Estimation of Large Dimensional Factor Models with an Unknown Number of Breaks

Shujie Ma
Department of Statistics, University of California-Riverside
and
Liangjun Su
School of Economics, Singapore Management University

July 18, 2015

Abstract

In this paper we study the estimation of a large dimensional factor model when the factor loadings exhibit an unknown number of changes over time. We propose a novel three-step procedure to detect the breaks if any and then identify their locations. In the first step, we divide the whole time span into subintervals and fit a conventional factor model on each interval. In the second step, we apply the adaptive fused group Lasso to identify intervals containing a break. In the third step, we devise a grid search method to estimate the location of the break on each identified interval. We show that with probability approaching one our method can identify the correct number of changes and estimate the break locations. Simulation studies indicate superb finite sample performance of our method. We apply our method to investigate Stock and Watson’s (2009) U.S. monthly macroeconomic data set and identify five breaks in the factor loadings, spanning 1959-2006.

Keywords: Break point; Convergence rates; Factor model; Fused Lasso; Group Lasso; Information criterion; High-dimension; Principal component; Structural change; Time-varying parameter.

Correspondence should be addressed to Liangjun Su. Shujie Ma is Assistant Professor, Department of Statistics, University of California-Riverside, Riverside, CA 92521; E-mail: shujie.ma@ucr.edu; Phone: +951 827 3730. Liangjun Su is Professor, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg; Phone: +65 6828 0386. Ma’s research was partially supported by NSF grant DMS 1306972 and Hellman Fellowship. Su gratefully acknowledges the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021.
1 Introduction

High-dimensional time series data are frequently encountered in modern statistical and econometric studies, and they may be one of the most common types of data in the “big data” era. Examples come from many fields including economics, finance, genomics, environmental study, medical study, meteorology, chemometrics, and so forth. Hence, there is a pressing need to develop effective statistical tools for their analysis. The celebrating large-dimensional factor models which allow both the sample size and the dimension of time series to go to infinity have become a popular method in analyzing high-dimensional time series data, and therefore have received considerable attention in statistics and econometrics since Stock and Watson (1998, 2002) and Bai and Ng (2002). We refer to Bai and Li (2012, 2014) and Lam and Yao (2012) for recent advancement in estimation methods and inference theory in large-dimensional factor modeling.

In large-dimensional factor models, it is assumed that a large number of time series are driven by low-dimensional latent factors. Most existing estimation and forecasting methods in factor models are based on the assumption of time-invariant factor loadings. However, with large-scale data spanning over a long period of time, more and more evidence shows that the factor loadings tend to exhibit structural changes over time. That is, some structural breaks may occur at some dates over a period of time in the study. Ignoring structural breaks generally lead to misleading estimation, inference, and forecasting (Hansen, 2001). Hence, it is prudent to identify structural breaks of the factor loadings before one relies on the conventional time-invariant factor models. Indeed, a growing number of researches have been devoted to studying structural changes in factor loadings recently. To the best of our knowledge, most existing works have focused on developing testing procedures to detect breaks or conducting forecasts by assuming the existence of breaks. For example, Stock and Watson (2009) examine the forecasting reliability by assuming that there is a structural break and we have a knowledge of the break date. Breitung and Eickmeier (2011), Chen et al. (2014), and Han and Inoue (2014) propose various tests for a one-time structural change in the factor loadings. Corradi and Swanson (2014) propose a test to check structural stability of both factor loadings and factor-augmented forecasting regression coefficients. Su and Wang (2015) consider estimation and testing in time-varying factor models and their
test allows for multiple breaks in the factor loadings.

Frequently, one can reject the null hypothesis of constant factor loadings in empirical applications. Despite this, methods for determining the number of breaks and for identifying the locations of the break dates in factor models still remain unavailable, due to great technical challenges in developing the asymptotic tools. In this paper, we propose a novel three-step structural break detection procedure, which can automatically check the existence of breaks and then identify the exact locations of breaks if any. The procedure is easy-to-implement and theoretically reliable. Specifically, in Step I, we divide the whole time span into subintervals and estimate a conventional factor model with time-invariant factor loadings on each interval by the means of principal component analysis (PCA) (Bai, 2002; Bai and Ng, 2003). Based on the piecewise constant PCA estimates on each subinterval, we propose a BIC-type information criterion to determine the number of common factors and show that our information criterion can identify the correct number of common factors with probability approaching one (w.p.a.1). Our method extends Bai and Ng’s (2002) method to allow for an unknown number of breaks in the data and is thus robust to the presence of structural breaks in factor models. In Step II, we adopt the adaptive group fused Lasso (AGFL, Tibshirani et al., 2005; Yuan and Lin, 2005; Zou, 2006) to find intervals that contain a break point. We apply an adaptive group fusion penalty to the successive differences of the normalized factor loadings, which can identify the correct number of breaks and the subintervals that the breaks reside in w.p.a.1. In step III, we devise a grid search method to find the break locations in the identified subintervals sequentially and show that w.p.a.1 we can estimate the break points precisely.

The above three-step method provides an automatic way to detect breaks in factor models, and it is computationally fast. The major challenges in the asymptotic analysis of the proposed three-step procedure are threefold. Firstly, some subintervals obtained in the first step may contain a break point in which case the conventional time-invariant factor model is a misspecified model. Hence, we need to develop asymptotic properties of the estimators of the factors and factor loadings in the misspecified factor models, which do not exist in the literature. We find that the properties depend on whether the break point lies in the interior or boundary region of such a time interval. Secondly, we consider this paper
as the first work to apply the AGFL procedure to the normalized factor loadings to identify whether a subinterval contains a break point or not, where the adaptive weights behave substantially different from the weights investigated in the adaptive Lasso literature (e.g., Zou 2006) due to the presence of misspecified factor models in the first step. In particular, the adaptive weights have distinct asymptotic behaviors when the break points occur in the interior or boundary region of a subinterval, which greatly complicates the analysis of the Lasso procedure. Thirdly, it is technically very challenging to establish the theoretic claim that the grid search in the third step identifies the true break points w.p.a.1., even after we find the subintervals that contain a break point. In fact, our grid search method appears to be the first method to estimate the break locations consistently in the presence of estimation errors in early stages.

We conduct a sequence of Monte Carlo simulations to evaluate the finite sample performance of our procedure. We find that our information criterion can determine the correct number of factors accurately and our three-step procedure can identify the true number of breaks and estimate the break dates precisely in large samples. We apply our method to Stock and Watson’s (2009) macroeconomic dataset and detect five breaks for the period of 1959m01-2006m12.

The rest of this paper is organized as follows. In Section 2, we introduce the three-step procedure for break points detection and estimation. In Section 3, we study the asymptotic theory. In Section 4, we study the finite sample performance of our method. Section 5 provides an empirical study. Section 6 concludes. All proofs are relegated to the appendix. Further technical details are contained in the online supplementary material.

2 The Factor Model and Estimation Procedure

In this section, we consider a large-dimensional factor model with an unknown number of breaks, and then propose a three-step procedure for estimation. We first introduce some notations which will be used throughout the paper. Let $\mu_{\text{max}}(B)$ and $\mu_{\text{min}}(B)$ denote the largest and smallest eigenvalues of a symmetric matrix $B$, respectively. We use $B > 0$ to denote that $B$ is positive definite. For an $m \times n$ real matrix $A$, we denote its transpose as $A^\top$, its Moore-Penrose generalized inverse as $A^+$, its rank as $\text{rank}(A)$, its Frobenius norm
as $\|A\| (\equiv [\text{tr}(AA^\top)]^{1/2})$, and its spectral norm as $\|A\|_{sp} (\equiv \sqrt{\mu_{\text{max}}(A^\top A)})$. Note that the two norms are equal when $A$ is a vector. We will frequently use the submultiplicative property of these norms and the fact that $\|A\|_{sp} \leq \|A\| \leq \|A\|_{sp} \text{rank}(A)^{1/2}$. Let $P_A \equiv A (A^\top A)^+ A^\top$ and $M_A \equiv I_m - P_A$, where $I_m$ denotes an $m \times m$ identity matrix. For any set $S$, we use $|S|$ to denote its cardinality. For any positive numbers $a_n$ and $b_n$, let $a_n \asymp b_n$ denote $\lim_{n \to \infty} a_n / b_n = c$, for a positive constant $c$, and let $a_n \gg b_n$ denote $a_n^{-1} b_n = o(1)$.

The operator $\rightarrow$ denotes convergence in probability and $\text{plim}$ denotes probability limit. We use $(N,T) \to \infty$ to denote that $N$ and $T$ pass to infinity jointly.

2.1 The Factor Model

We consider the time-varying factor model:

$$X_{it} = \lambda_{it}^\top F_t + e_{it}, \ i = 1, \ldots, N, \ t = 1, \ldots, T,$$

where $\lambda_{it}$ is an $R \times 1$ vector of time-dependent factor loadings, $F_t$ is an $R \times 1$ vector of unobserved common factors, $e_{it}$ is the idiosyncratic error term, and both $N$ and $T$ pass to infinity. For simplicity of technical proofs, we assume that $R$ does not depend on $N$ and $T$, but it is unknown. Hence we need to estimate $R$ from the data. Writing the above model in the vector form, we have

$$X_t = \lambda_t^\top F_t + e_t, \ t = 1, \ldots, T,$$

where $X_t = (X_{1t}, \ldots, X_{Nt})^\top$, $\lambda_t = (\lambda_{1t}, \ldots, \lambda_{Nt})^\top$, and $e_t = (e_{1t}, \ldots, e_{Nt})^\top$.

We assume that the factor-loadings $\{\lambda_1, \ldots, \lambda_T\}$ exhibit certain sparse nature such that the total number of distinct vectors in the set is given by $m+1$, where $m$ denotes the total number of break points in the process $\{\lambda_t\}$ and it satisfies $T \gg m$. When $m \geq 1$, let $\{t_1, \ldots, t_m\}$ denote the $m$ change-points satisfying

$$1 \equiv t_0 < t_1 < \cdots < t_m < t_{m+1} \equiv T + 1,$$

so that the whole time span is divided into $m+1$ regimes/segments, denoted by $I_\kappa = [t_\kappa, t_{\kappa+1})$ for $\kappa = 0, 1, \ldots, m-1$ and $I_m = [t_m, t_{m+1}]$. We assume that

$$\lambda_{it} = \alpha_{i\kappa} \text{ for all } t \in I_\kappa \text{ and } \kappa = 0, 1, \ldots, m.$$

When $m = 0$, we have $I_0 = I_m = [t_0, t_1) = [1, T]$ and $\lambda_{it} = \alpha_{i0}$ for all $t \in [1, T]$, so that no
break happens in this scenario. Let \( \alpha_\kappa = (\alpha_{1k}, \ldots, \alpha_{N\kappa})^T \) for \( \kappa = 0, 1, \ldots, m \). In practice, the number of breaks, \( m \), and the locations of the breaks are unknown if there are any breaks. Our target is to detect breaks, to find the number of breaks and identify their locations, and to estimate \( R, \alpha_{i\kappa}, F_t \). Let \( t_0^0, \alpha_0^0, F_t^0 \) denote the true values of \( t_\kappa, \alpha_{i\kappa}, \alpha_\kappa \) and \( F_t \), respectively.

### 2.2 A Three-step Procedure

We propose a three step procedure to automatically detect breaks, to determine the number of breaks if any, and to estimate their locations.

#### 2.2.1 Step I: Piecewise constant estimation

Noting that \( \lambda_t F_t = \lambda_t (H_t^{-1})^T H_t^T F_t \) for any \( R \times R \) nonsingular matrix \( H_t, \lambda_t \) and \( F_t \) are not separately identified, and their identification requires \( R^2 \) restrictions at each time point \( t \). For the estimation of \( \lambda_t \) and \( F_t \), following the lead of Bai and Ng (2002), we shall impose the following identification conditions:

\[
\lambda_t^T \lambda_t / N = I_R \quad \forall t,
\]

In this step, we propose to approximate \( \lambda_{it} \) by piecewise-constants, and then estimate \( \lambda_{it} \) and \( F_t \) accordingly. The procedure is described as follows. Let \( J = J(N, T) \) be a prescribed integer that depends on \( (N, T) \), satisfying \( T \gg J \gg m \). Divide \([1, T]\) into \((J + 1)\) subintervals \( S_j = [v_j, v_{j+1}) \) for \( j = 0, 1, \ldots, J - 1 \) and \( S_J = [v_J, T] \), where \( \{v_j\}_{j=1}^J \) are a sequence of “equally-spaced” interior knots given as \( v_0 \equiv 1 < v_1 < \cdots < v_J < T \equiv v_{J+1} \), where \( v_j = \lfloor Tj/(J + 1) \rfloor \) for \( j = 1, \ldots, J \) and \( \lfloor \cdot \rfloor \) denotes the integer part of \( \cdot \). Note that each interval contains \( T/(J + 1) \) observations, for \( j = 0, 1, \ldots, J - 1 \), when \( T/(J + 1) \) is an integer. For any \( t \in S_j \), \( \lambda_{it} \) is treated as a constant and can be approximated by \( \lambda_{it} \approx \delta_{ij} \), so that the identification condition that \( \lambda_t^T \lambda_t / N = I_R \forall t \) implies that \( \sum_{i=1}^N \delta_{ij} \delta_i^T / N = I_R \) for each \( j = 0, 1, \ldots, J \). Denote \( \Delta_j = (\delta_{ij}, \ldots, \delta_{Nj})^T \). Then we need

\[
\Delta_j^T \Delta_j / N = I_R \quad \forall j = 0, 1, \ldots, J.
\]

The estimators \( \hat{\Delta}_j \) and \( \hat{F}_t \) are obtained by minimizing

\[
\sum_{t \in S_j} (X_t - \Delta_j F_t)^T (X_t - \Delta_j F_t)
\]
subject to $\Delta_j \mathbf{X}_t/N = \mathbb{I}_R$ and $\mathbf{F}^\top_j \mathbf{S}_j \mathbf{F}_j = \text{diagonal}$, where $\mathbf{F}_j = (F_{ij}, t \in S_j)^\top = (F_{ij}, \ldots, F_{ij+1})^\top$. By concentrating out $F_t = (\Delta_j \mathbf{X}_t/N)^{-1}(\Delta_j \mathbf{X}_t/N) = \Delta_j \mathbf{X}_t/N$, the above objective function becomes

$$\sum_{t \in S_j} (\mathbf{X}_t - \Delta_j \mathbf{X}_t/N)^\top (\mathbf{X}_t - \Delta_j \mathbf{X}_t/N) = \sum_{t \in S_j} \mathbf{X}_t^\top \mathbf{X}_t - N^{-1}\text{tr}(\Delta_j \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t),$$

where $\mathbf{X}_S = (\mathbf{X}_t, t \in S)$ and we have used the restriction that $\Delta_j \mathbf{X}_t/N = \mathbb{I}_R$. Thus, the estimators $\hat{\Delta}_j = (\hat{\delta}_{1j}, \ldots, \hat{\delta}_{Nj})^\top$ can be obtained by maximizing

$$N^{-1}\text{tr}(\Delta_j \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t \mathbf{X}_t)$$

subject to $\Delta_j \mathbf{X}_t/N = \mathbb{I}_R$. When $\text{rank}(\mathbf{X}_S \mathbf{X}_S^\top) \geq R$ for every $j = 0, \ldots, J$, $\hat{\Delta}_j$ is $\sqrt{N}$ times the eigenvectors corresponding to the $R$ largest eigenvalues of the $N \times N$ matrix $\mathbf{X}_S \mathbf{X}_S^\top = \sum_{t \in S_j} \mathbf{X}_t \mathbf{X}_t^\top$, and $\hat{F}_t = \hat{\Delta}_j \mathbf{X}_t/N$ for $t \in S_j$.

