ANNALS OF ECONOMICS AND FINANCE **1**, 303–339 (2000)

Vector Autoregressive Models with Structural Changes in Regression Coefficients and in Variance-Covariance Matrices *

Jushan Bai

Department of Economics,
Boston College,
Chestnut Hill, MA 02467.
E-mail: jushan.bai@bc.edu

This paper analyzes vector autoregressive models (VAR) with multiple structural changes. One distinct feature of this paper is the explicit consideration of structural changes in the variance-covariance matrix, in addition to changes in the autoregressive coefficients. The model is estimated by the quasi-maximum likelihood method. It is shown that shifts in the covariance matrix help identify the change points. We obtain consistency, rate of convergence, and limiting distributions for the estimated change points and the estimated regression coefficients and variance-covariance matrix. We also show that the number of change points can be consistently estimated via the information criterion approach. The paper provides tools for constructing confidence intervals for change points in multiple time series. The result is also useful for analyzing volatility changes in economic time series.

Key Words: Structural change; Multiple change points; QMLE; VAR; BIC.
JEL Classification Numbers: C12; C22; C13; C52.

1. INTRODUCTION

The concept of a structural change (change in policy regimes, change in underlying structural relations in the economy, or change in particular reduced-form relationships) has widespread use in economics. The shifts of the Phillips curve over time serve as one illustration [Alogoskoufis and Smith (1991)]. As a result, structural changes have always been an important concern in econometric modeling. Earlier studies on this topic include Chow (1960) and Quandt (1960). More recent studies include, among others, Andrews (1993), Bai and Perron (1998), and Stock (1994).

* Financial support from the National Science Foundation under grants SBR-9414083 and SBR-9709508 is gratefully acknowledged.

1529-7373/2000

Correctly detecting and identifying a structural change can have profound effect on policy evaluation and recommendation. In their examination of the feasibility of using a monetary aggregate to influence the path of GDP, Feldstein and Stock (1993) found, using the structural-change techniques, that the relationship between M1 and GDP is highly unstable, while the relationship between M2 and GDP is relatively stable, suggesting the possibility of an active rule for modifying M2 growth to reduce the volatility of nominal GDP growth. When studying temporal variation in the interest-rate response to money-announcement surprises, Roley and Wheatley (1990), by estimating the shift points, were able to determine whether the response was immediate (as different policy regimes were adopted), or the response was gradual (reflecting the establishment of Federal Reserve credibility).

The purpose of this paper is to study the problem of structural changes in vector autoregressive models (VAR) with unknown change points. Structural changes are permitted in variance-covariance matrices as well as in regression coefficients. Our analysis focuses on the statistical properties of estimated parameters. In particular, we analyze the asymptotic behavior of break-point estimators. Consistency, rate of convergence, and limiting distributions of the estimated parameters and the estimated break points are established.

Structural change in VAR is an important problem for at least two reasons. The first is related to VAR's popularity as a modeling tool in macroeconomics. The second has to do with the finding of Bai, Lumsdaine, and Stock (1998). These authors show that the accuracy of the break-point estimators cannot be improved upon by acquiring longer data, but can be significantly improved upon by adding series that have common breaks. That is, analyzing multiple equations with common break points can yield more precise estimates of the change points. As a result, the structural change problem in VAR is of theoretical and practical importance. The results derived in this paper are, of course, also applicable for univariate time series. They are also applicable for cross-section regression models as well as for seemingly unrelated regressions.

This paper is closely related to Bai, Lumsdaine, and Stock (1998). However, the latter paper considers a single change point instead of multiple ones. Because a myriad of political and economic factors may alter the data generating process, multiple changes may be a more accurate characterization of economic time series. In addition, Bai, Lumsdaine, and Stock (1998) only consider changes in regression coefficients. In this paper, we allow for changes in variance-covariance matrices. The results of this paper can be useful in modeling changes in volatilities of economic time series.

2. MODEL AND ESTIMATION

This paper analyzes the following vector autoregressive (VAR) model with m regimes:

$$Y_t = \begin{cases} \mu_1 + A_{1,1}Y_{t-1} + A_{1,2}Y_{t-2} + \cdots + A_{1,p}Y_{t-p} + (\Sigma_1^0)^{1/2}\eta_t \\ \qquad\qquad (t = 1, ..., , k_1^0) \\ \mu_2 + A_{2,1}Y_{t-1} + A_{2,2}Y_{t-2} + \cdots + A_{2,p}Y_{t-p} + (\Sigma_2^0)^{1/2}\eta_t \\ \qquad\qquad (t = k_1^0 + 1, ..., k_2^0) \\ \qquad\qquad \vdots \\ \mu_{m+1} + A_{m+1,1}Y_{t-1} + A_{m+1,2}Y_{t-2} + \cdots + A_{m+1,p}Y_{t-p} + (\Sigma_{m+1}^0)^{1/2}\eta_t \\ \qquad\qquad (t = k_m^0 + 1, ..., T) \end{cases} \tag{1}$$

where Y_t and η_t are $r \times 1$, $A_{i,j}$ is $r \times r$. The disturbances η_t are unobservable random variables with $E\eta_t = 0$ and $Var(\eta_t) = I$, an identity matrix. The coefficients $A_{i,j}$, variance-covariance matrices Σ_i^0, and the change points k_i^0 are all unknown.

We assume at least one coefficient either in the autoregressive part or in the variance-covariance matrix has undergone a shift. That is, we assume either $(\mu_i, A_{i,1}, ..., A_{i,p}) \neq (\mu_{i+1}, A_{i+1,1}, ..., A_{i+1,p})$, or $\Sigma_i^0 \neq \Sigma_{i+1}^0$, or both $(i = 1, ..., m + 1)$. Thus the variance-covariance matrix is allowed to be different across regimes.

Although the literature on change point is large, most studies focus on a single shift in univariate series, see Picard (1985), where variance is not allowed to change. Ng and Vogelsang (1997) study structural changes in VAR models with shifts in the intercept only. In this paper, we consider multiple change points in multiple series.

The analysis of multiple changes is much more demanding than that of a single change. When studying a single change (two regimes), the regime boundaries are partially known. For example, the lower boundary for the first regime is known (the first observation) and the upper boundary for the second regime is also known (the last observation). In the case of multiple changes, however, a hypothesized regime may not overlap with the underlying true regime. This constitutes the source of difficulty and complexity associated with multiple breaks. Nevertheless, based on the earlier work of Bai and Perron (1998) and Bai, Lumsdaine and Stock (1998), we can tackle the problem in an elegant way.

Let $V_t' = (1, Y_{t-1}', ..., Y_{t-p}')$ and $\theta_i^0 = \text{Vec}(\mu_i, A_{i,1}, ..., A_{i,p})$. Then we can rewrite (1) as

$$Y_t = (V_t' \otimes I)\theta_i^0 + (\Sigma_i^0)^{1/2}\eta_t, \qquad t = k_{i-1}^0 + 1, ..., k_i^0 \tag{2}$$

for $i = 1, 2, ..., m$ with $k_0^0 = 0$ and $k_{m+1}^0 = T$.

Let $\beta = (\theta_1, ..., \theta_{m+1})$ and $\Gamma = (\Sigma_1, ..., \Sigma_{m+1})$. The true regression coefficients are denoted by $\beta^0 = (\theta_1^0, ..., \theta_{m+1}^0)$ and $\Gamma^0 = (\Sigma_1^0, ..., \Sigma_{m+1}^0)$. We shall consider quasi-Gaussian likelihood estimation based on observations $Y_{-p+1}, Y_{-p+2}, ..., Y_0, Y_1, ..., Y_T$.

Quasi-likelihood and likelihood ratio.

Define the quasi-likelihood for the whole sample $Y_1, Y_2, ..., Y_T$ conditional on $Y_{-p+1}, ..., Y_0$ as

$$\prod_{i=1}^{m+1} \prod_{t=k_{i-1}+1}^{k_i} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i, \Sigma_i)$$

where

$$f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i, \Sigma_i) = \frac{1}{(2\pi \det(\Sigma_i))^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2}[Y_t - (V_t' \otimes I)\theta_i]' \Sigma_i^{-1} [Y_t - (V_t' \otimes I)\theta_i] \Big\}$$

The parameters $(\beta, \Gamma, k_1, ..., k_m)$ are estimated by maximizing the quasi-likelihood function. Bai and Perron (1998, 1999) show that the optimization can be done with no more than $O(T^2)$ maximum likelihood estimations no matter how large the number of change points m is.

The theoretical analysis is performed using likelihood ratios defined below. Let $(\varepsilon_1, ..., \varepsilon_T) = ((\Sigma_1^0)^{1/2}\eta_1, ..., (\Sigma_{m+1}^0)^{1/2}\eta_T)$ denote the disturbances of model (1). Define the quasi-likelihood ratio for the entire sample as

$$\Lambda_T(k_1, ..., k_m, \beta, \Gamma) = \frac{\prod_{i=1}^{m+1} \prod_{t=k_{i-1}+1}^{k_i} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i, \Sigma_i)}{\prod_{i=1}^{m+1} \prod_{t=k_{i-1}^0+1}^{k_i^0} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i^0, \Sigma_i^0)}$$

$$= \frac{\prod_{i=1}^{m+1} \prod_{t=k_{i-1}+1}^{k_i} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i, \Sigma_i)}{\prod_{t=1}^{T} f(\varepsilon_t)}.$$

where $f(\varepsilon_t)$ is the density of $N(0, \Sigma_i^0)$ for $t \in [k_{i-1}^0 + 1, k_i^0]$ $(i = 1, ..., m)$. Note that no normality is assumed, although the density of multivariate normal is used.

Estimation. For a given set of integers $(k_1, ..., k_m)$, we let

$$\Lambda_T(k_1, ..., k_m) = \sup_{\beta,\Gamma} \Lambda_T(k_1, ..., k_m, \beta, \Gamma).$$

The change-point estimator is defined as the set of integers $(\hat{k}_1, ..., \hat{k}_m)$ which maximize the quasi-likelihood function. Namely,

$$(\hat{k}_1, ..., \hat{k}_m) = \text{argmax}_{k_1,...,k_m} \Lambda_T(k_1, ..., k_m). \tag{3}$$

We also define

$$\hat{\beta} = (\hat{\theta}_1, ..., \hat{\theta}_{m+1}), \quad \text{and} \quad \hat{\Gamma} = (\hat{\Sigma}_1, ..., \hat{\Sigma}_{m+1}) \tag{4}$$

where $(\hat{\theta}_i, \hat{\Sigma}_i)$ is the estimator of (θ_i^0, Σ_i^0) using the subsample $[\hat{k}_i + 1, \hat{k}_{i+1}]$. Thus, $(\hat{\beta}, \hat{\Gamma}, \hat{k}_1, ..., \hat{k}_m)$ is the quasi-Maximum likelihood estimator of $(\beta^0, \Gamma^0, k_1^0, ..., k_m^0)$.

For technical reasons, the supremum with respect to $(k_1, k_2, ..., k_m)$ in (3) is taken over a restricted set of partitions. For a small number $\nu > 0$, define, for $m \geq 2$,

$$K_\nu = \{(k_1, ..., k_m) : k_2 - k_1 \geq T\nu, ..., k_m - k_{m-1} \geq T\nu, 1 < k_i < T\} \tag{5}$$

The maximization of (3) is taken over K_ν. Note that we still allow arbitrary first and last regimes. In particular, when we only have a single break point, the search is taken over the entire set $[1, T]$. When two breaks are allowed, the search is taken over all pairs of (k_1, k_2) such that k_1 and k_2 are νT apart. Typically, νT is specified as a small integer in practice. The restricted search greatly lessens the technical difficulty in theoretical proofs. Since $\nu > 0$ can be arbitrarily small, the hypothesized change points, $k_1, ..., k_m$, are permitted to take on values within a single true regime such that $k_i^0 < k_1 < \cdots < k_m < k_{i+1}^0$.

3. ASSUMPTIONS AND ESTIMATION THEORY

Assumption A1. The $\{\eta_t, \mathcal{F}_t\}$ form a sequence of martingale differences, where $\mathcal{F}_t = \sigma$-field $\{Y_t, Y_{t-1}, ...\}$. That is, $E(\eta_t | \mathcal{F}_{t-1}) = 0$. In addition, $E(\eta_t \eta_t' | \mathcal{F}_{t-1}) = I$, and $\sup_t E(\|\eta_t\|^{4+\delta}) < \infty$.

Assumption A2. $(\theta_i^0, \Sigma_i^0) \neq (\theta_{i+1}^0, \Sigma_{i+1}^0)$. That is, there exists an entry in (θ_i^0, Σ_i^0) that is different from the corresponding entry in $(\theta_{i+1}^0, \Sigma_{i+1}^0)$. Each true regime parameter (θ_i^0, Σ_i^0) corresponds to that of a stationary process so that unit roots and explosive roots are ruled out.

Assumption A3. $k_i^0 = [T\tau_i^0]$ $(i = 1, ..., m)$ with $0 < \tau_1^0 < \tau_2^0 < \cdots \tau_m^0 < 1$.

Under A3, the number of observations in a single regime increases as the sample size increases. This is not the way that data are generated. This assumption allows for asymptotic theoretical analysis. Since in practice, T is finite, the assumption is always satisfied. The problem is whether the asymptotic distributions derived from such an assumption can approximate the finite-sample distributions well. Various simulation studies have shown that the asymptotic distribution delivers a satisfactory approximation. In fact, the precision of change-point estimators is not so much determined by the sample size but rather by the magnitude of shift. The idea is that sample size is not crucial for the behavior of change-point estimators.

Now we define the estimated fractions as $\hat{\tau}_i = \hat{k}_i/T$.

THEOREM 1. *Under A1-A3, we have*

$$T(\hat{\tau}_i - \tau_i^0) = O_p(1) \qquad (i = 1, ..., m),$$

$$\sqrt{T}(\hat{\theta}_i - \theta_i^0) = O_p(1) \qquad (i = 1, ..., m+1),$$

$$\sqrt{T}(\hat{\Sigma}_i - \Sigma_i^0) = O_p(1) \qquad (i = 1, ..., m+1).$$

In terms of the fractions of the sample size, the estimated break points converge rapidly to the true fraction. In terms of the real time index, the result is that $\hat{k}_i = k_i^0 + O_p(1)$. Thus with large probability, the estimated break point deviates from the true one by a finite number of observations no matter how large the sample is.

