Least squares estimation of a shift in linear processes

Bai, Jushan

Massachusetts Institute of Technology

16 February 1993

Least Squares Estimation of a Shift in Linear Processes

by
Jushan Bai
Massachusetts Institute of Technology

Feburary 16, 1993

Mailing Address:

Department of Economics
E52-274B
Massachusetts Institute of Technology
Cambridge, MA 02139
U.S.A.

jbai@athena.mit.edu

Abstract

This paper considers a mean shift with an unknown shift point in a linear process and estimates the unknown shift point (change point) by the method of least squares. Pre-shift and post-shift means are estimated concurrently with the change point. The consistency and the rate of convergence for the estimated change point are established. The asymptotic distribution for the change point estimator is obtained when the magnitude of shift is small. It is shown that serial correlation affects the variance of the change point estimator via the sum of the coefficients (impulses) of the linear process. When the underlying process is an ARMA, a mean shift causes overestimation of its order. A simple procedure is suggested to mitigate the bias in order estimation.

Keywords. Mean shift; linear processes; change point; rate of convergence; order estimation; generalized residuals.

1. Introduction and notations

The problem of a mean shift with an unknown shift point in an independent and identically distributed sequence has received considerable attention in the literature. Sen and Srivastava (1975a,1975b), Hawkins (1977), Worsley (1979, 1986), James, James, and Siegmund (1987), and Srivastava and Worsley (1986) proposed tests for testing a shift in a sequence of normal means. Hinkley (1970), Bhattacharya (1987), Yao (1987), and many others considered the estimation of the shift point in a sequence of independent variables. For serially correlated data, Picard (1985) estimated a shift in a Gaussian autoregressive process with a known order. These authors considered maximum likelihood estimation (MLE).

In this paper, we apply the least squares method (LS) to the estimation of a shift point. Unlike the MLE, the LS method does not need to specify the underlying error distribution function and is computationally simple. The least squares procedure also allows a broader specification of correlation structure in the data than MLE can typically permit. In particular, we assume that observations are drawn from a linear process of martingale differences, rendering ARMA processes as special cases. When the underlying process is assumed to have an ARMA representation, the orders of the process is not assumed to be known. This is important because of the following two reasons. First, in practice, orders of an ARMA process are rarely known and have to be estimated. Second, order determination via the AIC and BIC criteria tends to overestimate the order of an ARMA process if a shift exists, as was reported by MacNeill and Duong (1982). In this paper, a simple procedure is suggested to alleviate the bias caused by a mean shift when estimating the orders.

The model considered in this paper is as follows:

$$Y_t = \mu(t) + X_t \qquad (t = \cdots, -2, -1, 0, 1, 2, \cdots) \tag{1}$$

where $\mu(t)$ is a nonstochastic function in time and X_t is a linear stochastic process given by

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = a(B)\varepsilon_t \tag{2}$$

with $a(B) = \sum_{j=0}^{\infty} a_j B^j$, $B^l \varepsilon_t = \varepsilon_{t-l}$ $(l \geq 0)$, and ε_t being white noise.

We consider the simple case that $\mu(t)$ only takes two different values, μ_1 before time k_0 and μ_2 after time k_0. That is,

$$\mu(t) = \begin{cases} \mu_1 & \text{if } t \leq k_0 \\ \mu_2 & \text{if } t > k_0 \end{cases}$$

where μ_1, μ_2, and k_0 are unknown and k_0 is the change point.

The problem is to estimate μ_1, μ_2, and k_0 given T observations $Y_1, Y_2, \cdots, Y_T$. We assume that $k_0 = [T\tau]$ for some $\tau \in (0,1)$, where $[\cdot]$ is the integer-valued function. When X_t has an ARMA representation, we may also want to estimate its orders as well as its coefficients.

The least squares (LS) estimation of a shift is not new. Hawkins (1986) examined the LS method for a shift in an i.i.d. sequence. He proved that $T^{1/2-\delta}(\hat{\tau} - \tau) \to 0$ in probability for any $\delta > 0$, where $\hat{\tau}$ is the LS estimator of τ (defined below). Hawkin's rate of convergence is improved in this paper despite serial correlations in observations. We shall show that $T(\hat{\tau} - \tau) = O_p(1)$. We also show how serial correlation in data affects the variance of the change point estimator. In particular, we find that , when X_t is an ARMA process given by $\Psi(B)X_t = \Theta(B)\varepsilon_t$, the variance of the change point estimator is smaller than that of an i.i.d. sequence with a shift if $|\Theta(1)\Psi(1)^{-1}| < 1$.

Throughout this paper, we assume:

(A) the ε_t are i.i.d. with mean zero and variance σ^2, or

(A') the ε_t are martingale differences satisfying: $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, $E\varepsilon_t^2 = \sigma^2$, $n^{-1}\sum_{t=1}^{n} E(\varepsilon_t^2|\mathcal{F}_{t-1}) \to \sigma^2$, and there exists a $\delta > 0$ such that $\sup_t E|\epsilon_t|^{2+\delta} < \infty$, where $\mathcal{F}_t$ is the σ-field generated by ε_s, $s \leq t$.

We shall focus on the estimation of the change point. Once the change point is estimated, μ_1 and μ_2 can be estimated by using the estimated pre-change and post change subsamples. The least squares estimator $\hat{k}$ of the change point k_0 is defined as follows:

$$\hat{k} = \text{argmin}_k \left(\min_{\mu_1,\mu_2} \left\{ \sum_{t=1}^{k} (Y_t - \mu_1)^2 + \sum_{t=k+1}^{T} (Y_t - \mu_2)^2 \right\} \right). \tag{3}$$

Thus the shift point is estimated by minimizing the sum of squares of residuals among all possible sample splits. Statistical properties of this estimator will be examined in later sections.

We denote the mean of the first k observations by $\bar{Y}_k$ and the mean of the last $T-k$ observations by $\bar{Y}_k^*$. If the shift point is k, then $\bar{Y}_k$ and $\bar{Y}_k^*$ are the usual least squares estimators of μ_1 and μ_2, respectively. The corresponding sum of squares of residuals is

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

Thus $\hat{k} = \text{argmin}_k(S_k^2)$ and the LS estimators for μ_1 and μ_2 are $\hat{\mu}_1 = \bar{Y}_{\hat{k}}$ and $\hat{\mu}_2 = \bar{Y}_{\hat{k}}^*$, respectively. Next, write $\bar{Y} = \bar{Y}_T$, which is the overall mean of the given data. Since for each k $(1 \le k \le T-1)$,

$$\sum_{t=1}^{T} (Y_t - \bar{Y})^2 = S_k^2 + V_k^2$$

where

$$V_k = \left(\frac{k(T-k)}{T} \right)^{1/2} (\bar{Y}_k^* - \bar{Y}_k), \tag{4}$$

it follows that

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

As will be seen, the statistical properties of the change point estimator are obtained by studying the behavior of V_k and the argmax functional.

