Multiple Change Points Detection in Low Rank and Sparse High Dimensional Vector Autoregressive Models

Peiliang Bai, Abolfazl Safikhani, and George Michailidis, *Member, IEEE*

***Abstract*—Identifying change/break points in multivariate time series represents a canonical problem in signal processing, due to numerous applications related to anomaly detection problems. The underlying detection methodology heavily depends on the nature of the mechanism determining the temporal dynamics of the data. Vector auto-regressive models (VAR) constitute a widely used model in diverse areas, including surveillance applications, economics/finance and neuroscience. In this work, we consider piece-wise stationary VAR models exhibiting break points between the corresponding stationary segments, wherein the transition matrices that govern the model's temporal evolution are decomposed into a *common* low-rank component and time evolving sparse ones. Further, we assume that the number of available time points are smaller than the number of model parameters and hence we are operating in a *high-dimensional regime*. We develop a three-step strategy that accurately detects the number of change points together with their location and subsequently estimates the model parameters in each stationary segment. The effectiveness of the proposed procedure is illustrated on both synthetic and real data sets.**

***Index Terms*—Blocked fused lasso, vector auto-regression, detection, consistency.**

I. Introduction

DETECTING multiple changes in time series data constitutes a canonical problem with numerous applications in signal detection [6], economics and finance [20], quality control [41], risk analysis [33], surveillance and environmental monitoring [38], and neuroscience [29]. A change point represents a discontinuity in the parameters of the data generating process. The problem can be considered either in an (1) *offline* [7] setting, or (2) an *online* [17] one. In the first case, one is given a sequence of observations and questions of interest include: (i) whether there exist change (break) points and (ii) if there exist change points, identify their locations, as well as estimate the parameters of the data generating process (see the review paper [3]). In the online case, one sequentially obtains new observations and the main interest is in *quickest detection* of the change point (see e.g. [46], [21] and references therein).

The focus of this paper is on offline break point detection based on a vector autorgressive (VAR) data generation mechanism. The literature to date has focused on a number of univariate and multivariate statistical models, including constant signal plus noise ones [23], linear regression [31], Gaussian graphical models [30], [43], vector autoregressive models (VAR) [44], panel-type time series models [15], [16], and factor models [4], [5]. VAR models represent a canonical model with wide range of applications in economics [22], [25], functional genomics [35], [40], speech signal analysis [26], [45], smart cities [36] and neuroscience [2], [27], [28], [39]. There has been a lot of interest recently in their high dimensional counterparts assuming a (structured) sparse [9] and also low rank transition matrix [8] for *stationary* data.

However, in numerous application areas the assumption of stationarity does not hold for the entire data set, but only for relatively short segments of the available data (see e.g. discussion in [32] for a specific example of log-returns of stocks exhibiting structural breaks due to economic shocks, as well as [44] for occurrence of seizures and its effect on brain signal data). Due to the existence of several discontinuity points in the distribution of the data, on many occasions a good working model is to assume a piece-wise stationary model and then the problem becomes to identify the number of unknown break (change) points of the segments, locate them and finally estimate the model parameters within each segment.

This paper aims to develop a *fast/scalable* strategy for identifying change points in low-rank plus sparse high dimensional time series models and also provide *probabilistic guarantees* for the accuracy of their identification. Specifically, the focus is on VAR models whose transition matrices that capture their dynamic evolution can be decomposed into a *constant low-rank* component, plus a *sparse time evolving* one at selected (unknown) time points, thus inducing structural breaks in the system's evolution. This data generating process occurs in many

Manuscript received June 20, 2019; revised December 9, 2019 and March 15, 2020; accepted April 30, 2020. Date of publication May 11, 2020; date of current version May 27, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Elias Aboutanios. The work of George Michailidis was supported by the NSF under Grants DMS-1830175, IIS 1633158, and NIH 5-R01-GM114029. *(Corresponding author: George Michailidis.)*

Peiliang Bai is with the Department of Statistics, University of Florida, Gainesville, FL 32611 USA (e-mail: baipl92@ufl.edu).

Abolfazl Safikhani is with the Departments of Statistics and the Informatics Institute, University of Florida, Gainesville, FL 32611 USA (e-mail: a.safikhani@ufl.edu).

George Michailidis is with the Departments of Statistics and Computer and Information Sciences and Engineering and the Informatics Institute, University of Florida, Gainesville, FL 32611 USA (e-mail: gmichail@ufl.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes a detailed description of the algorithms presented, some additional numerical results and the proofs of some auxiliary lemmas. This material is six pages in size.

Digital Object Identifier 10.1109/TSP.2020.2993145

real world applications, including surveillance video data, where the background theme remains stationary over time, but some small portion of the frames changes at certain time points due to adding/removing objects (more details on this application are given in Section V-A). Other applications include environmental monitoring from sensor measurements, where the background can be described by a low rank stationary process, and monitoring of financial markets, as discussed in Section V-B.

Note that the presence of a *fixed* low-rank component combined with piece-wise constant sparse ones, makes the detection problem significantly more challenging than that of employing a sparse VAR model. Note that there are two types of signal in the data, -one coming from the fixed low-rank component and the other from the changing sparse one- thus requiring significant enhancements in the detection algorithm and even more importantly in the technical analysis for providing rigorous probabilistic guarantees on the detection accuracy and estimation of the model parameters, all successfully resolved in Section III. Specifically, we rigorously address the following issues: (1) estimate accurately the total number of break points in the data; (2) locate all break points consistently; (3) estimate accurately all model parameters including the low rank and sparse auto-regressive components. To do so, we propose a three-step procedure. In Step 1, the detection problem is reformulated as a variable selection one based on a regularized high dimensional linear regression framework, with a blocked fused lasso penalty. This step over-selects an initial set of *candidate* break points. In Step 2, based on a carefully defined *information criterion* that accounts for the low rank plus sparse structure of the transition matrices, we *screen out redundant* candidate break points obtained in Step 1 and establish that the remaining selected points are consistent estimates of the true break points (see Theorem 3). Finally, in Step 3, two different methods are developed to estimate all model parameters within all the identified stationary segments. Hence, key technical contributions of this work include:

- The introduction of suitable conditions to ensure identifiability of the low-rank and sparse components in piece-wise stationary VAR models that are also of independent interest.
- The development of an efficient three-step algorithm to estimate and locate the break points, as well as to estimate the model parameters within each segment.
- The introduction of a novel information criterion to select a consistent subset of break points obtained initially through a penalized high-dimensional regression model.
- The development of a blocked fused lasso regression estimator to accelerate the detection of break points, which is also of independent interest in the field of variable selection in high dimensional linear regression models.
- Recommending cross-validation type methods to select the key tuning parameters involved in our algorithm.

The remainder of the paper is organized as follows. The modeling framework together with identifiability issues are presented in Section II. Section III introduces the proposed 3-step detection strategy and establishes its asymptotic properties. An extensive evaluation analysis based on synthetic data is provided in Section IV. Finally, two real data sets (one on surveillance video data and another one on financial data) are analyzed using the developed algorithm and discussed in Section V. Finally, additional simulation scenarios, and proofs of the main results are given in the Appendix.

Notation: Throughout the paper, we denote with a superscript "$\star$" the true value of the corresponding model parameters. Further, for any $p \times p$ matrix we use $\|\cdot\|_2$, $\|\cdot\|_F$ and $\|\cdot\|_*$ to denote the spectral, Frobenius and nuclear norm of the matrix, respectively. For any matrix B, we use B' to denote its transpose, and finally we denote the ℓ_1, ℓ_0 and ℓ_∞ norms of its vectorized form as follows: $\|B\|_1$ for $\|\text{vec}(B)\|_1$, $\|B\|_0$ for $\text{Card}(\text{vec}(B))$ and $\|B\|_\infty$ for $\|\text{vec}(B)\|_\infty$.

II. Model Formulation

We start by considering a piece-wise structured stationary VAR(1) model; the extension to a VAR(d) model with d lags is briefly discussed in the Conclusions section. Specifically, suppose we have $n+1$ time points and there exist m_0 change points $0 = t_0 < t_1 < \cdots < t_{m_0} < t_{m_0+1} = n$, such that for $t_{j-1} \le t < t_j$, $j = 1, \ldots, m_0+1$, the structured VAR(1) process is given by

$$X_t = B_j' X_{t-1} + \epsilon_t \text{ and } B_j = L^\star + S_j^\star \tag{1}$$

where X_t is the p dimensional vector of observed time series at time t, B_j is the $p \times p$ transition matrix for the $j-$th segment that captures the lead-lag relationships among the time series under consideration; further, each transition matrix is assumed to be a superposition of a stable $L^\star$ low rank component and a time varying $S_j^\star$ sparse component. Finally, we assume that the $p-$dimensional noise process is normally distributed; i.e. $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}_p(0, \Sigma_\epsilon)$. Note that in principle, Σ_ϵ can vary over segments, but is considered fixed in our setting for ease of presentation. We further assume that the j-th sparse component $S_j^\star$ has sparsity density $\|S_j^\star\|_0 = d_j^\star$ with $d_j^\star \ll p^2$ and that the low rank component $L^\star$ has rank $r^\star$ with $r^\star \ll p$, respectively. Based on the decomposition of the transition materices B_j, it can be seen that the low rank component $L^\star$ captures *invariant cross-autocorrelation* structure across all p time series for the entire time period, while $S_j^\star$ reflects *time evolving* additional cross-sectional autocorrelations.

The objective is to detect the change points t_j, and obtain estimates of the transition matrices B_j's under a high-dimensional regime, wherein the number of parameters within each stationary segment exceeds the corresponding number of time points. Therefore, according to the formulation of the structured VAR(1) model above, it can be seen that the presence of change points is driven by changes in the sparse components $S_j^\star$.

However, there is a natural *identifiability* issue being masked by the posited low rank plus sparse structure of the transition matrices. Suppose the low rank component $L^\star$ provides most of the signal, while the sparse components $S_j^\star$ contribute only a small portion of the signal. In such a setting, detection of change points becomes impossible. Therefore, in order to identify the changes in the sparse components, the signal "originating" from the low rank component can not be dominant.

Further, this identifiability issue will also influence the probabilistic guarantees for accurately estimating the low rank and sparse components. Suppose the low rank component itself is $d_j^\star$ sparse, while the sparse components are of rank $r^\star$. Then, we can not expect to estimate $L^\star$ and $S_j^\star$'s separately, without imposing any further restrictions. In this case, a minimal condition for accurate recovery of the low rank and sparse components is that the former should not be too sparse and the latter should not be low rank.

In a recent paper [14], this issue has been rigorously addressed for independent and identically distributed data and resolved by imposing an *incoherence* condition, such a condition is sufficient for *exact* recovery of the low rank and the sparse component by solving a convex program. In [1], the authors considered a noisy setting and also to where a model parameter (e.g. a regression coefficient matrix) admits such a decomposition, wherein exact recovery of the two components is impossible. They proceeded to formulate a general measure for the *radius of non-identifiability* of the problem under consideration and established a non-asymptotic upper bound on the estimation error $\|\widehat{L} - L^\star\|_F^2 + \|\widehat{S}_j - S_j^\star\|_F^2$, which depends on this radius. In our work, we introduce the information ratio (see Section III-A, Assumption H2), which reflects similar constraints imposed on the radius of non-identifiability in [1], to constrain the signal strength originating from the low-rank component that will render changes in sparse components *detectable*.

III. The Change Point Detection Procedure and Its Properties

Our proposed strategy comprises of the following steps: (A) Solving a regularized regression problem, with a Block Fused Lasso (BFL) penalty to identify candidate change points; (B) Screening the obtained candidates by computing a novel information criterion; and (C) Estimating consistently the parameters of each transition matrix B_j.[1]

A. Step 1: Block Fused Lasso (BFL) Based Estimation

In our first step, we leverage a regularized regression problem with a BFL penalty to identify an initial set of candidate change points. Specifically, we partition the observed time points into blocks of size b_n and fix the model parameters within each block. In other words, each end point of a block corresponds to a *candidate* break point in this step. Therefore, BFL has $(\lceil \frac{n}{b_n} \rceil + 1)p^2$ parameters, compared to $2p^2$ when no break points are present. Note that in order to identify the change points consistently, we can not set b_n to be too large as explained below.