COROLLARY 1. *Under the assumptions of Theorem 1, the limiting distributions of $\hat{\beta}$ and $\hat{\Gamma}$ are the same as those of known $k_1^0, ..., k_m^0$.*

In the following, we characterize the limit distribution of the estimated break points. We introduce a random process defined on the set of integers.

Let

$$\begin{aligned} W_1^{(i)}(r) \;=\; & -\frac{r}{2}\Big(\log|\Sigma_i^0| - \log|\Sigma_{i+1}^0|\Big) \\ & -\frac{1}{2}\sum_{k_i^0+1}^{k_i^0+r} \varepsilon_t'\Big((\Sigma_i^0)^{-1} - (\Sigma_{i+1}^0)^{-1}\Big)\varepsilon_t \\ & -\Delta\theta_i' \sum_{k_i^0+1}^{k_i^0+r} (V_t \otimes I)(\Sigma_i^0)^{-1}\varepsilon_t - \frac{1}{2}\Delta\theta_i' \sum_{k_i^0+1}^{k_i^0+r} (V_t V_t' \otimes (\Sigma_i^0)^{-1})\Delta\theta_i \end{aligned} \tag{6}$$

$(r = 1, 2, ...)$, where $\Delta\theta_i = \theta_{i+1}^0 - \theta_i^0$. Let

$$W_2^{(i)}(r) = -\frac{r}{2}\Big(\log|\Sigma_{i+1}^0| - \log|\Sigma_i^0|\Big) - \frac{1}{2}\sum_{k_i^0+r}^{k_i^0} \varepsilon_t'\Big((\Sigma_{i+1}^0)^{-1} - (\Sigma_i^0)^{-1}\Big)\varepsilon_t \quad (7)$$
$$+\Delta\theta_i' \sum_{k_i^0+r}^{k_i^0} (V_t \otimes I)(\Sigma_{i+1}^0)^{-1}\varepsilon_t - \frac{1}{2}\Delta\theta_i' \sum_{k_i^0+r}^{k_i^0} (V_t V_t' \otimes (\Sigma_{i+1}^0)^{-1})\Delta\theta_i$$

$(r = -1, -2, ...)$. Finally, let

$$W^{(i)}(r) = \begin{cases} W_1^{(i)}(r) & r = 1, 2, ..., \\ 0 & r = 0 \\ W_2^{(i)}(r) & r = -1, -2, ... \end{cases}$$

The process $W^{(i)}(r)$ only depends on the true parameters of the model. The following theorem characterizes the limiting distribution under fixed magnitude of shift.

THEOREM 2. *Assume that A1-A3 hold and that η_t has a continuous distribution for all t, then*

$$\hat{k}_i - k_i^0 \xrightarrow{d} \text{argmax}_r W^{(i)}(r)$$

Three comments are in order.

1. When there is only a shift in the variance-covariance matrix, the third and forth terms of (6) and (7) vanish, the limiting distribution is determined by the first two terms. When there is only a shift in the autoregressive parameters, the first two terms disappear, yielding similar results as in Bai (1997) and Bai and Perron (1998).

2. The limiting distribution depends on the distribution of ε_t, thus not asymptotically distribution free.

3. It is difficult to derive an analytical solution for the density function of the change-point estimator because it depends on the distribution of ε_t, in view of the second comment. One approach to overcome this difficulty is to consider shrinking shifts. That is, the magnitude of shifts converges to zero as the sample size increases to infinity. Under shrinking shifts, the limiting distribution of the estimated break points has a limiting distribution free from nuisance parameters. In addition, the analytical density function can be derived so that feasible confidence intervals can be constructed. Simulations show that the confidence intervals obtained this way are also applicable for large shifts.

Assumption A4. (shrinking shifts) The magnitude of shift satisfies $(\theta^0_{i+1,T} - \theta^0_{i,T}) = v_T\delta_i$ and $(\Sigma^0_{i+1,T} - \Sigma^0_{i,T}) = v_T\Phi_i$, where $(\delta_i, \Phi_i) \neq 0$, not depending on T. Moreover, v_T is a sequence of positive numbers such that

$$v_T \to 0, \quad \sqrt{T}v_T/(\log T)^2 \to \infty. \tag{8}$$

Finally, $\Sigma^0_{i,T} \to \Sigma_0$ for every i.

Even though the magnitude of shift in variance is shrinking, we do not assume variances themselves become small. Otherwise this amounts to strong assumptions in terms of "information-noise ratio." For example, If we assume the variance is such that $\Sigma^0_{i,T} = v_T\Sigma$, then $\Sigma^0_{i,T} \to 0$ because $v_T \to 0$. This makes the problem less interesting because, in the limit, there is no noise in the system. In addition, if $\theta^0_{i,T} = v_T\theta$, then the result will be identical to a fixed magnitude of shift because the "information-noise ratio" $\|\Sigma^0_{i,T}\|/\|\theta^0_{i,T}\| = \|\Sigma\|/\|\theta\|$ is a constant. We thus assume that the variance of each regime converges to a common positive definite matrix, while the magnitude of shift (the difference of two consecutive regimes) in variance converges to zero. In this way, less information is assumed, making it a nontrivial task to identify break points. This is due to the fact that the smaller the magnitude of shift, the more difficult it will be to identify the breaks.

Under shrinking shifts we have

THEOREM 3. *Under A1-A4, we have*

$$Tv_T^2(\hat{\tau}_i - \tau_i^0) = O_p(1),$$

$$\sqrt{T}(\hat{\beta}_T - \beta_T^0) = O_p(1),$$

$$\sqrt{T}(\hat{\Gamma}_T - \Gamma_T^0) = O_p(1).$$

The convergence rate for the estimated break points is slower under shrinking magnitude of shift, but the rate for the estimated regression parameters and the variance-covariance matrices is the same.

We note that there is no need to estimate v_T. It is only used for theoretical purposes. The final result will be expressed in terms of the estimated parameters of $\theta^0_{i,T}$ and $\Sigma^0_{i,T}$ $(i = 1, ..., m+1)$. For notational simplicity, the subscript T attached to the parameters will be suppressed.

COROLLARY 2. *Under the assumptions of Theorem 2, the limiting distribution of $\hat{\beta}$ and $\hat{\Gamma}$ are the same as those of known $k_1^0, ..., k_m^0$.*

To derive the limiting distribution for the estimated break points, we need additional assumptions. Assume the following functional central limit theorem holds (note that $E(\eta_t\eta_t') = I$)

$$v_T \sum_{k_i^0+1}^{k_i^0+[vv_T^{-2}]} (\eta_t\eta_t' - I) \Rightarrow \xi_1(v) \tag{9}$$

where the weak convergence is in the space $D[0,M]^{q^2}$ for each fixed $M > 0$.[1] Each entry of the matrix $\xi_1(v)$ is a (nonstandard) Brownian motion process defined on $[0,\infty)$. We note that the partial sum involves $O(v_T^{-2})$ observations, thus the normalizing factor is v_T. Similarly, assume

$$v_T \sum_{k_i^0+[vv_T^{-2}]}^{k_i^0} (\eta_t\eta_t' - I) \Rightarrow \xi_2(-v) \tag{10}$$

for $v < 0$. Let $\xi(v) = \xi_1(v)$ for $v > 0$ and $\xi(v) = \xi_2(-v)$ for $v < 0$, and $\xi(0) = 0$, then each entry of $\xi(v)$ is a two-sided (nonstandard) Brownian motion on the real line. Let

$$\text{plim}\frac{1}{\Delta k_i^0} \sum_{k_i^0+1}^{k_i^0+\Delta k_i^0} (V_tV_t' \otimes (\Sigma_{i+1}^0)^{-1}) = Q. \tag{11}$$

where $\Delta k_i^0 = k_{i+1}^0 - k_i^0$, the length of regime $i+1$. Under the setup of shrinking shift, the limiting matrix Q does not depend on i. We assume the following functional central limit theorem:

$$v_T \sum_{k_i^0+1}^{k_i^0+[vv_T^{-2}]} [V_t \otimes (\Sigma_i^0)^{-1/2}]\eta_t \Rightarrow Q^{1/2}\zeta_1(v), \qquad v > 0 \tag{12}$$

where $\zeta_1(v)$ is vector of independent and standard Brownian motion processes. For $v < 0$, let the limit be denoted by $Q^{1/2}\zeta_2(-v)$. Define a vector of standard Brownian motion processes on the real line by $\zeta(v) = \zeta_1(v)$ for $v > 0$ and $\zeta(v) = \zeta_2(-v)$.

[1] This is equivalent to the weak convergence in the space of $D[0,\infty)^{q^2}$ under the topology of compacta, see Pollard (1984).

Let $(\xi^{(i)}, \zeta^{(i)})$ be independent copies of the stochastic processes (ξ, ζ), and define

$$\Lambda^{(i)}(v) = -|v|4^{-1}\mathrm{tr}(A_i^2) + 2^{-1}\mathrm{tr}(A_i\xi^{(i)}(v)) + \delta_i' Q^{1/2}\zeta^{(i)}(v) - \frac{1}{2}|v|\delta_i' Q\delta_i \tag{13}$$

where $A_i = \Sigma_0^{-1/2}\Phi_i\Sigma_0^{-1/2}$, and δ_i, Φ_i and Σ_0 are defined in Assumption A4. We have

THEOREM 4. *Under assumptions A1-A4 and (9)-(12), we have for each i, on compact set of $v \in [-M, M]$,*

$$\log \frac{\Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0 + [vv_T^{-2}], \hat{k}_{i+1}, ..., \hat{k}_m)}{\Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0, \hat{k}_{i+1}, ..., \hat{k}_m)} \Rightarrow \Lambda^{(i)}(v)$$

THEOREM 5. *Under the assumptions of Theorem 4,*

$$v_T^2(\hat{k}_i - k_i^0) \xrightarrow{d} \mathrm{argmax}_v\, \Lambda^{(i)}(v).$$

Throughout, we shall omit the superscript attached to ξ and ζ, in case of no confusion. Typically, $\xi(v)$ and $\zeta(v)$ are dependent. If we assume $E(\eta_{tk}\eta_{t\ell}\eta_{th}) = 0$ for all k, ℓ, h, and for every t, then $\xi(v)$ and $\zeta(v)$ will be independent. This is the case under normality assumption for η_t. For simplicity, we consider several useful results under the assumption of independence of ξ and ζ.

COROLLARY 3. *Under assumptions of Theorem 4 and $E(\eta_{tk}\eta_{t\ell}\eta_{th}) = 0$, we have*

$$\left[\frac{\left(2^{-1}\mathrm{tr}(A_i^2) + \delta_i' Q\delta_i\right)^2}{4^{-1}\mathrm{vec}(A_i)'\Omega\,\mathrm{vec}(A_i) + \delta_i' Q\delta_i}\right] v_T^2(\hat{k} - k_i^0) \xrightarrow{d} \mathrm{argmax}_v\{U(v) - |v|/2\}$$

where $\Omega = var(\bar{\xi}(1))$ and $\bar{\xi} = \mathrm{vec}(\xi)$, A_i and Q are defined earlier, and $U(v)$ is a two-sided standard Brownian motion process on the real line.

Remark 3.1. The term $\mathrm{vec}(A_i)'\Omega\,\mathrm{vec}(A_i)$ represents the variance of $\mathrm{tr}(A_i\xi(1))$. Because both A_i and ξ are symmetric matrices, $\mathrm{tr}(A_i\xi(1)) =$

$\text{vec}(A_i)'\text{vec}(\xi(1))$. It follows that the variance of $\text{tr}(A_i\xi(1))$ is given by $\text{vec}(A_i)'\Omega\,\text{vec}(A_i)$. Note that Ω is a singular matrix due to the symmetry of $\xi(1)$ ($\text{vec}(\xi)$ has many repeated variables). Alternatively, $\text{tr}(A_i\xi(1)) = \sum_h a_{hh}\xi_{hh} + 2\sum_{h<\ell} a_{hl}\xi_{h\ell}$, where $a_{h\ell}$ is the (h,ℓ)th entry of A_i and $\xi_{h\ell}$ is the (h,ℓ)th entry of $\xi(1)$. The variance of the trace can be easily derived in terms of the covariance matrix of the non-redundant elements of $\xi(1)$. But the final result is identical.

Let

$$V^* = \text{argmax}_v\{U(v) - |v|/2\}$$

This random variable has a known analytical density function, see Bai (1997) and the references therein.

In order to use Corollary 3 to construct confidence intervals for k_i^0, the unknown scaling factor of $(\hat{k}_i - k_i^0)$ in Corollary 3 must be estimated. Let $\hat{\Sigma}_i$ and $\hat{\theta}_i$ be the estimators given in (4). Let $\hat{B}_i = \hat{\Sigma}_i^{-1/2}(\hat{\Sigma}_{i+1} - \hat{\Sigma}_i)\hat{\Sigma}_i^{-1/2}$. Then $\hat{B}_i$ is an estimator of $A_i v_T$. Similarly, $(\hat{\theta}_{i+1} - \hat{\theta}_i)$ is an estimator of $\delta_i v_T$.

COROLLARY 4. *Assume the conditions of Corollary 3. Let $\hat{\Omega}$ be a consistent estimator for Ω and $\hat{Q}$ be a consistent estimator for Q. Then*

$$\left[\frac{\left(2^{-1}\text{tr}(\hat{B}_i^2) + (\hat{\theta}_{i+1} - \hat{\theta}_i)'\hat{Q}(\hat{\theta}_{i+1} - \hat{\theta}_i)\right)^2}{4^{-1}\text{vec}(\hat{B}_i)'\hat{\Omega}\text{vec}(\hat{B}_i) + (\hat{\theta}_{i+1} - \hat{\theta}_i)'\hat{Q}(\hat{\theta}_{i+1} - \hat{\theta}_i)}\right](\hat{k}_i - k_i^0) \xrightarrow{d} V^*$$

If the η_t's are iid, then Ω is equal to the covariance matrix of $\text{vec}(\eta_t\eta_t' - I)$. Thus we can estimate Ω by $\hat{\Omega} = T^{-1}\sum_{t=1}^T \text{vec}(\hat{\eta}_t\hat{\eta}_t' - I)\cdot\text{vec}(\hat{\eta}_t\hat{\eta}_t' - I)'$, where $\hat{\eta}_t$ is the estimated residual. There are a number of ways to estimating the matrix Q. The first approach is regime-length weighted estimation. It is given by

$$\hat{Q} = \frac{1}{T}\sum_{i=0}^{m}\sum_{k_i^0+1}^{k_i^0+\Delta k_i^0}(V_t V_t' \otimes \hat{\Sigma}_{i+1}^{-1}).$$

(cf. (11)). The second approach is equal-weighted estimation. Let $\hat{\Sigma} = (m+1)^{-1}\sum_{i=0}^m \hat{\Sigma}_i$ and define

$$\hat{Q} = \frac{1}{T}\sum_{t=1}^{T}(V_t V_t' \otimes \hat{\Sigma}^{-1}).$$

The third approach is regime-specific estimation, defined by

$$\hat{Q} = \frac{1}{\hat{k}_{i+1} - \hat{k}_i} \sum_{\hat{k}_i+1}^{\hat{k}_{i+1}} (V_t V_t' \otimes (\hat{\Sigma}_{i+1})^{-1}).$$

Bai and Perron (1999) discuss the pros and cons of various estimation procedures.