Denote the LS residuals by $\hat{X}_t$, which is defined as

$$\hat{X}_t = Y_t - \hat{\mu}_1 - (\hat{\mu}_2 - \hat{\mu}_1) I(t > \hat{k}),$$

where $I(\cdot)$ is the indicator function. We shall call $\hat{X}_t$ the generalized residuals in view of the presence of a change point estimator. If X_t is an ARMA process, then we can use the $\hat{X}_t$ to estimate the orders and other parameters in X_t. Monte Carlo experiments show that the order estimation based on $\hat{X}_t$ yields almost identical results as those based on X_t, thus providing a practical solution to the problem reported by MacNeill and Duong (1984). This two step procedure for estimating the parameters of the model is much simpler than MLE from the computation point of view.

Denote $\hat{\tau} = \hat{k}/T$. We shall establish the consistency, the rate of convergence, and the limiting distribution of $\hat{\tau}$ in the following several sections. We shall first, however, generalize the Hájek and Rényi inequality to serially correlated variables. This inequality is important to our results.

2. A generalization of the Hájek and Rényi inequality

Let $\varepsilon_1, \varepsilon_2, \cdots,$ be a sequence of martingale differences with $E\varepsilon_i^2 = \sigma^2$, and $\{c_k\}$ be a decreasing positive sequence of constants. Hájek and Rényi (1955) proved that

$$Pr\left(\max_{m\le k\le n} c_k|\sum_{i=1}^{k}\varepsilon_i| > \alpha\right) \le \frac{\sigma^2}{\alpha^2}\left(mc_m^2 + \sum_{i=m+1}^{n} c_i^2\right). \tag{5}$$

This inequality was initially stated in terms of i.i.d. random variables and was later generalized to martingales by Birnbaum and Marshall (1961). We now generalize this inequality to serially correlated variables. Let X_t be given by (2). We assume:

(B) $\sum_{j=0}^{\infty} j|a_j| < \infty$.

This condition is satisfied for stationary ARMA processes. Under assumptions (A or A′) and (B), the generalized Hájek and Rényi inequality takes the following form:

$$Pr\left(\max_{m\le k\le n} c_k|\sum_{i=1}^{k}X_i| > \alpha\right) \le A\frac{\sigma^2}{\alpha^2}\left(mc_m^2 + \sum_{i=m+1}^{n} c_i^2\right), \tag{6}$$

where $A < \infty$ is a constant only depending on the $a_j's$. The proof is given in the appendix. For $c_k = 1/k$, because $\sum_{k=m}^{\infty} k^{-2} = O(m^{-1})$, we have

$$Pr\left(\sup_{k\ge m}\frac{1}{k}|\sum_{i=1}^{k}X_i| > \alpha\right) \le \frac{A_1}{\alpha^2 m} \tag{7}$$

for some $A_1 < \infty$. The weak law of large numbers is an immediate consequence of the above inequality. Next consider the case $c_k = 1/\sqrt{k}$ and $m = 1$. By inequality (6),

$$Pr\left(\sup_{1\le k\le n} \frac{1}{\sqrt{k}}|\sum_{i=1}^{k} X_i| > \alpha\right) \le A\frac{\sigma^2}{\alpha^2}\sum_{k=1}^{n}\frac{1}{k} \le \frac{C\log n}{\alpha^2}$$

for some $C > 0$. This implies

$$\sup_{1\le k\le n} \frac{1}{\sqrt{k}}\left|\sum_{i=1}^{k} X_i\right| = O_p(\sqrt{\log n}) \tag{8}$$

Furthermore, the following invariance principle holds for X_t under assumptions (A or A′) and (B) (see Hall and Heyde 1980, Theorem 5.5, p. 141-146),

$$T^{-1/2}\sum_{t=1}^{[Ts]} X_t \Rightarrow (\sum_{j=0}^{\infty} a_j)\sigma B(s) \tag{9}$$

where $B(s)$ is a standard Brownian motion on [0,1].

3. The consistency of $\hat{\tau}$

The proof of consistency is almost standard. Recall how we prove, in general, the consistency of an estimator obtained by maximizing an objective function. We need to argue that the objective function converges uniformly in probability to a nonstochastic function of parameters and that the nonstochastic function has a unique global maximum. The objective function in our problem is $|V_k|$ $(k = 1, 2, \cdots, T-1)$. However, we will be able to work with V_k (without the absolute sign). This is because the expected values of V_k $(k = 1, 2, \cdots, T-1)$ do not change signs. We shall prove that the expected value of V_k has a unique maximum at k_0 and that $(V_k - EV_k)$ is uniformly small in k for large T.

First notice that

$$\begin{aligned} |V_k| - |V_{k_0}| &\le |V_k - EV_k| + |V_{k_0} - EV_{k_0}| + |EV_k| - |EV_{k_0}| && (10)\\ &\le 2\left(\sup_k |V_k - EV_k|\right) + |EV_k| - |EV_{k_0}|. && (11)\end{aligned}$$

For the sake of simplicity, we shall assume that $T\tau$ itself is an integer and is equal to k_0. Write $d = k/T$ and $\tau = k_0/T$. We next show that $|EV_k|$ achieves its maximum at $k = k_0$ by showing that

$$|EV_{k_0}| - |EV_k| \geq C_\tau |\lambda(d - \tau)| \tag{12}$$

for some $C_\tau > 0$, where $\lambda = \mu_2 - \mu_1$ is the magnitude of shift. We need only consider the case $k \leq k_0$ because of symmetry. We assume without loss of generality that $\lambda > 0$ (otherwise consider the series $-Y_t$). Then

$$EV_k = \frac{1-\tau}{1-d}\{d(1-d)\}^{1/2}\lambda > 0 \qquad \text{for } k \leq k_0. \tag{13}$$

In particular, $EV_{k_0} = \{\tau(1-\tau)\}^{1/2}\lambda$. It follows that

$$\begin{aligned} |EV_{k_0}| - |EV_k| &= \lambda\left(\{\tau(1-\tau)\}^{1/2} - \{d(1-d)\}^{1/2}(1-\tau)(1-d)^{-1}\right) \\ &= \lambda(1-\tau)\left[\left(\frac{\tau}{1-\tau}\right)^{1/2} - \left(\frac{d}{1-d}\right)^{1/2}\right]. \end{aligned}$$

Multiplying and dividing the above expression by $[\tau/(1-\tau)]^{1/2} + [d/(1-d)]^{1/2}$, we obtain

$$\begin{aligned} |EV_{k_0}| - |EV_k| &= \lambda\frac{\tau - d}{1-d}\left[\left(\frac{\tau}{1-\tau}\right)^{1/2} + \left(\frac{d}{1-d}\right)^{1/2}\right]^{-1} \\ &\geq \frac{1}{2}\lambda(\tau - d)\left(\frac{\tau}{1-\tau}\right)^{-1/2}. \end{aligned} \tag{14}$$