Define a sequence of time points $1 = r_0 < r_1 < \cdots < r_{k_n+1} = n$ corresponding to the end points of the blocks (i.e. $r_{i+1} - r_i = b_n$ and $k_n = \lceil \frac{n}{b_n} \rceil$). Subsequently, by using the same notation as in the model (1), we define the following block variables: $\mathbf{X}_{r_j} = [X_{r_{j-1}}, \ldots, X_{r_j - 1}]$, $\mathbf{Y}_{r_j} = [X_{r_{j-1}+1}, \ldots, X_{r_j}]$ and $\boldsymbol{\epsilon}_{r_j} = [\epsilon_{r_{j-1}+1}, \ldots, \epsilon_{r_j}]$, for the j-th block respectively, translated in matrix form notation as follows: let $\mathcal{X} = [\mathbf{X}_{r_1}, \ldots, \mathbf{X}_{r_{k_n+1}}]' \in \mathbb{R}^{n \times p}$, $\mathcal{Y} = [\mathbf{Y}_{r_1}, \ldots, \mathbf{Y}_{r_{k_n+1}}]' \in \mathbb{R}^{n \times p}$, $\mathcal{E} = [\boldsymbol{\epsilon}_{r_1}, \ldots, \boldsymbol{\epsilon}_{r_{k_n}}]' \in \mathbb{R}^{n \times p}$ and

$$\mathcal{Z} = \begin{bmatrix} \mathbf{X}'_{r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}'_{r_2} & \mathbf{X}'_{r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}'_{r_{k_n+1}} & \mathbf{X}'_{r_{k_n+1}} & \cdots & \mathbf{X}'_{r_{k_n+1}} \end{bmatrix} \in \mathbb{R}^{n \times pk_n}.$$

We then formulate the model (1) into the following linear regression problem

$$\mathcal{Y} = \mathcal{X}L^\star + \mathcal{Z}\Theta + \mathcal{E}, \tag{2}$$

wherein $\Theta = [\theta_1', \ldots, \theta_{k_n}']' \in \mathbb{R}^{pk_n \times p}$. We set $\theta_1 = S_1^\star$; for $i = 2, 3, \ldots, k_n$, and for the subsequent ones we set

$$\theta_i = \begin{cases} S_{j+1}^\star - S_j^\star, & \text{when } i = t_j \text{ for some } j, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

It should be noted that in this parameterization, $\theta_i \neq \mathbf{0}$ wherein $\mathbf{0}$ corresponds to the $p \times p$ zero matrix, indicates a change in the VAR transition matrix B_j. Therefore, for $j = 1, 2, \ldots, m_0$, the structural change points t_j can be detected as time points $i \geq 2$, whenever $\theta_i \neq \mathbf{0}$.

The linear regression representation in (2) implies that the model coefficients Θ and L can be estimated through the following restricted penalized least squares problem

$$(\widehat{\Theta}, \widehat{L}) = \underset{\Theta, L \in \Omega}{\arg\min} \frac{1}{n}\|\mathcal{Y} - \mathcal{X}L - \mathcal{Z}\Theta\|_2^2 + \lambda_{1,n}\|L\|_* + \lambda_{2,n}\|\Theta\|_1 + \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\theta_j\right\|_1. \tag{4}$$

In the objective function above, $\Omega \stackrel{\text{def}}{=} \{L \in \mathbb{R}^{p \times p} : \|L\|_\infty \leq \frac{\alpha}{p}\}$ corresponds to the set of $p \times p$ matrices whose elements do not exceed a threshold, thus limiting their "spikeness" and consequently limiting the radius of non-identifiability; $\lambda_{1,n}$, $\lambda_{2,n}$ and $\lambda_{3,n}$ are non-negative tuning parameters controlling the two regularization terms. The parameter α constrains the strength of the signal originating from the low rank component; in other words, it controls the degree of non-identifiability of the coefficients allowed in the model. Due to the assumption H2 presented below, we can derive a relationship between α and the information ratio γ, since $\gamma \propto \alpha^{-1}$. Hence, we obtain that $\Omega \stackrel{\text{def}}{=} \{L \in \mathbb{R}^{p \times p} : \|L\|_\infty \leq \frac{C_0}{p\gamma}\}$ for some constant $C_0 > 0$, and in all subsequent developments we work with γ instead of α.

The basic idea of adding a block fused lasso penalty in the objective function is to expand the space of feasible solutions to make the estimation step flexible enough, so as not to miss any true break points, when the tuning parameters are appropriately tuned; the latter need to be selected in such a manner, so as not to lead to too many false positives (wrongly estimated break points). Finding the appropriate/optimal tuning parameter rate is a crucial step in verifying the probabilistic guarantees in fused lasso based procedures [42]. Notice that the space of feasible solutions for problem (4) consists of all pairs (C, D)

[1]Code implementing the strategy is available at https://github.com/abolfazlsafikhani/LS-VAR-ChangePoint-Detection.

such that the square p-dim matrix C is low-rank and belongs to the space Ω, while the matrix $D \in \mathbb{R}^{pk_n \times p}$ is sparse. Based on Assumption H3, the number of blocks k_n is much larger than m_0. This expansion on the space of model parameters is a crucial development in Step 1.

Remark 1: The computational complexity of estimating the sparse components in (4) is of order $O(k_n p^2)$ [10]. If the size of the blocks is set to 1 (i.e. $b_n = 1$) the method would revert to a standard fused lasso penalty [42]. However, to speed up computations, we allow b_n to increase as a function of the sample size. On the other hand, larger values of b_n may lead to detection loss, in the presence of closely spaced true break points. Therefore, there is a trade-off between achieving faster computations vs detection accuracy, controlled by the block sizes and properly quantified in Assumption H3.

The estimator defined in (4) may not be a consistent estimator of the model parameters, since the design matrix $\mathcal{Z}$ does not satisfy the restricted eigenvalue assumption which is needed for verifying consistency [9]. Instead, this estimator exhibits the following two properties: (a) Prediction consistency; (b) Over-estimation of the number of break points. These two properties make this step suitable for obtaining an initial set of good candidate break points. To consistently identify the true ones, a screening step (presented below) is required.

Before stating our main results, we introduce the following assumptions:

H1 For all $j = 1, 2, \ldots, m_0 + 1$ we have $d_j^\star \ll p^2$, i.e. the $S_j^\star$ are sparse. Further, there exists a positive constant $M_S > 0$ such that

$$\max_{1 \le j \le m_0+1} \left\| S_j^\star \right\|_\infty \le M_S.$$

H2 Define the information ratio

$$\gamma = \frac{\left\| S_j^\star \right\|_\infty}{\left\| L^\star \right\|_\infty}, \quad \text{for } j = 1, 2, \ldots, m_0 + 1.$$

Then, with fixed γ, we obtain that $\|L^\star\|_\infty \le \gamma^{-1} M_S$ by H1. In this model, we recommend choosing γ in the range $\gamma \ge 1$.

H3 There exists a positive constant v such that

$$\min_{1 \le j \le m_0} \|S_{j+1}^\star - S_j^\star\|_2 \ge v > 0.$$

Moreover, letting $\Delta_n = \min_{1 \le j \le m_0} |t_{j+1} - t_j|$ and $d_n^\star = \sum_{j=1}^{m_0+1} d_j^\star$, there exists a vanishing positive sequence γ_n such that, as $n \to +\infty$,

$$\frac{\Delta_n}{n\gamma_n} \to +\infty, \ \limsup \frac{b_n}{n\gamma_n} \le C < \frac{1}{12},$$

$$\frac{d_n^\star \log p}{n\gamma_n} \to 0 \text{ and } \frac{r^\star p}{n\gamma_n} \to 0.$$

Assumption H1 is standard in the high-dimensional liner regression literature, while Assumption H2 ensures identifiablity of the model parameters, as discussed in Section II. Assumption H3 links the detection rate to the tuning parameters selected in the estimation step and the block sizes. This assumption also provides a minimum distance-type requirement on the elements of B_j across different segments, which can be regarded as the counterpart of Assumption A3 in [44], Assumptions A2 and A3 in [24], and Assumptions H2 and H3 in [13].

Theorem 1: Assume that H1 and H2 hold. Select $\lambda_{1,n} = 2C_1\sqrt{\frac{p}{n}}$, $\lambda_{2,n} = 2C_2\sqrt{\frac{\log n + 2\log p}{n}}$ for some $C_1, C_2 > 0$ and $\lambda_{3,n} = o((nd_n^\star)^{-1})$, and further assume $m_0 \le m_n$ with $m_n = o(\lambda_{2,n}^{-1})$. Then, by also imposing the restricted space Ω constraint, the optimal solution to (4) satisfies the following result with high probability, werein the positive constant C_0 is defined in (4)

$$\frac{1}{n}\left\| \mathcal{X}(\widehat{L} - L^\star) + \mathcal{Z}(\widehat{\Theta} - \Theta^\star) \right\|_2^2 \le 4C_1 \frac{C_0}{\gamma}\sqrt{\frac{r^\star}{n}} + 2M_S\lambda_{2,n}m_n \max_{1 \le j \le m_0+1} \left\{ d_j^\star + d_{j-1}^\star \right\} + o(1). \quad (5)$$

Theorem 1 establishes prediction consistency for the first step in the proposed strategy, assuming that the total number of break points allowed is upper bounded properly. The fused lasso tuning parameter in equation (4) is $\lambda_{2,n}$ and its optimal rate to establish prediction consistency is $\lambda_{2,n} = 2C_2\sqrt{\frac{\log n + 2\log p}{n}}$ for some $C_2 > 0$, as stated in Theorem 1. Higher rates for this tuning parameter will miss true break point and thus compromise prediction consistency, while lower rates may lead to having too many false positives which is going to be detrimental for change point detection and it will also increase the computation time for other steps in the proposed procedure.

The next theorem shows that under a suitable choice of the tuning parameters, the selected break points in this first step are an overestimate of the true number of break points in the model. Further, it asserts that no true break point is isolated, in the sense that there exists a candidate change point close by.

Before stating the next theorem, we need some additional definitions. Let $\mathcal{A}_n = \{t_1, \ldots, t_{m_0}\}$ be the set of true change points, and $\widehat{\mathcal{A}}_n = \{\widehat{t}_1, \ldots, \widehat{t}_{\widehat{m}}\}$ be the set of estimated candidate change points. Following [11] and [13], we define the Hausdorff distance between two countable sets on the real line as

$$d_H(A, B) = \max_{b \in B} \min_{a \in A} |b - a|.$$

Note that this definition is not symmetric and therefore not a real distance. Nevertheless, this version of function $d_H(A, B)$ is adequate for the result established in the next theorem.

Theorem 2: Suppose H1–H3 hold. Choose the tuning parameters as $\lambda_{1,n} = C_1\frac{b_n}{n}\sqrt{\frac{p}{n\gamma_n}}$, $\lambda_{2,n} = 2C_2\sqrt{\frac{\log n + 2\log p}{n}}$ and $\lambda_{3,n} = C_3\frac{b_n}{n}\sqrt{\frac{\log p}{n\gamma_n}}$ for some large constants $C_1, C_2, C_3 > 0$. Then, as $n \to +\infty$,

$$\mathbb{P}(|\widehat{\mathcal{A}}_n| \ge m_0) \ \to \ 1,$$

and

$$\mathbb{P}(d_H(\widehat{\mathcal{A}}_n, \mathcal{A}_n) \le n\gamma_n) \to 1.$$

Remark 2: In Theorem 2, We express the tuning parameters $\lambda_{1,n}$ and $\lambda_{3,n}$ in different forms. Note that under the setting in Theorem 2, the quantities $\frac{b_n}{n}\sqrt{\frac{p}{n\gamma_n}}$ and $nd_n^\star\lambda_{3,n}$ equal to $\frac{b_n}{n\gamma_n}\sqrt{\gamma_n}\sqrt{\frac{p}{n}}$ and $C_3 b_n\sqrt{d_n^\star}\sqrt{\frac{d_n^\star \log p}{n\gamma_n}}$, respectively.