### 2.2.2 Step II: Adaptive group fused Lasso penalization for break detection

Let $\tau_j = |S_j|$ be the cardinality of the set $S_j$. Let $V_{N \tau_j,j}$ denote the $R \times R$ diagonal matrix of the first $R$ largest eigenvalues of $\frac{1}{N \tau_j} \mathbf{X}_S \mathbf{X}_S^\top$ in descending order. For those time points in the same true regime ($I_k$, say), their factor loadings should be the same. By Proposition 1(ii) below, $\hat{\Delta}_j V_{N \tau_j,j}$ is a consistent estimator of $\alpha_0 \Sigma Q_\alpha^\top$ for all $j$ satisfying $S_j \subset I_k$, where $\Sigma_F$ is defined in Assumption A1 and $Q_\alpha$ is defined in Proposition 1, both of which do not depend on $j$. Note that $\alpha_0 \Sigma F Q_\alpha^\top$ remains unchanged if two consecutive intervals, say $S_j$ and $S_{j+1}$, belong to $I_k$. This motivates us to consider the following objective function by imposing an AGFL penalty to detect the breaks between segments:

$$\frac{1}{2N} \sum_{j=0}^J \frac{1}{\tau_j} \sum_{t \in S_j} (\mathbf{X}_t - \Delta_j \hat{F}_t)^\top (\mathbf{X}_t - \Delta_j \hat{F}_t) + \gamma \sum_{j=1}^J w_j^0 \|\Delta_j V_{N \tau_j,j} - \Delta_{j-1} V_{N \tau,j-1}\|, \quad (1)$$

where $\gamma$ is a tuning parameter and $w_j^0$'s are adaptive weights to be specified later.

Let $\hat{\Theta}_j = \hat{\Delta}_j V_{N \tau_j,j}$, $\Theta_j = N^{1/2} \Delta_j V_{N \tau_j,j}/\|\hat{\Theta}_j\|$, $\hat{Z}_{jt} = N^{-1/2} \|\hat{\Theta}_j\| V_{N \tau,j}^{-1} \hat{F}_t$ and $\hat{Z}_{S_j} = (\hat{Z}_{jt}, t \in S_j)$. We can rewrite the objective function in (1) in terms of $\Theta_j$

$$\frac{1}{2N} \sum_{j=0}^J \frac{1}{\tau_j} \|\mathbf{X}_S - \Theta_j \hat{Z}_{S_j}\|^2 + \gamma \sum_{j=1}^J w_j \|\Theta_j - \Theta_{j-1}\|, \quad (2)$$

where $w_j = w_j^0 \|\hat{\Theta}_j\|$. Note that (1) compares $\Delta_j V_{N \tau,j}$ with $\Delta_{j-1} V_{N \tau,j-1}$ while (2) contrasts their normalized versions. Let $\hat{\Theta}_{j,r}$ denote the $r$th column of $\hat{\Theta}_j$ for $r = 1, \ldots, R$. Let
\( \tilde{\rho}_{j,r} \) denote the sample Pearson correlation coefficient of \( \tilde{\Theta}_{j,r} \) and \( \tilde{\Theta}_{j-1,r} \) for \( j = 1, \ldots, J \). When the eigenvectors in \( \tilde{\Delta}_j \) are properly normalized to ensure the sign-identification, with Proposition 1 below we can show that \( \tilde{\rho}_{j,r} \overset{P}{\to} 1 \) when both \( S_j \) and \( S_{j-1} \) belong to \( I_\kappa \) and they may converge in probability to a value different from one otherwise. This motivates us to consider the following adaptive weights

\[
w_j = \left( 1 - R^{-1} \sum_{r=1}^{R} \tilde{\rho}_{j,r} \right)^{-\kappa},
\]

where \( \kappa \) is some fixed positive constant, e.g., 2. Let \( \tilde{\Theta}_j \) denote the penalized estimator of \( \Theta_j \) in (2). Then the penalized estimator of \( \Delta_j \) is given by \( \tilde{\Delta}_j = \tilde{\Theta}_j V_{N+j}^{-1} |\tilde{\Theta}_j| \).

We apply Boyd et al.’s (2011) alternating direction method of multipliers (ADMM) algorithm to obtain the penalized estimator \( \tilde{\Theta}_j \). Boyd et al. (2011) show that the ADMM algorithm has a good global convergence property. The detailed procedure is provided in Section C of the Supplementary Material. The tuning parameter \( \lambda \) is chosen by the BIC method as given in Section 3.5.

2.2.3 Step III: Grid search for the locations of the breaks

Let \( \tilde{\beta}_j \equiv \tilde{\Theta}_j - \tilde{\Theta}_{j-1} \) for \( j = 1, \ldots, J \). By step II, we are able to identify the subintervals containing the breaks. There are four situations that can happen for each subinterval \( S_j \): (1) when \( \tilde{\beta}_j \neq 0 \) and \( \tilde{\beta}_{j+1} \neq 0 \), the break happens in the interior of the interval \( S_j \); (2) when \( \tilde{\beta}_j \neq 0 \) and \( \tilde{\beta}_{j+1} = 0 \) and \( \tilde{\beta}_{j-1} = 0 \), the break may happen near the left end of \( S_j \) or the right end of \( S_{j-1} \); (3) when \( \tilde{\beta}_{j+1} \neq 0 \) and \( \tilde{\beta}_j = 0 \) and \( \tilde{\beta}_{j+2} = 0 \), the break may happen near the right end of \( S_j \) or the left end of \( S_{j+1} \); and (4) when \( \tilde{\beta}_j = 0 \) and \( \tilde{\beta}_{j+1} = 0 \), no break happens in \( S_j \). For case (1), we can conclude that an estimated break happens in the interval \( S_j \), and for cases (2) and (3), we have that an estimated break happens in the intervals \( S_{j-1}^* \) and \( S_j^* \), respectively, where, e.g., \( S_{j-1}^* \equiv [v_{j-1} + \lfloor \tau_{j-1}/2 \rfloor + 1, v_j + \lfloor \tau_{j}/2 \rfloor] \). Suppose that we have found \( \hat{m} \) intervals that contain a break point. We denote such \( \hat{m} \) intervals as \( \tilde{S}_{j_1}, \ldots, \tilde{S}_{j_{\hat{m}}} \). Note that \( \tilde{S}_{j_{\kappa}} \) coincides with either \( S_{j_{\kappa}} \) or \( S_{j_{\kappa}}^* \). Write \( \tilde{S}_{j_{\kappa}} = [t_{\kappa,1}^*, \ldots, t_{\kappa,\hat{r}_{j_{\kappa}}}^*] \) with \( \tilde{r}_{j_{\kappa}} = |\tilde{S}_{j_{\kappa}}| \) for \( \kappa = 1, \ldots, \hat{m} \). We discuss how to estimate these \( \hat{m} \) break points below.

To estimate the first break point, we conduct a grid search over the interval \( \tilde{S}_{j_1} \) by using as many observations as possible from both pre-\( \tilde{S}_{j_1} \) and post-\( \tilde{S}_{j_1} \) intervals. If the first break
point happens to be $t^*_l$ for some $l \in \{1, 2, \ldots, \tau_j\}$, we know that observations that occur before $t^*_l$ belong to the first regime w.p.a.1. Similarly, the observations that occur after $t^*_l$ but before the first observation in $\tilde{S}_{j_2}$ belong to the second regime w.p.a.1. But $t^*_l$ is unknown and has to be searched over all points in $\tilde{S}_{j_2}$. After obtaining the first break point, we can find subsequent break points analogously.

To state the algorithm, let $S^b = \{t : a \leq t \leq b\}$ and $F_{S^b} = (F_a, \ldots, F_b)^\top$ for any integers $a \leq b$. Let $\alpha_l = (\alpha_{1l}, \ldots, \alpha_{Nl})^\top$ for $l = 1, 2, \ldots$. The following procedure describes how we can find the locations of all $\hat{m}$ break points sequentially:

1. To search for the first break point $t_1$, we consider the following minimization problem:

$$
\min_{\{\alpha_1, \alpha_2, \{F_t\}\}} Q_1 (\alpha_1, \alpha_2, \{F_t\}; t_1) = \sum_{t \in S^\tau_{1-1}} \|X_t - \alpha_1 F_{t_1}\|^2 + \sum_{t \in S^\tau_{2-1}} \|X_t - \alpha_2 F_{t_2}\|^2
$$

subject to the constraints $N^{-1} \alpha_1^\top \alpha_1 = I_R$, $N^{-1} \alpha_2^\top \alpha_2 = I_R$, $\frac{1}{\tau_1-\tau_1} F_{S^\tau_{1-1}}^\top F_{S^\tau_{1-1}}$ is diagonal and $\frac{1}{\tau_1-\tau_1} F_{S^\tau_{1-1}}^\top F_{S^\tau_{1-1}}$ is diagonal. Denote the solution to the above minimization problem as $(\hat{\alpha}_1 (t_1), \hat{\alpha}_2 (t_1), \{\hat{F}_t (t_1)\})$. The first break point is estimated as

$$
\hat{t}_1 = \arg \min_{t_1 \in \tilde{S}_{j_1}} Q_1 (\hat{\alpha}_1 (t_1), \hat{\alpha}_2 (t_1), \{\hat{F}_t (t_1)\}; t_1).
$$

2. After obtaining the break points, $\hat{t}_1, \ldots, \hat{t}_{\kappa-1}$, we can search for the $\kappa$th break point $t_{\kappa}$ by considering the following minimization problem

$$
\min_{\{\alpha_{\kappa}, \alpha_{\kappa+1}, \{F_t\}\}} Q_{\kappa} (\alpha_{\kappa}, \alpha_{\kappa+1}, \{F_t\}; t_{\kappa})
$$

$$
= \sum_{t \in S^\tau_{\kappa+1-1}} \|X_t - \alpha_{\kappa} F_{t_\kappa}\|^2 + \sum_{t \in S^\tau_{\kappa+2-1}} \|X_t - \alpha_{\kappa+1} F_{t_{\kappa+1}}\|^2
$$

subject to the constraints $N^{-1} \alpha_{\kappa}^\top \alpha_{\kappa} = I_R$, $N^{-1} \alpha_{\kappa+1}^\top \alpha_{\kappa+1} = I_R$, $\frac{1}{\tau_{\kappa} - \tau_{\kappa-1}} F_{S^\tau_{\kappa+1-1}}^\top F_{S^\tau_{\kappa+1-1}}$ is diagonal and $\frac{1}{\tau_{\kappa} - \tau_{\kappa-1}} F_{S^\tau_{\kappa+1-1}}^\top F_{S^\tau_{\kappa+1-1}}$ is diagonal. Denote the solution to the above minimization problem as $(\hat{\alpha}_{\kappa} (t_{\kappa}), \hat{\alpha}_{\kappa+1} (t_{\kappa}), \{\hat{F}_t (t_{\kappa})\})$. The $\kappa$th break point is estimated as

$$
\hat{t}_{\kappa} = \arg \min_{t_{\kappa} \in \tilde{S}_{j_{\kappa}}} Q_{\kappa} (\hat{\alpha}_{\kappa} (t_{\kappa}), \hat{\alpha}_{\kappa+1} (t_{\kappa}), \{\hat{F}_t (t_{\kappa})\}; t_{\kappa}).
$$

3. Repeat the above step until we obtain all $\hat{m}$ estimated break points.

At last, after we find the locations of the break points, $\hat{t}_1, \ldots, \hat{t}_{\hat{m}}$, the whole time span is divided into $\hat{m} + 1$ regimes/segments, denoted by $\tilde{I}_\kappa = [\hat{t}_{\kappa}, \hat{t}_{\kappa+1})$. On each segment, we
estimate the factors and their loadings as
\[
(\hat{\alpha}_\kappa, \{\hat{F}_i\}) = \arg \min_{\alpha_\kappa, \{F_i\}} \sum_{t \in I_\kappa} (X_t - \alpha_\kappa F_t)\top(X_t - \alpha_\kappa F_t)
\]
subject to the constraints \(N^{-1}\alpha_\kappa\top\alpha_\kappa = \mathbb{I}_R\), and \[\frac{1}{|I_n|} F_{I_n}\top F_{I_n}\] = diagonal.

### 3 Asymptotic Theory

In this section, we study the asymptotic properties of our estimators.

#### 3.1 Theory for the Piecewise Constant Estimators

For each subinterval \(S_j\), we will establish the asymptotic property of \(\hat{\alpha}_j\) from the piecewise constant estimation in Step I. Denote \(S = \{0, 1, 2, \ldots, J\}\). Let \(\tau_{j1} = t^0_\kappa - v_j\) and \(\tau_{j2} = \tau_j - \tau_{j1}\) when \(S_j\) contains a true break point \(t^0_\kappa\) for some \(\kappa = \kappa(j)\). Define

\[
S_1 = \left\{ j \in S : S_j \subset I_\kappa \text{ for some } \kappa(j) \right\},
\]

\[
S_{2a} = \left\{ j \in S : S_j \text{ contains a break } t^0_\kappa \text{ for some } \kappa(j) \text{ such that } \lim_{T \to \infty} \tau_{j1}/\tau_j = 1 \right\},
\]

\[
S_{2b} = \left\{ j \in S : S_j \text{ contains a break } t^0_\kappa \text{ for some } \kappa(j) \text{ such that } \lim_{T \to \infty} \tau_{j1}/\tau_j = 0 \right\},
\]

\[
S_{2c} = \left\{ j \in S : S_j \text{ contains a break } t^0_\kappa \text{ for some } \kappa(j) \text{ such that } \lim_{T \to \infty} \tau_{j1}/\tau_j \in (0, 1) \right\}.
\]

Let \(S_2 = S_{2a} \cup S_{2b} \cup S_{2c}\). When no confusion arises, we will suppress the dependence of \(\kappa = \kappa(j)\) on \(j\). Noting that \(|S_2| = m \ll J\), we have \(|S_1|/J \to 1\).

**Case 1.** When no break occurs in the subinterval \(S_j\), i.e., \(S_j \subset I_\kappa\) for some segment \(I_\kappa\), then we have \(\lambda_{it} = \alpha^0_{it\kappa}\) for all \(t \in S_j\), where \(\alpha^0_{it\kappa}\) is the vector of the true factor loadings for the segment \(I_\kappa\). Let \(F^0_t\) be the vector of true factors for \(t \in S_j\). Then we have

\[
X_{it} = \alpha^0_{it\kappa} F^0_t + \epsilon_{it}, \quad i = 1, \ldots, N, \ t \in S_j.
\]

Let \(F^0_{S_j} = (F^0_t, t \in S_j)\top = (F^0_{e_j}, \ldots, F^0_{e_{j+1}-1})\top\) and \(\alpha^0_\kappa = (\alpha^0_{1\kappa}, \ldots, \alpha^0_{N\kappa})\top\). Denote \(\gamma_{N, s, t} = N^{-1}E(e_s e_t)\), \(\gamma_{N,F, s, t} = N^{-1}E(F^0_s e_s e_t F^0_t)\top\), \(\zeta_{st} = N^{-1}[e_s e_t - E(e_s e_t)]\), \(\zeta_{F,st} = N^{-1}[F^0_s e_s e_t F^0_t ]\top - E(F^0_s e_s e_t F^0_t)\top\), and \(\zeta_{it} (t_1, t_2) = \frac{1}{t_2 - t_1} \sum_{s = t_1}^{t_2} e_s e_{ls} - E(e_s e_{ls})\) Let \(\omega_{il, t} = E(\epsilon_{it} e_{ls})\) and \(\omega_{il, t} = \omega_{il, tt}\). Let \(C < \infty\) denote a positive constant that may vary from case to case.

We make the following assumptions.
Assumption A1. $E\|F_0^0\|^4 \leq C$ and $\frac{1}{t^2} F_{S_j}^{0t} F_{S_j}^0 = \Sigma_F + O_P((t - s)^{-1/2})$ for some $R \times R$ positive definite matrix $\Sigma_F$ and for any two points $t, s \in [1, T]$ satisfying $t - s \to \infty$.

Assumption A2. $\lambda_{it}$’s are nonrandom such that $\max_{1 \leq i \leq N, 1 \leq t \leq T} \|\lambda_{it}\| \leq C$ and $\frac{1}{N} \alpha_k^0 \alpha_k^0 = \Sigma_k + O \left( N^{-1/2} \right)$ for some $R \times R$ positive definite matrix $\Sigma_k$ for $\kappa = 0, 1, \ldots, m$.

Assumption A3. (i) $E(e_{it}) = 0$ and $\max_{1 \leq i \leq N, 1 \leq t \leq T} E(e_{it}^4) \leq C$.

(ii) $\max_{1 \leq i \leq T} \sum_{s=1}^T \|\gamma(s, t)\| \leq C$ and $\max_{1 \leq s \leq T} \sum_{t=1}^T \|\gamma(s, t)\| \leq C$ for $\gamma = \gamma_N, \gamma_{N,F}$, and $\gamma_{N,F}$. $\max_{1 \leq i \leq T} |\omega_{il,tt}| \leq \omega_{il}$ for some $\omega_{il}$ such that $\max_{1 \leq i \leq N} \sum_{t=1}^T |\omega_{il}| \leq C$.