This proves (12) for $C_\tau = [\tau/(1-\tau)]^{-1/2}/2$. By (11), (12), and $|V_{\hat{k}}| - |V_{k_0}| \geq 0$, we obtain immediately (replacing d by $\hat{\tau} = \hat{k}/T$),

$$|\hat{\tau} - \tau| \leq 2C_\tau^{-1}\lambda^{-1}\sup_k |V_k - EV_k|.$$

From (4),

$$V_k - EV_k = T^{-\frac{1}{2}}(k/T)(T-k)^{-1/2}\sum_{t=k+1}^{T} X_t - T^{-\frac{1}{2}}(1-k/T)k^{-1/2}\sum_{t=1}^{k} X_t,$$

hence

$$|V_k - EV_k| \leq T^{-\frac{1}{2}}\left((T-k)^{-1/2}\left|\sum_{t=k+1}^{T} X_t\right| + k^{-1/2}\left|\sum_{t=1}^{k} X_t\right|\right). \tag{15}$$

It follows from (8) that the above is $T^{-1/2}O_p(\log T)$ uniformly in k. Therefore

$$|\hat{\tau} - \tau| = (T^{-1/2}\lambda^{-1})O_p(\log T), \tag{16}$$

establishing the consistency. Notice that λ is kept on the right hand side in order to illustrate how the rate depends on the magnitude of change. In addition, this allows us to incorporate the case that λ varies with the sample size T. In fact, we will examine specifically the case of a small change in the sense that $\lambda = \lambda_T \to 0$. When λ is a fixed constant, (16) implies that $T^{1/2-\delta}(\hat{\tau} - \tau) \xrightarrow{p} 0$ for any $\delta > 0$, giving rise to the result of Hawkins (1986) for i.i.d. errors. This result is improved in the next section.

4. The rate of convergence

We shall establish the stronger result:

$$\hat{\tau} - \tau = O_p\left(\frac{1}{T\lambda^2}\right). \tag{17}$$

To this end, choose a $\delta > 0$ such that $\tau \in (\delta, 1-\delta)$. Since $\hat{k}/T$ is consistent for τ, for every $\epsilon > 0$, $Pr(\hat{k}/T \notin (\delta, 1-\delta)) < \epsilon$ when T is large. Thus we now only need to examine the behavior of V_k over those k for which $T\delta \le k \le T(1-\delta)$. To prove (17), we shall prove that $Pr\left(|\hat{\tau} - \tau| > M(T\lambda^2)^{-1}\right)$ is small when T and M are large. For every $M > 0$, define $D_{T,M} = \{k;\ T\delta \le k \le T(1-\delta), |k - k_0| > M\lambda^{-2}\}$. Then

$$\begin{aligned}
&Pr\left(|\hat{\tau} - \tau| > M(T\lambda^2)^{-1}\right) \\
&\quad \le Pr\left(\hat{\tau} \notin (\delta, 1-\delta)\right) + Pr\left(|\hat{\tau} - \tau| > M(T\lambda^2)^{-1}, \hat{\tau} \in (\delta, 1-\delta)\right) \\
&\quad \le \epsilon + Pr\left(\sup_{k \in D_{T,M}} |V_k| \ge |V_{k_0}|\right).
\end{aligned}$$

Because $|x| \ge |y|$ implies either $x - y \ge 0$ and $x + y \ge 0$ or $x - y \le 0$ and $x + y \le 0$, we have

$$\begin{aligned}
&Pr\left(\sup_{k \in D_{T,M}} |V_k| \ge |V_{k_0}|\right) \\
&\quad \le Pr\left(\sup_{k \in D_{T,M}} V_k - V_{k_0} \ge 0\right) + Pr\left(\sup_{k \in D_{T,M}} V_k + V_{k_0} \le 0\right) \stackrel{def}{=} P_1 + P_2.
\end{aligned}$$

We next argue that P_1 and P_2 are small when T and M are large. Define $b(k) = \{(k/T)(1-k/T)\}^{1/2}$ $(k = 1, 2, ...T)$. The following fact will be useful:

$$0 \le b(k) \le 1, \quad |b(k_0) - b(k)| \le B|k_0 - k|/T \quad \text{for some } B > 0 \tag{18}$$

Now consider P_2.

$$\begin{aligned}
P_2 &= Pr\left(\sup_{k \in D_{T,M}} \{EV_{k_0} - V_{k_0} - (V_k - EV_k)\} \ge EV_{k_0} + EV_k\right) \\
&\le P\left(2 \sup_{k \in D_{T,M}} |V_k - EV_k| \ge EV_{k_0}\right) \qquad\qquad \text{since } EV_k > 0 \\
&\le Pr\left(2 \sup_{T\delta \le k \le T(1-\delta)} b(k)\left|\frac{1}{T-k}\sum_{t=k+1}^{T} X_t - \frac{1}{k}\sum_{t=1}^{k} X_t\right| \ge EV_{k_0}\right) \\
&\le Pr\left(2 \sup_{k \le T(1-\delta)} \frac{1}{T-k}\left|\sum_{t=k+1}^{T} X_t\right| \ge \frac{1}{2}EV_{k_0}\right) + Pr\left(2 \sup_{k \ge T\delta} \frac{1}{k}\left|\sum_{t=1}^{k} X_t\right| \ge \frac{1}{2}EV_{k_0}\right).
\end{aligned}$$

Because $EV_{k_0} = \{\tau(1-\tau)\}^{1/2}\lambda$, inequality (7) implies that each of the two terms above converges to zero as T tends to infinity.

The argument for P_1 is more delicate. Note that

$$P_1 = Pr\left(\sup_{k \in D_{T,M}} \{V_k - EV_k - (V_{k_0} - EV_{k_0})\} \ge EV_{k_0} - EV_k\right).$$

However,

$$\begin{aligned}
&V_k - EV_k - (V_{k_0} - EV_{k_0}) \\
&\quad = b(k)\left(\frac{1}{T-k}\sum_{t=k+1}^{T} X_t - \frac{1}{k}\sum_{t=1}^{k} X_t\right) - b(k_0)\left(\frac{1}{T-k_0}\sum_{t=k_0+1}^{T} X_t - \frac{1}{k_0}\sum_{t=1}^{k_0} X_t\right) \\
&\quad = \left(b(k_0)\frac{1}{k_0}\sum_{t=1}^{k_0} X_t - b(k)\frac{1}{k}\sum_{t=1}^{k} X_t\right) + \left(\frac{b(k)}{T-k}\sum_{t=k+1}^{T} X_t - \frac{b(k_0)}{T-k_0}\sum_{t=k_0+1}^{T} X_t\right) \\
&\quad \stackrel{def}{=} G(k) + H(k).
\end{aligned} \tag{19}$$

Since $(EV_{k_0} - EV_k) \ge C_\tau \lambda (k_0 - k)/T$ by (12), we have

$$\begin{aligned}
P_1 &\le Pr\left(\sup_{k \in D_{T,M}} |G(k)| > \frac{1}{2}\lambda\frac{|k_0 - k|}{T}C\right) + Pr\left(\sup_{k \in D_{T,M}} |H(k)| > \frac{1}{2}\lambda\frac{|k_0 - k|}{T}C\right) \\
&\stackrel{def}{=} P_{1,1} + P_{1,2}
\end{aligned}$$

We now prove that $P_{1,1}$ is small if T and M are large. Again because of symmetry, we consider only the case of $k \leq k_0$ and $k \in D_{T,M}$. More precisely, we consider those $k's$ such that $T\delta \leq k \leq T\tau - M\lambda^{-2}$. By adding and subtracting terms, $G(k)$ can be written as:

$$G(k) = b(k_0)\frac{k-k_0}{kk_0}\sum_{t=1}^{k_0} X_t + \{b(k_0) - b(k)\}\frac{1}{k}\sum_{t=1}^{k} X_t + b(k_0)\frac{1}{k}\sum_{t=k}^{k_0} X_t \tag{20}$$