Further, we have a positive vanishing sequence $\{\gamma_n\}$ satisfying $\limsup \frac{b_n}{n\gamma_n} \leq C < \frac{1}{12}$ and $\frac{d_n^\star \log p}{n\gamma_n} \to 0$ in assumption H3, which yields to $\lambda_{1,n} \propto C_1\sqrt{\frac{p}{n}}$ and $\lambda_{3,n} = o((nd_n^\star)^{-1})$. These calculations confirm that the tuning parameters are of the same order in both Theorems 1 and 2.

B. Step 2: Screening

Since the set of estimated break points $\widehat{\mathcal{A}}_n$ is a superset of $\mathcal{A}_n$, we require another step to screen out redundant points in this set. For the screening step, we need to reformulate our model and further note that the parameters defined are different from those in the first step. Specifically, suppose that we have already selected m candidate change points based on the previous step: $1 = s_0 < s_1 < \cdots < s_m < s_{m+1} = n$. Define the following matrices: $\mathbf{X}_{s_j} = [X_{s_{j-1}}, \ldots, X_{s_j-1}]$, $\mathbf{Y}_{s_j} = [X_{s_{j-1}+1}, \ldots, X_{s_j}]$ for $j = 1, 2, \ldots, m+1$, respectively. Then, the combined matrices across all segments become $\mathcal{X} = [\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_m}]'$ and $\mathcal{Y} = [\mathbf{Y}_{s_1}, \ldots, \mathbf{Y}_{s_m}]'$. Further, the block diagonal design matrix is defined by $\mathcal{Z}_{s_1,\ldots,s_m} = \text{diag}(\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_{m+1}}) \in \mathbb{R}^{n\times(m+1)p}$, and the corresponding coefficient matrix is given by $\Theta_{s_1,\ldots,s_m} = [\theta'_{(1,s_1)}, \theta'_{(s_1,s_2)}, \ldots, \theta'_{(s_m,n)}]' \in \mathbb{R}^{(m+1)p\times p}$. Specifically, by using the notations we defined, we form the following linear regression

$$\mathcal{Y} = \mathcal{Z}_{s_1,\ldots,s_m}\Theta_{s_1,\ldots,s_m} + \mathcal{X}L + \Xi, \tag{6}$$

where $\Xi \overset{\text{def}}{=} [\xi_1, \xi_2, \ldots, \xi_n]' \in \mathbb{R}^{n\times p}$ is the error term.

Therefore, we estimate $\Theta_{s_1,\ldots,s_m}$ and L as the optimal solution of the following regularized optimization problem for all selected segments with different tuning parameters $\eta_{(s_{i-1},s_i)}$, for $i = 1, 2, \ldots, m+1$.

$$\begin{aligned}(\widehat{L}, \widehat{\Theta}_{s_1,\ldots,s_m}) &= \underset{L,\Theta_{s_1,\ldots,s_m}}{\arg\min} \sum_{i=1}^{m+1} \frac{1}{s_i - s_{i-1}} \left\|\mathbf{Y}_{s_i} - \mathbf{X}_{s_i}(\theta_{(s_{i-1},s_i)} + L)\right\|_2^2 \\ &+ \eta_{(s_{i-1},s_i)}\|\theta_{(s_{i-1},s_i)}\|_1 + \eta_L\|L\|_*. \end{aligned} \tag{7}$$

Next, we define the following objective function with tuning parameter vector $\eta_n \overset{\text{def}}{=} (\eta_{(s_0,s_1)}, \eta_{(s_1,s_2)}, \ldots, \eta_{(s_m,s_{m+1})})$:

$$\begin{aligned}L_n(\mathbf{s}; \eta_n) = &\left\|\mathcal{Y} - \mathcal{Z}_{s_1,\ldots,s_m}\widehat{\Theta}_{s_1,\ldots,s_m} - \mathcal{X}\widehat{L}\right\|_F^2 \\ &+ \sum_{i=1}^{m+1} \eta_{(s_{i-1},s_i)}\|\widehat{\theta}_{(s_{i-1},s_i)}\|_1 + \eta_L\|\widehat{L}\|_*, \end{aligned} \tag{8}$$

where $\mathbf{s} = (s_1, \ldots, s_m)$. Then, for a penalty sequence ω_n (which can be selected in accordance to assumption H4 below), we consider the following *information criterion*

$$\text{IC}(\mathbf{s}; \eta_n) = L_n(\mathbf{s}; \eta_n) + m\omega_n. \tag{9}$$

The second step of our strategy selects a subset of initial $\widehat{m}$ change points derived from (4) by solving

$$(\widetilde{m}, \widetilde{t}_j; j = 1, 2, \ldots, \widetilde{m}) = \underset{0\leq m\leq\widehat{m}, \mathbf{s}\in\widehat{\mathcal{A}}_n}{\arg\min} \text{IC}(\mathbf{s}; \eta_n). \tag{10}$$

To establish consistency properties of the screening procedure, we need the following two additional assumptions.

H4 Assume that

$$\frac{m_0 n\gamma_n(d_n^{\star 2} + r^{\star 2})}{\omega_n} \to 0 \text{ and } \frac{\Delta_n}{m_0\omega_n} \to +\infty.$$

H5 There exists a large positive constant $c > 0$ such that (a) if $|s_i - s_{i-1}| \leq n\gamma_n$, then $\eta_{(s_{i-1},s_i)} = c\sqrt{n\gamma_n \log p}$ and $\eta_L = c\sqrt{n\gamma_n p}$; (b) if there exists t_j and t_{j+1} such that $|s_{i-1} - t_j| \leq n\gamma_n$ and $|s_i - t_{j+1}| \leq n\gamma_n$, then, $\eta_{(s_{i-1},s_i)} = 2(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_n^\star \frac{n\gamma_n}{s_i - s_{i-1}})$ and $\eta_L = 2c\sqrt{\frac{p}{n\gamma_n}}$; (c) otherwise, $\eta_{(s_{i-1},s_i)} = 2(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_n^\star)$ and $\eta_L = 2c\sqrt{\frac{p}{n\gamma_n}}$.

Assumption H4 connects the penalty term ω_n defined in the information criterion to the minimum spacing allowed between break points. Assumption H5 specifies the magnitude (rate) of the tuning parameters used in the least squares problem given in (7). Note that assumptions on the rate of the tuning parameter of the penalty are needed even in lasso regression problems for independent and identically distributed data and without break points for (see e.g. [47]). In the presence of break points, one works with misspecified models and hence a more careful and complex selection of the various tuning parameters are required [12], [13], [44].

Theorem 3: Suppose assumptions H1–H5 hold. Then, as $n \to +\infty$, the minimizer $(\widetilde{m}, \widetilde{t}_j; j = 1, 2, \ldots, \widetilde{m})$ of (10) satisfies

$$\mathbb{P}(\widetilde{m} = m_0) \to 1.$$

Moreover, there exists a positive constant $B > 0$ such that

$$\mathbb{P}\left(\max_{1\leq j\leq m_0} |\widetilde{t}_j - t_j| \leq Bm_0 n\gamma_n(d_n^{\star 2} + r^{\star 2})\right) \to 1.$$

Remark 3: For the case of finite m_0, the sequence γ_n can be chosen as $\gamma_n = \frac{(rp + d_n^\star \log p)^{1+v/2}}{n}$ for some small $v > 0$. Assuming that the low-rank component and total degree of sparsity satisfy $d_n^{\star 2} + r^{\star 2} = o((rp + d_n^\star \log p)^{v/2})$, then the consistency rate for identifying the relative location of true break points -t_j/n- is of the order $(rp + d_n^\star \log p)^{1+v}/n$ in Theorem 3. Finally, in this setting, ω_n can be chosen as $(rp + d_n^\star \log p)^{1+2v}$ and the minimum spacing allowed between consecutive break points - Δ_n- must be at least of order $(rp + d_n^\star \log p)^{1+3v}$. Comparing the consistency rates with those in Theorem 3 in [44], we observe that the additional term rp captures the complexity introduced in the model due to the need to estimate the unknown low-rank component.

Remark 4: If $r = 0$ (no low-rank component present in the model), the consistency results are similar to those in [44]. Specifically, Theorem 3 could be seen as an extension of Theorem 3 in [44]. Further, whenever $r = 0$, the total number of time series components could be of order $o(e^n)$, while for $r \geq 1$, we must have $p = o(n)$ since the low-rank component in each transition matrix is potentially dense. This is similar to the stationary (no break points) case discussed in [8].

C. Step 3: Consistent Parameter Estimation

The main idea to consistently estimate the model parameters is that Theorem 2 and Theorem 3 indicate that removing the estimated change points together with an adequate R_n-radius neighborhood around them will also remove the true change points. Hence, the remainder time segments would be stationary. Theorem 2 points out that the radius R_n can be as small as $n\gamma_n$, while Theorem 3 establishes that this radius should be at least $Bm_0n\gamma_n(d_n^{\star 2}+r^{\star 2})$ for some large constant $B>0$, in order to drop out redundant change points.

Given the results in Theorem 3, suppose that we have selected m_0 change points using the screening procedure. Denote these estimated change points by $\widetilde{t}_1,\widetilde{t}_2,\ldots,\widetilde{t}_{m_0}$. Then, by Theorem 3, we have

$$\mathbb{P}\left(\max_{1\le j\le m_0}|\widetilde{t}_j-t_j|\le R_n\right)\to 1,$$

as $n\to+\infty$. Denote the neighborhood of $\widetilde{t}_j$ as $I_{j+1}=[r_{j2},r_{(j+1)1}]$ for $j=0,1,\ldots,m_0$, where $r_{j1}=\widetilde{t}_j-R_n-1$ and $r_{j2}=\widetilde{t}_j+R_n+1$ for $j=1,2,\ldots,m_0$ and let $r_{02}=1$ and $r_{(m_0+1)1}=n$. Then, we formulate a regularized linear regression on $\bigcup_{j=0}^{m_0}I_{j+1}$ and estimate the sparse and low rank components of VAR parameters.

Similar to Theorem 1, we consider estimating the transition matrices in each obtained segment *separately* through a regularized linear regression method. Specifically, for interval I_{j+1}, we can write the following linear regression

$$\mathcal{Y}_j=\mathcal{X}_j(S_j+L)+\epsilon_j, \tag{11}$$

where we analogously define the matrix variables $\mathcal{Y}_j=[X_{r_{j2}},\ldots,X_{r_{(j+1)1}}]'$, $\mathcal{X}_j=[X_{r_{j2}-1},\ldots,X_{r_{(j+1)1}-1}]'$ and ϵ_j is the corresponding error term. Let N_j be the length of the interval I_{j+1} for $j=0,1,\ldots,m_0$ and $N=\sum_{j=1}^{m_0}N_j$. Then, $\mathcal{X}_j$ and $\mathcal{Y}_j\in\mathbb{R}^{N_j\times p}$, S_j and $L\in\mathbb{R}^{p\times p}$. We simultaneously estimate the low rank and sparse components of the VAR transition matrices in each stationary interval I_{j+1} by solving the following restricted regularized optimization problem

$$\begin{aligned}(\widehat{L},\widehat{S}_j)=&\ \underset{L\in\Omega,S_j}{\arg\min}\frac{1}{N_j}\|\mathcal{Y}_j-\mathcal{X}_j(S_j+L)\|_F^2\\&+\rho_j\|S_j\|_1+\rho_L\|L\|_*.\end{aligned} \tag{12}$$

Then, the error bound for each estimated segment is:

Theorem 4: Suppose assumptions H1–H5 hold, m_0 is unknown and $R_n=Bm_0n\gamma_n(d_n^{\star 2}+r^{\star 2})$. Assuming that $\rho_j=C_1\sqrt{\frac{\log N_j+2\log p}{N_j}}+C_2\frac{\tau}{p\gamma}$ and $\rho_L=C_1'\max_j\sqrt{\frac{p}{N_j}}$ for some large enough constants $C_1,C_1',C_2>0$ and curvature parameter $\tau>0$ in the restricted strong convexity assumption [37] . Then, as $n\to+\infty$, the optima $(\widehat{L},\widehat{S}_j)$ of (12) satisfies

$$\|\widehat{S}_j-S_j^\star\|_F^2+\|\widehat{L}-L^\star\|_F^2=\mathcal{O}\left(\frac{r^\star p+d_j^\star\log p}{N_j}+\frac{d_j^\star}{p^2\gamma^2}\right).$$