For a univariate autoregressive process, the estimation can be simplified. Let

$$\begin{aligned} \hat{\sigma}_i^2 &= \frac{1}{\hat{k}_i - \hat{k}_{i-1}} \sum_{\hat{k}_{i-1}+1}^{\hat{k}_i} (Y_t - V_t'\hat{\theta}_i)^2, \\ \hat{H} &= \frac{1}{T} \sum_{t=1}^{T} V_t V_t', \\ \hat{\eta}_t &= (Y_t - V_t'\hat{\theta}_i)\hat{\sigma}_i^{-1}, \quad for \quad t \in [\hat{k}_{i-1}, \hat{k}_i] \end{aligned}$$

$i = 1, ..., m+1$. Then

$$\begin{aligned} \hat{B}_i &= (\hat{\sigma}_{i+1}^2 - \hat{\sigma}_i^2)/\hat{\sigma}_i^2, \\ \hat{Q} &= \hat{H}/\hat{\sigma}^2, \\ \hat{\Omega} &= \frac{1}{T} \sum_{t=1}^{T} (\hat{\eta}_t^2 - 1)^2. \end{aligned}$$

where $\hat{\sigma}^2 = T^{-1} \sum_{i=0}^{m} (\hat{k}_{i+1} - \hat{k}_i)\hat{\sigma}_{i+1}^2$. We note that $\hat{B}_i$ and $\hat{\Omega}$ are scalars.

COROLLARY 5. *For univariate series, if there is no change in the regression parameters but only changes in variance, then*

$$\left[\frac{(\hat{\sigma}_{i+1}^2 - \hat{\sigma}_i^2)^2}{\sigma_i^4 \hat{\Omega}}\right](\hat{k}_i - k_i^0) \xrightarrow{d} V^*.$$

If there is no change in the variance but only changes in regression parameters, then

$$\left[\frac{(\hat{\theta}_{i+1} - \hat{\theta}_i)'\hat{H}(\hat{\theta}_{i+1} - \hat{\theta}_i)}{\hat{\sigma}^2}\right](\hat{k}_i - k_i^0) \xrightarrow{d} V^*.$$

Under normality assumption, the results of Corollary 3 and Corollary 4 can be further simplified.

COROLLARY 6. *For VAR models with* $\eta_t \sim N(0, I)$. *Then*

$$\left[2^{-1}\mathrm{tr}(A_i^2) + \delta_i' Q \delta_i\right] v_T^2(\hat{k}_i - k_i^0) \xrightarrow{d} V^*$$

and

$$\left[2^{-1}\mathrm{tr}(\hat{B}_i^2) + (\hat{\theta}_{i+1} - \hat{\theta}_i)'\hat{Q}(\hat{\theta}_{i+1} - \hat{\theta}_i)\right](\hat{k}_i - k_i^0) \xrightarrow{d} V^*.$$

COROLLARY 7. *For univariate time series and* $\eta_t \sim N(0,1)$, *we have*

$$\left[\frac{1}{2}\Big(\frac{\hat{\sigma}_{i+1}^2 - \hat{\sigma}_i^2}{\hat{\sigma}_i^2}\Big)^2 + \frac{(\hat{\theta}_{i+1} - \hat{\theta}_i)'\hat{H}(\hat{\theta}_{i+1} - \hat{\theta}_i)}{\hat{\sigma}_i^2}\right](\hat{k}_i - k_i^0) \xrightarrow{d} V^*$$

The $\hat{\sigma}_i^2$ can be replaced by $\hat{\sigma}^2$ without affecting the limiting distribution.

Confidence intervals for change points. We can use Corollary 4 to construct confidence intervals. Denote by $\hat{a}$ the scaling factor of $(\hat{k}_i - k_i^0)$ in Corollary 4 (i.e., the expression inside the bracket). Then the $100(1-\alpha)\%$ confidence interval can be constructed as

$$[\hat{k}_i - h - 1, \hat{k}_i + h + 1]$$

where $h = [c/\hat{a}]$ represents the integer part of $c/\hat{a}$. The constant c can be found from the distribution of V^*. For example, for $\alpha = 0.05, c = 7.0$, see Bai (1997).

4. DETERMINING THE NUMBER OF BREAKS

The number of break points can be determined by the BIC criterion, as in Yao (1988), who first studies the problem for mean shifts in an otherwise iid setting under normality assumption. His analysis is designed for fixed shift only. Here we show how this criterion can be applied to both fixed and shrinking shifts. The penalty term in the criterion function can be quite flexible but still permits consistent estimation of the number of breaks. The reason is that overestimating the number of breaks can only increase the log likelihood by a magnitude of $O_p(\log T)$, whereas under estimating the number of breaks will decrease the log-likelihood by a magnitude of $O(T)$ for fixed shifts (by a magnitude of $O(Tv_T^2)$ for shrinking shifts). Thus any penalty that is in between $\log T$ and $O(Tv_T^2)$ will yield consistent estimates. Let $L(\hat{k}_1, ..., \hat{k}_m)$ denote the optimal likelihood function when m breaks are

allowed. That is,

$$L(\hat{k}_1, ..., \hat{k}_m) = \max_{k_1,...,k_m} \max_{\theta_i, \Sigma_i; 1 \le i \le m+1} \prod_{i=1}^{m+1} \prod_{t=k_{i-1}+1}^{k_i} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta_i, \Sigma_i)$$

The information criterion is defined as

$$\mathrm{BIC}(m) = -\log L(\hat{k}_1, ..., \hat{k}_m) + m\, g(T)$$

Suppose that $m^0 < M$ for a specified integer M. Let $\hat{m}$ be chosen such that the BIC criterion is minimized over $m \le M$. That is, $\hat{m} = \mathrm{argmin}_{m \le M} BIC(m)$.

THEOREM 6. *Assume A1-A4 hold. Furthermore, assume $v_T \equiv 1$ or $v_T \to 0$ but satisfying (8). Let $g(T)$ be a sequence of positive numbers such that $g(T)/(Tv_T^2) \to 0$ but $g(T)/\log T \to \infty$. Then $\hat{m} \to m_0$.*

The information criterion approach requires to estimate more than one change point even though only one or no change point is present for the underlying model. However, at most $O(T^2)$ quasi-maximum likelihood estimations are needed no matter how many changes are entertained. This is because there are at most $O(T^2)$ distinct segments. Once the likelihood value for each segment is computed, dynamic programming approach can be used to efficiently select the optimal change points. Bai and Perron (1998) give a detailed discussion on this point.

5. CONCLUSION

In this paper, we studied the problem of multiple structural changes in VAR models occurring at unknown times. We considered shifts in the variance-covariance matrices in addition to shifts in the regression coefficients. Various theoretical properties of the estimators were obtained. We gave a unified analysis for fixed and shrinking magnitude of shifts.

This paper provides a systematic treatment of multiple structural changes. We derive a number of results that are stated in a series of lemmas. The methodology and central idea are applicable for different estimation methods such as GMM and robust estimations. They are also applicable for different models, such as nonlinear models with structural changes. All that is needed is to derive the counterparts of these lemmas under the new setting, be it either a new estimation method, or a new model. If these lemmas can be proved, then the rate of convergence follows automatically. This is true because our proof of rate of convergence does not make any reference to VAR. The limiting distribution of the estimators varies with

estimation methods and with the underlying model. However, limiting distribution is much easier to obtain than rate of convergence in the context of structural change. The former is a local property of the objective function (likelihood function in the present paper) and the latter is a global property, which is much more difficult to analyze. We hope the methodology developed in this paper will lead to progress in analyzing structural changes for different models and different estimation methods.

APPENDIX A

Preliminary Results

To prove the main results, we first establish a series of properties of sequential quasi-likelihood ratios and sequential estimators to be defined below in the absence of structural change. We then show that the main results can be derived from these properties. To begin with, let

$$Y_t = \mu + A_1 Y_{t-1} + \cdots + A_{t-p} Y_{t-p} + \varepsilon_t$$

$$Y_t = (V_t' \otimes I)\theta_0 + \varepsilon_t$$

where $\varepsilon_t = \Sigma_0^{1/2}\eta_t$, $Var(\eta_t) = I$, $V_t' = (1, Y_{t-1}', ..., Y_{t-p}')$, $\theta_0 = \text{Vec}(\mu, A_1, ..., A_p)$ and the ε_t are martingale differences with variance Σ_0. This model is otherwise identical to model (1), but no change point is allowed here.

Let (θ_0, Σ_0) denote the true parameter. Consider the centered-likelihood ratio based on the first k observations,

$$\mathcal{L}(0,k;\theta,\Sigma) = \frac{\prod_{t=1}^{k} f(Y_t|Y_{t-1},\cdots;\theta_0 + T^{-\frac{1}{2}}\theta, \Sigma_0 + T^{-\frac{1}{2}}\Sigma)}{\prod_{t=1}^{k} f(Y_t|Y_{t-1},\cdots;\theta_0,\Sigma_0)} \tag{A.1}$$

$$= \frac{|\Sigma_0 + T^{-\frac{1}{2}}\Sigma|^{-k/2} \exp\left\{ -\frac{1}{2}\sum_{t=1}^{k} \varepsilon_t(\theta)'(\Sigma_0 + T^{-1/2}\Sigma)^{-1}\varepsilon_t(\theta)\right\}}{|\Sigma_0|^{-k/2}\exp\{-\frac{1}{2}\sum_{t=1}^{k}\varepsilon_t'\Sigma_0^{-1}\varepsilon_t\}}.$$

where $\varepsilon_t(\theta) = Y_t - (V_t' \otimes I)(\theta_0 + T^{-\frac{1}{2}}\theta) = \varepsilon_t - T^{-1/2}(V_t' \otimes I)\theta$. The first two arguments $(0,k)$ in $\mathcal{L}(\cdot)$ emphasizes the likelihood ratio is for observations $t \in (0,k]$. Likehilood ratio for an arbitrary segment will be introduced later. We shall call $\mathcal{L}(\cdot)$ the (centered) quasi-sequential likelihood ratio.

The first five lemmas are proved in Bai, Lumsdaine and Stock (1998) (hereafter, BLS). Denote by $\hat{\theta}_{(k)}$ and $\hat{\Sigma}_{(k)}$ the values of θ and Σ that $\mathcal{L}(0,k,\theta,\Sigma)$ achieves its maximum. Then we have

LEMMA 1. *For each* $\delta \in (0,1]$,

$$\sup_{T \geq k \geq T\delta} \left(\|\hat{\theta}_{(k)}\| + \|\hat{\Sigma}_{(k)}\| \right) = O_p(1),$$

$$\sup_{T\geq k\geq T\delta} \mathcal{L}(0,k;\hat{\theta}_{(k)},\hat{\Sigma}_{(k)}) = O_p(1)$$

This lemma says that the sequential likelihood ratios and the sequential estimators are bounded in probability if a positive fraction of observations are used. This result is a direct consequence of the functional central limit theorem for martingale differences.

Proof. See Property 1 of BLS. ▮

The following lemma is concerned with the supremum of the likelihood ratios over all k and over the whole parameter space.

LEMMA 2. *For each $\epsilon > 0$, there exists a $B > 0$ such that for all large T*

$$Pr\left(\sup_{1\leq k\leq T} T^{-B}\mathcal{L}(0,k;\hat{\theta}_{(k)},\hat{\Sigma}_{(k)}) > 1\right) < \epsilon$$

This property says that the log-valued quasi-sequential likelihood ratio has its maximum value bounded by $O_p(\log T)$.

Proof. see Property 2 of BLS. ▮

Let $S_T = \{(\theta,\Sigma); \|\theta\| \geq \log T \text{ or } \|\Sigma\| \geq \log T\}$. We assume that $\Sigma_0 + T^{-\frac{1}{2}}\Sigma$ is positive definite so that the likelihood ratio is well defined.

LEMMA 3. *For any $\delta \in (0,1)$, $D > 0$, $\epsilon > 0$, the following holds when T is large*

$$Pr\left(\sup_{T\geq k\geq T\delta}\sup_{(\theta,\Sigma)\in S_T} T^{D}\mathcal{L}(0,k;\theta,\Sigma) > 1\right) < \epsilon$$

Proof. see Property 3 of BLS. ▮

The following lemma will be used in the proof of Theorem 6, the consistency of BIC criterion.

