0)M/T$. 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 (33) with $k = k_2^0$, we see that the term in the bracket is no smaller than $\frac{\tau_2^0 - \tau_1^0}{1 - \tau_2^0}C^*/4$ for large T. Thus

$$ES_T(k) - ES_T(k_1^0) \geq (k - k_1^0)\frac{\tau_2^0 - \tau_1^0}{1 - \tau_1^0}C^*/8 \tag{34}$$

for all large T. Combining (30), (33), and (34), 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

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

This further implies that the absolute value of the left hand side of the above 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,

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

First note that

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

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

$$P\left(\sup_{k \in D_{T,M}} \left|\frac{T\{R_{1T}(k) - R_{1T}(k_1^0)\}}{k - k_1^0}\right| > \eta\right) < \epsilon. \tag{35}$$

We consider the case $k < k_1^0$. From (18),

$$\begin{aligned} &T\{R_{1T}(k) - R_{1T}(k_1^0)\} \\ &= 2\Big(a_{Tk} \sum_{t=k+1}^{k_1^0} X_t\Big) + 2\Big(\{b_{Tk} - b_T(k_1^0)\} \sum_{k_1^0+1}^{k_2^0} X_t\Big) + 2\Big(\{c_{Tk} - c_T(k_1^0)\} \sum_{k_2^0+1}^{T} X_t\Big) \\ &\quad -(k_1^0 - k)a_{Tk}A_{Tk}^* - (k_2^0 - k_1^0)\Big\{b_{Tk}A_{Tk}^* - b_T(k_1^0)A_T(k_1^0)^*\Big\} \\ &\quad -(T - k_2^0)\Big\{c_{Tk}A_{Tk}^* - c_T(k_1^0)A_T^*(k_1^0)\Big\} \\ &\quad +\Big[\frac{1}{k_1^0}\Big(\sum_{t=1}^{k_1^0} X_t\Big)^2 - \frac{1}{k}\Big(\sum_{t=1}^{k} X_t\Big)^2\Big] + \Big[(T - k_1^0)(A_T^*(k_1^0))^2 - (T-k)(A_{Tk}^*)^2\Big] \end{aligned} \tag{36}$$

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}, c_{Tk} are all uniformly bounded, with an upper bound say, L, the first term divided by $k_1^0 - k$ is bounded by $L|\frac{1}{k_1^0-k}\sum_{k+1}^{k_1^0} X_t|$, which is uniformly small in $k < k_1^0 - M$ for large M by the strong law of large numbers. For the rest of terms, we will use the following easily verifiable facts:

$$|b_{Tk} - b_T(k_1^0)| \leq \left|\frac{k_1^0 - k}{T - k}\right| C, \tag{37}$$

$$|c_{Tk} - c_T(k_1^0)| \leq \left|\frac{k_1^0 - k}{T - k}\right| C, \tag{38}$$

for some $C < \infty$, and

$$A_T^*(k_1^0) - A_T^*(k) = \frac{k_1^0 - k}{(T-k)(T-k_1^0)} \sum_{t=k_0}^{T} X_t - \frac{1}{T-k} \sum_{t=k+1}^{k_1^0} X_t. \tag{39}$$

In view of (37), the second term on the right hand side of (36) divided by $k_1^0 - k$ is bounded by

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

for some $C' < \infty$, which converges to zero in probability by the law of large numbers (note that $T - k \geq T(1 - \tau_2^0)$ 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

$$(k_2^0 - k_1^0)\{b_{Tk} - b_T(k_1^0)\}A_{Tk}^* + (k_2^0 - k_1^0)b_T(k_1^0)\{A_T^*(k_1^0) - A_T^*(k)\} \tag{40}$$