In order to consider all segments simultaneously, the length of estimated segments must be similar to each other, otherwise the error rate may not be optimal. In the next Theorem, we assume $\Delta_n>\delta n$ for some positive constant δ in order to ensure that all N_j's are of the same order n. Then, when considering all estimated segments simultaneously, (11) can be written into another matrix form as follows

$$\mathcal{Y}_r=\mathcal{X}_r(\mathbf{S}+\mathbf{1}_{m_0+1}\otimes L)+E_r,$$

where the coefficient matrix is $\mathbf{S}=[S_1',S_2',\ldots,S_{m_0+1}']'$ and $\mathbf{1}_{m_0+1}=[1,1,\ldots,1]'\in\mathbb{R}^{(m_0+1)\times 1}$; the design matrix is given by $\mathcal{X}_r=\mathrm{diag}(\mathcal{X}_1,\ldots,\mathcal{X}_{m_0+1})$, the response matrix is $\mathcal{Y}_r=[\mathcal{Y}_1',\ldots,\mathcal{Y}_{m_0+1}']'$ and the corresponding error matrix is defined as $E_r=[\epsilon_1',\ldots,\epsilon_{m_0+1}']'$. Let $N=\sum_{j=0}^{m_0}N_j$. Then, $\mathcal{X}_r\in\mathbb{R}^{N\times(m_0+1)p}$, $\mathcal{Y}_r\in\mathbb{R}^{N\times p}$, and $E_r\in\mathbb{R}^{N\times p}$; $\mathbf{S}\in\mathbb{R}^{(m_0+1)p\times p}$. Then, solving the following restricted regularized optimization problem

$$\begin{aligned}(\widehat{L},\widehat{\mathbf{S}})=&\ \underset{L\in\Omega,\mathbf{S}}{\arg\min}\frac{1}{N}\|\mathcal{Y}_r-\mathcal{X}_r(\mathbf{S}+\mathbf{1}_{m_0+1}\otimes L)\|_F^2\\&+\rho_n\|\mathbf{S}\|_1+\rho_L\|L\|_*.\end{aligned}$$

yields the desired estimates, for which we establish the following error bound.

Theorem 5: Suppose assumptions H1–H5 hold, m_0 is unknown and define $R_n=Bm_0n\gamma_n(d_n^{\star 2}+r^{\star 2})$. Assume that $\Delta_n>\delta n$ for some large positive constant δ, and $\rho_n=C_1\sqrt{\frac{\log N+2\log p}{N}}+C_2\frac{\tau}{p\gamma}$, $\rho_L=C_1'\sqrt{\frac{p}{N}}$ for some large enough constants $C_1,C_2,C_1'>0$ and curvature parameter $\tau>0$ in the restricted strong convexity assumption [37] . Then, as $n\to+\infty$, the optimal $(\widehat{L},\widehat{\mathbf{S}})$ satisfies

$$\begin{aligned}&\|\widehat{\mathbf{S}}-\mathbf{S}^\star\|_F^2+(m_0+1)\|\widehat{L}-L^\star\|_F^2\\&\quad=\mathcal{O}\left(\frac{r^\star pm_0+d_n^\star\log p}{N}+\frac{d_n^\star}{p^2\gamma^2}\right).\end{aligned}$$

Remark 5: The above Theorems provide a simultaneous error bound for the low-rank and sparse components. Note that a separate error bound for each component can not be derived, which is also the case for i.i.d. data and in the absence of a change point, as discussed in [1], or for stationary data in [8]. Therefore, as seen in the statement of Theorems 4 and 5, the error bound provided comprises of two key terms. The first term corresponds to the estimation error. For a given model, this term converges to zero as the sample size increases. The second term reflects the lack of exact identifiability of the model parameters, and only depends on the model size p, the total sparsity $d_n^\star$, and the information ratio γ and does not vanish even in the presence of infinite samples.

Remark 6: All optimization problems introduced in our methodology including (4), (7) and (12) are convex and can be solved by proximal gradient methods by combining algorithms developed in [8] and [44]. To speed up the detection procedure, new and fast algorithms are defined which approximate the minimizers in the three steps numerically (see details in the Supplement). These algorithms are implemented and used in our numerical experiments and the results in Section IV show their good performance.

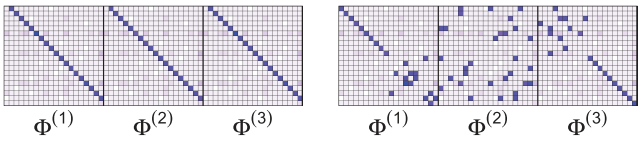


Fig. 1. Left: True structure of transition matrices for scenarios A-D and F. Right: True structure of transition matrices for scenario E.

IV. Performance Evaluation

Next, we present results from several numerical experiments that evaluate the performance of the proposed strategy for detecting change points and also estimating the VAR parameters of the posited model. The time series data $\{X_t\}$ with m_0 change points are generated from the model $X_t = B_j' X_{t-1} + \epsilon_t$, where $B_j = L^\star + S_j^\star$ and $t \in (t_{j-1}^\star, t_j^\star)$ for $j = 1, 2, \ldots, m_0$. We set the true rank $r^\star = \lfloor p/15 \rfloor + 1$ and the block size $b_n = \sqrt{n}$ for the BFL step unless otherwise specified. We also set the convergence tolerance to 10^{-1} for the BFL step to select candidate break points and 10^{-3} for the estimation (3 rd) step. We set the information ratio $\gamma = 4$ (defined in H2) for most settings (i.e. we set $\alpha = p/4$ in the constrained space Ω previously defined). We investigate smaller values for γ in scenario D and higher dimension p in scenario F as well.

There are a number of factors potentially influencing the performance of the strategy; in particular, the number of time series p, the sample size n, the location of change points, the rank of $L^\star$ and the information ratio γ. In this section, we mainly consider the following scenarios.

The transition matrices have the same structure, but different magnitudes. Fig. 1 illustrates the 1-off diagonal structure for transition matrices with values $-\gamma\|L^\star\|_\infty$, $\gamma\|L^\star\|_\infty$ and $-\gamma\|L^\star\|_\infty$, respectively. We set $\gamma = 4$ and the locations of two change points at $t_1^\star = \lfloor n/3 \rfloor$ and $t_2^\star = \lfloor 2n/3 \rfloor$.

A. Scenarios Examined

A. In the first scenario, the principle factor investigated is sample size and we examine three different sample sizes.
B. In this scenario, we investigate how different choices for rank influence performance. We consider both small and larger ranks.
C. In this scenario, we consider settings involving different number of change points. Specifically, we examine the following two cases: *(a)* $t_1^\star = \lfloor n/6 \rfloor$, $t_2^\star = \lfloor n/3 \rfloor$ and $t_3^\star = \lfloor 2n/3 \rfloor$; *(b)* $T = 600$ with $t_1^\star = \lfloor n/6 \rfloor$, $t_2^\star = \lfloor n/4 \rfloor$, $t_3^\star = \lfloor n/3 \rfloor$, $t_4^\star = \lfloor 2n/3 \rfloor$ and $t_5^\star = \lfloor 5n/6 \rfloor$. It should be noted that we adopt smaller block sizes $b_n = \sqrt{n}/2$ or $b_n = \sqrt{n}/5$ for the BFL step in order to obtain a better result in this experiment.
D. In this scenario, we investigate a lower information ratio $\gamma = 2$. As mentioned in the theory section, γ is a crucial factor for identifying and estimating the low rank and sparse components and hence detecting change points.
E. In this scenario, we consider a random sparse component, instead of 1-off diagonal sparse component. We also examine a combination of diagonal and random sparse structures and evaluate the performance under levels of sparsity for the latter components. The right panel in Fig. 1 depicts the random structure employed.
F. In this scenario, we examine the effect of the dimension p (number of time series). We consider three different dimension settings with two change points at locations $t_1^\star = \lfloor n/3 \rfloor$ and $t_2^\star = \lfloor 2n/3 \rfloor$.

TABLE I
Model Parameters Under Different Scenario Settings

	p	n	$t_j^\star/n$	$r^\star$	γ
A.1	20	150	$(0.3333, 0.6667)$	2	4
A.2	20	300	$(0.3333, 0.6667)$	2	4
A.3	20	600	$(0.3333, 0.6667)$	2	4
B.1	20	300	$(0.3333, 0.6667)$	5	4
B.2	20	300	$(0.3333, 0.6667)$	10	4
B.3	20	300	$(0.3333, 0.6667)$	15	4
C.1	20	300	$(0.1667, 0.3333, 0.6667)$	2	4
C.2	20	600	$(0.1667, 0.2500, 0.3333, 0.6667, 0.8333)$	2	4
D.1	20	300	$(0.3333, 0.6667)$	2	2
E.1	20	300	$(0.3333, 0.6667)$	2	4
F.1	50	600	$(0.3333, 0.6667)$	4	4
F.2	100	1000	$(0.3333, 0.6667)$	7	4
F.3	200	1000	$(0.3333, 0.6667)$	14	4

Table I summarizes all the model parameters in the various scenarios discussed above.

B. Tuning Parameter Selection

There are a number of tuning parameters in the developed three-step strategy: $\lambda_{1,n}$, $\lambda_{2,n}$, $\lambda_{3,n}$, $\boldsymbol{\eta}_n$, η_L, ω_n, R_n, ρ_L and ρ_j for $j = 1, 2, \ldots, m_0$. Although the theoretical rates for these tuning parameters are provided in the theory section, their selection in finite sample applications should be further discussed. Next, we provide guidelines for selecting them.

$\lambda_{1,n}$: We use fixed $\lambda_{1,n}$ in accordance to the nature of the application. In most cases, we manually choose $\lambda_{1,n}$ in the range $[\sqrt{\frac{p}{n}}, 10\sqrt{\frac{p}{n}}]$.
$\lambda_{2,n}$: We can select $\lambda_{2,n}$ through cross-validation. In the simulation study, we randomly select 20% of the blocks equally spaced with a random initial point. Denote the last time point in these selected blocks by $\mathcal{T}$. Data without observations in $\mathcal{T}$ can then be used in the first step of our procedure to estimate Θ for a range of values for $\lambda_{2,n}$. The parameters estimated in the first step are used to predict the time series at time points in $\mathcal{T}$. The value of $\lambda_{2,n}$ which minimizes the mean squared prediction error over $\mathcal{T}$ is the cross-validated choice of $\lambda_{2,n}$.
$\lambda_{3,n}$: As previously discussed, the rate for $\lambda_{3,n}$ vanishes fast as n increases. Thus for simplicity, we suggest setting $\lambda_{3,n}$ to zero. This choice was used in all of the numerical experiments in this paper and it gives satisfactory results.
η_L: This parameter is set to be the same as $\lambda_{1,n}$.
ρ_L: This parameter is suggested to be kept fixed; in practice, it was set in the range $[\sqrt{\frac{p}{n}}, 10\sqrt{\frac{p}{n}}]$.
ρ_j: Finally, we need to select the tuning parameters ρ_j for sparse estimation in each selected segment. We select ρ_j as the minimizer of the Bayesian Information Criterion (BIC) for the j-th segments. For $j = 0, 1, \ldots, \widetilde{m}$, We

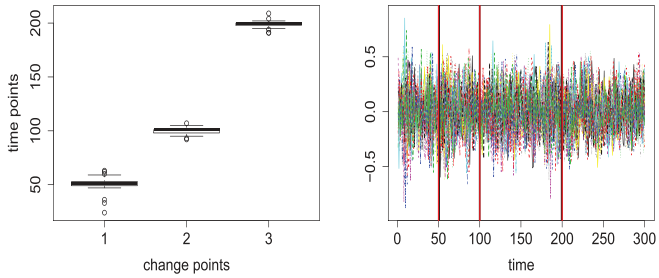


Fig. 2. Left panel: Box-plot for estimated change points for all 50 simulation replicates under scenario C.1. Right panel: Mean results for estimated change points (black lines) and true change points (red lines).