LEMMA 4. *Let a_T be a sequence of positive numbers such that $a_T \geq \log T$ and $a_T/\sqrt{T}$ is bounded. Then for every $\epsilon > 0$, there exists a $c > 0$ such that, for all large T,*

$$Pr\left(\sup_{k\geq T\delta}\sup_{\|\theta\|\geq a_T,\ \text{or}\|\Sigma\|\geq a_T} \log\mathcal{L}(0,k;\theta,\Sigma) > -ca_T^2\right) < \epsilon.$$

Proof. This lemma is implicitly proved by BLS. Let $b_T = a_T/\sqrt{T}$. The proof of Property 3 of BLS is valid for either $b_T = o(1)$ or $b_T = O(1)$. The equation immediately below (A.15) together with (A.19) of BLS says that $\log \mathcal{L}(0, k; \theta, \Sigma) < -(T\delta)cb_T^2 = -c\delta a_T^2$ with large probability. This implies the lemma. The details are omitted. ▮

As a corollary of this lemma we have for $a_T = a\sqrt{T}$ $(a > 0)$, there exists a $c > 0$ such that

$$Pr\left(\sup_{k \geq T\delta} \sup_{\|\theta\| \geq a\sqrt{T},\ \text{or} \|\Sigma\| \geq a\sqrt{T}} \log \mathcal{L}(0, k; \theta, \Sigma) > -cT\right) < \epsilon \qquad \text{(A.2)}$$

for all large T. For $a_T = a\sqrt{T}v_T$, where $v_T = 1$ or $v_T \to 0$ but satisfying $\sqrt{T}v_T \geq \log T$, then by Lemma 4 for any $\epsilon > 0$, there exists a $c > 0$, such that for all large T,

$$Pr\left(\sup_{k \geq T\delta} \sup_{\|\theta\| \geq a\sqrt{T}v_T,\ \text{or} \|\Sigma\| \geq a\sqrt{T}v_T} \log \mathcal{L}(0, k; \theta, \Sigma) > -cTv_T^2\right) < \epsilon. \quad \text{(A.3)}$$

LEMMA 5. *Let h_T and d_T be positive sequences such that h_T is non-decreasing, $d_T \to +\infty$, and $(h_T d_T^2)/T \to h > 0$, where $h < \infty$. Let $S_T = \{(\theta, \Sigma); \|\theta\| \geq d_T \text{ or } \|\Sigma\| \geq d_T\}$. Then for any $\epsilon > 0$, there exists an $A > 0$, such that when T is large*

$$Pr\left(\sup_{T \geq k \geq Ah_T} \sup_{(\theta, \Sigma) \in S_T} \mathcal{L}(0, k; \theta, \Sigma) > \epsilon\right) < \epsilon.$$

Remark A.1. The existence of a limit for $h_T d_T^2/T$ is not necessary. It is sufficient to have $\liminf_{T \to \infty} h_T d_T^2/T \geq h > 0$. In addition, if Property 5 holds for $h < \infty$, then it holds for $h = \infty$ (larger h corresponds to a smaller set S_T). Lemma 3, for example, is a case where $h = \infty$. The assumption that $h < \infty$ is convenient for proof and is also the actual case in the application of this lemma.

Proof. see Property 5 of BLS. ▮

LEMMA 6. *Assume $v_T = 1$ or $v_T \to 0$ but satisfying (8). Let $\Upsilon = \Sigma_1 - \Sigma_0$. For each given ϕ and $\Sigma_1 > 0$ such that $\|\phi\| \leq Mv_T$ and $\|\Upsilon\| \leq Mv_T$, with $M < \infty$ (an arbitrary given constant),*

$$\sup_{1 \leq k \leq \sqrt{T}v_T^{-1}} \sup_{\lambda, \Xi} \frac{\mathcal{L}(0, k; \sqrt{T}\phi + \lambda, \sqrt{T}\Upsilon + \Xi)}{\mathcal{L}(0, k; \sqrt{T}\phi, \sqrt{T}\Upsilon)} = O_p(1)$$

where the supremum with respect to λ and Ξ is taken over a compact set such that $\|\lambda\| \leq M$ and $\|\Xi\| \leq M$.

This lemma will be applied in the context that ϕ represents the magnitude of shift in regression parameters such as $\theta^0_{i+1} - \theta^0_i$, and Υ represents the magnitude of shift in covariance matrices such as $\Upsilon = \Sigma^0_{i+1} - \Sigma^0_i$. This lemma is useful in establishing $|\hat{k}_i - k^0_i| = O_p(\sqrt{T} v_T^{-1})$.

Proof. We prove that the logarithm of left hand side is $O_p(1)$. We can write

$$\log \mathcal{L}(0, k; \theta, \Sigma) = \mathcal{L}_{1,T}(0, k; \theta, \Sigma) + \mathcal{L}_{2,T}(0, k; \theta, \Sigma)$$

where

$$\mathcal{L}_{1,T} = -\frac{1}{2} k \log |I + \Psi_T| - \frac{1}{2} k \left[\frac{1}{k} \sum_{t=1}^{k} \eta_t' (I + \Psi_T)^{-1} \eta_t - \frac{1}{k} \sum_{t=1}^{k} \eta_t' \eta_t \right] \tag{A.4}$$

and

$$\begin{aligned} \mathcal{L}_{2,T} &= T^{-\frac{1}{2}} \theta' \Big(I \otimes (I + \Psi_T)^{-1} \Big) \sum_{t=1}^{k} (V_t \otimes \eta_t) \\ &- \frac{1}{2} \frac{k}{T} \theta' \left(\frac{1}{k} \sum_{t=1}^{k} V_t V_t' \otimes (I + \Psi_T)^{-1} \right) \theta \end{aligned} \tag{A.5}$$

with $\eta_t = \Sigma_0^{-1/2} \varepsilon_t$ and $\Psi_T = T^{-\frac{1}{2}} (\Sigma_0^{-1/2} \Sigma \Sigma_0^{-1/2})$. It suffices to show

$$\mathcal{L}_{1,T}(0, k, \sqrt{T}\phi + \lambda, \sqrt{T}\Upsilon + \Xi) - \mathcal{L}_{1,T}(0, k; \sqrt{T}\phi, \sqrt{T}\Upsilon) = O_p(1), \tag{A.6}$$

$$\mathcal{L}_{2,T}(0, k, \sqrt{T}\phi + \lambda, \sqrt{T}\Upsilon + \Xi) - \mathcal{L}_{2,T}(0, k; \sqrt{T}\phi, \sqrt{T}\Upsilon) = O_p(1). \tag{A.7}$$

Consider (A.7). Take Σ to be $\sqrt{T}\Upsilon + \Xi$, we have $I + \Psi_T = \Sigma_0^{-1/2} \Upsilon \Sigma_0^{-1/2} + T^{-1/2} \Sigma_0^{-1/2} \Xi \Sigma_0^{-1/2} \stackrel{def}{=} A + T^{-1/2} B$. And if we take Σ to be $\sqrt{T}\Upsilon$, then $I + \Psi_T = A$. Using $(A + T^{-1/2}B)^{-1} - A^{-1} = -T^{-1/2} A^{-1} B (A + T^{-1/2} B)^{-1}$,

we find the left hand side of (A.7) is equal to

$$\begin{aligned}
&\tfrac{1}{\sqrt{T}}\phi'\Big(I\otimes A^{-1}B(A+T^{-1/2}B)\Big)^{-1}\textstyle\sum_{t=1}^{k}(V_t\otimes\eta_t)\\
&-\tfrac{1}{\sqrt{T}}\lambda'\Big(I\otimes(A+T^{-1/2}B)^{-1}\Big)\textstyle\sum_{t=1}^{k}(V_t\otimes\eta_t)\\
&-\tfrac{k}{\sqrt{T}}\phi'\Big(\tfrac{1}{k}\textstyle\sum_{t=1}^{k}V_tV_t'\otimes A^{-1}B(A+T^{-1/2}B)^{-1}\Big)\phi\\
&-\tfrac{k}{\sqrt{T}}\phi'\Big(\tfrac{1}{k}\textstyle\sum_{t=1}^{k}V_tV_t'\otimes(A+T^{-1/2}B)^{-1}\Big)\lambda\\
&-\tfrac{k}{T}\lambda'\Big(\tfrac{1}{k}\textstyle\sum_{t=1}^{k}V_tV_t'\otimes(A+T^{-1/2}B)^{-1}\Big)\lambda
\end{aligned}\tag{A.8}$$

The first two and the last expressions are $O_p(1)$ uniformly in $k\in[1,T]$ and uniformly in bounded λ and Ξ. The third and fourth are $O_p(1)$ uniformly in $k\le\sqrt{T}v_T^{-1}$ and in bounded λ and Ξ. For example, $\|\frac{k}{\sqrt{T}}\phi\|\le\|v_T^{-1}\phi\|\le M$ since $\|\phi\|\le Mv_T$. Next consider (A.6). It can be written as

$$-\frac{k}{2}\Big(\log|A+T^{-1/2}B|-\log|A|\Big)-\frac{1}{2}\left[\sum_{t=1}^{k}\eta_t'(A+T^{-1/2}B)^{-1}-A^{-1})\eta_t\right]\tag{A.9}$$

From

$$(A+T^{-1/2}B)^{-1}-A^{-1}=-T^{-1/2}A^{-1}BA^{-1}+O(T^{-1})$$

the second term of (A.9) is, ignoring the factor $1/2$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{k}\eta_t'A^{-1}BA^{-1}\eta_t+\frac{k}{T}\Big(\frac{1}{k}\sum_{t=1}^{k}\eta_t'\eta_t\Big)O(1)$$

$$=-\frac{k}{\sqrt{T}}\mathrm{tr}(A^{-1}BA^{-1})+\mathrm{tr}\Big[A^{-1}BA^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{k}(\eta_t\eta_t'-I)\Big]+o_p(1)$$

where the $o_p(1)$ follows from $k/T=o(1)$ for $k\le\sqrt{T}v_T^{-1}$ and $\frac{1}{k}\sum_{t=1}^{k}|\eta_t\|^2=O_p(1)$. The second term on r.h.s. above is $o_p(1)$ because $\frac{1}{\sqrt{T}}\sum_{t=1}^{k}(\eta_t\eta_t'-I)=\frac{1}{(\sqrt{T}v_T)^{1/2}}\frac{1}{(\sqrt{T}v_T^{-1})^{1/2}}\sum_{t=1}^{k}(\eta_t\eta_t'-I)=\frac{1}{(\sqrt{T}v_T)^{1/2}}O_p(1)=o_p(1)$ uniformly in $k\le\sqrt{T}v_T^{-1}$. Thus the second term of (A.9) is equal to $kT^{-1/2}\mathrm{tr}(A^{-1}BA^{-1})+o_p(1)$. Next, by Taylor expansion, $\log|A+T^{-1/2}B|-\log|A|=T^{-1/2}\mathrm{tr}(A^{-1}B)+O(T^{-1})$. It follows that (A.9) can be written as

$$-\frac{k}{2\sqrt{T}}\Big(\mathrm{tr}(A^{-1}B)-\mathrm{tr}(A^{-1}BA^{-1})\Big)+o_p(1)$$

By the definition of A and B, $A^{-1}B - A^{-1}BA^{-1} = A^{-1}B\Sigma_0^{-1/2}\Upsilon\Sigma_0^{-1/2}$. Thus $\|\frac{k}{\sqrt{T}}\mathrm{tr}(A^{-1}B - A^{-1}BA^{-1})\| \le v_T^{-1}\|\Upsilon\| \le v_T^{-1}Mv_T = M$. This proves the lemma. ∎

The above lemma can be strengthened if we restrict the supremum with respect to k to be taken over a smaller range. The right hand side $O_p(1)$ can be replaced by $o_p(1)$ after taking logarithm. Without taking logarithm, $O_p(1)$ is replaced by $1 + o_p(1)$.

LEMMA 7. *Under the assumptions of the previous lemma. We have*

$$\sup_{1\le k\le Mv_T^{-2}} \sup_{\lambda,\Xi} \log \frac{\mathcal{L}(0,k;\sqrt{T}\phi+\lambda,\sqrt{T}\Upsilon+\Xi)}{\mathcal{L}(1,k;\sqrt{T}\phi,\sqrt{T}\Upsilon)} = o_p(1)$$

where the supremum with respect to λ and Ξ is taken over a compact set such that $\|\lambda\| \le M$ and $\|\Xi\| \le M$.

Proof. The proof is virtually identical to the proof of of Lemma 6. The details are omitted. ∎

LEMMA 8.

$$\sup_{T\ge k\ge [T\delta]} \sup_{\phi,\Upsilon,\lambda,\Xi} \log \frac{\mathcal{L}(0,k;\phi+T^{-1/2}\lambda,\Upsilon+T^{-1/2}\Xi)}{\mathcal{L}(0,k;\phi,\Upsilon)} = o_p(1)$$

where the supremum with respect to $\phi, \Upsilon, \lambda, \Xi$ are taken over an arbitrary bounded set.

Proof. The proof is the same as that of Lemma 6 with the following changes. Replace each of ϕ, Υ, λ, and Ξ by itself multiplied by $T^{-1/2}$. The rest of argument is similar (note that the range of k is enlarged to T. ∎

LEMMA 9. *Let $T_1 = [Ta]$ for some $a \in (0,1]$, and let $T_2 = [\sqrt{T}v_T^{-1}]$, where $v_T \equiv 1$ or $v_T \to 0$ but satisfying (8). Consider*

$$\begin{aligned} Y_t &= (V_t \otimes I)\phi_1^0 + (\Upsilon_1^0)^{1/2}\eta_t \qquad t = 1,2,...,T_1 \\ Y_t &= (V_t \otimes I)\phi_2^0 + (\Upsilon_2^0)^{1/2}\eta_t \qquad t = T_1+1,...,T_1+T_2 \end{aligned} \tag{A.10}$$

where $\|\phi_1^0 - \phi_2^0\| \le Mv_T$ and $\|\Upsilon_1^0 - \Upsilon_2^0\| \le Mv_T$ for some $M < \infty$. Let $n = T_1 + T_2$ be the size of the pooled sample. Let $(\hat{\phi}_n, \hat{\Upsilon}_n)$ be the estimator based on the pooled sample (treated as a single regime). Then

$$\hat{\phi}_n - \phi_1^0 = O_p(T^{-1/2}),$$

$$\hat{\Upsilon}_n - \Upsilon_1^0 = O_p(T^{-1/2}).$$

Proof. This lemma says that when pooled data from two regimes are used, the estimated parameters are close to those of the "dominating regime." This is of course obvious. But this lemma quantities the intuition in terms of the rate of convergence. The proof is trivial and thus is omitted. ▌

LEMMA 10. *Assume the same setup as in Lemma 9, but with* $T_2 = [Mv_T^{-2}]$. *Then*

$$\hat{\phi}_n - \phi_1^0 = O_p(T^{-1/2}),$$

$$\hat{\Upsilon}_n - \Upsilon_1^0 = O_p(T^{-1/2}),$$

$$\hat{\phi}_n - \hat{\phi}_1 = O_p(T^{-1}), \quad \text{and}$$

$$\hat{\Upsilon}_n - \hat{\Upsilon}_1 = O_p(T^{-1}).$$

where $(\hat{\phi}_1, \hat{\Upsilon}_1)$ *is the estimator of* (ϕ_1^0, Υ_1^0) *based on the first* T_1 *observations only.*

For $v_T \equiv 1$, T_2 is finite. In this case, the results of this lemma are easy to prove. This is because only a finite number of observations do not belong to the "dominating regime." When v_T converges to zero, T_2 converges to infinity. In this case, the second regime (i.e., the non-dominating regime) contains an increasing number of observations. However, the lemma is still true and the proof is still easy. The idea is to use the fact the magnitude of shift itself is also decreasing as T increases. This idea can be found in Bai (1994, 1997), where it is proved that the estimated regression coefficients have the same rate of convergence as if the change points are known (even though the estimated change point can be Mv_T^2 away from the true ones). The details will be omitted.