(iv) $(N \tau_j)^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s \in S_j} \sum_{s \in S_j} |\omega_{il,ts}| \leq C$.

(v) $\max_{1 \leq i, t \leq T} E \left( (t_2 - t_1)^{1/2} \zeta_{il} (t_1, t_2) \right)^4 \leq C$ for all $t_1 < t_2$ such that $t_2 - t_1 \to \infty$.

(vi) $\max_{1 \leq i, t \leq T} E \left\| \left( N^{1/2} \zeta^+_{st} \right) \right\|^4 \leq C$ for $\zeta^+_{st} = \zeta_{st}, \zeta_{F,st}$ and $\zeta_{FF,st}$, and $\max_{1 \leq t \leq T} E \|N^{-1/2} \alpha_k^0 e_t\|^4 \leq C$ for $\kappa = 0, 1, \ldots, m$.

Assumption A4. The eigenvalues of the $R \times R$ matrices $\Sigma_k^{1/2} \Sigma_F \Sigma_k^{1/2}$ are distinct for $\kappa = 0, 1, \ldots, m$.

Assumptions A1-A2 parallel Assumptions A and B in Bai (2003). A1 implies that $\frac{1}{\tau_j} F_{S_j}^{0t} F_{S_j}^0 = \Sigma_F + O_P(\tau_j^{-1/2})$ as $\tau_j \to \infty$ and A2 requires $\lambda_{it}$ to be nonrandom and uniformly bounded. A3(i) imposes moment conditions on $e_{it}$ and A3(ii)-(v) restricts the cross-sectional and serial dependence among $\{e_{it}, F_t\}$. Similar conditions are also imposed in the literature; see, Bai and Ng (2002) and Bai (2003). A4 is required to establish the convergence of certain eigenvector estimates.

Let $\eta_{N\tau_j} = \min \{ \sqrt{N}, \sqrt{\tau_j} \}$ and $H_j = H_{N\tau_j,j} = (\frac{1}{\tau_j} F_{S_j}^{0t} F_{S_j}^0) (\frac{1}{\eta_{N\tau_j}^2} \hat{\Delta}_j) V_{N\tau_j,j}^{-1}$. Following Bai (2003), we can readily obtain the following results:

Proposition 1. Suppose that Assumptions A1-A4 hold. Then as $(N, \tau_j) \to \infty$,

(i) $\frac{1}{N} \| \hat{\Delta}_j - \alpha_k^0 H_j \|^2 = O_P(\eta_{N\tau_j}^{-2})$ and $\| \hat{F}_t - H_j^{-1} F_t^0 \| = O_P(\eta_{N\tau_j}^{-1})$ for any $t \in S_j$ and $j \in S_1$,

(ii) $\| \frac{1}{N} \hat{\Delta}_j^\top \alpha_k^0 - Q_k \| = O_P(\eta_{N\tau_j}^{-1})$ and $\frac{1}{N} \| \hat{\Delta}_j V_{N\tau_j,j}^- - \alpha_k^0 \Sigma_F Q_k \|^2 = O_P(\eta_{N\tau_j}^{-2})$ for any $j \in S_1$, where the matrix $Q_k$ is invertible and is given by $Q_k = V_k^{1/2} \Upsilon_k \Sigma_F^{-1/2}, V_k = \text{diag}(v_{1\kappa}, \ldots, v_{R\kappa})$, $v_{1\kappa} > v_{2\kappa} \geq \cdots \geq v_{R\kappa} > 0$ are the eigenvalues of $\Sigma_F^{1/2} \Sigma_k \Sigma_F^{1/2}$, and $\Upsilon_k$ is the corresponding eigenvector matrix such that $\Upsilon_k^\top \Upsilon_k = I_R$.
Remark 3.1. The above result can be proved by modifying the arguments used in Bai (2003). Alternatively, they can be derived from the results in Proposition 2(ii) below.

Case 2. When a break point $t^0_k$ lies in the interval $S_j = [v_j, v_{j+1})$, we have

$$
\lambda_{it} = \left\{ \begin{array}{ll}
\alpha^0_{i,k-1} & \text{for } t \in [v_j, t^0_k) \\
\alpha^0_{i,k} & \text{for } t \in [t^0_k, v_{j+1})
\end{array} \right. \text{ for some } k = k(j).
$$

Let $F^0_{S_j,1} = (F^0_t, t \in [v_j, t^0_k))^\top$, $F^0_{S_j,2} = (F^0_t, t \in [t^0_k, v_{j+1}))^\top$, and $\alpha^*_k = (\alpha^0_{k-1}, \alpha^0_k)$. Let $F^*_t = (F^0_t 1_{jt}, F^0_t 1_{jt}^\top)^\top$ and $F^*_{S_j} = (F^*_v, ..., F^*_v)^\top$, where $1_{jt} = 1 \{v_j \leq t < t^0_k\}$, $\bar{1}_{jt} = 1 \{t^0_k \leq t < v_{j+1}\}$, and we suppress the dependence of $F^*_t$ on $j$. Let $H^*_j = \frac{1}{\tau_j} F^*_{S_j} F^*_m \frac{1}{\tau_j} \alpha^*_N \hat{\Delta}_j V_{N\tau_j,j}^{-1}$, $H_{j,1} = (\frac{1}{\tau_j} F^*_{S_j,1} F^*_{S_j,1}) (\frac{1}{\alpha^0_{k-1} \hat{\Delta}_j}) V_{N\tau_j,j}^{-1}$, and $H_{j,2} = (\frac{1}{\tau_j} F^*_{S_j,2} F^*_{S_j,2}) (\frac{1}{\alpha^0_{k} \hat{\Delta}_j}) V_{N\tau_j,j}^{-1}$.

The following proposition establishes the asymptotic property of $\hat{\Delta}_j$ in Case 2.

Proposition 2. Suppose that Assumptions A1-A4 hold. Then

(i) $\frac{1}{N} ||\hat{\Delta}_j - \alpha^0_{k-1} H_{j,1,j}||^2 = O_P(c^2_{j2a})$ and $\frac{1}{N} ||\hat{\Delta}_j V_{N\tau_j,j} - \alpha^0_{k-1} \Sigma_F Q_{k-1}^\top||^2 = O_P(c^2_{j2b})$ for all $j \in S_{2a}$;

(ii) $\frac{1}{N} ||\hat{\Delta}_j - \alpha^0_{k} H_{j,2,j}||^2 = O_P(c^2_{j2b})$ and $\frac{1}{N} ||\hat{\Delta}_j V_{N\tau_j,j} - \alpha^0_{k} \Sigma_F Q_{k}^\top||^2 = O_P(c^2_{j2b})$ for all $j \in S_{2b}$;

(iii) $\frac{1}{N} ||\hat{\Delta}_j - \alpha^*_k H^*_j||^2 = O_P(\eta_{N\tau_j}^{-2})$ and $\frac{1}{N} ||\hat{\Delta}_j V_{N\tau_j,j} - (N\tau_j)^{-1} (\alpha^0_{k-1} F^*_{S_j,1} F^*_{S_j,1} \alpha^0_{k-1} + \alpha^0_{k} F^*_{S_j,2} F^*_{S_j,2} \alpha^0_{k}) \hat{\Delta}_j||^2 = O_P(\eta_{N\tau_j}^{-2})$ for all $j \in S_{2c}$, where $c_{j2a} = \eta_{N\tau_j}^{-1} + \tau_j/\tau_j$ and $c_{j2b} = \eta_{N\tau_j}^{-1} + \tau_j/\tau_j$.

Remark 3.2. Proposition 2 indicates that the asymptotic property of $\hat{\Delta}_j$ depends on whether $j$ lies in $S_{2a}$, $S_{2b}$, or $S_{2c}$. Propositions 2(i) (resp. (ii)) says, when the observations in $S_j$ are mainly from regime $k-1$ (resp. $k$), the asymptotic property of $\hat{\Delta}_j$ mainly depends on $\alpha^0_{k-1}$ (resp. $\alpha^0_k$), in which case the probability limit of $\hat{\Delta}_j V_{N\tau_j,j}$ will be different from that of $\hat{\Delta}_{j+1} V_{N\tau_{j+1,j+1}}$ (resp. $\hat{\Delta}_{j-1} V_{N\tau_{j-1,j-1}}$), given by $\alpha^0_{k-1} \Sigma_F Q_{k-1}^\top$ (resp. $\alpha^0_{k} \Sigma_F Q_{k}^\top$). In the case where $j \in S_{2c}$, the limit of $\hat{\Delta}_j V_{N\tau_j,j}$ will be different from those of $\hat{\Delta}_{j-1} V_{N\tau_{j-1,j-1}}$ (which is given by $\alpha^0_{k-1} \Sigma_F Q_{k-1}^\top$) and $\hat{\Delta}_{j+1} V_{N\tau_{j+1,j+1}}$ (which is given by $\alpha^0_{k} \Sigma_F Q_{k}^\top$).

3.2 Determination of the Number of Factors

In the above analysis, we assume that the number of factors, $R$, is known. In practice, one has to determine $R$ from the data. Here we assume that the true value of $R$, denoted as
$R_0$, is bounded from above by a finite integer $R_{\text{max}}$. We propose a BIC-type information criterion to determine $R_0$.

Now, we use $\tilde{\Delta}_j(R)$ and $\hat{F}_t(R)$ to denote the estimators of $\Delta_j$ and $F_t$ by using $R$ factors defined in Section 2.2.1. Let $\tilde{\Delta}_j(R) = (N\tau_j)^{-1} X_{S_j}X_{S_j}^T \tilde{\Delta}_j(R)$ for $j = 0, 1, \ldots, J$. Define

$$V(R) = V(R, \{\tilde{\Delta}_j(R)\}) = \min_{\{F_1(R), \ldots, F_T(R)\}} (J + 1)^{-1} \sum_{j=0}^J (N\tau_j)^{-1} \sum_{t \in S_j} (X_t - \tilde{\Delta}_j(R)F_t(R))^T(X_t - \tilde{\Delta}_j(R)F_t(R)).$$

Following the lead of Bai and Ng (2002), we consider the following BIC-type information criterion to determine $R_0$:

$$IC(R) = \ln V \left(R, \{\tilde{\Delta}_j(R)\}\right) + \rho_{NT} R,$$

where $\rho_{NT}$ plays the role of $\ln(NT)/(NT)$ in the case of BIC. Let $\hat{R} = \arg\min_{R} IC(R)$. We add the following assumptions.

**Assumption A5.** (i) $\max_{1 \leq i \leq N} \max_{1 \leq t \leq T-s} \left\| \frac{1}{N} \sum_{t=r}^{r+s} F_t^0 e_{it} \right\| = O_P \left((s/\ln T)^{-1/2}\right)$ for any $s \to \infty$.

(ii) $\max_{1 \leq s, t \leq T} \frac{1}{N} \left| e_t^T e_s - E(e_t^T e_s) \right| = O_P \left((N/\ln T)^{-1/2}\right)$.

(iii) $\max_{0 \leq s \leq m} \max_{1 \leq s \leq T} \frac{1}{N} \left\| a_0^\tau e_t \right\| = O_P \left((N/\ln T)^{-1/2}\right)$.

(iv) $\max_{j \in S_1} \left\| E_{S_j} \right\|_{sp} = O_P(\max(\sqrt{N}, \sqrt{T}))$, where $E_{S_j} = (e_t, t \in S_j)$ and $\tau = \min_{0 \leq j \leq J} \tau_j$.

**Assumption A6.** As $(N, T) \to \infty$, $\rho_{NT} \to 0$ and $\rho_{NT} [J/m + \eta_{N\tau}^2] \to \infty$ where $\eta_{N\tau} = \min(\sqrt{N}, \sqrt{T})$.

A5 is used to obtain some uniform result and can be verified under certain primitive conditions. For example, under certain strong mixing and moment conditions on the process $\{F_t^0 e_{it}, t \geq 1\}$, A5(i) can be verified by a simple use of Bernstein inequality for strong mixing processes provided that $N$ and $T$ diverge to infinity at comparable rates. See Moon and Weidner (2015) for primitive conditions to ensure A5(iv) to hold. The conditions on $\rho_{NT}$ in A6 are typical conditions in order to estimate the number of factors consistently. The penalty coefficient $\rho_{NT}$ has to shrink to zero at an appropriate rate to avoid both overfitting and underfitting.

**Proposition 3.** Suppose that Assumptions A1-A6 hold. Then $P \left(\hat{R} = R_0\right) \to 1$ as $(N, T) \to \infty$. 

13
Remark 3.3. Proposition 3 indicates the class of information criteria defined by $IC(R)$ can consistently estimate $R_0$. To implement the information criterion, one needs to choose the penalty coefficient $\rho_{NT}$. Following the lead of Bai and Ng (2002), we suggest setting $\rho_{NT} = \frac{N+\bar{\tau}}{N\bar{\tau}} \ln \left( \frac{N\bar{\tau}}{N+\bar{\tau}} \right)$ or $\rho_{NT} = \frac{N+\bar{\tau}}{N\bar{\tau}} \ln \eta_{N\bar{\tau}}^2$ with $\eta_{N\bar{\tau}} = \min \{ \sqrt{\bar{\tau}}, \sqrt{N} \}$ and $\bar{\tau} = T/(J+1)$, and evaluate the performance of these two information criteria in our simulation studies. Define

$$\begin{align*}
IC_1(R) &= \log(V(R)) + R \frac{N+\bar{\tau}}{N\bar{\tau}} \ln \left( \frac{N\bar{\tau}}{N+\bar{\tau}} \right), \\
IC_2(R) &= \log(V(R)) + R \frac{N+\bar{\tau}}{N\bar{\tau}} \ln \eta_{N\bar{\tau}}^2.
\end{align*}$$

(4)

Let $\hat{R}_1 = \arg \min R IC_1(R)$ and $\hat{R}_2 = \arg \min R IC_2(R)$. When the number of breaks, $m$, is fixed, it appears that one can choose $J$ such that $J \asymp \bar{\tau}$, in which case $J/m + \eta_{N\bar{\tau}}^2 \asymp \eta_{N\bar{\tau}}^2$ provided $\bar{\tau} = O(N)$.

### 3.3 Identifying the Intervals That Contain a Break Point

Let $\Theta_j^*$ denote the elementwise probability limit of $\hat{\Theta}_j \equiv N^{1/2} \hat{\Delta}_j V_{N\tau,j}/\|\hat{\Delta}_j V_{N\tau,j}\|$, $j = 0, 1, ..., J$. In the absence of break points on the whole time interval $[1, T]$, we can readily show that $\Theta_j^* - \Theta_{j-1}^* = 0$ for $j = 1, ..., J$. In the general case, $\Theta_j^* - \Theta_{j-1}^*$ may be equal to or different from the zero matrix depending on whether the subinterval $S_j$ or $S_{j-1}$ contains a break point.

Let $|I_{\min}| = \min_{0 \leq \kappa \leq m} |I_{\kappa}|$. To state the next result, we add the following two assumptions.

**Assumption A7.** For $\kappa = 1, 2, ..., m$, $\frac{1}{N} \| \alpha_0^0 \Sigma_F Q_\kappa^T - \alpha_{\kappa-1}^0 \Sigma_F Q_{\kappa-1}^T \|^2 \to c_\kappa > 0$ as $(N, T) \to \infty$.

**Assumption A8.** (i) $\tau = O(N)$, $\tau \ln T = o(|I_{\min}|)$, and $m/J = o(1)$.

(ii) As $(N, T) \to \infty$, $(N\tau)^{1/2} \gamma = O(1)$ and $(N\tau)^{1/2} \gamma \eta_{N\tau}^\kappa \to \infty$.