Using (37), the first expression of (40) 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 (39)

$$\frac{k_2^0 - k_1^0}{(T-k)(T-k_1^0)} \sum_{t=k_0}^{T} X_t - \frac{k_2^0 - k_1^0}{T-k}\Big(\frac{1}{k_1^0 - k}\Big) \sum_{t=k+1}^{k_1^0} X_t \tag{41}$$

with the first term being $o_p(1)$ and the second term being small for large M. Thus the fifth term of (36)) 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 (36) divided by $k_1^0 - k$ can be arbitrarily small in probability provided that M and T are large. This proves the 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 $k \le k_1^0$, the lemma is implied by Lemma 3 which holds for $U(\tau_1^0) = U(\tau_2^0)$, see Remark 1(b). Next, consider for $k \in [k_1^0 + 1, k_2^0]$. From (31), (32), 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

$$\begin{aligned} & ES_T(k) - ES_T(k_1^0) \\ = \;& (k - k_1^0)\left(\frac{k_2^0}{k} - \frac{T-k_2^0}{T-k}\right)\frac{\tau_1^0}{\tau_2^0}(\mu_2 - \mu_1)^2 \\ & + (k - k_1^0)O(T^{-1}) + T\{ER_{2T}(k) - ER_{2T}(k_1^0)\} \end{aligned} \tag{42}$$

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 (42) 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 $P(S_T(\hat{k}_1) - S_T(\hat{k}_2) < 0) \to 1/2$ or equivalently, $P(T^{-1/2}\{S_T(\hat{k}_1) - S_T(\hat{k}_2)\} < 0) \to 1/2$. Because $\hat{k}_i = k_i^0 + O_p(1)$, $S_T(\hat{k}_i) = S_T(k_i^0) + O_p(1)$ (see the proof of Proposition 4). It suffices therefore to prove

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

The equality of $U(\tau_1^0)$ and $U(\tau_2^0)$ translates into an approximate equality of $ES_T(k_1^0)$ and $ES_T(k_2^0)$. More precisely, 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 $|ES_T(k_1^0) - ES_T(k_2^0)| < 2A$. 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 (22). Note that we have used the fact that $S_T(k)$, when $k = k_1^0$, can be represented by both (17) and (22) and we have used (22). Consequently, it suffices to prove

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

From (23),

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

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

$$\hat{\mu}_1^* = \frac{1}{k_1^0 + \ell}\sum_{t=1}^{k_1^0+\ell} Y_t \quad \text{and} \quad \hat{\mu}_2^* = \frac{1}{T - k_1^0 - \ell}\sum_{t=k_1^0+\ell+1}^{T} Y_t, \tag{43}$$

$$\hat{\mu}_1 = \frac{1}{k_1^0}\sum_{t=1}^{k_1^0} Y_t \qquad \text{and} \qquad \hat{\mu}_2 = \frac{1}{T-k_1^0}\sum_{t=k_1^0+1}^{T} Y_t. \tag{44}$$

Thus $\hat{\mu}_1^*$ is the least squares estimator of μ_1 using the first $k_1^0+\ell$ observations and $\hat{\mu}_2^*$ is the least squares estimator of a weighted average of μ_2 and μ_3 using the last $T-k_1^0-\ell$ 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:

$$\hat{\mu}_1^* - \mu_1 = O_p(T^{-1/2}) \qquad \text{and} \qquad \hat{\mu}_1 - \mu_1 = O_p(T^{-1/2}) \tag{45}$$

$$\hat{\mu}_2^* - \mu_2 - \frac{1-\tau_2^0}{1-\tau_1^0}(\mu_3-\mu_2) = O_p(T^{-1/2}) \qquad \text{and} \qquad \hat{\mu}_2 - \mu_2 - \frac{1-\tau_2^0}{1-\tau_1^0}(\mu_3-\mu_2) = O_p(T^{-1/2}) \tag{46}$$

$$\hat{\mu}_i^* - \hat{\mu}_i = O_p(T^{-1}) \quad (i=1,2) \tag{47}$$

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

$$S_T(k_1^0+\ell) = \sum_{t=1}^{k_1^0}(Y_t-\hat{\mu}_1^*)^2 + \sum_{t=k_1^0+1}^{k_1^0+\ell}(Y_t-\hat{\mu}_1^*)^2 + \sum_{k_1^0+\ell+1}^{T}(Y_t-\hat{\mu}_2^*)^2. \tag{48}$$

Similarly,

$$S_T(k_1^0) = \sum_{t=1}^{k_1^0}(Y_t-\hat{\mu}_1)^2 + \sum_{t=k_1^0+1}^{k_1^0+\ell}(Y_t-\hat{\mu}_2)^2 + \sum_{k_1^0+\ell+1}^{T}(Y_t-\hat{\mu}_2)^2. \tag{49}$$