TABLE II
RESULTS FOR CHANGE POINT SELECTION UNDER PARAMETERS SETTINGS IN TABLE I

	points	truth	mean	sd	selection rate
A.1	1	0.3333	0.3272	0.0138	1
	2	0.6667	0.6527	0.0270	1
A.2	1	0.3333	0.3332	0.0087	1
	2	0.6667	0.6583	0.0181	1
A.3	1	0.3333	0.3324	0.0098	1
	2	0.6667	0.6712	0.0100	1
B.1	1	0.3333	0.3413	0.0234	0.98
	2	0.6667	0.6665	0.0089	1
B.2	1	0.3333	0.3357	0.0128	1
	2	0.6667	0.6585	0.0139	1
B.3	1	0.3333	0.3291	0.0154	0.98
	2	0.6667	0.6629	0.0103	1
C.1	1	0.1667	0.1699	0.0221	0.90
	2	0.3333	0.3328	0.0097	1
	3	0.6667	0.6643	0.0096	1
C.2	1	0.1667	0.1651	0.0055	0.98
	2	0.2500	0.2502	0.0050	1
	3	0.3333	0.3349	0.0051	1
	4	0.6667	0.6653	0.0049	1
	5	0.8333	0.8049	0.0199	0.98
D.1	1	0.3333	0.3329	0.0117	1
	2	0.6667	0.6574	0.0141	1
E.1	1	0.3333	0.3340	0.0190	1
	2	0.6667	0.6610	0.0214	1
F.1	1	0.3333	0.3252	0.0089	1
	2	0.6667	0.6728	0.0097	1
F.2	1	0.3333	0.3372	0.0087	1
	2	0.6667	0.6660	0.0074	1
F.3	1	0.3333	0.3218	0.0587	1
	2	0.6667	0.6660	0.0090	1

define the BIC on the interval $I_{j+1} = [r_{j2}, r_{(j+1)1}]$ as follows:

$$\text{BIC}(\rho_j) = \log\det\widehat{\Sigma}_{\epsilon,j} + \frac{\log(r_{(j+1)1} - r_{j2})}{(r_{(j+1)1} - r_{j2})}\|\widehat{S}_{j+1}\|_0,$$

where $\widehat{\Sigma}_{\epsilon,j}$ is the residual sample covariance matrix with $\widehat{L}$ and $\widehat{S}_j$ estimated in (12), and $\|\widehat{S}_{j+1}\|_0$ is the number of non-zero elements in $\widehat{S}_{j+1}$.

The remaining tuning parameters can be selected based on the choices mentioned in [44].

C. Simulation Results

We evaluate the empirical performance of our algorithm by considering the mean and standard deviation of the estimated

TABLE III
PERFORMANCE EVALUATION OF LOW-RANK COMPONENT UNDER DIFFERENT MODEL SETTINGS

	$r^\star$	$\lceil\widehat{r}\rceil$	Error
A.1	2	$2_{(0.855)}$	$0.71_{(0.038)}$
A.2	2	$2_{(0.723)}$	$0.62_{(0.042)}$
A.3	2	$2_{(0.141)}$	$0.60_{(0.035)}$
B.1	5	$5_{(0.913)}$	$0.67_{(0.034)}$
B.2	10	$10_{(0.974)}$	$0.58_{(0.040)}$
B.3	15	$15_{(3.173)}$	$0.76_{(0.177)}$
C.1	2	$2_{(0.767)}$	$0.87_{(0.041)}$
C.2	2	$2_{(0.888)}$	$0.81_{(0.041)}$
D.1	2	$2_{(0.519)}$	$0.73_{(0.036)}$
E.1	2	$2_{(0.707)}$	$1.09_{(0.111)}$
F.1	4	$4_{(0.012)}$	$0.66_{(0.022)}$
F.2	7	$7_{(0.707)}$	$0.61_{(0.013)}$
F.3	14	$15_{(0.627)}$	$0.98_{(0.050)}$

TABLE IV
PERFORMANCE EVALUATION OF SPARSE COMPONENTS UNDER DIFFERENT MODEL SETTINGS

	SEG	SEN	SPC	Error
A.1	1	$0.99_{(0.016)}$	$0.94_{(0.031)}$	$0.31_{(0.072)}$
	2	$0.99_{(0.024)}$	$0.92_{(0.036)}$	$0.34_{(0.072)}$
	3	$0.99_{(0.024)}$	$0.92_{(0.040)}$	$0.34_{(0.101)}$
A.2	1	$1.00_{(0.000)}$	$0.95_{(0.023)}$	$0.24_{(0.043)}$
	2	$1.00_{(0.000)}$	$0.95_{(0.024)}$	$0.24_{(0.066)}$
	3	$1.00_{(0.000)}$	$0.95_{(0.024)}$	$0.23_{(0.076)}$
A.3	1	$1.00_{(0.000)}$	$0.96_{(0.015)}$	$0.18_{(0.030)}$
	2	$1.00_{(0.000)}$	$0.95_{(0.017)}$	$0.23_{(0.070)}$
	3	$1.00_{(0.000)}$	$0.96_{(0.015)}$	$0.16_{(0.020)}$
B.1	1	$0.99_{(0.023)}$	$0.91_{(0.063)}$	$0.32_{(0.138)}$
	2	$1.00_{(0.010)}$	$0.93_{(0.026)}$	$0.23_{(0.047)}$
	3	$1.00_{(0.013)}$	$0.94_{(0.020)}$	$0.23_{(0.041)}$
B.2	1	$0.99_{(0.018)}$	$0.98_{(0.011)}$	$0.37_{(0.072)}$
	2	$1.00_{(0.000)}$	$0.96_{(0.020)}$	$0.20_{(0.048)}$
	3	$1.00_{(0.000)}$	$0.98_{(0.012)}$	$0.35_{(0.080)}$
B.3	1	$0.94_{(0.065)}$	$0.99_{(0.012)}$	$0.50_{(0.146)}$
	2	$1.00_{(0.000)}$	$0.95_{(0.021)}$	$0.15_{(0.051)}$
	3	$0.96_{(0.043)}$	$0.99_{(0.009)}$	$0.48_{(0.135)}$
C.1	1	$0.95_{(0.054)}$	$0.97_{(0.030)}$	$0.50_{(0.124)}$
	2	$0.94_{(0.044)}$	$0.99_{(0.024)}$	$0.53_{(0.105)}$
	3	$1.00_{(0.000)}$	$0.93_{(0.023)}$	$0.24_{(0.041)}$
	4	$0.96_{(0.023)}$	$1.00_{(0.001)}$	$0.40_{(0.057)}$
C.2	1	$0.94_{(0.050)}$	$1.00_{(0.000)}$	$0.58_{(0.138)}$
	2	$0.92_{(0.142)}$	$0.98_{(0.027)}$	$0.59_{(0.122)}$
	3	$0.96_{(0.063)}$	$0.98_{(0.012)}$	$0.53_{(0.088)}$
	4	$1.00_{(0.000)}$	$1.00_{(0.001)}$	$0.32_{(0.047)}$
	5	$0.96_{(0.026)}$	$1.00_{(0.002)}$	$0.39_{(0.085)}$
	6	$0.96_{(0.026)}$	$1.00_{(0.006)}$	$0.49_{(0.150)}$
D.1	1	$0.99_{(0.018)}$	$0.98_{(0.011)}$	$0.40_{(0.052)}$
	2	$1.00_{(0.013)}$	$0.98_{(0.009)}$	$0.39_{(0.043)}$
	3	$0.99_{(0.016)}$	$0.97_{(0.018)}$	$0.38_{(0.055)}$
E.1	1	$0.94_{(0.042)}$	$0.92_{(0.021)}$	$0.57_{(0.050)}$
	2	$0.93_{(0.068)}$	$0.96_{(0.015)}$	$0.55_{(0.066)}$
	3	$0.93_{(0.057)}$	$0.91_{(0.038)}$	$0.62_{(0.084)}$
F.1	1	$1.00_{(0.000)}$	$0.98_{(0.006)}$	$0.20_{(0.030)}$
	2	$1.00_{(0.000)}$	$0.98_{(0.007)}$	$0.27_{(0.062)}$
	3	$1.00_{(0.000)}$	$0.98_{(0.005)}$	$0.19_{(0.017)}$
F.2	1	$1.00_{(0.000)}$	$0.95_{(0.015)}$	$0.17_{(0.042)}$
	2	$1.00_{(0.000)}$	$0.98_{(0.003)}$	$0.19_{(0.036)}$
	3	$1.00_{(0.000)}$	$0.96_{(0.008)}$	$0.14_{(0.013)}$
F.3	1	$1.00_{(0.016)}$	$0.93_{(0.019)}$	$0.29_{(0.093)}$
	2	$0.99_{(0.056)}$	$0.91_{(0.038)}$	$0.30_{(0.168)}$
	3	$1.00_{(0.000)}$	$0.93_{(0.015)}$	$0.25_{(0.034)}$

Fig. 3. The detected change points corresponding to the following times and events: $\widehat{t_1} = 115$ first man walks out of lobby; $\widehat{t_2} = 173$ two men walk in to lobby; $\widehat{t_3} = 231$ two men keep walking in to lobby; $\widehat{t_4} = 289$ two men stand together; $\widehat{t_5} = 347$ two men stand closer; $\widehat{t_6} = 405$ two men walk together; $\widehat{t_7} = 463$ two men walk out of lobby; $\widehat{t_8} = 521$ two men walk to the door; $\widehat{t_9} = 579$ two men walk through the door; $\widehat{t_{10}} = 637$ two men already exit; and $\widehat{t_{11}} = 695$ empty lobby.

Fig. 4. Selected segments and the corresponding sparse component of the time varying transition matrices. From left to right, we illustrate the 1st, 4th, 6th, and 11th estimated segments.

change point locations relative to the sample size, i.e. $\widetilde{t_j}/n$, and the percentage of simulation runs where change points are correctly detected. A detected change point is counted as a *success* for the j-th true change point, if it falls in the *selection interval:* $[t_{j-1} + \frac{t_j - t_{j-1}}{5}, t_j + \frac{t_{j+1} - t_j}{5}]$. Moreover, we use the estimated rank, sensitivity (SEN), specificity (SPC) and relative error in Frobenius norm (RE) (all defined next) as additional criteria to evaluate the performance of the estimates of low rank and sparse components of transition matrices.

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPC} = \frac{\text{TN}}{\text{FN} + \text{TN}},$$
$$\text{RE} = \frac{\|\text{Est.} - \text{Truth}\|_F}{\|\text{Truth}\|_F}.$$

As an illustration of the variability of the estimates, Fig. 2 depicts the estimated change points (left-panel boxplot of the estimates and right panel mean of the estimates) based on 50 replicates in scenario C.1. It shows that the proposed strategy estimates the change points with high accuracy.

The results in Table II illustrate the performance of change point detection for each of the settings considered in Table I. For most of the cases in scenarios A and B, the implemented algorithm provides a near perfect performance. In scenario C, we considered multiple changes. As expected, for those change points close to the boundary of the observation interval (or other change points), the selection rate exhibits a slight deterioration. In scenarios D, E and F, we still obtain a perfect selection rate even under the weaker sparse signal (scenario D) and the high

dimensional settings (scenario F). Overall, the results in Table II are highly satisfactory and clearly show that the proposed strategy is highly accurate in detecting both the number of change points and also their locations.