Proofs of Theorems 1-5

We introduce here additional notation to simplify some expressions.

Likelihood ratio of a segment. Denote by $D(k, \ell; \theta, \Sigma)$ the likelihood ratio of the segment $(k, \ell]$ (for observations from $k+1$ to ℓ, treated as a single regime). That is,

$$D(k, \ell; \theta, \Sigma) = \frac{\prod_{t=k+1}^{\ell} f(Y_t | Y_{t-1}, ..., Y_{t-p}; \theta, \Sigma)}{\prod_{t=k+1}^{\ell} f(\varepsilon_t)}$$

and its optimized value

$$D(k,\ell) = \sup_{\theta,\Sigma} D(k,\ell;\theta,\Sigma). \tag{A.11}$$

The likelihood ratio for the entire sample can be written as

$$\Lambda_T(k_1,...,k_m) = D(0,k_1)\cdot D(k_1,k_2)\cdots D(k_m,T) \tag{A.12}$$

Centered Likelihood Ratio of a Segment. Suppose that (θ_i^0,Σ_i^0) is the true parameter for the segment $(k,\ell]$ (i.e., $(k,\ell]\subset(k_{i-1}^0,k_i^0]$), we define the centered likelihood ratio as

$$\mathcal{L}(k,\ell;\theta,\Sigma) = D(k,\ell;\theta_i^0+T^{-\frac{1}{2}}\theta,\Sigma_i^0+T^{-\frac{1}{2}}\Sigma). \tag{A.13}$$

The centered likelihood ratio is only defined for segments that do not contain breaks. However, we can always express the likelihood ratio of a segment in terms of the centered ones even though the segment overlaps with more than one true regime. For example, suppose a segment $(k,\ell]$ overlaps with two regimes such that the segment contains portion of regime i and portion of regime $i+1$ such that $k_i^0\in[k+1,\ell-1]$, then

$$\begin{aligned} D(k,\ell;\phi,\Upsilon) = \;& \mathcal{L}(k,k_i^0;\sqrt{T}(\phi-\theta_i^0),\sqrt{T}(\Upsilon-\Sigma_i^0)) \\ & \cdot\; \mathcal{L}(k_i^0,\ell;\sqrt{T}(\phi-\theta_{i+1}^0),\sqrt{T}(\Upsilon-\Sigma_{i+1}^0)). \end{aligned}$$

The optimal values of centered and non-centered likelihood are the same. That is,

$$\sup_{\theta,\Sigma}\mathcal{L}(k,\ell;\theta,\Sigma) = D(k,\ell). \tag{A.14}$$

This fact will be useful.

We next give a unified proof for fixed and shrinking shifts.

PROPOSITION 1. *Assume A1-A4. For $v_T\equiv 1$ or for $v_T\to 0$ but satisfying (8), we have for every $\epsilon>0$,*

$$P(|\hat{k}_j-k_j^0|>\sqrt{T}v_T^{-1})<\epsilon \quad (j=1,...,m).$$

Proof. Let $N=[\sqrt{T}v_T^{-1}]$. Let $A_j=\{(k_1,...,k_m)\in K_\nu;|k_i-k_j^0|>N,\ i=1,...,m\}$ where K_ν is given in (5). Because $\Lambda_T(\hat{k}_1,...,\hat{k}_m)\geq$

$\Lambda_T(k_1^0,...,k_m^0) \geq \Lambda_T(k_1^0,...,k_m^0,\beta^0,\Gamma^0) = 1$, to show $(\hat{k}_1,...,\hat{k}_m) \notin A_j$, it suffices to show

$$P\Big(\sup_{(k_1,...,k_m)\in A_j} \Lambda_T(k_1,...,k_m) > \epsilon\Big) < \epsilon. \tag{A.15}$$

We now extend the definition of Λ_T to every subset $\{\ell_1,...,\ell_r\}$ of $\{1,2,...,T-1\}$ such that $\Lambda_T(\ell_1,...,\ell_r) = \Lambda_T(\ell_{(1)},...,\ell_{(r)})$ where $0 < \ell_{(1)} < \cdots < \ell_{(r)}$ are the ordered version of $\ell_1,...,\ell_r$. For every $(k_1,...,k_m) \in A_j$,

$$\Lambda_T(k_1,...,k_m) \leq \Lambda_T(k_1,...,k_m,k_1^0,...,k_{j-1}^0,k_j^0-N,k_j^0+N,k_{j+1}^0,...,k_m^0). \tag{A.16}$$

The right hand side above can be written as the product of at most $(2m+2)$ terms expressible as $D(\ell,k)$, see (A.12). There are at most $(2m+2)$ terms because k_i may coincide with k_ℓ^0 for some i and ℓ. One of these $(2m+2)$ terms is $D(k_j^0-N,k_j^0+N)$ and all the rest can be written as $D(\ell,k)$ with $(\ell,k] \subset [k_i^0+1,k_{i+1}^0]$ for some i. By Lemmas 1 and 2, $\log D(\ell,k) = O_p(\log T)$ uniformly in ℓ,k such that $k_i^0+1 \leq \ell < k \leq k_{i+1}^0$ with $|\ell-k| \geq T\nu$. That is, $D(k,\ell) = O_p(T^B)$ for some $B > 0$. Thus

$$\Lambda_T(k_1,...,k_m) \leq O_p(T^{(2m+1)B}) \cdot D(k_j^0-N,k_j^0+N). \tag{A.17}$$

We next show $D(k_j^0-N,k_j^0+N)$ is small. Introduce the reparameterization,

$$\mathcal{L}^*(k,\ell;\theta,\Sigma) = D(k,\ell;\theta_0+(\ell-k)^{-1/2}\theta,\Sigma_0+(\ell-k)^{-1/2}\Sigma)$$

assuming that (θ_0,Σ_0) is the true parameter of the segment $(k,\ell]$. (Note the difference between $\mathcal{L}^*$ and $\mathcal{L}$; the latter uses $T^{-1/2}$ rather than $(\ell-k)^{-1/2}$ in the reparameterization). We note that

$$\begin{aligned} D(k_j^0-N,k_j^0+N) &= \sup_{\theta,\Sigma}\Big[D(k_j^0-N,k_j^0;\theta,\Sigma)\cdot D(k_j^0,k_j^0+N;\theta,\Sigma)\Big] \\ &= \sup_{\theta,\Sigma}\Big[\mathcal{L}^*\Big(k_j^0-N,k_j^0;\sqrt{N}(\theta-\theta_j^0),\sqrt{N}(\Sigma-\Sigma_j^0)\Big) \\ &\times\ \mathcal{L}^*\Big(k_j^0,k_j^0+N;\sqrt{N}(\theta-\theta_{j+1}^0),\sqrt{N}(\Sigma-\Sigma_{j+1}^0)\Big)\Big]. \end{aligned} \tag{A.18}$$

The above follows from the definition of $\mathcal{L}^*$ and the fact that (θ_j^0,Σ_j^0) is the true parameter for the segment $[k_j^0-N,k_j^0]$ and $(\theta_{j+1}^0,\Sigma_{j+1}^0)$ is the true parameter for the segment $[k_j^0+1,k_j^0+N]$. From $\max\{\|x-z\|,\|y-z\|\} \geq \|x-y\|/2$ for all (x,y,z), we have for all θ and Σ,

$$\max\{\sqrt{N}\|\theta-\theta_j^0\|,\sqrt{N}\|\theta-\theta_{j+1}^0\|\} \geq \sqrt{N}\|\theta_j^0-\theta_{j+1}^0\|/2$$

$$\max\{\sqrt{N}\|\Sigma-\Sigma_j^0\|,\sqrt{N}\|\Sigma-\Sigma_{j+1}^0\|\}\geq\sqrt{N}\|\Sigma_j^0-\Sigma_{j+1}^0\|/2.$$

By A4, we either have $\sqrt{N}\|\theta_j^0-\theta_{j+1}^0\|/2\geq\log N$ or $\sqrt{N}\|\Sigma_j^0-\Sigma_{j+1}^0\|/2\geq\log N$. This follows from if $\|\theta_j^0-\theta_{j+1}^0\|\geq v_TC$ for some $C>0$, then $N^{1/2}\|\theta_j^0-\theta_{j+1}^0\|/2=(\sqrt{T}v_T^{-1})^{1/2}v_TC=C(\sqrt{T}v_T)^{1/2}\geq\log T\geq\log N$. Now suppose that $\sqrt{N}\|\theta_j^0-\theta_{j+1}^0\|/2\geq\log N$ (the case for which $\sqrt{N}\|\Sigma_j^0-\Sigma_{j+1}^0\|/2\geq\log N$ is the same). Then we have either (i) $\sqrt{N}\|\theta-\theta_j^0\|\geq\log N$ or (ii) $\sqrt{N}\|\theta-\theta_{j+1}^0\|\geq\log N$. For case (i), we can apply Lemma 3 to the first term inside the bracket of (A.18) to obtain

$$\mathcal{L}^*\Big(k_j^0-N,k_j^0;\sqrt{N}(\theta-\theta_j^0),\sqrt{N}(\Sigma-\Sigma_j^0)\Big)=O_p(N^{-A})$$

for every $A>0$ (The lemma is applied with T replaced by N and with $\delta=1$). Moreover, by Lemma 2, the second term inside the bracket of (A.18) is bounded by $O_p(\log T)$. Similarly, for case (ii), we can apply Lemma 3 to show that the second term of (A.18) is $O_p(N^{-A})$ and the first term is bounded by $O_p(\log T)$. So for each case, we have

$$D(k_j^0-N,k_j^0+N)=(\log T)\,O_p(N^{-A})$$

for an arbitrary $A>0$. But $N^{-A}\leq T^{-A/2}$ since $N\geq T^{1/2}$ for all large T. Thus from (A.17),

$$\Lambda_T(k_1,...,k_m)\leq O_p(T^{(2m+1)B-\frac{1}{2}A})\log T\xrightarrow{p}0$$

for a large A. This proves (A.15) and thus the proposition. ∎

PROPOSITION 2. *Assume A1-A4. For every $\epsilon>0$, there exists a $C>0$ such that*

$$P(|\hat{k}_j-k_j^0|>Cv_T^{-2})<\epsilon\qquad(j=1,...,m)$$

Proof. For concreteness, we shall prove the proposition for $j=2$ (the case of $j=1$ or $j=m$ is simpler. Other cases are the same as $j=2$). Let $A_2=\{(k_1,...,k_m);|k_i-k_i^0|\leq\sqrt{T}v_T^{-1},\forall i\}$ and let $A_2(C)$ be a subset of A_2 such that $A_2(C)=\{(k_1,...,k_m)\in A_2:|k_2-k_2^0|>Cv_T^{-2}\}$. By Proposition 1, $P(\hat{k}_1,...,\hat{k}_m)\in A_2)\to1$. From

$$\frac{\Lambda_T(\hat{k}_1,\hat{k}_2,...,\hat{k}_m)}{\Lambda_T(\hat{k}_1,k_2^0,\hat{k}_3,...,\hat{k}_m)}\geq1,$$

to show $|\hat{k}_2 - k_2^0| \leq Cv_T^{-2}$ or, equivalently, $(\hat{k}_1, ..., \hat{k}_m) \notin A_2(C)$ for large C with large probability, it suffices to establish that

$$P\left(\sup_{A_2(C)} \frac{\Lambda_T(k_1, k_2, ..., k_m)}{\Lambda_T(k_1, k_2^0, k_3, ..., k_m)} > \epsilon\right) < \epsilon. \tag{A.19}$$

Canceling common terms, we find

$$\begin{aligned} \frac{\Lambda_T(k_1, k_2, ..., k_m)}{\Lambda_T(k_1, k_2^0, k_3, ..., k_m)} &= \frac{D(k_1, k_2)D(k_2, k_3)}{D(k_1, k_2^0)D(k_2^0, k_3)} \\ &\leq \frac{D(k_1, k_2)}{D(k_1, k_2^0;\, \theta_2^0, \Sigma_2^0)} \cdot \frac{D(k_2, k_3)}{D(k_2^0, k_3;\, \theta_3^0, \Sigma_3^0)} \end{aligned} \tag{A.20}$$

where the inequality follows from $D(k_1, k_2^0) \geq D(k_1, k_2^0; \theta_2^0, \Sigma_2^0)$ (cf. (A.11)). For concreteness, we assume $k_1 \leq k_1^0$, $k_2 < k_2^0 - Cv_T^{-2}$, and $k_3 \leq k_3^0$. Other cases are similar. Let $\rho_2^0 = (\theta_2^0, \Sigma_2^0)$ and $\rho_3^0 = (\theta_3^0, \Sigma_3^0)$ denote the true parameters of regime 2 and regime 3, respectively. Suppose $\hat{\rho}_2 = (\hat{\theta}_2, \hat{\Sigma}_2)$ maximizes $D(k_1, k_2, \theta, \Sigma)$. We can write

$$D(k_1, k_2) = D(k_1, k_2; \hat{\rho}_2) = D(k_1, k_1^0; \hat{\rho}_2) \cdot D(k_1^0, k_2; \hat{\rho}_2) \tag{A.21}$$

Similarly,

$$D(k_1, k_2; \rho_2^0) = D(k_1, k_1^0; \rho_2^0) \cdot D(k_1^0, k_2; \rho_2^0) = D(k_1, k_1^0; \rho_2^0) \tag{A.22}$$

We have used the fact that the likelihood ratio $D(k_1^0, k_2; \rho_2^0) = 1$ because $[k_1^0 + 1, k_2] \subset [k_1^0 + 1, k_2^0]$ and ρ_2^0 is the true parameter for this segment. Thus, from (A.21) and (A.22),

$$\frac{D(k_1, k_2)}{D(k_1, k_2; \rho_2^0)} = \frac{D(k_1, k_1^0; \hat{\rho}_2)}{D(k_1, k_1^0; \rho_2^0)} D(k_1^0, k_2; \hat{\rho}_2) \tag{A.23}$$

By Lemma 9, $\sqrt{T}(\hat{\rho}_2 - \rho_2^0) = O_p(1)$ (apply the lemma with $T_1 = k_2 - k_1^0$ and $T_2 = k_1^0 - k_1$. Note that $T_1 \geq aT$ for some $a > 0$ and $T_2 \leq \sqrt{T}v_T^{-1}$ by the definition of A_j). From the relationship between D and $\mathcal{L}$ (cf. (A.13)) and in view of ρ_1^0 being the true parameter for the segment $(k_1, k_1^0]$, we can write

$$\frac{D(k_1, k_1^0; \hat{\rho}_2)}{D(k_1, k_1^0; \rho_2^0)} = \frac{\mathcal{L}(k_1, k_1^0; \sqrt{T}(\rho_2^0 - \rho_1^0) + \sqrt{T}(\hat{\rho}_2 - \rho_2^0))}{\mathcal{L}(k_1, k_1^0; \sqrt{T}(\rho_2^0 - \rho_1^0))}.$$

Because $|k_1 - k_1^0| \leq T^{1/2}v_T^{-1}$, we can apply Lemma 6 to the above expression, applied with $(\phi, \Upsilon) = (\rho_2^0 - \rho_1^0)$, and $(\lambda, \Xi) = \sqrt{T}(\hat{\rho}_2 - \rho_2^0) = O_p(1)$.