The difference between the two first terms on the right hand side of (48) and (49), respectively, is

$$\sum_{t=1}^{k_1^0}(Y_t-\hat{\mu}_1^*)^2 - \sum_{t=1}^{k_1^0}(Y_t-\hat{\mu}_1)^2 = k_1^0(\hat{\mu}_1^*-\hat{\mu}_1)^2 = O_p(T^{-1}). \tag{50}$$

Similarly, the difference between the two third terms on the right hand sides of (48) and (49), respectively, is also $O_p(T^{-1})$. Next consider the difference between the two middle terms. For $t \in [k_1^0+1, \ell]$, $Y_t = \mu_2 + X_t$. Hence

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

$$= 2\{\mu_2-\hat{\mu}_1^*-(\mu_2-\hat{\mu}_2)\}\sum_{t=k_1^0+1}^{k_1^0+\ell} X_t + \ell\{(\mu_2-\hat{\mu}_1^*)^2-(\mu_2-\hat{\mu}_2)^2\}. \tag{51}$$

From (45) and (46), 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 (51) is equal to

$$2(\mu_2 - \mu_1)(1 + \lambda_1) \sum_{t=k_1^0+1}^{k_1^0+\ell} X_t + \ell(\mu_2 - \mu_1)^2(1 - \lambda_1^2) + O_p(T^{-1/2}). \tag{52}$$

Under strict stationarity, $\sum_{t=k_1^0+1}^{k_1^0+\ell} X_t$ has the same distribution as $\sum_{t=1}^{\ell} X_t$. Thus (52) or, equivalently, (51) converges in distribution to $(1 + \lambda_1)W_2^{(1)}(\ell, \lambda_1)$. This implies that $S_T(k_1^0 + \ell) - S_T(k_1^0)$ converges in distribution to $(1 + \lambda_1)W_2^{(1)}(\ell, \lambda_1)$ for $\ell > 0$. It remains to consider $\ell < 0$. We replace ℓ by $-\ell$ and still consider a positive ℓ. In particular, $\hat{\mu}_1^*$ and $\hat{\mu}_2^*$ are defined with $-\ell$ in place of ℓ. Then (48) and (49) are replaced, respectively, by

$$S_T(k_1^0 - \ell) = \sum_{t=1}^{k_1^0-\ell}(Y_t - \hat{\mu}_1^*)^2 + \sum_{t=k_1^0-\ell+1}^{k_1^0} (Y_t - \hat{\mu}_2^*)^2 + \sum_{k_1^0+1}^{T}(Y_t - \hat{\mu}_2^*)^2 \tag{53}$$

and

$$S_T(k_1^0) = \sum_{t=1}^{k_1^0-\ell}(Y_t - \hat{\mu}_1)^2 + \sum_{t=k_1^0-\ell+1}^{k_1^0} (Y_t - \hat{\mu}_1)^2 + \sum_{k_1^0+1}^{T}(Y_t - \hat{\mu}_2)^2. \tag{54}$$

The major distinction between (48) and (53) lies in the change of $\hat{\mu}_1^*$ to $\hat{\mu}_2^*$ for the middle terms on the right hand. One can observe a similar change for (49) and (54). Similar to (50), the difference between the two first terms on the right hand of (53) and (54) is $O_p(T^{-1})$. The same is true for the difference between the two third terms on the right hand. Using $Y_t = \mu_1 + X_t$ for $t \le k_1^0$, we have

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

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

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 (46). The details will not be presented to avoid repetition. □

Proof of Lemma 8. From the 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

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

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; 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. □

Proof of Proposition 9. 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 9.

Proof of (a). First consider $k \le k_1^0$. From (17),

$$\left|U_T(k/T) - EU_T(k/T) - T^{-1}\sum_{t=1}^{T}(X_t^2 - EX_t^2)\right| = \left|R_{1T}(k) - ER_{1T}(k)\right| \qquad (55)$$

The first two terms of $R_{1T}(k)$ [see (18)] 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 $k \le k_1^0$. This proves the lemma for $k \le k_1^0$. The proof for $k \ge k_1^0$ is the same and follows from $R_{iT}(k) - ER_{iT}(k) = O_p(T^{-1/2}v_T) \quad (i = 2, 3)$.