A7 ensures that parameters of interest in neighboring segments are distinct from each other. Note that $Q_\kappa = V_\kappa^{1/2} \Upsilon_\kappa^{1/2} \Sigma_F^{-1/2}$, where $V_\kappa$ and $\Upsilon_\kappa$ collect the eigenvalues and normalized eigenvectors of $\Sigma_F^{-1/2} \Sigma_F^2$, and $\Sigma_\kappa$ denotes the limit of $\frac{1}{N} \alpha_0^0 \kappa^0 \alpha_\kappa^0$. If $\alpha_\kappa^0 = \alpha_{\kappa-1}^0$, then $\alpha_\kappa^0 \Sigma_F Q_\kappa^T = \alpha_{\kappa-1}^0 \Sigma_F Q_{\kappa-1}^T$. When $\alpha_\kappa^0$ and $\alpha_{\kappa-1}^0$ are distinct from each other such that $\frac{1}{N} \| \alpha_\kappa^0 - \alpha_{\kappa-1}^0 \|^2 \to c_\kappa$ for some $c_\kappa > 0$, we generally expect A7 to be satisfied. A8(i) ensures that $\eta_{N\tau,j} = O(\tau_j^{-1/2})$ and each interval $S_j$, $j = 0, 1, ..., J$, contains at most one break.
Remark 3.5. Proposition 4(ii) requires that \( \gamma \) converge to zero at a suitable rate, which is required to identify all intervals that do not contain a break point.

The next proposition is crucial for identifying the intervals that contain the break points.

**Proposition 4.** Suppose that Assumptions A1-A4 and A7-A8 hold. Then

(i) \( N^{-1} \| \hat{\Theta}_j - \Theta_j^* \|^2 = O_p(a_j^2) \) for all \( j \in S \),

(ii) \( \Pr \left\{ \| \hat{\Theta}_j - \hat{\Theta}_{j-1} \| = 0 \text{ for all } j, j - 1 \in S_1 \right\} \rightarrow 1 \text{ as } (N,T) \rightarrow \infty \),

where \( a_j = \eta_{N \tau_j}^{-1} \) if \( j \in S_1 \cup S_{2c} \), \( a_j = c_{j2a} \) if \( j \in S_{2a} \), and \( a_j = c_{j2b} \) if \( j \in S_{2b} \), and \( c_{j2a} \) and \( c_{j2b} \) are defined in Proposition 2.

**Remark 3.4.** Proposition 4(i) establishes the mean square convergence rates of the penalized estimators \( \hat{\Theta}_j \) which depend on whether \( j \in S_1, S_{2a}, S_{2b}, \) and \( S_{2c} \). Proposition 4(ii) establishes the selection consistency of our AGFL method; it says that w.p.a.1 all the zero matrices \( \{ \Theta_j^* - \Theta_{j-1}^* \}, j, j - 1 \in S_1 \} \) must be estimated as exactly zeros by the AGFL method. On the other hand, we notice that \( \Theta_j^* - \Theta_{j-1}^* = 0 \) if \( j - 1 \in S_1 \) and \( j \in S_{2a} \), or \( j - 1 \in S_{2b} \) and \( j \in S_1 \). In the latter two cases, the estimate \( \hat{\Theta}_j - \hat{\Theta}_{j-1} \) of \( \Theta_j^* - \Theta_{j-1}^* \) may be zero or nonzero, depending on whether we allow \( (N \tau)^{1/2} \gamma (\tau_{j2}/\tau_j)^{-2\kappa} \) to pass to infinity in the case where \( j - 1 \in S_1 \) and \( j \in S_{2a} \), and \( (N \tau)^{1/2} \gamma (\tau_{j2}/\tau_j)^{-2\kappa} \) to pass to infinity in the case where \( j - 1 \in S_1 \) and \( j \in S_{2b} \). If the latter two conditions are satisfied, a close examination of the proof of Proposition 4(ii) indicates that \( \Theta_j^* - \Theta_{j-1}^* \) will also be estimated by exactly zero in large samples when \( j - 1 \in S_1 \) and \( j \in S_{2a} \), or \( j - 1 \in S_{2b} \) and \( j \in S_1 \). On the other hand, by (i), we know that the matrices \( \Theta_j^* - \Theta_{j-1}^* \) can be consistently estimated by \( \hat{\Theta}_j - \hat{\Theta}_{j-1} \). Putting these two results together, Proposition 4 implies that the AGFL is capable of identifying the intervals among \( \{ S_j, j = 0, 1, ..., J \} \) that might contain an unknown break point. Recall that we use \( \hat{m} \) to denote the estimated number of break points. A direct implication of Proposition 4 is that

\[
\Pr (\hat{m} = m) \rightarrow 1 \text{ as } (N,T) \rightarrow \infty.
\] (5)

**Remark 3.5.** In order to see whether a subinterval \( S_j, j = 1, ..., J - 1 \), contains a break point (say, \( \theta_{\kappa}^0 \)) or not, we need to compare \( \Theta_j^* \) with both \( \Theta_{j-1}^* \) and \( \Theta_{j+1}^* \) at the population level or compare \( \hat{\Theta}_j \) with both \( \hat{\Theta}_{j-1} \) and \( \hat{\Theta}_{j+1} \) at the sample level. At the population level, we have four scenarios: (1) \( \Theta_{j-1}^* \neq \Theta_j^* \neq \Theta_{j+1}^* \) when \( j \in S_{2c} \), (2) \( \Theta_{j-2}^* = \Theta_{j-1}^* \neq \Theta_j^* = \Theta_{j+1}^* \)
when \( j \in S_{2h} \) or \( j - 1 \in S_{2a}, \) (3) \( \Theta^*_j = \Theta^*_j \neq \Theta^*_{j+1} = \Theta^*_{j+2} \) when \( j \in S_{2a} \) or \( j + 1 \in S_{2b}, \) (d) \( \Theta^*_j = \Theta^*_j = \Theta^*_{j+1} \) when \( j \in S_2. \) In case (1), we can conclude that we have an estimated break point in the interval \( S_j, \) and for cases (2) and (3), we can conclude that a break point happens in \( S^*_j - 1 \) and \( S^*_j, \) respectively (see Section 2.2.3 for the definitions of \( S^*_j - 1 \) and \( S^*_j \)). The sample case has been discussed at the beginning of Section 2.2.3. In addition, under the condition that \( |I_{\min}| \gg T/J, \) any finite fixed number of consecutive intervals (e.g., \( S_{j-1}, S_j, \) and \( S_{j+1} \)) can contain at most one break, and \( S_0 \) and \( S_T \) cannot contain any break. Such information is useful to prove the result in Proposition 4.

### 3.4 Estimation of the Break Dates

**Assumption A9.** \( \frac{1}{\sqrt{N}} \left\| (\alpha^0_{\kappa} - \alpha^0_{\kappa-1}) F^0_{t_{\kappa-\ell}} \right\|^2 \gg c_{NT} \) for \( \ell = 0, 1 \) and \( \kappa = 1, ..., m, \) where \( c_{NT} = |I_{\min}|^{-1/2} (\ln T)^{3/2} + (\tau \ln T / |I_{\min}|)^{1/2}. \)

Assumption A9 is needed to consistently estimate all \( m \) break points. To understand this, we focus on the case where \( D_{N, \kappa} \equiv \frac{1}{N} (\alpha^0_{\kappa} - \alpha^0_{\kappa-1})^\top (\alpha^0_{\kappa} - \alpha^0_{\kappa-1}) \to D_{\kappa} > 0. \) In this case,

\[
\frac{1}{N} \left\| (\alpha^0_{\kappa} - \alpha^0_{\kappa-1}) F^0_{t_{\kappa-\ell}} \right\|^2 = \operatorname{tr} \left( D_{N, \kappa} F^0_{t_{\kappa-\ell}} F^0_{t_{\kappa-\ell}}^\top \right) \geq \mu_{\min} (D_{N, \kappa}) \left\| F^0_{t_{\kappa-\ell}} \right\|^2 \gg c_{NT} \text{ almost surely.}
\]

The next proposition establishes the consistency of the estimators of the break points.

**Proposition 5.** Suppose that Assumptions A1-A9 hold. Then \( \Pr(\hat{t}_1 = t_1^0, ..., \hat{t}_m = t_m^0 \mid \hat{m} = m) \to 1 \) as \( (N, T) \to \infty. \)

In conjunction with (5), the above proposition indicates that we can estimate the break points precisely w.p.a.1.

### 3.5 Choice of the Tuning Parameter

We select the tuning parameter \( \lambda \) in the fused penalization procedure described in Section 2.2.2 by minimizing the BIC-type information criterion:

\[
\text{BIC}(\lambda) = \log \left( (J + 1)^{-1} \sum_{j=0}^J (N \tau_j)^{-1} \sum_{t \in S_j} (X_t - \tilde{\Delta}_j \hat{F}_t)^\top (X_t - \tilde{\Delta}_j \hat{F}_t) \right) + \frac{\log(NT/(J + 1))}{NT/(J + 1)} R(\hat{m} + 1),
\]

where \( \hat{m} = \hat{m}(\lambda) \) denotes the number of breaks identified by the penalization procedure.
4 Monte Carlo Simulations

In this section, we conduct simulation studies to assess the finite-sample performance of our proposed break detection procedure.

4.1 Data Generating Processes

We generate data under the framework of high dimensional factor models with \( R = 2 \) common factors:

\[
X_{it} = \lambda_{it}^\top F_t + e_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,
\]

where \( F_t = (F_{1t}, F_{2t})^\top \), \( F_{1t} = 0.6F_{1,t-1} + u_{1t} \), \( u_{1t} \) are i.i.d. \( N(0, 1-0.6^2) \), \( F_{2t} = 0.3F_{2,t-1} + u_{2t} \), \( u_{2t} \) are i.i.d. \( N(0, 1-0.3^2) \) and independent of \( u_{1t} \). We consider the following setups for the factor loadings \( \lambda_{it} \) and error terms \( e_{it} \).

**DGP1: (Single structural break)**

\[
\lambda_{it} = \begin{cases} 
\alpha_{i1} & \text{for } t = 1, 2, \ldots, t_1 - 1 \\
\alpha_{i2} & \text{for } t = t_1, t_1 + 1, \ldots, T
\end{cases}
\]

where \( \alpha_{i1} \) are from i.i.d. \( N((0.5b, 0.5b)^\top, ((1,0)^\top, (0,1)^\top)) \) and \( \alpha_{i2} \) are from i.i.d. \( N((b,b)^\top, ((1,0)^\top, (0,1)^\top)) \) and independent of \( \alpha_{i1} \). The error terms \( e_{it} \) are generated in two ways: (1) (IID) \( e_{it} \) are i.i.d. \( N(0, 2) \), and (2) (CHeter) \( e_{it} = \sigma_i v_{it} \), where \( \sigma_i \) are i.i.d. \( U(0.5, 1.5) \), \( v_{it} \) are from i.i.d. \( N(0, 2) \), and CHeter denotes cross-sectional heterogeneity in the error terms. Let \( b = 1, 2 \).

**DGP2: (Multiple structural breaks)**

\[
\lambda_{it} = \begin{cases} 
\alpha_{i1} & \text{for } t = 1, 2, \ldots, t_1 - 1 \\
\alpha_{i2} & \text{for } t = t_1, t_1 + 1, \ldots, t_2 - 1 \\
\alpha_{i3} & \text{for } t = t_2, t_2 + 1, \ldots, T
\end{cases}
\]

where \( \alpha_{i1} \) are from i.i.d. \( N((0.5b, 0.5b)^\top, ((1,0)^\top, (0,1)^\top)) \), \( \alpha_{i2} \) are from i.i.d. \( N((b,b)^\top, ((1,0)^\top, (0,1)^\top)) \), \( \alpha_{i3} \) are from i.i.d. \( N((1.5b,1.5b)^\top, ((1,0)^\top, (0,1)^\top)) \), and they are mutually independent of each other. The error terms \( e_{it} \) are generated in two ways: (1) (IID) \( e_{it} \) are i.i.d. \( N(0, 2) \), and (2) (AR(1)) \( e_{it} = 0.2e_{it-1} + u_{it} \), where \( u_{it} \) are i.i.d. \( N(0, 2(1-0.2^2)) \). Let \( b = 1, 2 \).

**DGP3: (No breaks)** \( \lambda_{it} = \alpha_i \) and \( \alpha_i \) are i.i.d. \( N((1,1)^\top, ((1,0)^\top, (0,1)^\top)) \). The error
terms $e_{it}$ are i.i.d. $N(0,2)$.

For each DGP, we simulate 1000 data sets with sample sizes $T = 250, 500$ and $N = 50$. Since the factor loadings are assumed to be nonrandom, we generate them once and fix them across the 1000 replications. We use $J + 1 = 10$ subintervals for $T = 250$ and use $J + 1 = 10, 15$ subintervals for $T = 500$ in the piecewise constant estimation in Step I.

In DGP1, we consider two cases:

(Case 1) we set the break date $t_1 = T/2$, so that $t_1 = 125$ and 250 for $T = 250$ and 500, respectively;

(Case 2) we set $t_1 = T/2 + \lfloor 0.5T/(J + 1) \rfloor$, so that $t_1 = 137$ for $T = 250$ and $t_1 = 275, 266$ for $J + 1 = 10, 15$ and $T = 500$.

It is worth noting that when $T = 250$, $t_1 = 125$ is in the boundary of some subinterval and $t_1 = 137$ is located in the interior of the subinterval. When $T = 500$, $t_1 = 250$ and 266 are in the boundary of some subinterval, respectively, for $J + 1 = 10$ and 15, and $t_1 = 275$ and 250 are in the interior of some subinterval, respectively, for $J + 1 = 10$ and 15.

In DGP2, we consider two cases:

(Case 1) we set the breaks $t_1 = 0.3T$ and $t_2 = 0.7T$, so that $t_1 = 75$ and 150 for $T = 250$ and 500, and $t_2 = 175$ and 350 for $T = 250$ and 500;

(Case 2) we let $t_1 = 0.3T$ and $t_2 = 0.6T + \lfloor 0.5T/(J + 1) \rfloor$, so that $t_2 = 162$ for $T = 250$ and $t_2 = 325, 316$ for $J + 1 = 10, 15$ and $T = 500$.

Similarly to DGP1, some breaks are located in the boundary of an interval and some are in the interior of an interval.

4.2 Determination of the Number of Factors

First, we assume that the true number of factors is unknown. We select the number of factors by the two information criteria $IC_1(R)$ and $IC_2(R)$ given in (4) of Section 3.2. Table 1 presents the average selected number of factors (AVE) and the empirical probability of correct selection (PROB) by the two information criteria for DGP1-3 with $b = 1$. We
observe that the AVE is equal to or close to two, which is the true number of factors, and the PROB is equal to or close to one for all cases. The results in Table 1 demonstrate the selection consistency of the two information criteria established in Section 3.2.

To illustrate the relationship between the IC values and the number of factors, Figure 1 shows the average value of IC$_1(R)$ (thin line) and IC$_2(R)$ (thick line) among 1000 replications against the number of factors for (a) DGP1-Case1 with $T = 250$ and cross-sectionally heteroscedastic error terms; (b) DGP2-Case1 with $T = 250$ and autoregressive error terms; and (c) DGP3. We observe that the average IC value reaches its minimum at $R = 2$ in these three plots. In addition, we find that IC$_2(R)$ has steeper slope than IC$_1(R)$ when $R > 2$ so that it helps to avoid overselecting the number of factors.

### 4.3 Estimation of the Break Points

We set $\kappa = 2$ and 4 to determine the adaptive weight in the adaptive fused Lasso penalty given in Section 2.2.2. For a larger value of $\kappa$, more sparsity is induced. To examine the break detection performance, we calculate the percentages of correct estimation (C) of $m$, and conditional on the correct estimation of $m$, the accuracy of break date estimation, which is measured by average Hausdorff distance of the estimated and true break points divided by $T$ (HD/$T$). Let $D(A, B) \equiv \sup_{b \in B} \inf_{a \in A} |a - b|$ for any two sets $A$ and $B$. The Hausdorff distance between $A$ and $B$ is defined as $\max\{D(A, B), D(B, A)\}$.

The results for DGP 1-2 are shown in Tables 2 and 3 for $\kappa = 2$ and 4, respectively. All figures in the tables are in percentages (%). We observe that the percentage of correct estimation is closer to 100% for the larger signal of $b = 2$. By using the same number of subintervals with $J + 1 = 10$, the C value for $T = 500$ is larger than that for $T = 250$, and the HD/$T$ value for $T = 500$ is smaller than that for $T = 250$ for all cases. Moreover, for the same $T = 500$, the break detection procedure performs better by using $J + 1 = 10$ subintervals than $J + 1 = 15$ subintervals by observing larger C values. Furthermore, the HD/$T$ value for breaks located at the boundaries of the subintervals is smaller than that for breaks in the interior of the subintervals. For example, for DGP1-IID with $T = 500$, for $J + 1 = 10$, the HD/$T$ value for Case 1 (0.032) is smaller than that (0.283) for Case 2, since the break is in the boundary for Case 1 and it is in the interior of some subinterval.
for Case 2. However, the result is reversed for $J + 1 = 15$ by observing 0.365 and 0.041, respectively, for Case 1 and Case 2, since the break is in the boundary for Case 2 for this scenario.