Note that $\|\phi\| \leq Mv_T$ by Assumption A4. The same is true for Υ. Thus we have by Lemma 6

$$\frac{D(k_1, k_1^0; \hat{\rho}_2)}{D(k_1, k_1^0; \rho_2^0)} = O_p(1). \tag{A.24}$$

Furthermore,

$$D(k_1^0, k_2; \hat{\rho}_2) \leq D(k_1^0, k_2) = O_p(1) \tag{A.25}$$

by Lemma 1 because $[k_1^0+1, k_2]$ involves a positive fraction of observations from a single true regime (i.e., $|k_2 - k_1^0| \geq aT$ for some $a > 0$ and $[k_1^0 + 1, k_2] \subset [k_1^0 + 1, k_2^0]$). Equations (A.23)-(A.25) imply that the first factor on the right hand side of (A.20) is $O_p(1)$.

Next, consider the second factor on the right hand side of (A.20). Let $\hat{\rho}_3 = (\hat{\theta}_3, \hat{\Sigma}_3)$ maximize $D(k_2, k_3; \theta, \Sigma)$. Then

$$D(k_2, k_3) = D(k_2, k_2^0; \hat{\rho}_3) \cdot D(k_2^0, k_3; \hat{\rho}_3)$$

By Lemma 9, $\sqrt{T}(\hat{\rho}_3 - \rho_3^0) = O_p(1)$ because the regime misspecification is bounded by $O(\sqrt{T}v_T^{-1})$ observations (the dominating regime is $(k_2^0, k_3^0]$, and $|k_j - k_j^0| \leq \sqrt{T}v_T^{-1}$ for $j = 2, 3$).

Note that for the segment $(k_2, k_2^0]$, the true parameter is ρ_2^0. Thus

$$\begin{aligned} D(k_2, k_2^0; \hat{\rho}_3) &= D(k_2, k_2^0; \rho_2^0 + (\rho_3^0 - \rho_2^0) + (\hat{\rho}_3 - \rho_3^0)) \\ &= \mathcal{L}(k_2, k_2^0; \sqrt{T}(\rho_3^0 - \rho_2^0) + \sqrt{T}(\hat{\rho}_3 - \rho_3^0)) \\ &\leq \sup_{|u| \leq M} \mathcal{L}(k_2, k_2^0; \sqrt{T}(\rho_3^0 - \rho_2^0) + u) \end{aligned}$$

with large probability for large M because $\sqrt{T}(\hat{\rho}_3 - \rho_3^0) = O_p(1)$. For $|u| \leq M$, by assumption A3 and A4, $\|\sqrt{T}(\rho_3^0 - \rho_2^0) + u\| \geq \|\sqrt{T}(\rho_3^0 - \rho_2^0)\| - \|u\| \geq \frac{1}{2}\|\sqrt{T}(\rho_3^0 - \rho_2^0)\| \geq b\sqrt{T}v_T$ for some $b > 0$. By the definition of $A_2(C)$, $k_2^0 - k_2 \geq Cv_T^{-2}$. Apply Lemma 5 (with reversed data order) with $h_T = v_T^{-2}$, $A = C$, $d_T = b\sqrt{T}v_T$, and $(\theta, \Sigma) = \sqrt{T}(\rho_3^0 - \rho_2^0) + u$, we see that for every $\epsilon > 0$, there exists $C > 0$ such that,

$$P\Big(\sup_{k_2 \leq k_2^0 - Cv_T^2} D(k_2, k_2^0; \hat{\rho}_3) > \epsilon \Big) < \epsilon. \tag{A.26}$$

Finally, because $D(k_2^0, k_3; \hat{\rho}_3) \leq D(k_2^0, k_3) = O_p(1)$ by Lemma 1 (because $k_3 - k_2^0 \geq Ta$ for some $a > 0$). Thus the second factor on the right hand side of (A.20) is $O_p(1) \cdot D(k_2, k_2^0; \hat{\rho}_3)$. Since the first factor

is already shown to be $O_p(1)$, we see that (A.20) is bounded by $O_p(1) \cdot D(k_2, k_2^0; \hat{\rho}_3)$. In view of (A.26), we prove (A.19) and thus the proposition. ▮

Proof of Theorem 1 and Theorem 3. The rate convergence is implied by Proposition 2, which holds for fixed v_T as well as shrinking v_T. The rate of convergence of the estimated regression parameters and covariance matrices is a consequence of the fast rate of convergence of the estimated break points. See Bai (1997) and Bai and Perron (1998) for details. ▮

Proof of Theorem 2. Notice

$$\hat{k}_i = \text{argmax}_\ell \ \Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, \ell, \hat{k}_{i+1}, ..., \hat{k}_m). \tag{A.27}$$

or equivalently,

$$\hat{k}_i - k_i^0 = \text{argmax}_\ell \ \Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0 + \ell, \hat{k}_{i+1}, ..., \hat{k}_m). \tag{A.28}$$

$$= \text{argmax}_\ell \ D(\hat{k}_{i-1}, k_i^0 + \ell) \cdot D(k_i^0 + \ell, \hat{k}_{i+1})$$

$$= \text{argmax}_\ell \ \frac{D(\hat{k}_{i-1}, k_i^0 + \ell) \cdot D(k_i^0 + \ell, \hat{k}_{i+1})}{D(\hat{k}_{i-1}, k_i^0) \cdot D(k_i^0, \hat{k}_{i+1})}.$$

The behavior of the above expression as a function of ℓ will be examined. We focus on the case of $\ell > 0$. The case of $\ell < 0$ is similar.

LEMMA 11. *For v_T fixed or $v_T \to 0$ but satisfying (8), we have uniformly in ℓ $(0 \le \ell \le M v_T^{-2})$*

$$\frac{D(\hat{k}_{i-1}, k_i^0 + \ell) \cdot D(k_i^0 + \ell, \hat{k}_{i+1})}{D(\hat{k}_{i-1}, k_i^0) \cdot D(k_i^0, \hat{k}_{i+1})} = \frac{D(k_i^0, k_i^0 + \ell; \rho_i^0)}{D(k_i^0, k_i^0 + \ell; \rho_{i+1}^0)}(1 + o_p(1)). \tag{A.29}$$

Proof. Let $\hat{\rho}_i$ and $\hat{\rho}_{i+1}$ be the estimators of ρ_i^0 and ρ_{i+1}^0 based on the subsamples $(\hat{k}_{i-1}, k_i^0 + \ell]$ and $(k_i^0 + \ell, \hat{k}_{i+1}]$, respectively (these estimators depend on ℓ. But for notational simplicity, the dependence is suppressed). Similarly, let $\hat{\rho}_i^*$ and $\hat{\rho}_{i+1}^*$ be the estimators of ρ_i^0 and ρ_{i+1}^0 based on the subsamples $(\hat{k}_{i-1}, k_i^0]$ and $(k_i^0, \hat{k}_{i+1}]$, respectively. Break up the segmented-likelihood ratio $D(\hat{k}_{i-1}, k_i^0 + \ell)$ into two segments both evaluated at $\hat{\rho}_i$ such that

$$D(\hat{k}_{i-1}, k_i^0 + \ell) = D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i) \cdot D(k_i^0, k_i^0 + \ell; \hat{\rho}_i)$$

and, similarly

$$D(k_i^0, \hat{k}_{i+1}) = D(k_i^0, k_i^0 + \ell; \hat{\rho}_{i+1}^*) \cdot D(k_i^0 + \ell, \hat{k}_{i+1}; \hat{\rho}_{i+1}^*).$$

In addition, by definition, $D(\hat{k}_{i-1}, k_i^0) = D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i^*)$ and $D(k_i^0+\ell, \hat{k}_{i+1}) = D(k_i^0 + \ell, \hat{k}_{i+1}; \hat{\rho}_{i+1})$. Thus the left side of (A.29) can be rewritten as

$$\frac{D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i)}{D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i^*)} \cdot \frac{D(k_i^0, k_i^0 + \ell; \hat{\rho}_i)}{D(k_i^0, k_i^0 + \ell; \hat{\rho}_{i+1}^*)} \cdot \frac{D(k_i^0 + \ell, \hat{k}_{i+1}; \hat{\rho}_{i+1})}{D(k_i^0 + \ell, \hat{k}_{i+1}; \hat{\rho}_{i+1}^*)}. \qquad \text{(A.30)}$$

Next consider the first term of (A.30). By Lemma 10, uniformly in $|\ell| \leq Mv_T^{-2}$,

$$\hat{\rho}_j - \rho_j^0 = O_p(T^{-1/2}) \quad (j = i, i+1),$$

$$\hat{\rho}_j^* - \rho_j^0 = O_p(T^{-1/2}) \quad (j = i, i+1),$$

$$\hat{\rho}_j - \hat{\rho}_j^* = O_p(T^{-1}) \quad (j = i, i+1).$$

Using these results, we will show the first term of (A.30) is $1 + o_p(1)$. Suppose $\hat{k}_{i-1} \geq k_{i-1}^0$, then we can use Lemma 8 to show it is $1 + o_p(1)$. To see this, the denominator can be written as $\mathcal{L}(\hat{k}_{i-1}, k_i^0, \sqrt{T}(\hat{\rho}_i^* - \rho_i^0))$ and the numerator as $\mathcal{L}(\hat{k}_{i-1}, k_i^0, \sqrt{T}(\hat{\rho}_i^* - \rho_i^0) + T^{-1/2}[T(\hat{\rho}_i - \hat{\rho}_i^*)])$. (Note the true parameter for this segment is ρ_i^0.) Now take $(\phi, \Upsilon) = \sqrt{T}(\hat{\rho}_i^* - \rho_i^0) = O_p(1)$ and $(\lambda, \Xi) = T(\hat{\rho}_i - \hat{\rho}_i^*) = O_p(1)$, the desired result follows readily from Lemma 8 (note that Lemma 8 is stated in terms of log-valued likelihood. Without taking logarithm, it is $1 + o_p(1)$). Now suppose that $\hat{k}_{i-1} \leq k_{i-1}^0$. Then

$$\frac{D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i)}{D(\hat{k}_{i-1}, k_i^0; \hat{\rho}_i^*)} = \frac{D(\hat{k}_{i-1}, k_{i-1}^0, \hat{\rho}_i)}{D(\hat{k}_{i-1}, k_{i-1}^0; \hat{\rho}_i^*)} \cdot \frac{D(k_{i-1}^0, k_i^0; \hat{\rho}_i)}{D(k_{i-1}^0, k_i^0; \hat{\rho}_i^*)}.$$

We can again apply Lemma 8 to the second term on the right. But for the first term on the right, we can apply Lemma 7 twice after dividing both the numerator and the denominator by $D(\hat{k}_{i-1}, k_{i-1}^0; \rho_i^0)$ and conclude the resulting two ratios are each $1 + o_p(1)$ so that itself is also $1 + o_p(1)$ (this idea is elaborated below for other terms).

The entire analysis above is also applicable for the third term of (A.30) and so it is also $1 + o_p(1)$.

Next, consider the middle term of (A.30), which can be rewritten as

$$\frac{D(k_i^0, k_i^0+\ell; \rho_i^0)}{D(k_i^0, k_i^0+\ell; \rho_{i+1}^0)} \cdot \frac{D(k_i^0, k_i^0+\ell; \hat{\rho}_i)}{D(k_i^0, k_i^0+\ell; \rho_i^0)} \Big/ \frac{D(k_i^0, k_i^0+\ell; \hat{\rho}_{i+1}^*)}{D(k_i^0, k_i^0+\ell; \rho_{i+1}^0)} \tag{A.31}$$

The last two terms of above are each $1+o_p(1)$ by Lemma 7. To see this, consider the middle term. The denominator is equal to $\mathcal{L}(k_i^0, k_i^0+\ell; \sqrt{T}(\rho_i^0 - \rho_{i+1}^0))$ and the numerator is equal to $\mathcal{L}(k_i^0, k_i^0+\ell; \sqrt{T}(\rho_i^0 - \rho_{i+1}^0) + \sqrt{T}(\hat{\rho}_i - \rho_i^0))$ (note the true parameter for the segment $(k_i^0, k_i^0+\ell]$ is ρ_{i+1}^0). Now take $(\phi, \Upsilon) = (\rho_i^0 - \rho_{i+1}^0)$, and $(\lambda, \Xi) = \sqrt{T}(\hat{\rho}_i - \rho_i^0)$, then the conditions of Lemma 7 are satisfied. Thus the middle term of (A.31) is $1+o_p(1)$ uniformly in $0 \le \ell \le M v_T^{-2}$. Summarizing these results, we obtain (A.29). This proves Lemma 11. ∎

For fixed v_T, $\hat{k}_i - k_i^0 = O_p(1)$. Thus to prove Theorem 2, it suffices to consider $|\ell| \le M$. Now, for $\ell > 0$, we have

$$\begin{aligned} &\frac{D(k_i^0, k_i^0+\ell; \rho_i^0)}{D(k_i^0, k_i^0+\ell; \rho_{i+1}^0)} \\ &= \frac{|\Sigma_i^0|^{-\ell/2} \exp(-\frac{1}{2}\sum_{k_i^0+1}^{k_i^0+\ell} [Y_t - (V_t' \otimes I)\theta_i^0]'(\Sigma_i^0)^{-1}[Y_t - (V_t' \otimes I)\theta_i^0]}{|\Sigma_{i+1}^0|^{-\ell/2} \exp(-\frac{1}{2}\sum_{k_i^0+1}^{k_i^0+\ell} \varepsilon_t'(\Sigma_{i+1}^0)^{-1}\varepsilon_t)}. \end{aligned}$$