By (18) and $k \geq T\delta$, we have

$$|G(k)| \leq \frac{k_0 - k}{T\delta k_0}\left|\sum_{t=1}^{k_0} X_t\right| + B\frac{k_0-k}{T}\frac{1}{T\delta}\left|\sum_{t=1}^{k} X_t\right| + \frac{k_0-k}{T\delta}\frac{1}{k_0-k}\left|\sum_{t=k+1}^{k_0} X_t\right|. \tag{21}$$

It follows that

$$\begin{aligned}
P_{1,1} \quad &\leq \quad Pr\left(\frac{1}{T\tau}|\sum_{t=1}^{T\tau} X_t| > \frac{1}{2}\delta\lambda C\right) \\
&+ \quad Pr\left(\sup_{1\leq k\leq T}\frac{1}{T}|\sum_{t=1}^{k} X_t| > \frac{1}{2}\delta\lambda C B^{-1}\right) \\
&+ \quad Pr\left(\sup_{k\leq T\tau - M\lambda^{-2}}\frac{1}{T\tau - k}|\sum_{t=k}^{T\tau} X_t| > \frac{1}{2}\delta\lambda C\right) \\
&\leq \quad A\left(\frac{2\sigma}{\delta C\tau}\right)^2\frac{1}{T\lambda^2} + A\left(\frac{2B\sigma}{\delta C}\right)^2\frac{1}{T\lambda^2} + \frac{4A_1}{\delta^2 C^2 M}
\end{aligned}$$

by inequality (6) and (7). The last three terms are negligible when T and M are large. The proof for $P_{1,2}$ is similar.

5. The limiting distribution

In this section, we aim to derive the asymptotic distribution of $\hat{\tau}$ when the sample size increases to infinity. The limiting distribution provides a way for constructing confidence intervals for the change point. The limiting distribution also provides some qualitative aspects on how the estimated change point is related to other parameters in the model. We now assume that λ depends on T and it diminishes as T increases. When λ is a constant not depending on T, the results of Hinkely (1971a, 1971b) for the i.i.d. case indicate that the limiting distribution of $\hat{\tau}$ depends on the underlying distribution of the innovations ε_t and also on λ in quite an intricate way. Thus confidence intervals can not be easily constructed. In addition, when λ is large (relative to the variance of the innovations), the estimation of the change point is quite precise. Thus it might be more important to be able to construct confidence intervals for small changes. Furthermore, a confidence interval based on the limiting distribution for small λ is expected to cover the corresponding interval when λ is actually large and thus can always be used as a more conservative confidence interval even if λ is large.

We shall use $V(k)$ and V_k interchangeably in this section. Now denote λ by λ_T. If λ_T is not too small in the sense that $T^{1/2}\lambda_T/\log T \to \infty$, then the estimator $\hat{\tau}$ is still consistent, as can be seen from (16). The consistency in turn leads to (17).

Let us now assume that, for some $0 < \alpha < 1/2$

(C) $\quad \lambda_T \to 0$, $T^{1/2-\alpha}\lambda_T \to \infty$.

This assumption is sufficient for $\hat{\tau}$ to be consistent and is used by Picard (1985). Also note that $\lambda_T \gg T^{-1/2}$. From (17), we have $\hat{\tau} - \tau = O_p(T^{-1}\lambda_T^{-2})$, or equivalently,

$$\hat{k} - k_0 = O_p(\lambda_T^{-2}). \tag{22}$$

Notice

$$\hat{k} = \text{argmax}_k(V_k^2) = \text{argmax}_k T\left(V_k^2 - V_{k_0}^2\right).$$

Given the rate of convergence in (22), to study the limiting distribution, we only need to examine the behavior of $T(V_k^2 - V_{k_0}^2)$ for those k in the neighborhood of k_0 such that $k = [k_0 + v\lambda_T^{-2}]$, where v varies in an arbitrary bounded interval. We shall obtain some weak convergence result for $T(V_k^2 - V_{k_0}^2)$ and then apply the continuous mapping theorem for the argmax functional. The idea is similar to that of Picard (1985) and Yao (1987). To this end, connect by linear segments the points $(k,\ V_k^2 - V_{k_0}^2) \in R^2$, $(k = 1, 2, \cdots, T-1)$ and define:

$$\Lambda_T(v) = T\left\{V([k_0 + v\lambda_T^{-2}])^2 - V(k_0)^2\right\}.$$

We will find the limiting process of Λ_T on $|v| \le M$ for every given $M > 0$. Let $C[-M, M]$ be the space of continuous functions on $[-M, M]$ with the uniform metric.

Theorem 1. *Under assumptions (A), (B), and (C), then for every $M < \infty$, $\Lambda_T(v)$ converges weakly in $C[-M, M]$ to*

$$\Lambda(v) = 2\left(a(1)\sigma W(v) - \frac{1}{2}|v|\right) \tag{23}$$

and

$$T\lambda_T^2(\hat{\tau} - \tau) \xrightarrow{d} a(1)^2\sigma^2 \text{argmax}_v \left(W(v) - \frac{1}{2}|v|\right) \tag{24}$$

where $a(1) = \sum_{j=0}^{\infty} a_j$ and $W(v)$ is a two-sided Brownian motion on R. A two-sided Brownian motion $W(v)$ is defined as $W(v) = W_1(-v)$ for $v < 0$ and $W(v) = W_2(v)$ for $v \geq 0$ where $W_i(v)$ $(i = 1, 2)$ are two independent Brownian motions defined on the non-negative half real line.

As a consequence of Theorem 1, we have

Theorem 2. *When X_t is a stationary ARMA(p,q) process such that*

$$X_t = \rho_1 X_{t-1} + \cdots + \rho_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

then

$$T\lambda_T^2(\hat{\tau} - \tau) \xrightarrow{d} \left(\frac{1 + \theta_1 + \cdots + \theta_q}{1 - \rho_1 - \cdots - \rho_p}\right)^2 \sigma^2 \,\text{argmax}_v \left(W(v) - \frac{1}{2}|v|\right). \tag{25}$$

A few comments are in order. First, if $a(1) = 0$, Theorem 1 implies that $T\lambda_T^2(\hat{\tau} - \tau) \xrightarrow{p} 0$, a degenerate limiting distribution. Thus the convergence rate will be faster than $T\lambda_T^2$. A simple example of $a(1) = 0$ is given by the first order moving average process $X_t = \varepsilon_t - \varepsilon_{t-1}$. Second, in Theorem 2, when $\theta_1 = \cdots \theta_q = 0$ and $\rho_1 = \cdots = \rho_p = 0$, the right hand side of (25) gives the limiting distribution corresponding to the case of i.i.d. normal variables with a shift in mean and the result is consistent with that of Bhttacharya (1987) and Yao (1987). Third, (25) agrees with the result of Picard (1985) for a pure autoregressive process ($\theta_1 = \cdots \theta_q = 0$) under normality assumption with maximum likelihood estimation. Thus it is reasonable to expect that (25) is also the limiting distribution of the maximum likelihood estimator for ARMA model under normality. Fourth, comparing with an i.i.d. sequence with a shift in mean, we find that the effect of serial correlation in X_t can be beneficial or detrimental to the precision of the change point estimates depending on whether

$$\left|\frac{1+\theta_1+\cdots+\theta_q}{1-\rho_1-\cdots-\rho_p}\right|<1.$$

For example, in the ARMA(1,1) case (stationary and invertible), a negative ρ_1 and a negative θ_1 should help us in locating the change point. On the other hand, when ρ_1 is near the (positive) unit root, the change point estimator has a large variance since $(1-\rho)^{-1}$ can be very large. These qualitative results are all confirmed by Monte Carlo simulations.