To further evaluate the three-step break detection procedure for DGP1 with one break point, we calculate the frequency for all identified break points among 1000 replications. Since the percentage of correct estimation for $\kappa = 4$ is higher than that for $\kappa = 2$ for each case, in the following we just report the results for $\kappa = 2$ to save spaces. Figures 2-4 show the plots of the frequency of the identified breaks among 1000 replications for DGP1 and for $T = 250$ and $J + 1 = 10$, and $T = 500$ and $J + 1 = 10, 15$, respectively. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$. For plots (a) and (b) of Figure 2, the true break is at $t_1 = 125$, and for plots (c) and (d) of Figure 2, the true break is at $t_1 = 137$. We see that the height of the frequency bar around the true break is close to 1000. This indicates that the three-step procedure can identify the true break or some neighborhood point as a break with a high chance. For the stronger signal with $b = 2$, the identified breaks are more concentrated around the true break than those for the weaker signal with $b = 1$. Moreover, by using the same number of subintervals with $J + 1 = 10$, when we increase the $T$ value from 250 to 500, the frequency bar around the true break is closer to 1000 as shown in Figure 3. For $T = 500$, when we increase $J + 1$ from 10 to 15, more points are identified as breaks, especially for the weaker signal with $b = 1$ as shown in Figures 3 and 4. Figures 5-7 show the plots of the frequency of the identified breaks among 1000 replications for DGP2. We see that the two true breaks can be identified well. We can observe similar patterns as the frequency plots for DGP1. For example, for larger $T$ value, the frequency bars around the true breaks have height closer to 1000.

For DGP3 with no breaks, the false detection proportion among 1000 replication by using $\kappa = 4$ is 0.021, 0.000 and 0.008, respectively, for the three cases: $T = 250$ with $J + 1 = 10$, $T = 500$ with $J + 1 = 10$, and $T = 500$ with $J + 1 = 15$. There is no break detected for $T = 500$ and $J + 1 = 10$, while the false detection proportion is close to zero for the other two cases. This result indicates that our method works well when no break exists in the model.
5 Application

In this section, we apply our proposed method to the U.S. Macroeconomic Data Set (Stock and Watson, 2009) to detect possible structural breaks in the underlying factor model. The data set consists of $N = 108$ monthly macroeconomic time-series variables including real economic activity measures, prices, interest rates, money and credit aggregates, stock prices, exchange rates, etc. for the United States, spanning 1959m01-2006m12. Following the literature, we transform the data by taking the first order difference, so that we obtain a total of $T = 575$ monthly observations for each macroeconomic variable. The data have been centered and standardized for the analysis. We refer to Stock and Watson (2009) for the detailed data description.

We use $J + 1 = 10$ subintervals for the piecewise constant estimation, since as demonstrated in the simulation studies that the method works well for $T = 500$ by using $J + 1 = 10$ subintervals. We let $\kappa = 4$ in the fused penalization procedure. We first determine the appropriate number of common factors. We select the number of factors by the information criteria $IC_2(R)$ given in (4) of Section 3.2. As a result, the number of selected factors is 6. In Figure 8, we plot the values of $IC_2(R)$ against the number of factors. We observe that the IC value reaches its minimum at $R = 6$.

Next, we apply our proposed break detection procedure with the numbers of factors of $R = 6$. The tuning parameter in the fused penalization procedure is selected by the BIC method described in Section 3.5. Our method is able to identify five break dates in 1979m01, 1984m07, 1990m03, 1995m06, and 2002m01, respectively. The year of 1984 was considered as a potential break date by Stock and Watson (2009). As shown in a recent paper of Chen et al. (2014), their Sup-Wald test detected one break date around 1979-1980 (second oil price shock). This break date is also found by our proposed method. They mentioned that one possible explanation could be the impact on monetary policy in the US by the Iranian revolution in the beginning of 1979. Moreover, by using the U.S. labor productivity time-series data, Hansen (2001) plotted the sequence of Chow statistics for testing structural changes as a function of candidate break dates as shown in Figure 1 of page 120. It shows that the curve of the Chow test statistic has two peaks around the years of 1991 and 1995 which indicates that breaks may happen at these time points if any.
By using our proposed method, we detected two break dates in 1990m03 and 1995m06, respectively. For the break date in the year of 2002, it may be attributed to the early 2000s recession (Kliesen, 2003).

6 Conclusion

In this paper, we propose a novel three-step procedure by utilizing nonparametric local estimation, shrinkage methods and grid search to determine the number of breaks and to estimate the break locations in large dimensional factor models. Based on the first-stage piecewise constant estimation of the factor loadings, we also propose a BIC-type information criterion to determine the number of factors. The proposed procedure is easy to implement, computationally efficient, and theoretically reliable. We show that the information criterion can consistently estimate the number of factors and our three-step procedure can consistently estimate the number of breaks and the break locations. Simulation studies demonstrate good performance of the proposed method. An application to U.S. macroeconomic dataset further illustrates the usefulness of our method.

Mathematical Appendix

This appendix provides the proofs of Propositions 2-4 in Section 3. The proof of Proposition 5 as well as that of some technical lemmas are available in the online supplementary material.

A Proofs of The Propositions in Section 3

A.1 Proof of Proposition 2

Let $V_\kappa$, $\Upsilon_\kappa$, and $Q_\kappa$ be as defined in Proposition 1. We first state the following three lemmas that are used in proving Proposition 2. The proofs of these three lemmas are provided in the online Supplementary Material.
Lemma A.1. Suppose that Assumptions A1-A4 hold. Suppose that $S_j$ contains a break point $t^0_\kappa$ for some $\kappa = \kappa(j)$ and $j \in \mathbb{S}_{2a}$. Then

(i) $N^{-1} \hat{\Delta}_j^T \left[ (N\tau_j)^{-1} X_{S_j} X_{S_j}^T \right] \hat{\Delta}_j = V_{N\tau_j, j} = V_{\kappa - 1} + O_P(\eta_{N\tau_j}^{-1} + \tau_{j2}/\tau_j)$,

(ii) $N^{-1} \hat{\Delta}_j^T \alpha_{\kappa - 1}^0 = Q_{\kappa - 1} + O_P(\eta_{N\tau_j}^{-1} + \tau_{j2}/\tau_j)$,

(iii) $H_{j,1} = \Sigma_F^{1/2} \kappa_1 V_{\kappa - 1}^{-1/2} + O_P(\eta_{N\tau_j}^{-1} + \tau_{j2}/\tau_j)$,

(iv) $\frac{1}{N} \left\| \hat{\Delta}_j - \alpha_{\kappa - 1}^0 H_{j,1} \right\|^2 = \frac{1}{N} \sum_{i=1}^N \left\| \hat{\Delta}_{ij} - H_{i,1}^T \alpha_{i,\kappa - 1}^0 \right\|^2 = O_P(\eta_{N\tau_j}^{-2} + (\tau_{j2}/\tau_j)^2)$,

(v) $\hat{F}_t = \frac{1}{N} \hat{\Delta}_j^T \alpha_\kappa^* F_t^* + O_P(N^{-1/2} + \eta_{N\tau_j}^{-1} (\eta_{N\tau_j}^{-1} + \tau_{j2}/\tau_j))$ for each $t \in S_j$,

(vi) $\frac{1}{\tau_j} \left\| \hat{F}_{S_j} - F_{S_j}^T H_{j,1}^T \right\|^2 = O_P(N^{-1} + \eta_{N\tau_j}^{-2} (\eta_{N\tau_j}^{-1} + \tau_{j2}/\tau_j)^2)$.

Lemma A.2. Suppose that Assumptions A1-A4 hold. Suppose that $S_j$ contains a break point $t^0_\kappa$ for some $\kappa = \kappa(j)$ and $j \in \mathbb{S}_{2b}$. Then

(i) $N^{-1} \hat{\Delta}_j^T \left[ (N\tau_j)^{-1} X_{S_j} X_{S_j}^T \right] \hat{\Delta}_j = V_{N\tau_j, j} = V_{\kappa} + O_P(\eta_{N\tau_j}^{-1} + \tau_{j1}/\tau_j)$,

(ii) $N^{-1} \hat{\Delta}_j^T \alpha_{\kappa}^0 = Q_\kappa + O_P(\eta_{N\tau_j}^{-1} + \tau_{j1}/\tau_j)$,

(iii) $H_{j,2} = \Sigma_F^{1/2} \kappa_2 V_{\kappa - 1}^{-1/2} + O_P(\eta_{N\tau_j}^{-1} + \tau_{j1}/\tau_j)$,

(iv) $\frac{1}{N} \left\| \hat{\Delta}_j - \alpha_{\kappa}^0 H_{j,2} \right\|^2 = \frac{1}{N} \sum_{i=1}^N \left\| \hat{\Delta}_{ij} - H_{i,2}^T \alpha_{i,\kappa}^0 \right\|^2 = O_P(\eta_{N\tau_j}^{-2} + (\tau_{j1}/\tau_j)^2)$,

(v) $\hat{F}_t = \frac{1}{N} \hat{\Delta}_j^T \alpha_\kappa^* F_t^* + O_P(N^{-1/2} + \eta_{N\tau_j}^{-1} (\eta_{N\tau_j}^{-1} + \tau_{j1}/\tau_j))$ for each $t \in S_j$,

(vi) $\frac{1}{\tau_j} \left\| \hat{F}_{S_j} - F_{S_j}^T H_{j,2}^T \right\|^2 = O_P(N^{-1} + \eta_{N\tau_j}^{-2} (\eta_{N\tau_j}^{-1} + \tau_{j1}/\tau_j)^2)$.

Lemma A.3. Suppose that Assumptions A1-A4 hold. Suppose that $S_j$ contains a break point $t^0_\kappa$ for some $\kappa = \kappa(j)$ and $j \in \mathbb{S}_{2c}$. Then

(i) $N^{-1} \hat{\Delta}_j^T \left[ (N\tau_j)^{-1} X_{S_j} X_{S_j}^T \right] \hat{\Delta}_j = V_{N\tau_j, j} = V_{\kappa^*} + O_P(\eta_{N\tau_j}^{-1})$,

(ii) $N^{-1} \hat{\Delta}_j^T \alpha_{\kappa^*}^* = Q_{\kappa^*} + O_P(\eta_{N\tau_j}^{-1})$,

(iii) $H_{j^*} = \Sigma_F^{1/2} \kappa_2 V_{\kappa^*}^{-1/2} + O_P(\eta_{N\tau_j}^{-1})$,

(iv) $\frac{1}{N} \left\| \hat{\Delta}_j - \alpha_{\kappa^*}^* H_{j^*} \right\|^2 = \frac{1}{N} \sum_{i=1}^N \left\| \hat{\Delta}_{ij} - H_{i,j}^T \alpha_{i,\kappa^*} \right\|^2 = O_P(\eta_{N\tau_j}^{-2})$,

(v) $\hat{F}_t = \frac{1}{N} \hat{\Delta}_j^T \alpha_{\kappa^*}^* F_t^* + O_P(N^{-1/2} + \tau_{j^{-1}})$ for each $t \in S_j$ and $j \in \mathbb{S}_{2c}$,

(vi) $\frac{1}{\tau_j} \left\| \hat{F}_{S_j} - F_{S_j}^T \frac{1}{N} \alpha_{\kappa^*} \hat{\Delta}_j \right\|^2 = O_P(N^{-1} + \tau_{j^{-2}})$,

where $V_{\kappa^*}$ is the diagonal matrix consisting of the $R$ largest eigenvalues of $\Sigma_{F_k}^{1/2} \Sigma_{\kappa^*}^{1/2} \Sigma_{F_k}^{1/2}$ in descending order, $\kappa_{\kappa^*}$ being the corresponding (normalized) $2R \times R$ eigenvector matrix, $Q_{\kappa^*} = V_{\kappa^*}^{1/2} \kappa_{\kappa^*} \Sigma_{F_k}^{-1/2}$, $\Sigma_{F_k} = \text{diag}(c_j \Sigma_F, (1 - c_j) \Sigma_F)$, $c_j = \tau_{j1}/\tau_j$, and $\Sigma_{\kappa^*} = \lim_{N \to \infty} N^{-1} \alpha_{\kappa^*} \alpha_{\kappa^*}^T$.

The first part of Proposition 2(i) follows from Lemma A.1(iv). For the second part of Proposition 2(i), we have by the Cauchy-Schwarz inequality and the submultiplicative
property of Frobenius norm,

\[
\frac{1}{N}\|\hat{\Delta}_j V_{N\tau,j} - \alpha^0_{\kappa-1} \Sigma_F Q_{\kappa-1}^\top\|^2 \\
\leq \frac{2}{N}\|\hat{\Delta}_j - \alpha^0_{\kappa-1} H_{j,1}\| V_{N\tau,j}\|^2 + \frac{2}{N}\|\alpha^0_{\kappa-1} (H_{j,1} V_{N\tau,j} - \Sigma_F Q_{\kappa-1}^\top)\|^2 \\
\leq \frac{2}{N}\|\hat{\Delta}_j - \alpha^0_{\kappa-1} H_{j,1}\|^2 \|V_{N\tau,j}\|^2 + \frac{2}{N}\|\alpha^0_{\kappa-1}\|^2 \|H_{j,1} V_{N\tau,j} - \Sigma_F Q_{\kappa-1}^\top\|^2
\]

\[= O_P(\eta_{N\tau,j}^2 + (\tau_{j-2}/\tau_j)^2),\]

where the last equality follows from Lemma A.1(i), (iii) and (iv). Analogously, we can apply Lemma A.2 to prove Proposition 2(ii).

The first part of Proposition 2(iii) follows from Lemma A.3(iv). For the second part of Proposition 2(iii), noting that \(\alpha^*_{\kappa} H^*_j = (N\tau_j)^{-1}(\alpha^0_{\kappa-1} F^\top S_{j,1} F_{S_{j,1}} \alpha^0_{\kappa-1} + \alpha^0_{\kappa} F^\top S_{j,2} F_{S_{j,2}} \alpha^0_{\kappa}) \hat{\Delta}_j V_{N\tau,j}^{-1}\) by the definitions of \(\alpha^*_{\kappa}\) and \(H^*_j\), we have for any \(t \in S_j\) and \(j \in S_{2c}\)

\[
\frac{1}{N}\left\|\hat{\Delta}_j V_{N\tau,j} - (N\tau_j)^{-1}(\alpha^0_{\kappa-1} F^\top S_{j,1} F_{S_{j,1}} \alpha^0_{\kappa-1} + \alpha^0_{\kappa} F^\top S_{j,2} F_{S_{j,2}} \alpha^0_{\kappa}) \hat{\Delta}_j V_{N\tau,j}^{-1}\right\|^2 \\
\leq \frac{1}{N}\left\|\hat{\Delta}_j - \alpha^*_{\kappa} H^*_j\right\|^2 \|V_{N\tau,j}\|^2 = O_P(\eta_{N\tau,j}^{-2})
\]

by Lemmas A.3(i) and (iv).