For $t > k_i^0$, the true parameter is $(\theta_{i+1}^0, \Sigma_{i+1}^0)$. Thus $Y_t - (V_t' \otimes I)\theta_i^0 = \varepsilon_t + (V_t' \otimes I)(\theta_{i+1}^0 - \theta_i^0)$. From this, expanding and taking logarithm (logarithm transformation does not alter the value of the argmax functional), we obtain $W^{(i)}(\ell)$ given in (6). The case of $\ell < 0$ corresponds to $W_2^{(i)}(\ell)$ given in (7). In summary and in view of Lemma 11, we have

$$\log \frac{D(\hat{k}_{i-1}, k_i^0+\ell) \cdot D(k_i^0+\ell, \hat{k}_{i+1})}{D(\hat{k}_{i-1}, k_i^0) \cdot D(k_i^0, \hat{k}_{i+1})} \Rightarrow W^{(i)}(\ell) \tag{A.32}$$

on bounded set of ℓ. The assumption of continuous distribution of ε_t guarantees the uniqueness of the maximum value of $W^{(i)}(r)$. This implies that

$$\hat{k}_i - k_i^0 = \operatorname{argmax}_\ell \Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0+\ell, \hat{k}_{i+1}, ..., \hat{k}_m) \xrightarrow{d} \operatorname{argmax}_\ell W^{(\ell)}(\ell).$$

The detailed argument for the last claim can be found in Bai (1997). The proof of Theorem 2 is now complete. ∎

Proof of Theorem 4. From $\hat{k}_i = \operatorname{argmax}_k \Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k, \hat{k}_{i+1}, ..., \hat{k}_m)$ and $v_T^2(\hat{k}_i - k_i^0) = O_p(1)$, we consider the following parameterization:

$$\Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0 + [v v_T^{-2}], \hat{k}_{i+1}, ..., \hat{k}_m)$$

for v on an arbitrary bounded set. Consider

$$\begin{aligned}
&\log \frac{\Lambda_T(\hat{k}_1,...,\hat{k}_{i-1},k_i^0+[vv_T^{-2}],\hat{k}_{i+1},...,\hat{k}_m)}{\Lambda_T(\hat{k}_1,...,\hat{k}_{i-1},k_i^0,\hat{k}_{i+1},...,\hat{k}_m)} \\
&= \log \frac{D(\hat{k}_{i-1},k_i^0+[vv_T^{-2}])\cdot D(k_i^0+[vv_T^{-2}],\hat{k}_{i+1})}{D(\hat{k}_{i-1},k_i^0)\cdot D(k_i^0,\hat{k}_{i+1})} \\
&= W^{(i)}([vv_T^{-2}])+o_p(1),
\end{aligned}$$

The second equality follows from Lemma 11 (also cf. (A.32)). To prove Theorem 4, it is sufficient to prove that $W^{(i)}([vv_T^{-2}]) \Rightarrow \Lambda^{(i)}(v)$. The latter process is defined in the main text.

Let $r=[vv_T^{-2}]$ for $v\geq 0$. Then $r\geq 0$. The first term on the right side of (6) is

$$\begin{aligned}
-\frac{r}{2}\Big(\log|\Sigma_i^0|-\log|\Sigma_{i+1}^0|\Big) &= \frac{r}{2}\log\Big|[\Sigma_i^0+(\Sigma_{i+1}^0-\Sigma_i^0)](\Sigma_i^0)^{-1}\Big| \\
&= \frac{r}{2}\log|[\Sigma_i^0+v_T\Phi_i](\Sigma_i^0)^{-1}| \\
&= \frac{r}{2}\log|I+v_T\Phi_i(\Sigma_i^0)^{-1}| \qquad \text{(A.33)} \\
&= \frac{r}{2}v_T\,\mathrm{tr}(\Phi_i(\Sigma_i^0)^{-1})-\frac{r}{4}v_T^2\,\mathrm{tr}([\Phi_i(\Sigma_i^0)^{-1}]^2)+o(v_T^2).
\end{aligned}$$

Next, consider the second term on the right side of (6). Note $E\varepsilon_t\varepsilon_t'=\Sigma_{i+1}^0$ for $t\in(k_i^0,k_{i+1}^0]$. Subtracting and adding Σ_{i+1}^0 and noting that $[(\Sigma_i^0)^{-1}-(\Sigma_{i+1}^0)^{-1}]\Sigma_{i+1}^0=(\Sigma_i^0)^{-1}(\Sigma_{i+1}^0-\Sigma_i^0)=v_T(\Sigma_i^0)^{-1}\Phi_i$, we have

$$\begin{aligned}
&-\frac{1}{2}\sum_{k_i^0+1}^{k_i^0+r}\varepsilon_t'\Big((\Sigma_i^0)^{-1}-(\Sigma_{i+1}^0)^{-1}\Big)\varepsilon_t \\
&= -\frac{1}{2}\mathrm{tr}\left(\sum_{k_i^0+1}^{k_i^0+r}\Big[(\Sigma_i^0)^{-1}-(\Sigma_{i+1}^0)^{-1}\Big]\Big[\Sigma_{i+1}^0+(\varepsilon_t\varepsilon_t'-\Sigma_{i+1}^0)\Big]\right) \qquad \text{(A.34)} \\
&= -\frac{r}{2}\cdot v_T\,\mathrm{tr}(\Phi_i(\Sigma_i^0)^{-1}) \\
&- \frac{1}{2}tr\Big[(\Sigma_{i+1}^0)^{1/2}\Big((\Sigma_i^0)^{-1}-(\Sigma_{i+1}^0)^{-1}\Big)(\Sigma_{i+1}^0)^{1/2}\sum_{k_i^0+1}^{k_i^0+r}(\eta_t\eta_t'-I)\Big] \\
&= -\frac{r}{2}\cdot v_T\,\mathrm{tr}(\Phi_i(\Sigma_i^0)^{-1}) \\
&- \frac{1}{2}tr\Big[(\Sigma_{i+1}^0)^{1/2}(\Sigma_i^0)^{-1}\Phi_i(\Sigma_{i+1}^0)^{-1/2}v_T\sum_{k_i^0+1}^{k_i^0+r}(\eta_t\eta_t'-I)\Big]
\end{aligned}$$

The first two terms on the right side of (6) is equal to, by combining (A.33) and (A.34)

$$- \frac{r}{4} v_T^2 \operatorname{tr}([\Phi_i(\Sigma_i^0)^{-1}]^2)$$

$$- \frac{1}{2} tr\Big[(\Sigma_{i+1}^0)^{1/2}(\Sigma_i^0)^{-1}\Phi_i(\Sigma_{i+1}^0)^{-1/2} v_T \sum_{k_i^0+1}^{k_i^0+r}(\eta_t\eta_t' - I)\Big]$$

Since $\Sigma_i^0 \to \Sigma_0$ for all i, we have $\operatorname{tr}([\Phi_i(\Sigma_i^0)^{-1}]^2) \to \operatorname{tr}([\Phi_i(\Sigma_0)^{-1}]^2) = \operatorname{tr}(A_i^2)$. In addition, $(\Sigma_{i+1}^0)^{1/2}(\Sigma_i^0)^{-1}\Phi_i(\Sigma_{i+1}^0)^{-1/2} \to A_i$. Furthermore, $r\, v_T^2 = [vv_T^{-2}]v_T^2 \to v$ uniformly over bounded v. Finally, from (9), $v_T \sum_{k_i^0+1}^{k_i^0+r}(\eta_t\eta_t' - I) \Rightarrow \xi_1(v)$ for $r = [vv_T^{-2}]$. Combining these results and noting that $\xi_1(v)$ and $-\xi_1(v)$ have the same distribution, we obtain the first two expressions of $\Lambda^{(i)}(v)$ defined in (13). Next, consider the last two terms of (6). From $\Delta\theta_i = v_T\delta_i$,

$$\Delta\theta_i' \sum_{k_i^0+1}^{k_i^0+r}(V_t' \otimes I)(\Sigma_i^0)^{-1}\varepsilon_t = \delta_i' v_T \sum_{k_i^0+1}^{k_i^0+[vv_T^{-2}]}[V_t \otimes (\Sigma_i^0)^{-1/2}]\eta_t$$

$$\Rightarrow \delta_i' Q^{1/2}\zeta_1(v), \qquad v > 0$$

by (12). Next, by (11),

$$\Delta\theta_i' \sum_{k_i^0+1}^{k_i^0+r}[V_tV_t' \otimes (\Sigma_i^0)^{-1}]\Delta\theta_i = \delta_i' v_T^2 \sum_{k_i^0+1}^{k_i^0+[vv_T^{-2}]}[V_tV_t' \otimes (\Sigma_i^0)^{-1}]\delta_i \longrightarrow v\,\delta_i' Q \delta_i$$

Combining these results, we have

$$W_1^{(i)}([vv_T^{-2}]) \Rightarrow \Lambda^{(i)}(v) \qquad , v > 0.$$

The proof for $v < 0$ is similar. ▌

Proof of Theorem 5. Let $\hat{v} = \operatorname{argmax}_v \Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0 + [vv_T^{-2}], \hat{k}_{i+1}, ..., \hat{k}_m)$. This implies that $\hat{k}_i - k_i^0 = [\hat{v}v_T^{-2}]$. From $|[x] - x| \le 1$, $|v_T^2(\hat{k}_i - k_i^0) - \hat{v}| = |v_T^2[\hat{v}v_T^{-2}] - \hat{v}| \le v_T^2 \to 0$. Thus we have $v_T^2(\hat{k}_i - k_i^0) = \hat{v} + o_p(1)$. Note the identity

$$\hat{v} = \operatorname{argmax}_v \log \frac{\Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0 + [vv_T^{-2}], \hat{k}_{i+1}, ..., \hat{k}_m)}{\Lambda_T(\hat{k}_1, ..., \hat{k}_{i-1}, k_i^0, \hat{k}_{i+1}, ..., \hat{k}_m)}. \tag{A.35}$$

This is because the denominator does not depend on v and $\log(x)$ is an increasing function. Theorem 5 then follows from Theorem 4 and the continuous mapping theorem. We point out that the argmax functional generally is not a continuous functional. But it becomes continuous on the set of continuous functions defined on a bounded set with each function having a unique maximum. Because we already established the fact that $\hat{v} = O_p(1)$, we can limit the domain of the argmax functional on functions defined on bounded sets. This illustrates the importance of establishing the rate of convergence for $\hat{v}$. The sample path of $\Lambda^{(i)}(v)$ is continuous and has a unique maximum with probability one. Bai (1992) gives a thorough discussion on the argmax functional. ▮

Proof of Corollary 3. We use the fact that if $X \sim N(0, A)$, then $b'X \sim (b'Ab)^{1/2}N(0,1)$ (this is true even if A is a singular matrix). It follows that $\mathrm{tr}(A_i\xi(v)) = \mathrm{vec}(A_i)'\mathrm{vec}(\xi(v)) \stackrel{d}{=} (\bar{a}'\Omega\bar{a})^{1/2}U(v)$ and $\delta_i'Q^{1/2}\zeta(v) \stackrel{d}{=} (\delta_i'Q\delta_i)^{1/2}V(v)$, where $U(v)$ and $V(v)$ are standard two-sided Brownian motions and $\Omega = E\mathrm{vec}(\xi(1))\mathrm{vec}(\xi(1))'$. Note that $E\zeta(1)\zeta(1)' = I$, an identity matrix because $\zeta(v)$ is a vector of independent standard Brownian motions. The assumption on the third moment implies the independence of $\xi(v)$ and $\zeta(v)$. Thus, $2^{-1}\mathrm{tr}(A_i\xi(v)) + \delta_i'Q^{1/2}\zeta(v) \stackrel{d}{=} bB(v)$ where $b = [4^{-1}\bar{a}'\Omega\bar{a} + \delta_i'Q\delta_i]^{1/2}$, and $B(v)$ is also a standard two-sided Brownian motion. Thus $\Lambda^{(i)}(v)$ is equal (in distribution) to $-|v|2^{-1}c + bB(v)$ where $c = 2^{-1}\mathrm{tr}(A_i^2) + (\delta_i'Q\delta_i)$. By a change in variable, it can be shown that $\mathrm{argmax}_v\{-|v|2^{-1}c + bB(v)\} \stackrel{d}{=} (b^2/c^2)\mathrm{argmax}_s\{-|s|2^{-1} + B(s)\}$. This implies that

$$v_T^2(\hat{k}_i - k_i^0) \stackrel{d}{\longrightarrow} (b^2/c^2)\mathrm{argmax}_s\{-|s|2^{-1} + B(s)\}.$$

Equivalently, $(c^2/b^2)v_T^2(\hat{k}_i - k_i^0) \stackrel{d}{\longrightarrow} \mathrm{argmax}_s\{-|s|2^{-1} + B(s)\}$. This proves Corollary 3. ▮

Proof of Corollary 4. From Corollary 3, it is seen that the scaling factor of $\hat{k}_i - k_i^0$ (multiplying both the numerator and the denominator by v_T^2) can be written as

$$\frac{\left(2^{-1}\mathrm{tr}([A_iv_T]^2) + v_T\delta_i'Q\delta_iv_T\right)^2}{4^{-1}\mathrm{vec}(A_iv_T)'\Omega\,\mathrm{vec}(A_iv_T) + v_T\delta_i'Q\delta_iv_T}$$

Because $\Phi_iv_T = \Sigma_{i+1}^0 - \Sigma_i^0$ and $\delta_iv_T = \theta_{i+1}^0 - \theta_i^0$, this scaling factor can be rewritten as

$$\frac{\left(2^{-1}\mathrm{tr}(B_i^2) + (\theta_{i+1}^0 - \theta_i^0)'Q(\theta_{i+1}^0 - \theta_i^0)\right)^2}{4^{-1}\mathrm{vec}(B_i)'\Omega\,\mathrm{vec}(B_i) + (\theta_{i+1}^0 - \theta_i^0)'Q(\theta_{i+1}^0 - \theta_i^0)} \tag{A.36}$$

where $B_i = \Sigma_0^{-1/2}(\Sigma_{i+1}^0 - \Sigma_i^0)\Sigma_0^{-1/2}$. This follows from $A_i v_T = \Sigma_0^{-1/2}\Phi_i\Sigma_0^{-1/2}v_T = B_i$. Let $\hat{\Sigma}_i$ and $\hat{\theta}_i$ be estimators such that $\hat{\Sigma}_i - \Sigma_i^0 = O_p(T^{-1/2})$ and $\hat{\theta}_j - \theta_j^0 = O_p(T^{-1/2})$. Define $\hat{B}_i = \hat{\Sigma}_i^{-1/2}(\hat{\Sigma}_{i+1} - \hat{\Sigma}_i)\hat{\Sigma}_i^{-1/2}$ then $\hat{B}_i$ is consistent for B_i. In addition, if $\hat{\Omega}$ and $\hat{Q}$ are consistent estimators of Ω and Q, respectively, then we have

$$(\hat{\alpha}_i - \alpha_i)(\hat{k}_i - k_i^0) = o_p(1), \tag{A.37}$$

where α_i denotes the whole expression of (A.36) and $\hat{\alpha}$ is the estimated version. Note that it is not sufficient to have $\hat{\alpha}_i - \alpha_i = o_p(1)$ because $\hat{k}_i - k_i^0 = O_p(v_T^{-2})$, which converges to infinity when v_T converges to zero. However, it is also true that $\theta_{i+1}^0 - \theta_i^0 = O(v_T)$ and $\Sigma_{i+1}^0 - \Sigma_i^0 = O(v_T)$, which converge to zero. Together with the rate of convergence of $\hat{\Sigma}_i$ and $\hat{\theta}_i$, equation (A.37) can be proved in a routine fashion (by adding and subtracting terms). In fact, it can be shown that the left hand side of (A.37) is of $O_p(1/(\sqrt{T}v_T))$. The proof of Corollary 4 is complete. ▌