Proof of Theorem 1. We only consider the case of $v \leq 0$ because of symmetry. Define the set $K_T(M) = \{k;\ k \text{ is the integer part of } k_0 + v\lambda_T^{-2} \text{ for all } |v| \leq M\}$. Notice

$$\begin{aligned} T(V_k^2 - V_{k_0}^2) &= 2TV_{k_0}(V_k - V_{k_0}) + T(V_k - V_{k_0})^2 \\ &= 2T\,(EV_{k_0})(V_k - V_{k_0}) + 2T(V_{k_0} - EV_{k_0})(V_k - V_{k_0}) + T(V_k - V_{k_0})^2. \end{aligned} \tag{26}$$

Let us first prove that the last two terms on the right are negligible on $K_T(M)$. Since $T^{1/2}(V_{k_0} - EV_{k_0})$ is stochastically bounded due to (15), it is enough to show that $T^{1/2}(V_k - V_{k_0})$ is negligible on $K_T(M)$. Because,

$$\begin{aligned} T^{1/2}|V_k - V_{k_0}| &\leq T^{1/2}|V_k - EV_k - V_{k_0} + EV_{k_0}| + T^{1/2}|EV_k - EV_{k_0}| \\ &= T^{1/2}|G(k) + H(k)| + T^{1/2}|EV_k - EV_{k_0}|, \end{aligned} \tag{27}$$

where $G(k)$ and $H(k)$ are defined in (19), it suffices to show that each of the two terms on the right converges to zero uniformly on $K_T(M)$. It is clear that when $k \in K_T(M)$, there exists a $\delta > 0$ such that $k \geq T\delta$. Thus the upper bound for $G(k)$ given in (21) is valid. This bound consists of three terms. Consider the first term of (21) multiplied by $T^{1/2}$,

$$T^{1/2}\frac{k_0-k}{T\delta k_0}\left|\sum_{t=1}^{k_0} X_t\right| \leq \frac{M}{\delta\tau T\lambda_T^2}\left|\frac{1}{T^{1/2}}\sum_{t=1}^{k_0} X_t\right| = \frac{1}{T\lambda_T^2}O_p(1) = o_p(1) \tag{28}$$

uniformly for $k \in K_T(M)$. The second term of (21) can be treated similarly. Consider the third term of (21) multiplied by $T^{1/2}$,

$$T^{1/2}\frac{1}{T\delta}\left|\sum_{t=k+1}^{k_0} X_t\right| = \frac{1}{T^{1/2}\lambda_T}\lambda_T\left|\sum_{t=k+1}^{k_0} X_t\right| \tag{29}$$

Because $\lambda_T\left|\sum_{t=k+1}^{k_0} X_t\right| = O_p(1)$ uniformly on $K_T(M)$ by the invariance principle (the number of elements in $K_T(M)$ is no larger than $2M\lambda_T^{-2}$), (29) converges to zero in probability uniformly. Similarly, we can show that $T^{1/2}H(k)$ is negligible on $K_T(M)$. Next consider the second term of (27). Notice that

$$\begin{aligned} 0 \le T^{1/2}(EV_{k_0} - EV_k) &= T^{1/2}\left(b(k_0)\lambda_T - b(k)\frac{T-k_0}{T-k}\lambda_T\right) \\ &\le T^{1/2}\{b(k_0) - b(k)\}\lambda_T + T^{1/2}b(k)\frac{|k_0-k|}{T-k}\lambda_T. \end{aligned}$$

From (18) and $k \in K_T(M)$, we can easily show that $T^{1/2}(EV_{k_0} - EV_k)$ is bounded by $C(T^{1/2}\lambda_T)^{-1}$ for some $C > 0$, which converges to zero. We now prove that for $k = [k_0 + v\lambda_T^{-2}]$

$$2\,T\,(EV_{k_0})(V_k - V_{k_0}) = 2\,\{\tau(1-\tau)\}^{1/2}T\,\lambda_T\{V(k_0 + v\lambda_T^{-2}) - V(k_0)\} \tag{30}$$

has the stated limiting distribution. For the sake of simplicity, we shall assume that $k_0 + v\lambda_T^2$ and thus $v\lambda_T^2$ are integers. By (19) we have

$$T\lambda_T(V_k - V_{k_0}) = T\lambda_T\{G(k) + H(k)\} - T\lambda_T(EV_{k_0} - EV_k). \tag{31}$$

As in proving (28), we can easily show that the first two terms in (20) multiplied by $T\lambda_T$ are $o_p(1)$ uniformly on $K_T(M)$. However the product of $T\lambda_T$ and the third term of (20) does not vanish as T increases and can be written as:

$$T\lambda_T\,b(k_0)\frac{1}{k}\sum_{t=k+1}^{k_0} X_t = b(k_0)\frac{T}{k}\left(\lambda_T\sum_{t=0}^{|v|\lambda_T^{-2}} X_{k_0-t}\right)$$

for $k = k_0 + v\lambda_T^{-2}$. By the invariance principle of (9), $(\lambda_T\sum_{t=0}^{|v|\lambda_T^{-2}} X_{k_0-t})$ converges weakly in $C[-M, M]$ to $a(1)\sigma W_1(-v)$, where $W_1(\cdot)$ is a Brownian motion on $[0, \infty)$.