Proof of (b). Consider first $k \le k_1^0$. By the second equality of (28), 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

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

where C is given by (29) 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 10. As in the proof of Lemma 4, it suffices to show that for every $\eta > 0$, there exists an $M > 0$ such that

$$P\left(\sup_{k \in D_{T,M}^*} \left|\frac{T\{R_{1T}(k) - R_{1T}(k_1^0)\}}{k - k_1^0}\right| > \eta v_T^2\right) < \epsilon. \qquad (56)$$

The above is similar to (35) 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 (36) 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 (1994a): there exists a $C_1 < \infty$ such that for each $\ell > 0$,

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

Now consider the first term on the right hand side of (36). Note that $|a_{Tk}| \le v_T L$ for some $L < \infty$. Thus it is enough to show

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

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

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

The above probability is small if M is large. The proof of Lemma 4 demonstrates that all other terms are of lower 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 11. The proof is similar to that of Proposition 4. We only outline the major distinction. In view of the rate of convergence, we consider the process $\Lambda_T(s)=S_T(k_1^0+[sv_T^{-2}])-S_T(k_1^0)$, indexed by a real number s. We shall derive the limiting process for $|s|\le M$ for an arbitrary given $M<\infty$. Let $D[-M,M]$ denote the space of cadlag functions endowed with the Skorohod metric, see Pollard (1984). We shall show that $\Lambda_T(s)$ converges weakly in $D[-M,M]$ to a pertinent limiting process. First consider $s>0$. Let $\ell=[sv_T^{-2}]$. Define $\hat{\mu}_i^*$ and $\hat{\mu}_i$ as in (43) and (44), respectively. Then (45)-(47) still hold with μ_j interpreted as μ_{jT}. For example,

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

This follows because, from $|\ell|\le Mv_T^{-2}$, the first term on the right hand side is of $O(1/(\sqrt{T}v_T))$ which converges to zero, and the second term is $O_p(1)$. Equations (48)-(49) 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 (48) and (49) converge to zero in probability. Equation (52) in the present case is reduced to

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

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 $\ell=[sv_T^{-2}]$, using the functional central limit theorem for linear processes [e.g., Phillips and Solo (1992)]

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

in the space $D[0,M]$, where $B_2(s)$ is a Brownian motion process on $[0,\infty)$, and

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

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 $\text{argmin}_s\Gamma(s,\lambda) \stackrel{d}{=} a(1)^2\sigma_\epsilon^2\text{argmin}_s\Lambda(s,\lambda)$. Now because $cB_i(s)$ has the same distribution as $B_i(c^2 s)$, we have

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

This implies that

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

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

References

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

[2] Andrews, D.W.K. and W. Ploberger (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383-1414.

[3] Bai, J. (1994a) Least Squares Estimation of a Shift in Linear Processes. *Journal of Time Series Analysis* 15, 453-472.

[4] Bai, J. (1994b) Estimation of Structural Change based on Wald Type Statistics. Working Paper No. 94-6, Department of Economics, M.I.T.

[5] Bai, J. (1994c) GMM Estimation of Multiple Structural Changes. NSF Proposal.

[6] Bai, J. and P. Perron (1994) Testing for and Estimation of multiple structural changes. Manuscript, Department of Economics, M.I.T.

[7] Chong, T. T-L. (1994) Consistency of Change-Point Estimators When the Number of Change-Points in Structural Models is Underspecified. Manuscript, Department of Economics, University of Rochester.

[8] Cooper, S.J. (1995) Multiple regimes in U.S. output Fluctuations. Manuscript, Kennedy School of Government, Harvard University.

[9] Garcia, R. and P. Perron (1994) An Analysis of the Real Interest Rate under Regime Shifts. Forthcoming in *Review of Economics and Statistics.*

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

[11] Phillips, P. C. B. and Solo, V. (1992) Asymptotics for linear processes. *The Annals of Statistics* 20, 971-1001.

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

[13] Yao, Y-C. (1988) Estimating the Number of Change-Points via Schwarz' Criterion. *Statistics and Probability Letters* 6, 181-189.