Proof of Corollary 5. This corollary is a special case of Corollary 4. ▌

Proof of Corollary 6. We first prove that under normality, the scaling factor of $v_T^2(\hat{k}_i^0 - k_i^0)$ in Corollary 3 is simplified to $2^{-1}\mathrm{tr}(A_i^2) + \delta_i' Q \delta_i'$. It is sufficient to show $\mathrm{vec}(A_i)'\Omega\mathrm{vec}(A_i) = 2\mathrm{tr}(A_i^2)$. If this is the case, the numerator will be the square of the denominator and the desired simplification follows. Let a_{kl} denote the (k,l)th entry of A_i. Then $\mathrm{tr}(A_i^2) = \sum_k a_{kk}^2 + 2\sum_{k<l} a_{kl}^2$. Recall that $\mathrm{vec}(A_i)'\Omega\mathrm{vec}(A_i)$ is the variance of $\mathrm{tr}(A_i\xi(1))$. Let ψ_{kl} denote the (k,l)th entry of $\xi(v)$ for $v = 1$. Because $\xi(v)$ is the limiting distribution of $v_T\sum(\eta_t\eta_t - I)$, $\mathrm{vec}(\xi(1))$ has the same covariance matrix as $\mathrm{vec}(\eta_t\eta_t' - I)$. Now $\mathrm{tr}(A_i\xi(1)) = \sum a_{kk}\psi_{kk} + 2\sum_{k<l} a_{kl}\psi_{kl}$. So $\mathrm{vec}(A_i)'\Omega\mathrm{vec}(A_i) = E(\sum a_{kk}\psi_{kk} + 2\sum_{k<l} a_{kl}\psi_{kl})^2 = E[\sum a_{kk}(\eta_{tk}^2 - 1) + 2\sum_{k<l} a_{kl}\eta_{tk}\eta_{tl}]^2 = 2\sum a_{kk}^2 + 4\sum_{k<l} a_{kl}^2 = 2\mathrm{tr}(A_i^2)$. We have used the fact that under normality $(\{\eta_{tk}^2 - 1\}_{k=1}^q, \{\eta_{tk}\eta_{tl}\}_{k<l})$ is a vector of uncorrelated random variables with $E(\eta_{tk}^2 - 1)^2 = 2$ and $E(\eta_{tk}^2\eta_{tl}^2) = 1$. This proves the desired simplification under normality. As for the second part of the corollary, the argument is identical to the proof of Corollary 4. ▌

Proof of Corollary 7. This corollary is a special case of Corollary 6. ▌

To prove Theorem 6, we need additional results.

LEMMA 12. *Let h_T, d_T and S_T be the same as in Lemma 5, then for every given $A > 0$,*

$$\sup_{1\le k\le Ah_T}\ \sup_{(\theta,\Sigma)\in S_T^c} \mathcal{L}(0,k;\theta,\Sigma) = O_p(1).$$

where S_T^c is the complement set of S_T.

Proof. See Property 6 of BLS. ▮

LEMMA 13. *Let m^0 be the true number of change points, and let $(\hat{k}_1, ..., \hat{k}_{m^0})$ be the estimator defined in Section 3. Then*

$$\Lambda_T(\hat{k}_1, ..., \hat{k}_{m^0}) = O_p(1).$$

This lemma says that the optimal likelihood ratio is stochastically bounded. Because $\Lambda_T(\hat{k}_1, ..., \hat{k}_{m^0}) \geq 1$, the log-valued optimal likelihood ratio is also stochastically bounded.

Proof. From $\Lambda_T(\hat{k}_1, ..., \hat{k}_{m^0}) = D(0, \hat{k}_1)D(\hat{k}_1, \hat{k}_2) \cdots D(\hat{k}_{m^0}, T)$, we shall prove each of $D(\cdot, \cdot)$ is $O_p(1)$. Consider $D(\hat{k}_1, \hat{k}_2)$ for concreteness. Let $\hat{\rho}_2 = (\hat{\theta}_2, \hat{\Sigma}_2)$ be the estimator of (ρ_2^0, Σ_2^0) based on the segment $(\hat{k}_1, \hat{k}_2]$. The estimator is root-T consistent by Theorem 3. Suppose that $\hat{k}_1 \leq k_1^0$ and $\hat{k}_2 \geq k_2^0$, which is the most complicated situation. Then

$$D(\hat{k}_1, \hat{k}_2) = D(\hat{k}_1, k_1^0; \hat{\rho}_2) \cdot D(k_1^0, k_2^0; \hat{\rho}_2) \cdot D(k_2^0, \hat{k}_2; \hat{\rho}_2).$$

The middle term on the right hand side is $O_p(1)$ because $D(k_1^0, k_2^0; \hat{\rho}_2) \leq D(k_1^0, k_2^0) = O_p(1)$ by Lemma 1 (single true regime and positive fraction of observations). Consider the third term.

$$D(k_2^0, \hat{k}_2; \hat{\rho}_2) = \mathcal{L}(k_2^0, \hat{k}_2; \sqrt{T}(\hat{\rho}_2 - \rho_2^0) + \sqrt{T}(\rho_2^0 - \rho_3^0)). \tag{A.38}$$

Now $\|\sqrt{T}(\hat{\rho}_2 - \rho_2^0) + \sqrt{T}(\rho_2^0 - \rho_3^0)\| \leq 2\|\sqrt{T}(\rho_2^0 - \rho_3^0)\| \leq C\sqrt{T}v_T$ by Assumption A4. In addition, $\hat{k}_2 - k_2^0 \leq Mv_T^{-2}$ by Theorem 3. By Lemma 12, applied with $h_T = v_T^{-2}$, $d_T = C\sqrt{T}v_T$, and $A = M$ (note that $\sqrt{T}(\hat{\rho}_2 - \rho_2^0) + \sqrt{T}(\rho_2^0 - \rho_3^0)$ is in the set S_T^c), we see the right hand side of (A.38) is $O_p(1)$. Similarly, $D(\hat{k}_1, k_1^0; \hat{\rho}_2) = O_p(1)$. Thus $D(\hat{k}_1, \hat{k}_2) = O_p(1)$. ▮

Proof of Theorem 6. We first show that $P(\hat{m} < m^0) \to 0$. It is sufficient to show $P(\min_{m<m^0} \text{BIC}(m) - \text{BIC}(m^0) \leq 0) \to 0$ as T increases. Let $m < m^0$, and let $(\hat{k}_1^*, ..., \hat{k}_{m^0}^*)$ be the optimal estimator of the break points with m^0 known. We have

$$\text{BIC}(m) - \text{BIC}(m^0) = -\log L(\hat{k}_1, ..., \hat{k}_m) + \log L(\hat{k}_1^*, ..., \hat{k}_{m^0}^*) + (m - m^0)g(T)$$

$$= -\log \Lambda_T(\hat{k}_1, ..., \hat{k}_m) + \log \Lambda_T(\hat{k}_1^*, ..., \hat{k}_{m^0}^*) + (m - m^0)g(T).$$

The second equality follows from by adding and subtracting $\log \Pi_{t=1}^T f(\varepsilon_t)$ and by the definition of Λ_T. The second term on the right hand side does

not depend on m and it is $O_p(1)$ by Lemma 13. Thus

$$\mathrm{BIC}(m) - \mathrm{BIC}(m^0) = -\log \Lambda_T(\hat{k}_1, ..., \hat{k}_m) + (m - m^0)g(T) + O_p(1).$$

When $m < m^0$, there must exist at least one change point that cannot be consistently estimated. This implies that the model parameters for some regime cannot be consistently estimated. That is, there exists a segment $(k, \ell]$ which satisfies (i) $\ell - k \geq T\delta$ for some $\delta > 0$, (ii) $(k, \ell]$ is a subset of the intersection $(\hat{k}_{h-1}, \hat{k}_h] \cap (k^0_{i-1}, k^0_i]$ for some h and i, and (iii) $\sqrt{T}\|\hat{\theta}_h - \theta^0_i\| \geq b\sqrt{T}v_T$ or $\sqrt{T}\|\hat{\Sigma} - \Sigma^0_i\| \geq b\sqrt{T}v_T$ for some $b > 0$. By Lemma 4 or (A.3), the likelihood ratio for this segment

$$\log D(k, \ell; \hat{\theta}_h, \hat{\Sigma}_h) = \log \mathcal{L}(k, \ell; \sqrt{T}(\hat{\theta}_h - \theta^0_i), \sqrt{T}(\hat{\Sigma}_h - \Sigma^0_i))$$

is less than $-cTv_T^2$ for some $c > 0$ with large probability. By Lemma 2, the maximum value of the log-likelihood ratio for all other segments is at most $O_p(\log T)$. Thus $\mathrm{BIC}(m) - \mathrm{BIC}(m^0) \geq cTv_T^2 - |O_p(\log T)| - |m^0 - m|g(T) - |O_p(1)| \geq cTv_T^2 - m^0 g(T) - |O_p(\log T)| \to +\infty$ with probability tending to 1. This implies that $P(\hat{m} < m^0) \to 0$.

Next, we show $P(\hat{m} > m^0) \to 0$. Suppose $m > m^0$. By Lemma 2, when adding a break point in estimation, when there is in fact no break point, the log-likelihood ratio is increased at most by $O_p(\log T)$. Thus, for $m > m^0$ and $m \leq M$,

$$\mathrm{BIC}(m) - \mathrm{BIC}(m^0) = -\log \Lambda(\hat{k}_1, ..., \hat{k}_m) + (m - m^0)g(T) + O_p(1)$$

$$\geq -\log \Lambda(\hat{k}_1, ..., \hat{k}_M) + g(T) + O_p(1) \geq -|O_p(\log T)| + g(T) \to +\infty.$$

This implies that $P(\hat{m} > m^0) \to 0$. The proof of Theorem 6 is complete. ▌

APPENDIX: COMPUTATION

Assume m is known. For each set of hypothesized change points $(k_1, ..., k_m)$, let $\hat{\theta}_i = \hat{\theta}_i(k_1, ..., k_m)$ be the estimator of θ^0_i using the segment $(k_{i-1}, k_i]$ (cf. model (2)). Define

$$\hat{\Sigma}_i(k_1, ..., k_m) = \frac{1}{k_i - k_{i-1}} \sum_{t=k_{i-1}}^{k_i} [Y_t - (V_t \otimes \hat{\theta}_i)][Y_t - (V_t \otimes \hat{\theta}_i)]'$$

Then the log-valued quasi-likelihood as a function of $(k_1, ..., k_m)$ is simply

$$\log L(k_1, ..., k_m) = \sum_{i=1}^{m+1} (k_i - k_{i-1}) \log \det[\hat{\Sigma}_i(k_1, ..., k_m)]$$

where $\det(B)$ is the determinant of matrix B. Thus the change point estimator is

$$(\hat{k}_1, ..., \hat{k}_m) = \text{argmin}_{k_1,...,k_m} \log L(k_1, ..., k_m).$$

When $m > 1$, efficient algorithm based on dynamic programming is available. Note that $\hat{\theta}_i(k_1, ..., k_m)$ and $\hat{\Sigma}_i(k_1, ..., k_m)$ actually only depend on k_{i-1} and k_i. The dynamic programming algorithm requires the calculation of the following number for each segment $(k, \ell]$,

$$(\ell - k) \log \det \hat{\Sigma}(k, \ell).$$

Once this is computed for all segments of $(k, \ell]$ (there are at most $T(T-1)/2$ distinct segments with at least two observations), the algorithm can quickly search the optimal change points based on the principle of optimality. Details are given in Bai and Perron (1999) for univariate series with least squares estimation. But the idea is applicable for quasi-maximum likelihood method. Their computer code is also available upon request.

REFERENCES

Alogoskoufis, G. and R. Smith, 1991, The Phillips curve, the persistence of inflation, and the Lucas critique: Evidence from exchange rate regimes. *American Economic Review* **81**, 1254-1275.

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

Bai, J., 1992, Estimation of Structural Change in Econometric Models. (Unpublished Ph.D. dissertation, University of California, Berkeley).

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

Bai, J., 1997, Estimation of a change point in multiple regression models. *Review of Economics and Statistics* **79**, 551-563.

Bai, J. and P. Perron, 1998, Estimating and testing for multiple structural changes in linear models. *Econometrica* **66**, 47-78.

Bai, J. and P. Perron, 1999, Computations and analysis of multiple structural changes. Unpublished manuscript, Department of Economics, Boston University.

Bai, J., R. Lumsdaine, and J. Stock, 1998, Testing for and dating common breaks in multivariate time series. *Review of Economic Studies* **65**, 395-432.

Chow, G., 1960, Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28**, 591-605.

Feldstein, M. and J. H. Stock, 1993, The use of monetary aggregate to target nominal GDP. *NBER Working Paper* No. 4304.

Ng, S. and T. Vogelsang, 1997, Vector autoregressive and mean shifts. Department of Economics, Boston College.

Picard, D., 1985, Testing and estimating change-points in time series. *Advances in Applied Probability* **176**, 841-867.

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

Quandt, R. E., 1960, Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* **55**, 332-330.

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

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

Stock, J. H., 1994, Unit roots, structural breaks, and trends. In *Handbook of Econometrics*. Edited by Engle, R. and D. McFadden. **Vol IV**, 2740-2843. Amsterdam: Elsevier.