\section{Proof of Proposition 3}

Let \(S, S_1, S_{2a}, S_{2b},\) and \(S_{2c}\) be as defined in Section 3.1. Recall that \(\alpha^*_{\kappa} = (\alpha^0_{\kappa-1}, \alpha^0_{\kappa}),\)
\(F^*_{S_{j}} = (F^*_{v_1}, \ldots, F^*_{v_{1+1}})^\top,\) and \(F^*_{F_{l}} = (F^0_{l,1}, F^0_{l,1})^\top,\) where \(1 = 1 \{v_j \leq t < \hat{\tau}_k\}, \hat{\tau}_1 = 1 \{t \leq t < \hat{\tau}_j\}.\) Define \(H^R_{j} \equiv (\frac{1}{\tau_j} \hat{\Delta}_{j}^R)^{\top} \left(\frac{1}{N} \alpha^0_{\kappa} \hat{\Delta}_{j}^R\right),\) an \(R \times R\) matrix, and \(H^R_{j,\ell} \equiv \left(\frac{1}{\tau_j} \hat{\Delta}_{j}^R\right)^{\top} \left(\frac{1}{N} \alpha^0_{\kappa+\ell-2} \hat{\Delta}_{j}^R\right),\) an \(2R \times R\) matrix. Similarly, let \(H^R_{j,\ell} \equiv (\frac{1}{\tau_j} \hat{\Delta}_{j}^R)^{\top} \left(\frac{1}{N} \alpha^0_{\kappa+\ell-2} \hat{\Delta}_{j}^R\right),\) for \(\ell = 1, 2.\) Let \(J_1 = J + 1\) and \(\tau = \min_{0 \leq \ell \leq J} \tau_j.\) Define

\[
\bar{\Delta}_{j}^R = \begin{cases} 
\alpha^0_{\kappa} H^R_{j} & \text{if } j \in S_1 \\
\alpha^0_{\kappa-1} H^R_{j,1} & \text{if } j \in S_{2a} \\
\alpha^0_{\kappa} H^R_{j,2} & \text{if } j \in S_{2b} \\
\alpha^0_{\kappa} H^R_{j} & \text{if } j \in S_{2c}
\end{cases}
\]

and \(\Delta_{j}^0 = \begin{cases} 
\alpha^0_{\kappa} & \text{if } j \in S_1 \\
\alpha^0_{\kappa-1} & \text{if } j \in S_{2a} \\
\alpha^0_{\kappa} & \text{if } j \in S_{2b} \\
\alpha^0_{\kappa} & \text{if } j \in S_{2c}
\end{cases}\)

for some \(\kappa = \kappa(j).\)

To prove Proposition 3, we need the following three lemmas. More precisely, Lemmas A.4 and A.5 are used in the proof of Lemma A.6, which is used to prove Proposition 3.
The proofs of these three lemmas are provided in the on-line Supplementary Material.

**Lemma A.4.** Suppose that Assumptions A1-A4 hold. Then for any $R \geq 1$, there exist $R_0 \times R$ matrices $\{H_j^{(R)}, H_{j,1}^{(R)}, H_{j,2}^{(R)}\}$ and $2R_0 \times R$ matrices $\{\hat{H}_j^{(R)}\}$ with $\text{rank}(H_j^{(R)}) = \min \{ R, R_0 \}$, $\text{rank}(\hat{H}_j^{(R)}) = \min \{ R, R_0 \}$ with $\ell = 1, 2$, and $\text{rank}(H_j^{(R)}) = \min \{ R, 2R_0 \}$ such that

\[
(i) \sum_{j \in S_1} N^{-1} \left\| \hat{\Delta}_j^{(R)} - \Delta_j^{(R)} \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right),
\]

\[
(ii) \max_{j \in S_1} N^{-1} \left\| \hat{\Delta}_j^{(R)} - \Delta_j^{(R)} \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right),
\]

\[
(iii) \max_{j \in S_1} \left\| N^{-1} \hat{\Delta}_j^{(R)} - N^{-1} \Delta_j^{(R)} \right\|^2 = O_P \left( \eta_{N \tau}^{-1} \right),
\]

where $H_j^{(R)}$, $H_{j,1}^{(R)}$, $H_{j,2}^{(R)}$ and $\hat{H}_j^{(R)}$ are implicitly defined in $\hat{\Delta}_j^{(R)}$ in (A.1).

**Lemma A.5.** Suppose that Assumptions A1-A4 hold and $R > R_0$. Write the Moore-Penrose generalized inverse of $H_j^{(R)}$ as $H_j^{(R)+} = \begin{pmatrix} H_j^{(R)+} (1) \\ H_j^{(R)+} (2) \end{pmatrix}$, where $H_j^{(R)+} (1)$ and $H_j^{(R)+} (2)$ are $R_0 \times R_0$ and $(R - R_0) \times R_0$ matrices, respectively. Let $V_{N \tau,j}^{(R)}$ denote an $R \times R$ diagonal matrix consisting of the $R$ largest eigenvalues of the $N \times N$ matrix $(N \tau_j)^{-1} X S_j X^\top$, where the eigenvalues are ordered in decreasing order along the main diagonal line. Write $\hat{\Delta}_j^{(R)} = [\hat{\Delta}_j^{(R)} (1) \  \hat{\Delta}_j^{(R)} (2)]$ and $H_j^{(R)} = [H_j^{(R)} (1) \  H_j^{(R)} (2)]$, where $\hat{\Delta}_j^{(R)} (1)$, $\hat{\Delta}_j^{(R)} (2)$, $H_j^{(R)} (1)$, and $H_j^{(R)} (2)$ are $N \times R_0$, $N \times (R - R_0)$, $R_0 \times R_0$, and $R_0 \times (R - R_0)$ matrices, respectively. Write $V_{N \tau,j}^{(R)} = \text{diag} \left( V_{N \tau,j}^{(R)} (1), V_{N \tau,j}^{(R)} (2) \right)$, where $V_{N \tau,j}^{(R)} (1)$ denotes the upper left $R_0 \times R_0$ submatrix of $V_{N \tau,j}^{(R)}$. Then

\[
(i) \max_{j \in S_1} N^{-1} \left\| \hat{\Delta}_j^{(R)} (1) - \alpha_0^0 H_j^{(R)} (1) V_{N \tau,j}^{(R)} (1)^{-1} \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right) \text{ and} \max_{j \in S_1} \left\| H_j^{(R)} (2) \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right) \text{ and} \max_{j \in S_1} \left\| H_j^{(R)+} (2) \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right),
\]

\[
(ii) \max_{j \in S_1} \left\| H_j^{(R)+} (1) \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right) \text{ and} \max_{j \in S_1} \left\| H_j^{(R)+} (2) \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right) \text{ and} \max_{j \in S_1} \left\| H_j^{(R)+} (2) \right\|^2 = O_P \left( \eta_{N \tau}^{-2} \right),
\]

\[
(iii) \sum_{j \in S_1} N^{-1} \left\{ (N \tau_j)^{-1} \text{tr} \left( F_{S_j}^0 H_j^{(R)+} (1) \hat{\Delta}_j^{(R)} - \alpha_0^0 H_j^{(R)} (1) V_{N \tau,j}^{(R)} (1)^{-1} \right) \right\} = O_P \left( \eta_{N \tau}^{-2} \right),
\]

\[
(iv) \sum_{j \in S_1} N^{-1} \left\{ \left( \hat{\Delta}_j^{(R)} - \alpha_0^0 H_j^{(R)} \right) H_j^{(R)+} (1) + F_{S_j}^0 \right\}^2 = O_P \left( \eta_{N \tau}^{-2} \right).
\]

**Lemma A.6.** Suppose that Assumptions A1-A4 hold. Then

\[
(i) V \left( \hat{\Delta}_j^{(R)} \right) - V \left( \Delta_j^{(R)} \right) = O_P \left( \eta_{N \tau}^{-1} \right) \text{ for each } R \text{ with } 1 \leq R \leq R_0,
\]

\[
(ii) V \left( \hat{\Delta}_j^{(R)} \right) - V \left( H_j^{(R)+} \right) = O_P \left( \eta_{N \tau}^{-1} \right) \text{ for each } R \text{ with } 1 \leq R \leq R_0.
\]
(ii) there exists a $c_R > 0$ such that \( \liminf_{(N,T) \to \infty} \left[ V \left( R, \{ \hat{\Delta}^{(R)}_j \} \right) - V \left( R, \{ \Delta^0_j \} \right) \right] \geq c_R \) for each $R$ with $1 \leq R < R_0$.

(iii) $V \left( R, \{ \Delta^0_j \} \right) - V \left( R_0, \{ \hat{\Delta}^{(R_0)}_j \} \right) = O_p \left( m J^{-1} + \eta_{N,T}^{-2} \right)$ for each $R$ with $R \geq R_0$, where $\Delta^0_j, j = 0, 1, \ldots, J$, are defined in (A.1).

**Proof of Proposition 3.** Let $V \left( R \right) = V(R, \{ \hat{\Delta}^{(R)}_j \} )$ for all $R$. Note that $IC \left( R \right) - IC \left( R_0 \right) = \ln \left[ V \left( R \right) / V \left( R_0 \right) \right] + \left( R - R_0 \right) \rho_{NT}$. We discuss two cases: (i) $R < R_0$, and (ii) $R > R_0$.

In case (i), $V \left( R \right) / V \left( R_0 \right) > 1 + \epsilon_0$ for some $\epsilon_0 > 0$ w.p.a.1 by Lemmas A.6(i) and (ii). It follows that $\ln \left[ V \left( R \right) / V \left( R_0 \right) \right] \geq \epsilon_0 / 2$ w.p.a.1. Noting that $(R - R_0) \rho_{NT} \to 0$ under Assumption A6, this implies that $IC \left( R \right) - IC \left( R_0 \right) \geq \epsilon_0 / 4$ w.p.a.1. Consequently, we have $P \left( IC \left( R \right) - IC \left( R_0 \right) > 0 \right) \to 1$ for any $R < R_0$ as $(N,T) \to \infty$. In case (ii), we apply Lemma A.6(iii) and Assumption A6 to obtain

$$P \left( IC \left( R \right) - IC \left( R_0 \right) > 0 \right) = P \left\{ \ln \left[ V \left( R \right) / V \left( R_0 \right) \right] + \left( R - R_0 \right) \rho_{NT} \right\} \to 1$$

for any $R > R_0$ as $(N,T) \to \infty$. Consequently, the minimizer of $IC \left( R \right)$ can only be achieved at $R = R_0$ w.p.a.1. That is, $P(\hat{R} = R_0) \to 1$ for any $R \in [1, R_{\max}]$ as $(N,T) \to \infty$.

**A.3 Proof of Proposition 4**

Recall $\hat{Z}_{jt} = N^{-1/2} \| \hat{\Delta}_j V_{N_{\tau,j}} \| V_{N_{\tau,j}}^{-1} \hat{F}_t$, and $\hat{Z}_{S_j} = (\hat{Z}_{jt}, t \in S_j) = N^{-1/2} \hat{F}_{S_j} V_{N_{\tau,j}}^{-1} \| \hat{\Delta}_j V_{N_{\tau,j}} \|$ (a $\tau_j \times R$ matrix) where $\hat{F}_{S_j} = (\hat{F}_t, t \in S_j)$. Let $a_j$ be defined as in Proposition 4. Let

$$\bar{\Delta}_j = \begin{cases} \alpha^0_k \Sigma_F Q_k^\top & \text{if } j \in S_1 \\ \alpha^0_k \Sigma_F Q_{k-1}^\top & \text{if } j \in S_{2a} \\ \alpha^0_k \Sigma_F Q_k^\top & \text{if } j \in S_{2b} \\ \alpha_k^\kappa H_k V_{\kappa^*} & \text{if } j \in S_{2c} \end{cases}, \quad \bar{V}_j = \begin{cases} V_k & \text{if } j \in S_1 \\ V_{k-1} & \text{if } j \in S_{2a} \\ V_k & \text{if } j \in S_{2b} \\ V_{\kappa^*} & \text{if } j \in S_{2c} \end{cases}$$

for some $\kappa = \kappa \left( j \right)$. 

$$\bar{H}_j = \begin{cases} \Sigma_F Q_k^\top V_k^{-1} & \text{if } j \in S_1 \\ \Sigma_F Q_{k-1}^\top V_{k-1}^{-1} & \text{if } j \in S_{2a} \\ \Sigma_F Q_k^\top V_k^{-1} & \text{if } j \in S_{2b} \end{cases}, \quad \bar{H}_{\kappa^*} = \Sigma_F^\frac{1}{2} Y_{\kappa^*} V_{\kappa^*}^{-1/2} = \Sigma_F^\frac{1}{2} Q_{\kappa^*} V_{\kappa^*}^{-1} \text{ if } j \in S_{2c}$$
Note that $\tilde{\Delta}_j$ and $\tilde{V}_j$ denote the probability limits of $\Delta_j V_{N\tau_{j,j}}$ and $V_{N\tau_{j,j}}$, respectively. Let

$$Z_{ij} = (Z_{ij}, t \in S_j) = \begin{cases} N^{-1/2} F_{ij} H_j^T + \tilde{V}_j^{-1} \| \Delta_j \| & \text{if } j \in S_1 \cup S_{2a} \cup S_{2b} \\ N^{-1/2} F_{ij} Q_{\kappa^*, \kappa^*} V_{\kappa^*, \kappa^*}^{-1} \| \Delta_j \| & \text{if } j \in S_{2c} \end{cases},$$

where $Z_{ij} = N^{-1/2} \tilde{V}_j^{-1} H_j^T \| \Delta_j \|$ if $j \in S_1 \cup S_{2a} \cup S_{2b}$, and $= N^{-1/2} Q_{\kappa^*, \kappa^*} F_{ij} \| \Delta_j \|$ if $j \in S_{2c}$. Let $\Theta_j^* \equiv N^{-1/2} \Delta_j / \| \Delta_j \|$.

To prove Proposition 4, we need a lemma.

**Lemma A.7.** Let $E_{ij}^* = X_j - \Theta_j^* Z_{ij}^T$. Let $\vartheta_j = (\vartheta_{j,1}, ..., \vartheta_{j,R})$, an $N \times R$ matrix, for $j = 0, 1, ..., J$. Let $\vartheta_j = N^{-1/2} \text{vec}(\vartheta_j)$. Suppose that the conditions in Proposition 4 hold. Then for each $j \in S$, we have

(i) $\frac{1}{\tau_j a_j^2} \left\| \hat{Z}_{ij} - Z_{ij} \right\|^2 = o_P(1)$,

(ii) $\frac{1}{N^{1/2} \tau_j a_j^2} \left\| E_{ij}^* (\hat{Z}_{ij} - Z_{ij}) \right\| = o_P(1)$,

(iii) $\frac{1}{N^{1/2} \tau_j a_j^2} \left\| E_{ij}^* Z_{ij} \right\| = o_P(1)$,

(iv) $\frac{1}{N \tau_j a_j} \text{tr} \left[ \vartheta_j \left( \hat{Z}_{ij} Z_{ij}^T - Z_{ij} \hat{Z}_{ij}^T \right) \vartheta_j^T \right] = o_P(1) \| \vartheta_j \|^2$,

(v) $\frac{1}{N \tau_j a_j} \text{tr} \left[ \left( X_j - \Theta_j^* Z_{ij}^T \right) \vartheta_j \hat{Z}_{ij}^T - \left( X_j - \Theta_j^* Z_{ij}^T \right) \vartheta_j Z_{ij}^T \right] = o_P(1) \| \vartheta_j \|^2$.

**Proof of Proposition 4.** (i) Let $\Theta_j = \Theta_j^* + a_j \vartheta_j$. Let

$$\Gamma_{NT, \gamma} \left( \{ \Theta_j \} \right) = \frac{1}{2N} \sum_{j=0}^{J} \frac{1}{\tau_j} \sum_{t \in S_j} (X_t - \Theta_j \hat{Z}_{jt})^T (X_t - \Theta_j \hat{Z}_{jt}) + \gamma \sum_{j=0}^{J} w_j \| \Theta_j - \Theta_{j-1} \|$$

$$= \frac{1}{2N} \sum_{j=0}^{J} \frac{1}{\tau_j} \text{tr} \left[ (X_j - \Theta_j \hat{Z}_{ij}^T)^T (X_j - \Theta_j \hat{Z}_{ij}^T) \right] + \gamma \sum_{j=0}^{J} w_j \| \Theta_j - \Theta_{j-1} \|.$$

Let $c_j, j = 0, 1, ..., J$ be arbitrary positive constants that do not depend on $(N, T)$. Our aim is to show that for any given $\epsilon > 0$, there exists a large constant $L$ such that for sufficiently large $(N, T)$ we have

$$P \left\{ \inf_{N^{-1/2} \| \vartheta_j \| = c_j \text{L}, j = 0, 1, ..., J} \Gamma_{NT, \gamma} \left( \{ \Theta_j^* + a_j \vartheta_j \} \right) > \Gamma_{NT, \gamma} \left( \{ \Theta_j^* \} \right) \right\} \geq 1 - \epsilon. \quad (A.2)$$

This implies that w.p.a.1 there is a local minimum $\{ \hat{\Theta}_j \}$ such that the estimator $\{ \hat{\Theta}_j \}$ lies inside the ball $\{ \{ \Theta_j^* + a_j \vartheta_j \} : N^{-1/2} \| \vartheta_j \| \leq c_j \text{L} \}$. Then we have $N^{-1/2} \left\| \hat{\Theta}_j - \Theta_j^* \right\| = O_P(a_j)$ for $j = 0, 1, ..., J$.