Moreover, $b(k_0) = \{\tau(1-\tau)\}^{1/2}$ and $T/k \to \tau^{-1}$ for $k \in K_T(M)$, therefore $T\lambda_T G(k_0 + v\lambda_T^{-2})$ converges weakly to

$$\tau^{-1}\{\tau(1-\tau)\}^{1/2}a(1)\sigma W_1(-v).$$

Similarly, we can prove that

$$T\lambda_T H(k_0 + v\lambda_T^{-2}) = o_p(1) + b(k_0)\frac{T}{T-k}\left(\lambda_T \sum_{t=0}^{|v|\lambda_T^{-2}} X_{k_0-t}\right).$$

As $b(k_0)T/(T-k) \to (1-\tau)^{-1}\{\tau(1-\tau)\}^{1/2}$, we have

$$\begin{aligned} T\lambda_T\{G(k_0+v\lambda_T^{-2}) + H(k_0+v\lambda_T^{-2})\} &\Rightarrow (\tau^{-1} + (1-\tau)^{-1})\{\tau(1-\tau)\}^{1/2}a(1)\sigma W_1(-v) \\ &= \{\tau(1-\tau)\}^{-1/2}a(1)\sigma W_1(-v). \qquad (32) \end{aligned}$$

From (14),

$$\begin{aligned} T\lambda_T(EV_{k_0} - EV_k) &= T\lambda_T^2\frac{\tau-d}{1-d}\left[\left(\frac{\tau}{1-\tau}\right)^{1/2} + \left(\frac{d}{1-d}\right)^{1/2}\right]^{-1} \\ &\to |v|\{\tau(1-\tau)\}^{-1/2}/2 \qquad (33) \end{aligned}$$

since $d = k/T \to \tau$ uniformly on $K_T(M)$ and $T\lambda_T^2(\tau - d) = -v = |v|$. Combining (30) to (33), we obtain for $v \le 0$,

$$\Lambda_T(v) \Rightarrow 2\left(a(1)\sigma W_1(-v) - \frac{1}{2}|v|\right).$$

Similarly, working with the case for $v > 0$, we find that

$$\Lambda_T(v) \Rightarrow 2\left(a(1)\sigma W_2(v) - \frac{1}{2}|v|\right),$$

where $a(1)\sigma W_2(v)$ is the limiting process of $\lambda_T \sum_{t=k_0+1}^{k} X_t = \lambda_T \sum_{t=1}^{v\lambda_T^2} X_{k_0+t}$. It remains to show that $W_1(\cdot)$ and $W_2(\cdot)$ are independent. From (A.1) in the appendix,

$$\lambda_T \sum_{t=k+1}^{k_0} X_t = \lambda_T \sum_{t=k+1}^{k_0} a(1)\varepsilon_t - \lambda_T X_{k_0}^* + \lambda_T X_k^*$$

but $\sup_{k \in K_T(M)} \lambda_T |X_k^*| = o_p(1)$. This is because the number of elements in $K_T(M)$ is not larger than $2M\lambda_T^{-2}$ and $\lambda_T \to 0$ (see Chung 1968, p. 95). Thus $W_1(\cdot)$ is determined by ε_t with $t \leq k_0$. Similar arguments show that $W_2(\cdot)$ is determined by ε_t with $t > k_0$. Thus W_1 and W_2 are determined by non-overlapping sequences of $\{\varepsilon_t\}$ and hence W_1 and W_2 are independent. This finishes the proof of (23). To prove (24), define $C_{\max}[-M, M]$ to be the subset of $C[-M, M]$ such that each function has a unique maximum. It is straightforward to show that the *argmax* functional is continuous on the set $C_{\max}[-M, M]$. This together with (22) permits us to invoke the continuous mapping theorem, which implies

$$T\lambda_T^2(\hat{\tau} - \tau) \xrightarrow{d} \text{argmax}_v \Lambda(v).$$

For a rigorous treatment of the continuous mapping theorem for argmax functionals, see Kim and Pollard (1990). Since $bW(v) \stackrel{d}{=} W(b^2 v)$ for every $b \in R$, a change in variable leads to $\text{argmax}_v \Lambda(v) \stackrel{d}{=} a(1)^2\sigma^2 \text{argmax}_s (W(s) - |s|/2)$, which is (24). The proof of Theorem 1 is now complete.

Given the rate of convergence of $\hat{\tau}$ in Section 4, it is an easy matter to obtain the limiting distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$. Recall that $\hat{\mu}_1 = \bar{Y}_{\hat{k}}$ and $\hat{\mu}_2 = \bar{Y}_{\hat{k}}^*$.

Proposition 1.

$$T^{1/2}(\hat{\mu}_1 - \mu_1) \xrightarrow{d} N(0, \tau^{-1} a(1)^2 \sigma^2), \tag{34}$$

$$T^{1/2}(\hat{\mu}_2 - \mu_2) \xrightarrow{d} N(0, (1-\tau)^{-1} a(1)^2 \sigma^2). \tag{35}$$

Thus the limiting distributions are the same as if k_0 is known. The proof is given in the appendix.

6. Monte Carlo simulation

In this section, we assess through Monte Carlo simulations some qualitative aspects of the change point estimator predicted by the theory. We also examine the effect of a mean shift on the order identification of an ARMA process via the AIC criterion. The basic conclusion is that a shift in mean causes over-estimation of orders of the model. This finding is consistent with that of MacNeill and Duong (1982) who considered a shift in the AR coefficients rather than in the mean of an AR process. Since we can obtain a consistent estimator of the change point, we can remove to some extent the shifted mean and then use the generalized residuals $\hat{X}_t$ in place of X_t to identify the orders. The simulation results suggest that this is quite a satisfactory procedure.

6.1. Estimation of the change point

In this simulation, the series is generated according to

$$Y_t = \mu + \lambda I(t \geq k_0) + X_t \qquad (t = 1, 2, \cdots, T)$$

where X_t is an ARMA(1,1) process $X_t = \rho X_{t-1} + \varepsilon_t + \theta\varepsilon_t$ and $I(\cdot)$ is the indicator function. Experiments are carried out for $T = 100$, $k_0 = .5T$, $\lambda = 2$, $\rho = -.8, -.6, -.4, 0.0, .4, .6, .8$, and $\theta = -.5, .5$. The ε_t are i.i.d. $N(0,1)$. The least squares estimator of k_0 is defined by (3). The change point is estimated by maximizing V_k^2 defined in (4).

Theorem 2 indicates that the asymptotic variance of the estimated change point is smaller for a negative ρ than a positive ρ even if the magnitude of change λ is the same. A similar conclusion applies to the sign of θ. More precisely, the asymptotic variance is a decreasing function of $|(1+\theta)/(1-\rho)|$. Thus the range of the change point estimates in the simulation is expected to be smaller for a negative value of ρ and a negative value of θ than any positive values. Indeed these theoretical results are all confirmed by simulations. Figure 1 plots the histograms of the estimated change points for $\theta = .5$ and for different values of ρ. It is seen that as ρ varies from -.8 to .8, the range of $\hat{k}$ is becoming larger. The case for $\theta = -.5$ is graphed in Figure 2. By examining the seven sub-figures in Figure 2, one finds that the range of the estimated change points is smaller than each of its counterparts in Figure 1, as predicted by Theorem 2.

6.2. Effect of a mean shift on order estimation

ARMA models have been proved to be a powerful tool in time series analysis. In practice, however, the order of a series is unknown and has to be estimated. In fact, order determination often turns out to be the most challenging part in time series modeling. Akaike (1974) identifies the order by the AIC criterion which, for an autoregressive process, is defined as:

$$AIC(k) = T \log \hat{\sigma}^2(k) + 2k$$

where $\hat{\sigma}^2(k)$ is the residual variance when an AR model of order k is fitted and T is the sample size. The AIC criterion is an automatic procedure and is found in many computing packages. We thus investigate the effect of a mean shift on the order selection via the AIC procedure. The findings presumably shed light on other order selection criteria as well. It is interesting to compare our simulated result with that of MacNeill and Duong (1982) who considered a shift in AR coefficients rather than in mean. The mean shift seems to have more adverse effect on order estimation.