Tables III and IV present the performance of the estimation step. It is worth mentioning that for most of the simulation results, less than 20 iterations were needed to obtain the minimizers of the corresponding regularized optimization problems. The regularization parameters are selected based on the guidelines previously provided. The results strongly support the effectiveness of the strategy and the algorithms used in each step. One can easily see that all parameters are estimated with a high degree of accuracy. As expected, when the rank increases, a greater portion of the signal strength is absorbed into the low rank component and thus the estimation of the sparse components becomes less accurate. This is illustrated in the B.1, B.2 and B.3 settings of Table IV. Another interesting experiment is setting D.1, in which sparse components do not provide a strong signal; therefore, the estimation results for the sparse components under D.1 exhibit less accuracy compared to scenario A.2. The result for scenario E.1 demonstrates that our strategy and algorithm have high sensitivity and specificity for the sparse estimates and accurate estimation of the rank for the low rank component on the random sparse pattern as well. The last three results for scenario F illustrate the performance for larger size models; the results indicate that the proposed strategy is robust under higher dimensional settings.

V. Real Data Applications

A. Surveillance Video Data Set

The proposed detection algorithm is applied to a surveillance video data set obtained from the CAVIAR project.[2] A number of video clips record different actions by people in diverse settings, including walking alone, meeting with others, entering and exiting a room, etc. The resolution of each image is based on the half-resolution PAL standard (384×288 pixels, 25 frames per second). We analyzed the *Two other people meet and walk together* data set, comprising of 827 images.

We first re-sized the original images from 384×288 pixels to 32×24 pixels and used a gray-scaled scheme instead of the original colored image to accelerate computations. Therefore, the resulting data matrix has $n = 837$ time points and $p = 32 \times 24 = 768$ features.

The proposed model is perfectly suited for this task, since there is a non-changing low-rank component corresponding to the *stationary background* of the space surveyed, while the changing sparse component corresponds to movement of people in and out of the space in the *evolving foreground*.

Fig. 3 depicts the selected change points by the algorithm and the nature of the change is illustrated by a representative frame from the original video. Given the complexity of the background, a rank 18 component was selected to capture it. In Fig. 4, we also show the most *significantly* changing pixels captured in the sparse component of transition matrix for the 1^{st}, 4^{th}, 6^{th}, 11^{th} estimated segments, respectively. Specifically, for the j-th estimated interval, the (k, l) elements in $\widehat{S}_j$ reflect the partial autocorrelation between pixels k and l in the original image. Therefore, we selected the largest 20 elements in $\widehat{S}_j$ and mapped the pixels to the original image.

[2]http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

TABLE V
Detected CPs by the L+S VAR and a Factor Model

No. of CPs	L+S model	Factor model
1	2/12/02	12/17/02
2	9/3/02	4/8/03
3	3/18/03	7/24/07
4	7/17/07	8/7/07
5	2/12/08	7/15/08
6	8/26/08	9/8/09
7	3/10/09	8/17/10
8	10/19/10	-
9	1/28/14	-
10	9/8/15	-

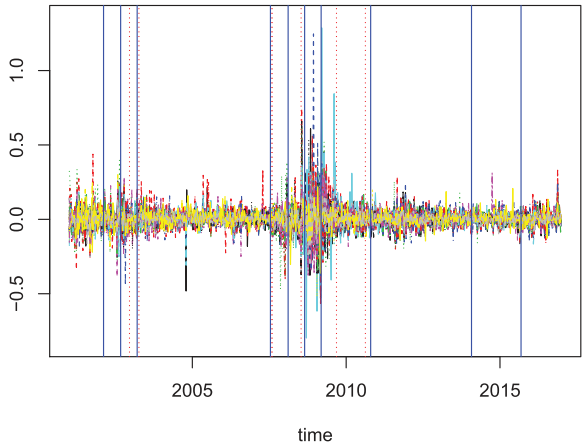


Fig. 5. Detected change points in the log-returns data during the 2001-2016 period. Red dashed lines are change points selected by Factor Analysis Model [5]; blue solid lines indicate the change points selected by our model.

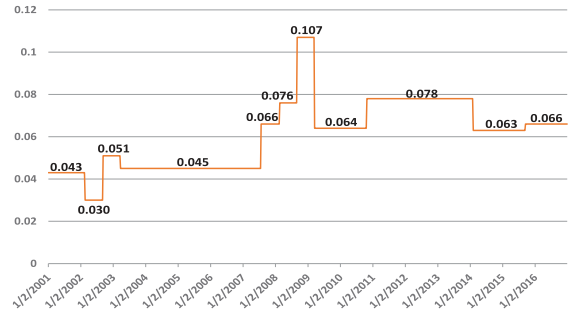


Fig. 6. Connectivity for each estimated sparse components in different selected time periods. 1/2/01-2/12/02 period: 4.3%; 2/12/02 - 9/3/02 period: 3.0%; 9/3/02 - 3/18/03 period: 5.1%; 3/18/03 - 7/17/07 period: 4.5%; 7/17/07 - 2/12/08 period: 6.6%; 2/12/08 - 8/26/08 period: 7.6%; 8/26/08 - 3/10/09 period: 10.7%; 3/10/09 - 10/19/10 period: 6.4%; 10/19/10 - 1/28/14 period: 7.8%; 1/28/14 - 9/8/15 period: 6.3%; 9/8/15 - 12/27/16 period: 6.6%.

B. Stock Data

Next, we employ the proposed detection strategy to identify change points in *weekly* financial stock price data, covering

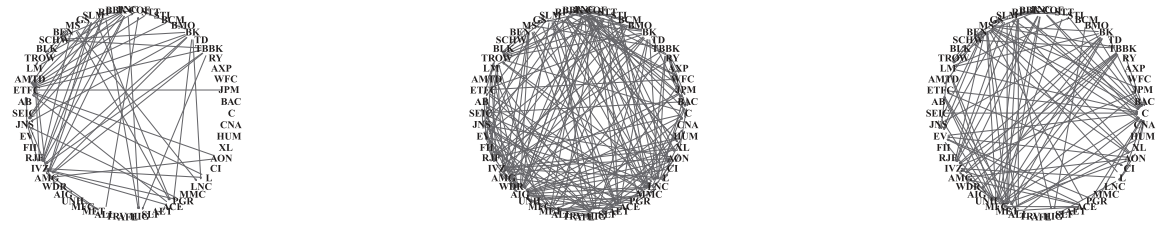


Fig. 7. Estimated connectivity based on selected time periods. Left: Structure of the connections in the pre-crisis period and has 62 edges; Middle: Structure of the connections among selected companies during the crisis period and has 228 edges; and Right: Structure of the connections in the post-crisis period and has 114 edges.

the 2001–2016 period. Extensive work in asset price theory indicates that stock log-returns can be accounted for by a few stable factors (either extracted through a statistical factor (low-rank) model [34], or constructed from a large scale diverse portfolio [18], [19]). The stocks in our analysis correspond to 52 stocks of banks, insurance companies and stock brokers that have complete data in the aforementioned time period.

We compared our model with factor analysis model proposed in [5]. Table V illustrates the ten change points selected by our strategy, along with seven change points identified by a competing procedure based on a factor analysis model [5]. Fig. 5 provides an overall picture of the selected change points based on the simplified version of the change point detection algorithm to select candidates in the first step (blue lines) compared with those detected by the factor analysis model (dashed red lines).

The overall (normalized) density of the time varying sparse component based on a 3-factor model is plotted in Fig. 6. The decision to use 3-factors was based on an examination of the singular values; for a 3-factor model they were 1.60, 0.094 and 0.054, while for a 5-factor model they were 2.29, 0.30, 0.11, 0.05, 0.04. It can be seen that the even for the 5-factor model, the first three singular values capture 95% of the total variance, while the last two contribute very little. We also conducted a residual analysis for model selection. Specifically, after obtaining the estimated $\widehat{L}$ and $\widehat{S}_j$ for $\widehat{m}+1$ segments, we derive the residuals for the j-th segment: $e_t = X_t - \widehat{X}_t$, for $t \in [\widehat{t}_j, \widehat{t}_{j+1} - 1]$. Then, the sum of squared residuals is given by $\sum_{j=1}^{\widehat{m}+1} \frac{1}{\widehat{t}_{j+1} - \widehat{t}_j} \sum_{t=\widehat{t}_j}^{\widehat{t}_{j+1}-1} \|e_t\|_2^2$. Naturally, a model is more suitable if this quantity is smaller. For the rank 3 and rank 5 component models the corresponding values are 1.569 and 1.582, respectively. This result indicates that a rank 3 component is preferable.

Our model identifies ten change points corresponding to major economic/financial shocks that occurred during the period under consideration and impacted in particular the performance of financial stocks. Specifically, the two 2002 change points cover the period when the telecommunications bubble popped following that of the dot-com crash and drove the NASDAQ index significantly lower, thus markedly affecting market sentiment. The first change point in 2008 precedes the collapse of Bear Sterns (early March 2008), while the second one that of Lehman Brothers (mid-September 2008), while the first one in 2009 marks the end of the sharp market downturn following the Great Recession. The next three change points capture shocks (affecting in particular financial stocks) related to the European sovereign debt crisis that involved significant downgrades of the debt of several European Union countries, bailouts and recapitalization of banks and in general a lot of market distress. Finally, the September 2015 change point captures the severe market downturn spanning most of late 2014 and beginning of 2015 time period. In contrast, the factor analysis based model, detects only seven change points (in fact, the third and fourth are too close to be identified as two independent change points), and does not identify any change points after 2010. Further, the location of the 2008 change point is two months before the collapse of Lehman Brothers, whereas our strategy identifies one three weeks before that event. Further, our model and strategy identify the turn of the market in early March of 2009, which coincides with the bottom that various stocks indices hit, while the factor model locates it in early September of 2009.

Fig. 7 provides the significant connectivity for the following three different time periods: March 2003–July 2007, August 2008–March 2009, October 2010 – January 2014 which correspond to instances before the financial crisis of 2008 (pre-crisis period), the apex of the crisis and the post-crisis period, respectively.

VI. Conclsion

In this paper, we developed a three-step strategy to detect the (unknown) break points and estimate the transition matrices in a high-dimensional VAR model in which the latter are assumed to be a superposition of a low-rank and a sparse component. The fixed, but unknown low-rank component introduces algorithmic challenges, since it needs to be estimated together with the other dynamically evolving parameters. From a technical perspective, the estimation of the low-rank part impacts the sum of squared error terms (SSEs), which is quantified in the consistency rates developed in Section III. Note that the developed methodology can be extended to VAR(d) with $d > 1$ in a similar way as discussed in [8]. Extension of the current framework to cases where the low-rank part could also change over time in a piece-wise manner constitutes an interesting future research direction which not only complicates the detection problem, but also requires a thorough investigation of associated identifiablity issues.