Figure 1: Histograms of the estimated break points for Model (I): (a) Sequential method; (b) Repartition method; (c) Simultaneous Method.

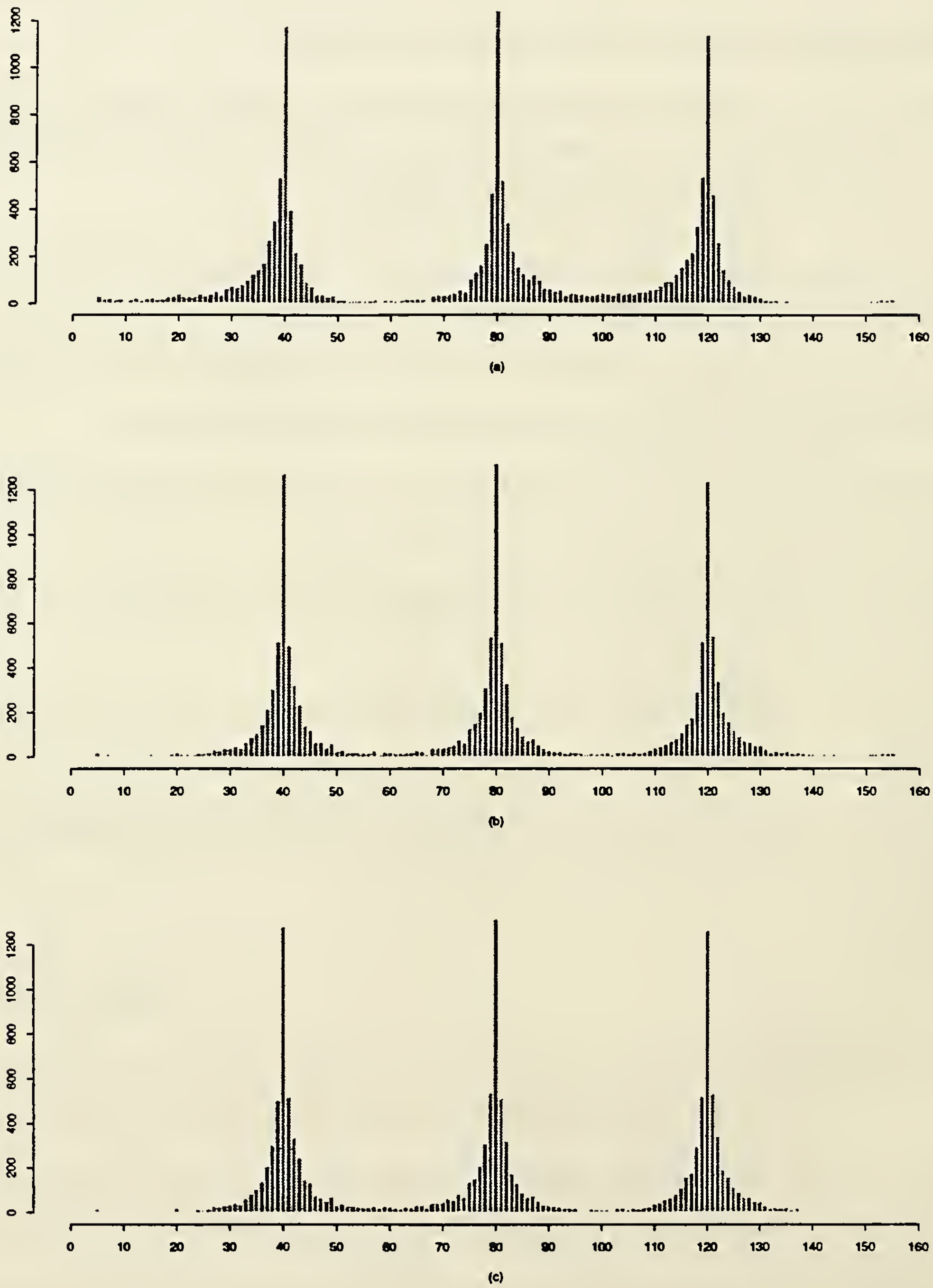

Figure 2: Histograms of the estimated break points for Model (II): (a) Sequential method; (b) Repartition method; (c) Simultaneous Method.