Let $D(\{ \vartheta_j \}) = \Gamma_{NT, \gamma}(\{ \Theta_j^* + a_j \vartheta_j \}) - \Gamma_{NT, \gamma}(\{ \Theta_j^* \})$. Noting that $X_j - \Theta_j \hat{Z}_{ij}^T = \...
\[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top) - a_j \vartheta_j \hat{Z}_{S_j}^\top, \] we have

\[
D (\{ \vartheta_j \}) = \frac{1}{2N} \sum_{j=0}^{J-1} \frac{1}{\tau_j} \text{tr} \left[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)(X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top) \right] \\
+ \gamma \sum_{j=1}^{J-1} w_j \left\{ \| \Theta_j - \Theta_{j-1} \| - \| \Theta_j^* - \Theta_{j-1}^* \| \right\}
\]

\[
= \frac{1}{2N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ \vartheta_j Z_{S_j}^\top Z_{S_j} \vartheta_j^\top \right] - \frac{1}{N} \sum_{j=0}^{J} \frac{a_j}{\tau_j} \text{tr} \left[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top \right] \\
+ \gamma \sum_{j=1}^{J-1} w_j \left\{ \| \Theta_j - \Theta_{j-1} \| - \| \Theta_j^* - \Theta_{j-1}^* \| \right\}
\]

\[
= \frac{1}{2N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ \vartheta_j Z_{S_j}^\top Z_{S_j} \vartheta_j^\top \right] - \frac{1}{N} \sum_{j=0}^{J} \frac{a_j}{\tau_j} \text{tr} \left[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top \right] \\
+ \frac{1}{2N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ \vartheta_j \left( \hat{Z}_{S_j}^\top Z_{S_j} - Z_{S_j}^\top Z_{S_j} \right) \vartheta_j^\top \right] \\
- \frac{1}{N} \sum_{j=0}^{J} \frac{a_j}{\tau_j} \text{tr} \left[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top - (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top \right] \\
+ \gamma \sum_{j=1}^{J-1} w_j \left\{ \| \Theta_j - \Theta_{j-1} \| - \| \Theta_j^* - \Theta_{j-1}^* \| \right\}
\]

\[ = D_1 (\{ \vartheta_j \}) - D_2 (\{ \vartheta_j \}) + D_3 (\{ \vartheta_j \}) - D_4 (\{ \vartheta_j \}) + D_5 (\{ \vartheta_j \}), \text{ say.} \]

Recall that \( \vartheta_j = N^{-1/2} \text{vec}(\vartheta_j) \) and \( \bm{E}_{S_j}^* = X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top \). Let \( \bm{A}_j = \frac{1}{\tau_j} Z_{S_j}^\top Z_{S_j} \) and \( \bm{B}_j = \frac{1}{N \tau_j a_j} \text{vec}(\bm{E}_{S_j}^* Z_{S_j}) \). Apparently, \( \| \bm{A}_j \| = O_P (1) \). By Lemma A.7(iii), \( \| \bm{B}_j \| = O_P (1) \) for \( j \in S \). Noting that \( \text{tr}(B_1B_2) = \text{vec}(B_1^\top)^\top \text{vec}(B_1) \) and \( \text{tr}(B_1B_2B_3) = \text{vec}(B_1^\top) (B_2 \otimes I) \text{vec}(B_3^\top) \) for any conformable matrices \( B_1, B_2, B_3 \) and an identity matrix \( I \) (see, e.g., Bernstein (2005, p.247 and p.253)), we have

\[
D_1 (\{ \vartheta_j \}) = \frac{1}{2N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ \vartheta_j Z_{S_j}^\top Z_{S_j} \vartheta_j^\top \right] = \frac{1}{2} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ (\bm{A}_j \otimes I_R) \vartheta_j \right], \text{ and}
\]

\[
D_2 (\{ \vartheta_j \}) = \frac{1}{N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j a_j} \text{tr} \left[ \vartheta_j Z_{S_j}^\top Z_{S_j} \vartheta_j^\top \right] = \sum_{j=0}^{J} \frac{a_j^2}{\tau_j a_j} \text{tr} \left[ \vartheta_j \right] \text{tr} \left[ \bm{B}_j \vartheta_j \right].
\]

By Lemmas A.7(iv)-(v),

\[
D_3 (\{ \vartheta_j \}) = \frac{1}{2N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \text{tr} \left[ \vartheta_j \left( \hat{Z}_{S_j}^\top Z_{S_j} - Z_{S_j}^\top Z_{S_j} \right) \vartheta_j^\top \right] = \sum_{j=0}^{J} \frac{a_j^2}{\tau_j} \left\| \vartheta_j \right\|^2, \text{ and}
\]

\[
D_4 (\{ \vartheta_j \}) = \frac{1}{N} \sum_{j=0}^{J} \frac{a_j^2}{\tau_j a_j} \text{tr} \left[ (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top - (X_{S_j} - \Theta_j^* \hat{Z}_{S_j}^\top)^\top \vartheta_j \hat{Z}_{S_j}^\top \right] \\
= \sum_{j=0}^{J} O_P \left( \frac{a_j^2}{\tau_j} \right) \left\| \vartheta_j \right\|.
\]

To study \( D_5 (\{ \vartheta_j \}) \), we define the event \( E_{NT} = \{ j - 1, j \in S : j - 1 \in S_2 \text{ and } j \in S_2 \} \). Let \( E_{NT}^c \) denote the complement of \( E_{NT} \). Noting that \( T / (J + 1) \ll I_{\text{min}} \), we have \( P (E_{NT}^c) \rightarrow 1 \)
as \((N, T) \to \infty\). Conditional on the event \(\mathcal{E}^c_{NT}\),

\[
D_5 \left( \{ \vartheta_j \} \right) = \gamma \sum_{j \in S_1, j \notin S_1} w_j \| \Theta_j - \Theta_{j-1} \| + \gamma \left\{ \sum_{j \in S_1, j \notin S_{2a} \cup S_{2c}} + \sum_{j \in S_1, j \notin S_{2b}} \right\} w_j \left\{ \| \Theta_j - \Theta_{j-1} \| - \| \Theta_j - \Theta_{j-1} \| \right\}
\]

\[
\equiv D_{5,1} \left( \{ \vartheta_j \} \right) + \sum_{l=2}^5 D_{5,l} \left( \{ \vartheta_j \} \right), \quad \text{say,}
\]

where, e.g., \(\sum_{j \in S_1, j \notin S_1} = \sum_{j=1, j \notin S_1, j \notin S_1}^J\). Apparently, \(D_{5,1} \left( \{ \vartheta_j \} \right) \geq 0\). Noting that when \(j \in S_1\) and \(j-1 \in S_{2b}\), \(\Theta_j = \Theta_{j-1}\) and \(D_{5,3} \left( \{ \vartheta_j \} \right) = \gamma \sum_{j \in S_1, j \notin S_{2b}} w_j \| \Theta_j - \Theta_{j-1} \| \geq 0\). Similarly, when \(j \in S_1\) and \(j \in S_{2a}\), \(\Theta_j = \Theta_{j-1}\) and

\[
D_{5,4} \left( \{ \vartheta_j \} \right) = \gamma \sum_{j \in S_{2a}, j \notin S_1} w_j \| \Theta_j - \Theta_{j-1} \| \geq 0.
\]

When \(j \in S_1\), \(j-1 \in S_{2a} \cup S_{2c}\), \(\Theta_j - \Theta_{j-1} \neq 0\) and

\[
|D_{5,2} \left( \{ \vartheta_j \} \right)| \leq \gamma \sum_{j \in S_1, j \notin S_{2a} \cup S_{2c}} a_j w_j \| \vartheta_j - \vartheta_{j-1} \|
\]

\[
\leq 2 \gamma \sum_{j \in S_1, j \notin S_{2a} \cup S_{2c}} a_j w_j \sum_{j=0}^J a_j \| \text{vec} (\vartheta_j) \| = O_P((N\tau)^{1/2}) \sum_{j=0}^J a_j^2 \| \vartheta_j \|,
\]

where we use the fact that \(\max_{j \in S_1, j \notin S_{2a} \cup S_{2c}} w_j = O_P(1)\) and \(a_j = O (\tau^{1/2})\). By the same token, we can show that \(|D_{5,5} \left( \{ \vartheta_j \} \right)| \leq O_P((N\tau)^{1/2}) \sum_{j=0}^J a_j^2 \| \vartheta_j \|.
\]

Consequently, we have shown that

\[
D \left( \{ \vartheta_j \} \right) \geq \frac{1}{2} \sum_{j=0}^J a_j^2 \| \vartheta_j \|^2 \left( A_j \otimes I_R \right) \vartheta_j - \sum_{j=0}^J a_j^2 B_j \vartheta_j
\]

\[
- O_P((N\tau)^{1/2} + 1) \sum_{j=0}^J a_j^2 \| \vartheta_j \| + \text{s.m.}
\]

\[
\geq \sum_{j=0}^J a_j^2 \left\{ \frac{1}{2} \mu_{\min} (A_j) \| \vartheta_j \|^2 - [\| B_j \| + O_P(1)] \| \vartheta_j \| \right\} + \text{s.m.,}
\]

where s.m. denotes terms that are of smaller order than the preceding displayed terms. Noting that \(\mu_{\min} (A_j) \geq c > 0\) and \(\| B_j \| = O_P(1)\), by allowing \(\| \vartheta_j \| = N^{-1/2} \| \vartheta_j \|\) sufficiently large, the linear term \([\| B_j \| + O_P(1)] \| \vartheta_j \|\) will be dominated by the quadratic term \(\frac{1}{2} \mu_{\min} (A_j) \| \vartheta_j \|^2\). This implies that \(N^{-1/2} \| \vartheta_j \|\) has to be stochastically bounded for each \(j\) in order for \(D \left( \{ \vartheta_j \} \right)\) to be minimized. That is, (A.2) must hold for some large positive constant \(L\) and \(N^{-1/2} \left\| \Theta_j - \Theta_j^* \right\| = O_P(a_j)\) for \(j = 0, 1, \ldots, J\).

(ii) Define \(S = \left\{ j \in S : \Theta_j \neq \Theta_j^* \right\}\) and \(S^c = \left\{ j \in S : \Theta_j^* = \Theta_j^* \right\}\). We focus on the case where \(|S| \geq 1\) which implies that \([1, T]\) contains at least one break. We will
show that
\[
\Pr \left\{ \| \tilde{\Theta}_j - \tilde{\Theta}_{j-1} \| = 0 \text{ for all } j, j - 1 \in S_1 \right\} \to 1 \text{ as } (N, T) \to \infty. \tag{A.3}
\]
Suppose that to the contrary, \( \tilde{\beta}_j = \tilde{\Theta}_j - \tilde{\Theta}_{j-1} \neq 0 \) for some \( j \) such that \( j, j - 1 \in S_1 \) for sufficiently large \( (N, T) \). Then exists \( r \in \{1, 2, \ldots, R\} \) such that \( \| \beta_{j,r} \| = \max \left\{ \| \beta_{j,l} \| , l = 1, \ldots, R \right\} \), where \( \beta_{j,r} \) denotes the \( r \)th column of \( \beta_j \). Without loss of generality, we assume that \( r = R \).

Then \( \| \tilde{\beta}_{j,R} \| / \| \tilde{\beta}_j \| \geq 1/\sqrt{R} \). To consider the first order condition (FOC) with respect to (wrt) \( \Theta_j, j = 1, \ldots, J \), we distinguish three cases: (a) \( 2 \leq j \leq J - 1 \), (b) \( j = J \), and (c) \( j = 1 \).

In case (a), we consider two subcases: (a1) \( j + 1 \in S_{2b} \cup S_{2c} \) and (a2) \( j + 1 \in S_1 \cup S_{2a} \). In either subcase, the FOC wrt \( \Theta_{j,R} \) is given by
\[
0 = \frac{a_j}{\sqrt{N}} \left( X_{S_j} - \tilde{\Theta}_j Z_{S_j}^\top \right) \tilde{Z}_{S_j,R} + a_j \tau_j N^{1/2} \gamma w_j \frac{\| \tilde{\beta}_{j,R} \|}{\| \tilde{\beta}_j \|} - a_j \tau_j N^{1/2} \gamma w_j + 1 \rho_{j+1} + a_j \frac{\| \tilde{\beta}_j \|}{\| \tilde{\beta}_{j,R} \|} \tag{A.4}
\]

\[
= \frac{a_j}{\sqrt{N}} \left( (\Theta_j^* - \tilde{\Theta}_j) Z_{S_j}^\top Z_{S_j,R} + \frac{a_j}{\sqrt{N}} \tilde{E}_{S_j}^* Z_{S_j,R} + \frac{a_j}{\sqrt{N}} \left( (\Theta_j^* - \tilde{\Theta}_j) Z_{S_j}^\top \left( \tilde{Z}_{S_j,R} - Z_{S_j,R} \right) \right) + \frac{a_j}{\sqrt{N}} \tilde{E}_{S_j}^* \left( \tilde{Z}_{S_j,R} - Z_{S_j,R} \right) + a_j \tau_j N^{1/2} \gamma w_j \right) \frac{\| \tilde{\beta}_{j,R} \|}{\| \tilde{\beta}_j \|} - a_j \tau_j N^{1/2} \gamma w_j + 1 \rho_{j+1,R} \equiv B_{1,j} + B_{2,j} + B_{3,j} + B_{4,j} + B_{5,j} + B_{6,j} - B_{7,j}, \text{ say,}
\]
where \( \tilde{Z}_{S_j,R} \) and \( Z_{S_j,R} \) denote the \( R \)th columns of \( \tilde{Z}_{S_j} \) and \( Z_{S_j} \), respectively, \( \rho_{j+1} = \tilde{\beta}_{j+1,R} / \| \tilde{\beta}_{j+1} \| \) if \( \| \tilde{\beta}_{j+1} \| \neq 0 \) for \( r = 1, \ldots, R \), and \( \rho_{j+1} = \left( \rho_{j+1,1}, \ldots, \rho_{j+1,R} \right) \) satisfies \( \| \rho_{j+1} \| \leq 1 \) otherwise. By part (i), \( \| B_{1,j} \| \leq \| \frac{\tau_j a_j^2 \left( N^{-1/2} a_j \right) (\Theta_j^* - \tilde{\Theta}_j) \|}{\| \tilde{\beta}_j \|} \frac{1}{\tau_j} Z_{S_j}^\top Z_{S_j,R} \| = O_P(1) \) where we use the fact \( a_j = \eta_{N,j}^{-1} = O(\tau_j^{-1/2}) \) for \( j \in S_1 \) under Assumption A8(i). Similarly, by Lemma A.7(iii), \( \| B_{2,j} \| = O_P(1) \). By the submultiplicative property of the
Frobenius norm, part (i) and Lemmas A.7(i)-(ii), we have that for \( j \in S_1 \),
\[
\|B_{3,j}\| \leq (\tau_j a_j^2)(N^{-1/2}a_j^{-1})\|\Theta_j^* - \tilde{\Theta}_j\| \tau_j^{-1/2} \|\mathbf{Z}_{S_j}\| \tau_j^{-1/2}a_j^{-1} \|\tilde{\mathbf{Z}}_{S_j,R} - \mathbf{Z}_{S_j,R}\|
\]
\[
= O_P(\tau_j a_j^3) = o_P(1),
\]
\[
\|B_{4,j}\| \leq (\tau_j a_j^2)N^{-1/2}\|\tilde{\Theta}_j\| \tau_j^{-1/2} \|\mathbf{Z}_{S_j,R}\| \tau_j^{-1/2}a_j^{-1} \|\mathbf{Z}_{S_j} - \tilde{\mathbf{Z}}_{S_j}\| = O_P(\tau_j a_j^2) = O_P(1),
\]
\[
\|B_{5,j}\| \leq (\tau_j a_j^2)N^{-1/2}\|\tilde{\Theta}_j\| \tau_j^{-1/2}a_j^{-1} \|\mathbf{E}_{S_j}^*(\tilde{\mathbf{Z}}_{S_j,R} - \mathbf{Z}_{S_j,R})\| = O_P(\tau_j a_j^2) = O_P(1).
\]
In addition,
\[
\|B_{6,j}\| \geq a_j\tau_jN^{1/2}\gamma w_j \|\tilde{\beta}_{j,R}\| / \|\tilde{\beta}_j\| \geq a_j\tau_jN^{1/2}\gamma w_j/\sqrt{R}
\]
which is explosive in probability under Assumption A8(ii) (i.e., \((N\tau)^{1/2}\gamma \eta_{N,T} \to \infty \) as \((N,T) \to \infty\)).