APPENDIX

Proof of Theorem 1: By the definition of $\widehat{\Theta}$ and $\widehat{L}$, we obtain that

$$\begin{aligned}
&\frac{1}{n}\|\mathcal{Y}-\mathcal{X}\widehat{L}-\mathcal{Z}\widehat{\Theta}\|_2^2+\lambda_{1,n}\|\widehat{L}\|_*+\lambda_{2,n}\|\widehat{\Theta}\|_1\\
&+\ \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\widehat{\theta}_j\right\|_1\le\frac{1}{n}\|\mathcal{Y}-\mathcal{X}L^\star-\mathcal{Z}\Theta^\star\|_2^2\\
&+\ \lambda_{1,n}\|L^\star\|_*+\lambda_{2,n}\|\Theta^\star\|_1+\lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\theta_j^\star\right\|_1. \qquad (13)
\end{aligned}$$

Denote by $\mathcal{A}=\{t_0,t_1,\ldots,t_{m_0}\}$ the set of true change points and also define $\widehat{\Delta}_L=\widehat{L}-L^\star$, $\widehat{\Delta}_\Theta=\widehat{\Theta}-\Theta^\star$. Using the conditions on $\lambda_{1,n}$, $\lambda_{2,n}$ and $\lambda_{3,n}$ to obtain:

$$\begin{aligned}
&\frac{1}{n}\left\|\mathcal{X}\widehat{\Delta}_L+\mathcal{Z}\widehat{\Delta}_\Theta\right\|_2^2\le\frac{2}{n}\widehat{\Delta}_L'\mathcal{X}'\mathcal{E}+\frac{2}{n}\widehat{\Delta}_\Theta'\mathcal{Z}'\mathcal{E}\\
&+\ \lambda_{1,n}(\|L^\star\|_*-\|L^\star+\widehat{\Delta}_L\|_*)+\lambda_{2,n}(\|\Theta^\star\|_1\\
&-\ \|\Theta^\star+\widehat{\Delta}_\Theta\|_1)+\lambda_{3,n}\sum_{l=1}^{k_n}\left(\left\|\sum_{j=1}^{l}\theta_j^\star\right\|_1-\left\|\sum_{j=1}^{l}\widehat{\theta}_j\right\|_1\right)\\
&\le 2\left\|\frac{\mathcal{X}'\mathcal{E}}{n}\right\|_{\text{op}}\|\widehat{\Delta}_L\|_*+2\left\|\frac{\mathcal{Z}'\mathcal{E}}{n}\right\|_\infty\|\widehat{\Delta}_\Theta\|_1\\
&\quad+\ \lambda_{1,n}(\|L^\star\|_*-\|L^\star+\widehat{\Delta}_L\|_*)-\lambda_{2,n}\sum_{i\in\mathcal{A}^c}\|\widehat{\theta}_i\|_1\\
&\quad+\ \lambda_{2,n}\sum_{i\in\mathcal{A}}\left(\|\theta_i^\star\|_1-\|\widehat{\theta}_i\|_1\right)+\lambda_{3,n}\sum_{l=1}^{k_n}b_n\|S_j^\star\|_1\\
&\le\lambda_{1,n}\|\widehat{\Delta}_L\|_*+\lambda_{1,n}(\|L^\star\|_*-\|L^\star+\widehat{\Delta}_L\|_*)\\
&\quad+\ 2\lambda_{2,n}\sum_{i\in\mathcal{A}}\|\theta_i^\star\|_1+\lambda_{3,n}(nd_n^\star)\\
&\le 2\lambda_{1,n}\|L^\star\|_*+2\lambda_{2,n}\sum_{i\in\mathcal{A}}\|\theta_i^\star\|_1+o(1)\\
&\le 4C_1\sqrt{r^\star p}C_0(p\gamma)^{-1}\sqrt{\frac{p}{n}}\\
&\quad+\ 4C_2m_n\max_{1\le j\le m_0+1}\left\{d_j^\star+d_{j-1}^\star\right\}\times\\
&M_S\sqrt{\frac{\log n+2\log p}{n}}+o(1) \qquad (14)
\end{aligned}$$

According to the definition of the information ratio γ and recalling the selection of $\lambda_{2,n}$, we can obtain the posited result. ■

Before we prove Theorem 2, we introduce the following two sets of sub-spaces $\{\mathcal{I},\mathcal{I}^c\}$ and $\{\mathcal{L}_A,\mathcal{L}_B\}$ corresponding to some generic sparse matrix $S\in\mathbb{R}^{p_1\times p_2}$ and some generic low-rank matrix $L\in\mathbb{R}^{p_1\times p_2}$, in which the ℓ_1 norm and nuclear norm are decomposable [37]. First, let $\mathcal{J}$ be the set of indices in which the sparse matrix S is non-zero. We then define the following sub-spaces

$$\begin{aligned}
\mathcal{I}&:=\{P\in\mathbb{R}^{p\times p}\mid P_{ij}=0\text{ for }(i,j)\notin\mathcal{J}\},\\
\mathcal{I}^c&:=\{P\in\mathbb{R}^{p\times p}\mid P_{ij}\neq 0\text{ for }(i,j)\in\mathcal{J}\}.
\end{aligned}$$

Then, for an arbitrary matrix M, $M|_{\mathcal{I}}\in\mathcal{I}$ is obtained by assigning the entries of M whose indices are not in $\mathcal{J}$ to 0, and $M|_{\mathcal{I}^c}\in\mathcal{I}^c$ is obtained by assigning the entries of M whose indices are in $\mathcal{J}$ to 0. Then, the following decomposition for the ℓ_1 norm holds

$$\|M\|_1=\|(M|_{\mathcal{I}}+M|_{\mathcal{I}^c})\|_1=\|M\|_{1,\mathcal{I}}+\|M\|_{1,\mathcal{I}^c},$$

where $\|A\|_{1,\mathcal{I}}=\sum_{(j,k)\in\mathcal{J}}|a_{jk}|$. Let the singular value decomposition of L be $L=U\Sigma V'$ with U and V being orthogonal matrices. Analogously, we define the following subspaces

$$\begin{aligned}
\mathcal{L}_A&:=\{\Delta\in\mathbb{R}^{p_1\times p_2}\mid\text{row}(\Delta)\subset V^r\text{ and col}(\Delta)\subset U^r\},\\
\mathcal{L}_B&:=\{\Delta\in\mathbb{R}^{p_1\times p_2}\mid\text{row}(\Delta)\perp V^r\text{ and col}(\Delta)\perp U^r\},
\end{aligned}$$

where $r=\text{rank}(L)$, and U^r and V^r denote the first r columns of U and V corresponding to the first r singular values of L. Therefore, we consider the restricted sub-matrices on the subspaces $\{\mathcal{L}_A,\mathcal{L}_B\}$ given by

$$L^A=U\begin{bmatrix}\widetilde{L}_{11}&\widetilde{L}_{12}\\ \widetilde{L}_{21}&O\end{bmatrix}V'\quad\text{and}\quad L^B=U\begin{bmatrix}O&O\\ O&\widetilde{L}_{22}\end{bmatrix}V',$$

where $\widetilde{L}_{11}\in\mathbb{R}^{r\times r}$. Then, we have $L^A+L^B=L$ and the following decomposition for the nuclear norm holds

$$\|L\|_*=\|(L^A+L^B)\|_*=\|L^A\|_*+\|L^B\|_*.$$

Proof of Theorem 2: First, we focus on the second part. Suppose for some $j=1,2,\ldots,m_0$, $|\widehat{t}_j-t_j|\ge n\gamma_n$. Then, there exists a true break point t_{j_0} which is isolated from all the estimated points, i.e., $\min_{1\le j\le m_0}|\widehat{t}_j-t_{j_0}|>n\gamma_n$. The idea is to show the estimated AR parameter $\widehat{S}_j$ in the interval $[t_{j_0-1}\vee\widehat{t}_j,t_{j_0+1}\wedge\widehat{t}_{j+1}]$ converges in ℓ_2 to both $S_j^\star$ and $S_{j+1}^\star$ which contradicts assumption H3.

Due to the definition of $(\widehat{L},\widehat{\Theta})$ in (4), the value of the function defined in (4) is minimized exactly at $(\widehat{L},\widehat{\Theta})$. Denote the closest r_i to the right side of t_{j_0-1} by s_{j_0-1} and the closest r_i to the left side of t_{j_0} by s_{j_0} similarly. First, we consider the interval $[s_{j_0-1}\vee\widehat{t}_j,s_{j_0}]$. Define a new parameter sequence ψ_k's, $k=1,2,\ldots,n$ with $\psi_k=\widehat{\theta}_k$ except for two time points $k=\widehat{t}_j$ and $k=s_{j_0}$. For these two points we assign $\psi_{\widehat{t}_j}=S_{j_0}^\star-\widehat{S}_j$ and $\psi_{s_{j_0}}=\widehat{S}_{j+1}-S_{j_0}^\star$ where $\widehat{S}_j=\sum_{k=1}^{s_{j_0-1}\vee\widehat{t}_j-1}\widehat{\theta}_k$ and $\widehat{S}_{j+1}=\sum_{k=1}^{s_{j_0-1}\vee\widehat{t}_j}\widehat{\theta}_k$, thus, $\widehat{\theta}_{s_{j_0}\vee\widehat{t}_j}=\widehat{S}_{j+1}-\widehat{S}_j$. Denoting $\Psi=[\psi_1',\psi_2',\ldots,\psi_{k_n}']'\in\mathbb{R}^{pk_n\times p}$, we obtain

$$\begin{aligned}
&\frac{1}{n}\|\mathcal{Y}-\mathcal{X}\widehat{L}-\mathcal{Z}\widehat{\Theta}\|_2^2+\lambda_{1,n}\|\widehat{L}\|_*+\lambda_{2,n}\|\widehat{\Theta}\|_1\\
&+\ \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\widehat{\theta}_j\right\|_1\\
&\le\frac{1}{n}\|\mathcal{Y}-\mathcal{X}L^\star-\mathcal{Z}\Psi\|_2^2+\lambda_{1,n}\|L^\star\|_*+\lambda_{2,n}\|\Psi\|_1\\
&\quad+\ \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\psi_j\right\|_1. \qquad (15)
\end{aligned}$$

According to the definition of ψ_k, we can define the differences between estimated coefficients and their true values $\widehat{\Delta}_L=$

$\widehat{L} - L^\star$ and $\widehat{\Delta}_S = \widehat{S}_{j+1} - S^\star_{j_0}$. For the specific interval, since we only consider the observations within this interval, and due to the fact that the length of the interval is large enough, we can verify the restricted eigenvalue and deviation bound inequalities (see [8]). We use $\widetilde{\mathcal{X}} = [X_{s_{j_0-1}\vee\widehat{t}_j}, X_{s_{j_0-1}\vee\widehat{t}_j+1}, \ldots, X_{s_{j_0}-1}]' \in \mathbb{R}^{(s_{j_0}-s_{j_0-1}\vee\widehat{t}_j)\times p}$ to denote the observations under consideration, while $\widetilde{\mathcal{E}}$ is the corresponding noise term. Then, a rearrangement of inequality (15) leads to

$$\begin{aligned}
&\frac{1}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\|\widetilde{\mathcal{X}}(\widehat{\Delta}_L+\widehat{\Delta}_S)\|_2^2 \\
&\le \frac{2\langle\widehat{\Delta}_L+\widehat{\Delta}_S, \widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\rangle}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j} + \frac{n\lambda_{1,n}\left(\|L^\star\|_* - \|\widehat{L}\|_*\right)}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j} \\
&\quad + \frac{n\lambda_{2,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\left(\|S^\star_{j_0}-\widehat{S}_{j+1}\|_1 + \|S^\star_{j_0}-\widehat{S}_j\|_1\right. \\
&\quad \left. - \|\widehat{S}_{j+1}-\widehat{S}_j\|_1\right) + \frac{n\lambda_{3,n}}{b_n}\left(\|S^\star_{j_0}\|_1 - \|\widehat{S}_{j+1}\|_1\right) \\
&\le \frac{2}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\langle\widehat{\Delta}_L+\widehat{\Delta}_S, \widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\rangle \\
&\quad + \frac{n\lambda_{1,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\left(\|\widehat{\Delta}^A_L\|_* - \|\widehat{\Delta}^B_L\|_*\right) \\
&\quad + \frac{2n\lambda_{2,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\|\widehat{\Delta}_S\|_1 + \frac{n\lambda_{3,n}}{b_n}\left(\|\widehat{\Delta}_S\|_{1,\mathcal{I}}\right. \\
&\quad \left. - \|\widehat{\Delta}_S\|_{1,\mathcal{I}^c}\right) + \frac{2n\lambda_{1,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\sum_{j=r+1}^{p}\sigma_j(L^\star), \qquad (16)
\end{aligned}$$

where the matrix pair (A, B) are from the sub-spaces $\{\mathcal{L}_A, \mathcal{L}_B\}$, respectively. The second inequality holds due to the decomposition of the ℓ_1-norm, the nuclear norm in [1] and an application of the triangle inequality.