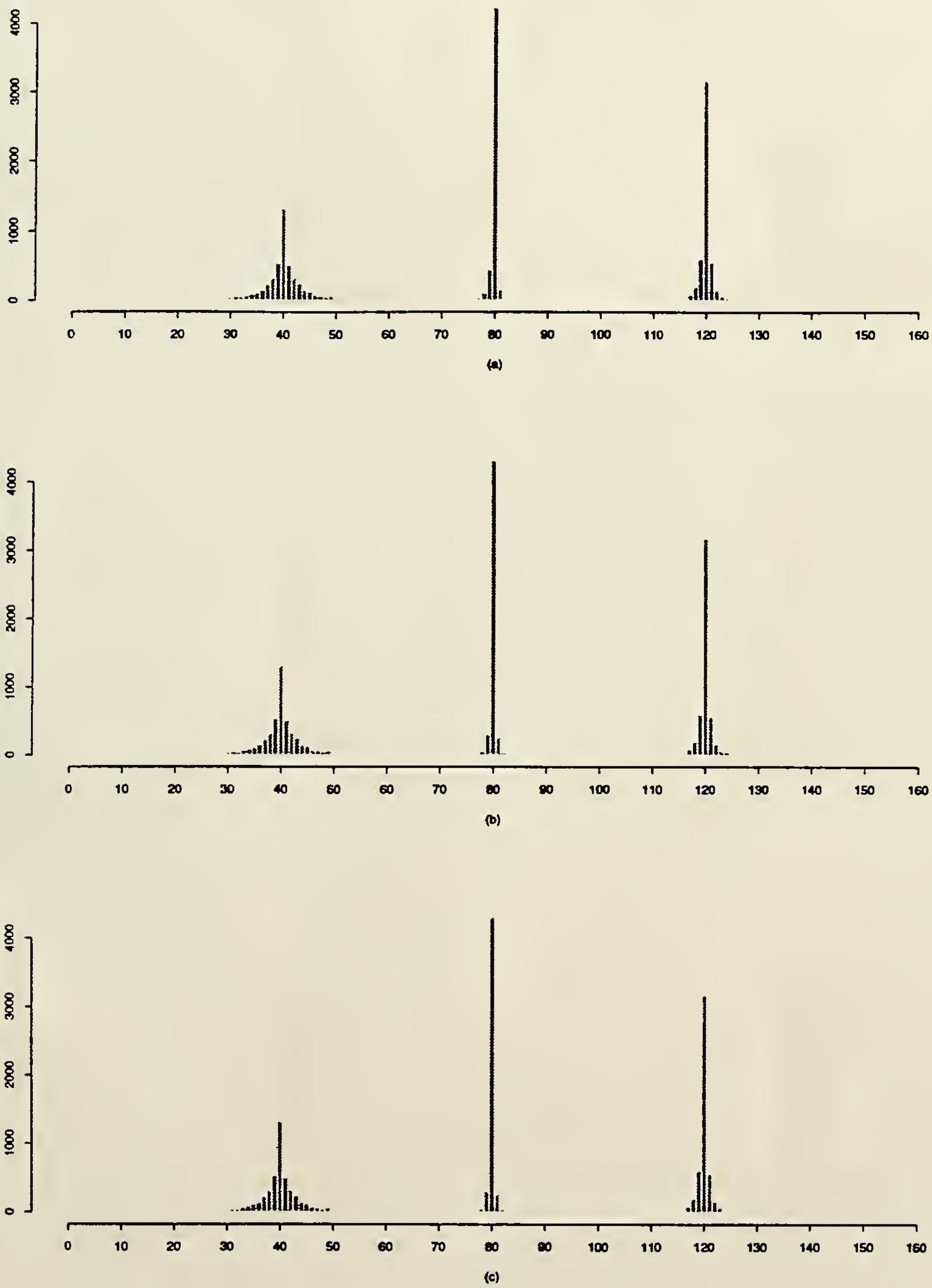

Figure 3: Histograms of the estimated numbers (of breaks) for models (I), (II) and (III). Sequential estimation: $(a), (b), (c)$. BIC criterion: $(a'), (b'), (c')$