The model employed in the simulation is the following AR(1) with a mean shift:

$$\begin{aligned} Y_t &= \mu + \lambda I(t \geq k_0) + X_t \\ X_t &= \rho X_{t-1} + \varepsilon_t \qquad (t = 1, 2, \cdots, T) \end{aligned}$$

Simulations have been performed with $T = 50, 100, 500$, $\rho = -.8, -.6, -.4, 0, .4, .6, .8$ and $\lambda = 1, 2$ in various combinations. The change point k_0 is chosen to be $T/2$. For each combination, 100 series are generated. The white noise ε_t are normally distributed with zero mean and unit variance. For comparison purposes, the data are generated in such a way that across different combinations of parameters, the white noise series is the same. In searching for an order, the upper bound is limited to 10. The autoregressive parameters are estimated using the Yule-Walker method. Note that parameter μ plays no role because the overall mean is subtracted from the series before fitting a model.

It is expected that a change in parameter, when ignored, introduces distortions in order identification, consequently rendering incorrect order selection. In other words, misspecification causes bias in order estimation. This is indeed the case and is verified by Monte Carlo simulations. Table 1 summarizes the simulation results in terms of the mean and standard deviation of the selected orders. Means and standard deviations are computed based on 100 repetitions for each combination. Standard deviations are provided in parentheses. The rows for $\lambda = 0$ correspond to models with no shift. By inspecting the table, one concludes that a change in mean causes over estimation of orders. The over estimation is particularly significant for negative $\rho's$. For positive and large $\rho's$, order selection is virtually unaffected by a shift. For example, for $\rho = -.8$, the average value of selected orders is 1.56 with no shift ($\lambda = 0$) and 3.56 with a shift ($\lambda = 1$); whereas for $\rho = .8$, the corresponding values are 1.55 and 1.54 respectively. The bias in order selection becomes larger as the number of observations (T) increases. Consider the column for $\rho = -.8$, for instance, when the number of observations varies from 50 to 100 then to 500, the average selected orders are 3.56, 5.23, and 9.38 respectively ($\lambda = 1$). To see how the selected orders are distributed, We plot histograms of the selected orders for two cases. Figure 3 plots the selected orders when no shift exists ($\lambda = 0$). Figure 4 plots the counterpart when a shift does exist ($\lambda = 1$).

6.3. Order estimation based on generalized residuals

In the previous subsection, the shift is ignored and the order determination is based on a misspecified model which leads to over estimation. Thus if a high order is identified in practice, one should consider a possible parameter change. Tests should be performed to test the parameter constancy. Many tests have been proposed in the literature for testing parameter stability for time series models, for example, Andrews (1990), Bai, Lumsdaine, and Stock (1991), Hansen (1990), MacNeill and Duong (1982), and Ploberger, Kramer, and Alt (1989).

For our present model, order identification could be based on generalized residuals

Table 1: Mean and Standard Deviation of Selected AR Orders

	$\rho =$	-0.8	-0.6	-0.4	0	0.4	0.6	0.8
$T = 50$	$\lambda = 0$	1.56	1.48	1.55	0.61	1.44	1.66	1.55
		(1.34)	(1.16)	(1.45)	(1.41)	(1.49)	(1.44)	(1.21)
	$\lambda = 1$	3.56	3.21	2.64	1.57	1.55	1.45	1.54
		(1.48)	(1.81)	(2.01)	(1.84)	(1.35)	(1.04)	(1.27)
$T = 100$	$\lambda = 0$	1.56	1.59	1.63	0.72	1.75	1.73	1.64
		(1.44)	(1.51)	(1.47)	(1.79)	(1.59)	(1.63)	(1.56)
	$\lambda = 1$	5.23	5.01	4.72	3.19	2.05	1.74	1.71
		(1.54)	(1.71)	(1.84)	(2.12)	(1.81)	(1.61)	(1.67)
	$\lambda = 2$	4.61	4.74	4.86	4.59	2.89	2.01	1.61
		(1.11)	(1.38)	(1.51)	(1.78)	(1.92)	(1.76)	(1.57)
$T = 500$	$\lambda = 0$	1.63	1.85	2.01	1.13	1.77	1.94	1.96
		(1.42)	(1.75)	(1.89)	(2.14)	(1.57)	(1.82)	(1.85)
	$\lambda = 1$	9.38	9.33	9.27	8.81	5.81	2.88	2.08
		(0.75)	(0.87)	(0.94)	(1.45)	(2.88)	(2.73)	(1.99)

once a change point is estimated. Recall that the generalized residuals are defined as

$$\hat{X}_t = Y_t - \hat{\mu}_1 - (\hat{\mu}_2 - \hat{\mu}_1)I(t > \hat{k}).$$

Because μ_1, μ_2, and τ can be consistently estimated we expect that order determination using the generalized residuals will remove, to some extent, the bias caused by a shift. Figure 5 displays the selected orders based on the generalized residuals $\hat{X}_t$. Comparing with Figure 3, which is based on X_t, we find that the results are almost the same.

Figure 1: Histogram of Estimated Change Points. Data is Generated According to: $Y_t = \mu + \lambda I(t \geq k_0) + X_t$, with $X_t = \rho X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$, $\lambda = 2$, $\theta = .5$, $T = 100$, and $k_0 = .5T$ (From 100 Repetitions).

Figure 2: Histogram of Estimated Change Points. Data is Generated According to: $Y_t = \mu + \lambda I(t \geq k_0) + X_t$, with $X_t = \rho X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$, $\lambda = 2$, $\theta = -.5$, $T = 100$, and $k_0 = .5T$ (From 100 Repetitions).

Figure 3: Selected Orders Via AIC Criterion, the Case of no Shift. Data is Generated According to: $X_t = \rho X_{t-1} + \varepsilon_t$, $T = 100$ (100 Repetitions).

Figure 4: Selected Orders Via AIC Criterion When a shift is Ignored. Data is Generated According to: $Y_t = \mu + \lambda I(t \geq k_0) + X_t$, with $X_t = \rho X_{t-1} + \varepsilon_t$, $\lambda = 1$, $T = 100$, and $k_0 = .5T$ (100 Repetitions).

Figure 5: Selected Orders Via AIC Criterion Using Generalized Residuals $\hat{X}_t$. Data is Generated According to: $Y_t = \mu + \lambda I(t \geq k_0) + X_t$, with $X_t = \rho X_{t-1} + \varepsilon_t$, $\lambda = 1$, $T = 100$, and $k_0 = .5T$ (100 Repetitions).