To determine the probability order of \( B_{7,j} \), we consider two subcases. In subcase (a1) \( j + 1 \in S_{2b} \cup S_{2c} \), \( N^{-1/2}\|\tilde{\beta}_{j,1+}\| \overset{P}{\to} \lim_{N \to \infty} N^{-1/2}\|\Theta_{j+1}^* - \Theta_j^*\| \neq 0 \), implying that \( w_{j+1} = O_P(1) \) and \( B_{7,j} \leq a_j\tau_jN^{1/2}\gamma w_{j+1} = O_P((N\tau)^{1/2}\gamma) = O_P(1) \). Consequently, \( B_{6,j} \gg \sum_{i=1}^5 B_{6,i} \| + B_{7,j} \| \) and (A.4) cannot hold for sufficiently large \((N,T)\). Then we conclude that w.p.a.1 \( \tilde{\beta}_j = \tilde{\Theta}_j - \tilde{\Theta}_{j-1} \) must lie in a position where \( \|\Theta_j - \Theta_{j-1}\| \) is not differentiable with respect to \( \Theta_j \) in subcase (a1) and we must have \((N\tau)^{1/2}\gamma w_j \|\tilde{\theta}_j\| = O_P(1) \) in order for the FOC wrt \( \Theta_j \) to hold.

In subcase (a2) \( j + 1 \in S_1 \cup S_{2a} \). First, we observe that in order for the FOC wrt \( \Theta_j \) in (A.4) to hold, \( B_{7,j} = a_j\tau_jN^{1/2}\gamma w_{j+1} \|\tilde{\theta}_{j+1,R}\| \) must be explosive at the same rate as \( B_{6,j} \). Next, considering the FOC wrt \( \Theta_{j+1} \) and noting that \( B_{6,j+1} = B_{7,j} \), this implies that both \( B_{6,j+1} \) and \( B_{7,j+1} \) must explode at the same rate if \( j + 2 \in S_1 \cup S_{2a} \). Deducting this way until \( j + i \in S_1 \cup S_{2a} \) but \( j + i + 1 \in S_{2b} \cup S_{2c} \) for some \( i \geq 1 \). By assumption, if the interval \( S_{j+1} \) contains a break (so that \( j + i + 1 \in S_{2b} \cup S_{2c} \)), then the intervals \( S_{j+1} \) and \( S_{j+i} \) cannot contain a break (so that we must have \( j + i - 1, j + i \in S_1 \)). But when \( j + i - 1, j + i \in S_1 \), and \( j + i + 1 \in S_{2b} \cup S_{2c} \), the analysis in subcase (a1) applies to the FOC wrt \( \Theta_{j+i} \), which forces \( a_{j+i}\tau_{j+i}N^{1/2}\gamma w_{j+1} \|\tilde{\theta}_{j+i+1}\| = O_P(1) \). In short, a contradiction would arise unless there is no point after \( j + 1 \) that belongs to \( S_{2b} \cup S_{2c} \).

Similarly, if there is a point in \( \{j + 1, ..., J\} \) that belongs to \( S_{2a} \), we denote it as \( j + i \) for some \( i \geq 1 \). Then by assumption, \( j + i - 2, j + i - 1, j + i + 1, j + i + 2 \in S_1 \), and we can apply arguments as used in subcase (a1) to derive a contradiction based on the FOC.
wrt $\Theta_{j+1}$. Hence $S_{j+1}, \ldots, S_J$ cannot contain any break. Third, considering the FOC wrt $\Theta_{j-1}$ and noting that $B_{6,j} = B_{7,j-1}$, $\|B_{6,j-1}\|$ and $\|B_{7,j-1}\|$ must explode the same rate if $j-2 \in S_1 \cup S_2$. Deducing this way until $j - i \in S_1 \cup S_2$ but $j - i - 1 \in S_{2a} \cup S_{2c}$ for some $i \geq 2$. Again, when $j - i - 1 \in S_{2a} \cup S_{2c}$, the interval $S_{j-i-1}$ contains a break so that the neighboring intervals $S_{j-i}$ and $S_{j-i+1}$ cannot contain a break. So the FOC wrt $\Theta_{j-i}$ suggests that $\|B_{6,j-i}\|$ and $\|B_{7,j-i}\|$ are explosive. Similarly, the FOC wrt $\Theta_{j-1}$ suggests that in the latter case $\|B_{6,j-i-1}\| = O_P((N\tau)^{1/2} \gamma) = O_P(1)$ but $\|B_{7,j-i-1}\| = \|B_{6,j-i}\|$ is explosive. So the FOC in this last case cannot be satisfied and a contradiction would arise unless there is no break point before $j - 1$ for $j \geq 2$. But if there is no point before $j - 1$ and after $j$ that belongs to $S$, there will be no break point in the time interval $[1, T]$, contradicting to the requirement that we have at least one break contained in $[1, T]$. Consequently, w.p.a.1 $\tilde{\beta}_j = \Theta_j - \tilde{\Theta}_{j-1}$ must lie in a position where $\|\Theta_j - \Theta_{j-1}\|$ is not differentiable with respect to $\Theta_j$ in subcase (a2).

Now, we consider case (b). Note that only one term in the penalty component which is $\gamma \sum_{j=1}^{J} w_j \|\Theta_j - \Theta_{j-1}\|$ is involved with $\Theta_J$. Suppose that $\tilde{\beta}_J \neq 0$ for sufficiently large $(N, T)$ (note that $J \in S_1$ under our assumption). Then the FOC wrt $\Theta_J$ is given by

$$0 = \frac{a_J}{\sqrt{N}} (X_{S_J} - \tilde{\Theta}_J \hat{Z}_{S_J}) \hat{Z}_{S_J} \tau + a_J \tau J N^{1/2} \gamma w_J \tilde{\beta}_{j,R} / \|\tilde{\beta}_j\|$$

$$= B_{1,J} + B_{2,J} + B_{3,J} + B_{4,J} + B_{5,J} + B_{6,J}.$$  

As in case (a), we can readily show that $\sum_{i=1}^{5} \|B_{i,J}\| = O_P(1)$ and $\|B_{6,J}\|$ is explosive in probability at the rate $(N\tau)^{1/2} \gamma \eta \beta_j$. So the above FOC cannot hold and $\tilde{\beta}_J = \tilde{\Theta}_J - \tilde{\Theta}_{J-1}$ must be in a position where $\|\Theta_J - \Theta_{J-1}\|$ is not differentiable with respect to $\Theta_J$. Analogously, we can show that in case (c), $\tilde{\beta}_1 = \tilde{\Theta}_1 - \tilde{\Theta}_0$ must be in a position where $\|\Theta_1 - \Theta_0\|$ is not differentiable with respect to $\Theta_0$.

In the case where $|S| = 0$ so that $[1, T]$ contains no break, following the above analysis for case (b), we can first conclude $B_{6,J} = B_{7,J} = O_P(1)$ and $\Pr \left\{ \|\tilde{\Theta}_J - \tilde{\Theta}_{J-1}\| = 0 \right\} \to 1$ as $(N, T) \to \infty$. Now, considering the FOC wrt $\Theta_{J-1}$ and utilizing the fact that $B_{7,J-1} = B_{6,J} = O_P(1)$, we can derive that $B_{7,J} = O_P(1)$ in order for such a FOC to hold and $\tilde{\beta}_{J-1} = \tilde{\Theta}_{J-1} - \tilde{\Theta}_{J-2}$ must be in a position where $\|\Theta_{J-1} - \Theta_{J-2}\|$ is not differentiable with respect to $\Theta_{J-1}$. Deducing this way until $j = 1$, we can conclude that for all $j = J, J - 1, \ldots, 1$, $\tilde{\beta}_j = \tilde{\Theta}_j - \tilde{\Theta}_{j-1}$ must be in a position where $\|\Theta_j - \Theta_{j-1}\|$ is not differentiable.
with respect to $\Theta_{j-1}$ and thus (A.3) also holds in this case.

REFERENCES


Table 1: Performance of the two information criteria in determining the number of factors: DGPs 1-3 with $b = 1$.

<table>
<thead>
<tr>
<th>$T, J + 1$</th>
<th>IC$_1$</th>
<th>IC$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(250, 10)</td>
<td>(500, 10)</td>
</tr>
</tbody>
</table>

### Average selected number of factors

<table>
<thead>
<tr>
<th>DGP1-IID</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP1-CHeter</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 2</td>
<td>2.003</td>
<td>2.001</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP2-IID</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>DGP3</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Empirical probability of correct selection

<table>
<thead>
<tr>
<th>DGP1-IID</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP1-CHeter</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 2</td>
<td>0.997</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP2-IID</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>DGP3</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Percentage of correct detection of the number of breaks (C) and accuracy of break-point estimation (100×HD/T): DGP1-2 with $\kappa = 2$.

<table>
<thead>
<tr>
<th>(T, J + 1)</th>
<th>(250, 10)</th>
<th>(500, 10)</th>
<th>(500, 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C  HD/T</td>
<td>C  HD/T</td>
<td>C  HD/T</td>
</tr>
<tr>
<td>b = 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP1-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>66.0 0.072</td>
<td>79.6 0.032</td>
<td>56.0 0.365</td>
</tr>
<tr>
<td>Case 2</td>
<td>64.2 0.627</td>
<td>76.7 0.283</td>
<td>53.0 0.041</td>
</tr>
<tr>
<td>DGP1-CHeter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>66.3 0.102</td>
<td>76.9 0.034</td>
<td>55.3 0.397</td>
</tr>
<tr>
<td>Case 2</td>
<td>64.0 0.698</td>
<td>75.0 0.265</td>
<td>52.2 0.043</td>
</tr>
<tr>
<td>DGP2-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>84.1 0.159</td>
<td>89.2 0.044</td>
<td>75.2 0.585</td>
</tr>
<tr>
<td>Case 2</td>
<td>82.0 0.338</td>
<td>88.3 0.061</td>
<td>74.6 0.468</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>74.6 0.287</td>
<td>85.0 0.081</td>
<td>61.8 0.292</td>
</tr>
<tr>
<td>Case 2</td>
<td>68.6 0.372</td>
<td>81.8 0.095</td>
<td>62.7 0.238</td>
</tr>
<tr>
<td>b = 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP1-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>93.4 0.018</td>
<td>98.2 0.007</td>
<td>95.1 0.142</td>
</tr>
<tr>
<td>Case 2</td>
<td>93.5 0.331</td>
<td>98.1 0.065</td>
<td>94.5 0.008</td>
</tr>
<tr>
<td>DGP1-CHeter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>94.5 0.026</td>
<td>98.4 0.008</td>
<td>94.2 0.150</td>
</tr>
<tr>
<td>Case 2</td>
<td>93.9 0.372</td>
<td>98.2 0.080</td>
<td>94.4 0.010</td>
</tr>
<tr>
<td>DGP2-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>95.9 0.053</td>
<td>97.3 0.012</td>
<td>95.0 0.357</td>
</tr>
<tr>
<td>Case 2</td>
<td>92.6 0.209</td>
<td>97.0 0.024</td>
<td>94.9 0.261</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>83.8 0.158</td>
<td>94.2 0.035</td>
<td>88.1 0.194</td>
</tr>
<tr>
<td>Case 2</td>
<td>78.3 0.242</td>
<td>92.0 0.046</td>
<td>87.2 0.151</td>
</tr>
</tbody>
</table>
Table 3: Percentage of correct detection of the number of breaks (C) and accuracy of break-point estimation (100×HD/T): DGP1-2 with $\kappa = 4$.

<table>
<thead>
<tr>
<th>(T, J + 1)</th>
<th>(250, 10)</th>
<th>(500, 10)</th>
<th>(500, 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C  HD/T</td>
<td>C  HD/T</td>
<td>C  HD/T</td>
</tr>
<tr>
<td>$b = 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP1-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>85.8 0.033</td>
<td>93.9 0.016</td>
<td>75.1 1.506</td>
</tr>
<tr>
<td>Case 2</td>
<td>77.8 1.779</td>
<td>92.1 1.014</td>
<td>75.3 0.021</td>
</tr>
<tr>
<td>DGP1-CHeter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>84.9 0.041</td>
<td>94.1 0.014</td>
<td>75.4 3.190</td>
</tr>
<tr>
<td>Case 2</td>
<td>75.2 1.900</td>
<td>90.6 0.988</td>
<td>75.4 0.023</td>
</tr>
<tr>
<td>DGP2-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>93.6 0.062</td>
<td>95.4 0.018</td>
<td>83.3 1.856</td>
</tr>
<tr>
<td>Case 2</td>
<td>88.0 1.933</td>
<td>95.6 0.947</td>
<td>82.7 1.506</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>91.7 0.099</td>
<td>95.7 0.033</td>
<td>82.3 1.482</td>
</tr>
<tr>
<td>Case 2</td>
<td>88.9 1.266</td>
<td>95.4 0.765</td>
<td>82.3 1.256</td>
</tr>
<tr>
<td>$b = 2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP1-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>99.5 0.011</td>
<td>100.0 0.006</td>
<td>94.0 0.817</td>
</tr>
<tr>
<td>Case 2</td>
<td>91.5 0.917</td>
<td>99.2 0.663</td>
<td>99.7 0.006</td>
</tr>
<tr>
<td>DGP1-CHeter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>99.3 0.014</td>
<td>100.0 0.007</td>
<td>94.1 0.718</td>
</tr>
<tr>
<td>Case 2</td>
<td>91.0 0.977</td>
<td>99.5 0.710</td>
<td>99.5 0.008</td>
</tr>
<tr>
<td>DGP2-IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>99.7 0.024</td>
<td>100.0 0.008</td>
<td>93.1 1.497</td>
</tr>
<tr>
<td>Case 2</td>
<td>95.4 1.458</td>
<td>99.3 1.124</td>
<td>94.3 1.327</td>
</tr>
<tr>
<td>DGP2-AR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>99.4 0.047</td>
<td>100.0 0.020</td>
<td>98.0 1.355</td>
</tr>
<tr>
<td>Case 2</td>
<td>96.3 1.323</td>
<td>99.9 0.868</td>
<td>97.2 1.180</td>
</tr>
</tbody>
</table>
Figure 1: Plots of the IC_1 (thin line) and IC_2 (thick line) against the number of factors with $b = 1$ for (a) DGP1-Case1 with $T = 250$ and cross-sectional heteroscedastic errors; (b) DGP2-Case1 with $T = 250$ and autoregressive errors; and (c) DGP3.

Figure 2: Plots of the frequency of the estimated breaks among 1000 replications for DGP1 and $T = 250$ and for (a) Case 1 and IID errors, (b) Case 1 and CHeter errors, (c) Case 2 and IID errors, and (d) Case 2 and CHeter errors. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$. 

38
Figure 3: Plots of the frequency of the estimated breaks among 1000 replications for DGP1 and $T = 500$, $J + 1 = 10$ and for (a) Case 1 and IID errors, (b) Case 1 and CHeter errors, (c) Case 2 and IID errors, and (d) Case 2 and CHeter error. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$.

Figure 4: Plots of the frequency of the estimated breaks among 1000 replications for DGP1 and $T = 500$, $J + 1 = 15$ and for (a) Case 1 and IID errors, (b) Case 1 and CHeter errors, (c) Case 2 and IID errors, and (d) Case 2 and CHeter errors. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$. 
Figure 5: Plots of the frequency of the estimated breaks among 1000 replications for DGP2 and $T = 250$ and for (a) Case 1 and IID errors, (b) Case 1 and AR errors, (c) Case 2 and IID errors, and (d) Case 2 and AR errors. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$.

Figure 6: Plots of the frequency of the estimated breaks among 1000 replications for DGP2 and $T = 500$, $J + 1 = 10$ and for (a) Case 1 and IID errors, (b) Case 1 and AR errors, (c) Case 2 and IID errors, and (d) Case 2 and AR errors. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$. 
Figure 7: Plots of the frequency of the estimated breaks among 1000 replications for DGP2 and $T = 500$, $J + 1 = 15$ and for (a) Case 1 and IID errors, (b) Case 1 and AR errors, (c) Case 2 and IID errors, and (d) Case 2 and AR errors. The blue shaded line with angle=135 is for $b = 1$ and the red shaded line with angle=45 is for $b = 2$.

Figure 8: Plots of the values of $IC_2$ against the number of factors for the real data application.