A Appendix

Proof of (6). Write $a_j^* = \sum_{k\ge j+1} a_k$ and $X_t^* = \sum_{j=0}^{\infty} a_j^* \varepsilon_{t-j}$. Under assumptions (A or A′) and (B), X_t^* is second-order stationary. Let $\sigma_1^2 = E(X_t^*)^2 = \sigma^2 \sum_{j=0}^{\infty}(a_j^*)^2$. Then we have

$$X_t = a(1)\varepsilon_t - X_t^* + X_{t-1}^*, \tag{A.1}$$

thus

$$c_k \sum_{i=1}^{k} X_i = a(1)c_k \sum_{i=1}^{k} \varepsilon_i + c_k X_0^* - c_k X_k^*,$$

where $a(1) = \sum_{j=0}^{\infty} a_j$. Since $c_k \le c_m$ for $k \ge m$ by assumption, we have

$$\begin{aligned}
&Pr\left(\max_{m\le k\le n} c_k|\sum_{i=1}^{k} X_i| > \alpha\right) \le Pr\left(\max_{m\le k\le n} c_k|a(1)\sum_{i=1}^{k}\varepsilon_i| > \alpha/3\right) \\
&+ Pr\left(c_m|X_0^*| > \alpha/3\right) \;+\; Pr\left(\max_{m\le k\le n} c_k|X_k^*| > \alpha/3\right) \\
&\le \frac{9\sigma^2 a(1)^2}{\alpha^2}\left(mc_m^2 + \sum_{i=m+1}^{n} c_i^2\right) \qquad \text{by Hájek and Rényi inequality} \\
&+ \frac{9\sigma_1^2}{\alpha^2}\left(c_m^2 + \sum_{i=m}^{n} c_i^2\right) \qquad \text{by Chebyshev inequality.}
\end{aligned}$$

Hence if we choose $A = 9\max\{a(1)^2, \sum_{j=0}^{\infty}(a_j^*)^2\}$, then inequality (6) follows. Equation (A.1) is called the Beveridge-Nelson decomposition in the econometrics literature. Phillips and Solo (1991) provide many interesting applications of this decomposition.

Proof of Proposition 1. Write $\hat\mu_1(k) = \frac{1}{k}\sum_{t=1}^{k} Y_t$, then $\hat\mu_1 = \hat\mu_1(\hat k)$. If k_0 is known, the LS estimation for μ_1 is $\hat\mu_1(k_0)$. To prove (34), consider

$$\begin{aligned}
&T^{1/2}\left(\hat\mu_1(\hat k) - \hat\mu_1(k_0)\right) = T^{1/2}\left(\frac{1}{\hat k}\sum_{t=1}^{\hat k} Y_t - \frac{1}{k_0}\sum_{t=1}^{k_0} Y_t\right) \\
&= I(\hat k \le k_0)\left(T^{1/2}\frac{k_0 - \hat k}{k_0 \hat k}\sum_{t=1}^{k_0} X_t + T^{1/2}\frac{1}{\hat k}\sum_{t=\hat k}^{k_0} X_t\right) \\
&+ I(\hat k > k_0)\left(T^{1/2}\frac{k_0 - \hat k}{k_0 \hat k}\sum_{t=1}^{k_0} X_t - T^{1/2}\frac{1}{\hat k}\sum_{t=k_0}^{\hat k} X_t + \lambda_T\frac{\hat k - k_0}{\hat k}\right)
\end{aligned}$$

Now use $k_0 = T\tau$, $\hat k = k_0 + O_p(\lambda_T^{-2})$, and $T\lambda_T^2 \to \infty$, we find that the above is $(T^{1/2}\lambda_T)^{-1}O_p(1)$, which converges to zero in probability. Thus $\hat\mu_1(\hat k)$ and $\hat\mu_1(k_0)$ have the same limiting distribution. The latter has a limiting distribution given by the right hand side of (34). The proof of (35) is similar.

Acknowledgments. This paper is based on Chapter 4 of my dissertation at the University of California at Berkeley. I am grateful to my advisor, Professor Thomas Rothenberg, for his encouragement and guidance throughout my study at Berkeley. I am indebted to Professors Deborah Nolan and James Stock for their advice. I thank professor Peter Bickel for invaluable discussions. Financial support from the Alfred P. Sloan Foundation Dissertation Fellowship is gratefully acknowledged.

References

[1] Akaike, H. (1974) A new look at the statistical model identification. *I.E.E.E. Trans. Auto. Contr.* Ac-19, 716-723.

[2] Andrews, D. W. K. (1990) Tests for parameter instability and structural change with unknown change point. Manuscript, Cowles Foundation, Yale University.

[3] Bai, J., Lumsdaine, R. L., and Stock, J. H. (1991) Testing for and dating breaks in integrated and cointegrated time series. Manuscript, Kennedy School of Government, Harvard University.

[4] Bhattacharya, P. K. (1987) Maximum likehood estimation of a change-point in the distribution of independent random variables: General Multiparameter case. *J. of Multivariate Analysis* 23, 183-208.

[5] Birnbaum, Z. W. and Marshall, A. W. (1961) Some multivariate Chebyshev inequalities with extensions to continuous parameter processes. *Ann. Math. Stat.* 32, 687-703.

[6] Chung, K. L. (1968) *A Course in Probability Theory.* Harcourt, Brace & World Inc.

[7] Hájek, J. and Rényi, A. (1955) Generalization of an inequality of Kolmogorov. *Acta. Math. Acad. Sci.* Hungar. 6, 281-283.

[8] Hall, P. and Heyde C. C. (1980) Martingale Limit Theory and its Applications. New York: Academic Press.

[9] Hansen, B.E. (1990). Testing for structural change of unknown form in models with non-stationary regressors. manuscript, University of Rochester.

[10] Hawkins, D. L. (1986) A simple least square method for estimating a change in mean. *Commun. Statist.-Simul.* 15, 655-679.

[11] Hawkins, D. M. (1977) Testing a sequence of observations for a shift in location. *J. of American Statistical Asso.* 72, 180-186.

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

[13] James, B., James, K., and Siegmund, D. (1987) Tests for a change point. *Biometrika* 74, 71-83.

[14] Kim, J. and Pollard, D. (1990) Cube root asymptotics. *Annals of Statistics* 18, 191-219.

[15] MacNeill, I. B. and Duong, Q. P. (1982) Effect on AIC-type criteria of change of parameters at unknown time. In *Time Series Analysis: Theory and Practice 1*, edited by O.D. Anderson. North-Holland Publishing Company.

[16] Picard, D. (1985) Testing and estimating change-points in time series. *J. Applied Prob.* 14, 411-415.

[17] Phillips, P. C. B. and Solo, V. (1989) Asymptotics for linear processes. Cowles Foundation Discussion Paper, No 932, Yale University.

[18] Ploberger, W. K., Kramer W., and Alt R. (1989) A modification of the CUSUM test in the linear regression model with lagged dependent variables. In W. Kramer (ed.) *Econometrics of Structural Change*. Physica-Verlag, Heidelberg.

[19] Sen, A. and Srivastava, M. S. (1975a) On tests for detecting change in mean. *Ann. stat.* 3, 96-103.

[20] ——— (1975b) Some one-sided tests for change in level. *Technometrics* 17, 61-64.

[21] Srivastava, M. S. and Worsley, K.J. (1986) Likelihood ratio tests for a change in the multivariate normal mean. *J. Amer. Stat. Asso.* 81, 199-204.

[22] Worsley, K. J. (1979) On the likehood ratio test for a shift in locations of normal populations. *J. of the American Statistical Association* 72, 180-186.

[23] ——— (1986) Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* 73, 91-104.

[24] Yao, Y. C. (1987) Approximating the distribution of the ML estimate of the change-point in a sequence of independent r.v.'s. *Annals of Statistics* 3, 1321-1328.