According to Hölder's inequality, the first term of the right hand side of the second inequality in (16) implies the following inequality

$$\begin{aligned}
&\langle\widehat{\Delta}_L+\widehat{\Delta}_S, \widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\rangle \le \|\widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\|_{\text{op}}\|\widehat{\Delta}_L\|_* + \|\widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\|_\infty\|\widehat{\Delta}_S\|_1 \\
&= \|\widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\|_{\text{op}}\left(\|\widehat{\Delta}^A_L\|_* + \|\widehat{\Delta}^B_L\|_*\right) \\
&\quad + \|\widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\|_\infty\left(\|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \|\widehat{\Delta}_S\|_{1,\mathcal{I}^c}\right) \qquad (17)
\end{aligned}$$

Substituting (17) into (16) and considering the conditions for $\lambda_{1,n}$, $\lambda_{2,n}$ and $\lambda_{3,n}$, we have

$$\begin{aligned}
&\frac{1}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\|\widetilde{\mathcal{X}}(\widehat{\Delta}_L+\widehat{\Delta}_S)\|_2^2 \\
&\le \frac{3n\lambda_{1,n}}{2b_n}\|\widehat{\Delta}^A_L\|_* + \frac{3n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{2n\lambda_{3,n}}{b_n}\|S^\star_{j_0}\|_{1,\mathcal{I}^c} \\
&\quad + \left(\frac{2n\lambda_{2,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j} + C\sqrt{\frac{\log p}{n\gamma_n}}\right)\|\widehat{\Delta}_S\|_1 \\
&\quad + \frac{2n\lambda_{1,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\sum_{j=r+1}^{p}\sigma_j(L^\star) \\
&\le \frac{3n\lambda_{1,n}}{2b_n}\|\widehat{\Delta}^A_L\|_* + \frac{3n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{2n\lambda_{3,n}}{b_n}\|S^\star_{j_0}\|_{1,\mathcal{I}^c} \\
&\quad + \frac{n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_1 + \frac{2n\lambda_{1,n}}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\sum_{j=r+1}^{p}\sigma_j(L^\star) \\
&= \frac{3n\lambda_{1,n}}{2b_n}\|\widehat{\Delta}^A_L\|_* + \frac{3n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_1. \qquad (18)
\end{aligned}$$

The first inequality holds with high probability converging to 1 due to part (a) in Lemma 2 and the fact that $s_{j_0} - s_{j_0-1}\vee\widehat{t}_j \ge \frac{1}{2}n\gamma_n$ and $b_n \le \frac{1}{4}n\gamma_n$ by assumption H3. The second inequality is based on triangle inequality and the selection for $\lambda_{2,n}$ and $\lambda_{3,n}$. The last equality holds by the definition of decomposition properties of the ℓ_1 and nuclear norm, respectively.

On the other hand, by the restricted strong convexity condition [8], there exists a constant $\tau > 0$ such that

$$\begin{aligned}
&\frac{1}{s_{j_0}-s_{j_0-1}\vee\widehat{t}_j}\|\widetilde{\mathcal{X}}(\widehat{\Delta}_L+\widehat{\Delta}_S)\|_2^2 \ge \frac{\tau}{2}\|\widehat{\Delta}_L+\widehat{\Delta}_S\|_2^2 \\
&\ge \frac{\tau}{2}\left(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2 - 2|\langle\widehat{\Delta}_L, \widehat{\Delta}_S\rangle|\right) \\
&\ge \frac{\tau}{2}\left(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2 - 2\|\widehat{\Delta}_L\|_\infty\|\widehat{\Delta}_S\|_1\right) \\
&\ge \frac{\tau}{2}\left(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2\right) - \frac{n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_1 \qquad (19)
\end{aligned}$$

Inserting the inequality (19) into (18), we have

$$\begin{aligned}
&\frac{\tau}{2}\left(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2\right) \le \frac{3n\lambda_{1,n}}{2b_n}\|\widehat{\Delta}^A_L\|_* + \frac{5n\lambda_{3,n}}{2b_n}\|\widehat{\Delta}_S\|_1 \\
&\le \left(\frac{3n\lambda_{1,n}}{2b_n}\sqrt{2r^\star}\right)\|\widehat{\Delta}_L\|_2 + \left(\frac{5n\lambda_{3,n}}{2b_n}\sqrt{d^\star_n}\right)\|\widehat{\Delta}_S\|_2 \\
&\le \sqrt{\left(\frac{3n\lambda_{1,n}}{2b_n}\sqrt{2r^\star}\right)^2 + \left(\frac{5n\lambda_{3,n}}{2b_n}\sqrt{d^\star_n}\right)^2}\sqrt{\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2}
\end{aligned}$$

Further, combining with our tuning parameters assumption, we obtain

$$\begin{aligned}
&\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2 \le \frac{4}{\tau^2}\left(\frac{9n^2\lambda^2_{1,n}}{2b_n^2}r^\star + \frac{25n^2\lambda^2_{3,n}}{4b_n^2}d^\star_n\right) \\
&= \frac{4}{\tau^2}\left(\frac{9C_1^2}{2}\frac{r^\star p}{n\gamma_n} + \frac{25C_3^2}{4}\frac{d^\star_n\log p}{n\gamma_n}\right) \qquad (20)
\end{aligned}$$

This result shows that

$$\|\widehat{L}-L^\star\|_2^2 + \|S^\star_{j_0}-\widehat{S}_{j+1}\|_2^2 = o_p\left(\frac{r^\star p + d^\star_n\log p}{n\gamma_n}\right), \qquad (21)$$

which indicates that $\|\widehat{L}-L^\star\|_2^2 + \|S^\star_{j_0}-\widehat{S}_{j+1}\|_2^2$ converges to zero in probability based on assumption H3. Similarly, we can perform the same procedure to the interval $[s_{j_0}, s_{j_0+1}\wedge\widehat{t}_{j+1}]$ to get that $\|\widehat{L}-L^\star\|_2^2 + \|S^\star_{j_0+1}-\widehat{S}_{j+1}\|_2^2$ converges to zeros as well, which leads to that $\|S^\star_{j_0}-\widehat{S}_{j+1}\|_2^2 - \|S^\star_{j_0}-\widehat{S}_j\|_2^2$ converges to zero as well, and this implies to a contradiction to the

first part of assumption H3. Therefore, we proved the second part of the theorem.

The first part can be proved as follows. We assume that $|\widehat{\mathcal{A}}_n| < m_0$, which implies that there exists an isolated true change point, denoted by s_{j_0}. Then, we can separately apply the same procedure as in establishing the second part to the intervals $[s_{j_0}, s_{j_0+1} \wedge \widehat{t}_{j+1}]$ and $[s_{j_0-1} \vee \widehat{t}_j, s_{j_0}]$ which can lead to $\|S^\star_{j+1} - S^\star_j\|_2$ converges to zero and therefore contradicts with assumption H3. ■

Proof of Theorem 3: To prove the first part, we need to consider the equivalent two parts (a) $\mathbb{P}(\widetilde{m} < m_0) \to 0$ and (b) $\mathbb{P}(\widetilde{m} > m_0) \to 0$ respectively.

For case (a), we can directly obtain from Theorem 2 that there exist points $\widehat{t}_j \in \widehat{\mathcal{A}}_n$ satisfying that $\max_{1\le j\le m_0} |\widehat{t}_j - t_j| \le n\gamma_n$. According to the arguments in Lemma 3, we get that there exists a constant $K > 0$ such that

$$L_n(\widehat{t}_1, \ldots, \widehat{t}_{m_0}; \eta_n) \le \sum_{t=1}^{n} \|\epsilon_t\|_2^2 + Km_0 n\gamma_n(d_n^{\star\,2} + r^{\star 2}). \tag{22}$$

To prove (22), we only need to consider one of the estimated segments. Suppose $s_{i-1} < t_j < s_i$ with $|t_j - s_{i-1}| \le n\gamma_n$. We use $\widehat{\theta}$ to denote the estimated sparse component in the segment (s_{i-1}, s_i) and we use $\widehat{L}$ to denote the estimated low-rank component. Moreover, let $\widehat{\Delta}_L = \widehat{L} - L^\star$ and $\widehat{\Delta}_\theta = \widehat{\theta} - S^\star_{j+1}$. Then, similar to the proof of Lemma 3 case (b), we have

$$\begin{aligned}
&\sum_{t=t_j}^{s_i-1} \|X_t - (\widehat{\theta} + \widehat{L})' X_{t-1}\|_2^2 \\
&\le \sum_{t=t_j}^{s_i-1} \|\epsilon_t\|_2^2 + c_3|s_i - t_j| \|\widehat{\Delta}_\theta + \widehat{\Delta}_L\|_2^2 \\
&\quad + c' \left(\sqrt{|s_i - t_j| \log p} \|\widehat{\Delta}_\theta\|_1 + \sqrt{|s_i - t_j| p} \|\widehat{\Delta}_L\|_* \right) \\
&\equiv \sum_{t=t_j}^{s_i-1} \|\epsilon_t\|_2^2 + J_1 + J_2.
\end{aligned} \tag{23}$$

Now, according to the convergence rate of the error in Lemma 3 case (b), we obtain

$$\begin{aligned}
J_1 &\le c_3|s_i - t_j| \left(\|\widehat{\Delta}_\theta\|_2^2 + \|\widehat{\Delta}_L\|_2^2 \right) \\
&\le c_3|s_i - t_j| \frac{256}{\tau^2} (d_n^\star \eta_{(s_{i-1}, s_i)} + 2r\eta_L^2) \\
&= O_p\left(n\gamma_n (d_n^{\star\,2} + r^{\star 2}) \right),
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
J_2 &= c'|s_i - t_j| \left(\sqrt{\frac{\log p}{s_i - t_j}} \|\widehat{\Delta}_\theta\|_1 + \sqrt{\frac{p}{s_i - t_j}} \|\widehat{\Delta}_L\|_* \right) \\
&\le c'|s_i - t_j| \left(\eta_{(s_{i-1}, s_i)} \|\widehat{\Delta}_\theta\|_1 + \eta_L \|\widehat{\Delta}_L\|_* \right) \\
&\le 4c'|s_i - t_j| \left(\eta^2_{(s_{i-1}, s_i)} d_n^\star + \eta_L^2 r^\star \right) \\
&= O_p\left(n\gamma_n (d_n^{\star\,2} + r^{\star 2}) \right).
\end{aligned} \tag{25}$$

Using a similar procedure to the smaller sub-segment (s_{i-1}, t_j), we obtain

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{t_j-1} \|X_t - (\widehat{\theta} + \widehat{L})' X_{t-1}\|_2^2 \\
&\le \sum_{t=s_{i-1}}^{t_j-1} \|\epsilon_t\|_2^2 + c_3|t_j - s_{i-1}| \|(\widehat{\theta} - S_j^\star) + \widehat{\Delta}_L\|_2^2 \\
&\quad + c' \left(\sqrt{|t_j - s_{i-1}| \log p} \|\widehat{\theta} - S_j^\star\|_1 \right. \\
&\quad \left. + \sqrt{|t_j - s_{i-1}| p} \|\widehat{\Delta}_L\|_* \right) \\
&\le \sum_{t=s_{i-1}}^{t_j-1} \|\epsilon_t\|_2^2 + 2c_3|t_j - s_{i-1}| \left(\|\widehat{\Delta}_\theta + \widehat{\Delta}_L\|_2^2 \right. \\
&\quad \left. + \|(S^\star_{j+1} - S^\star_j) + \widehat{\Delta}_L\|_2^2 \right) \\
&\quad + c' \left(\sqrt{|t_j - s_{i-1}| \log p} \left(\|\widehat{\Delta}_\theta\|_1 + \|S^\star_{j+1} - S^\star_j\|_1 \right) \right. \\
&\quad \left. + \sqrt{|t_j - s_{i-1}| p} \|\widehat{\Delta}_L\|_* \right) \\
&\le \sum_{t=s_{i-1}}^{t_j-1} \|\epsilon_t\|_2^2 + O_p\left(n\gamma_n (d_n^{\star\,2} + r^{\star 2}) \right).
\end{aligned} \tag{26}$$

Since we have

$$\begin{aligned}
&\eta_{(s_{i-1}, s_i)} \|\widehat{\theta}\|_1 + \frac{s_i - s_{i-1}}{n} \eta_L \|\widehat{L}\|_* \\
&\le \eta_{(s_{i-1}, s_i)} \left(\|\widehat{\Delta}_\theta\|_1 + \|S^\star_{j+1}\|_1 \right) \\
&\quad + \frac{s_i - s_{i-1}}{n} \eta_L \left(\|\widehat{\Delta}_L\|_* + \|L^\star\|_* \right) \\
&= O_p(d_n^\star + r).
\end{aligned} \tag{27}$$

Therefore, combining (23) to (27) yields

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\widehat{\theta} + \widehat{L})' X_{t-1}\|_2^2 + \eta_{(s_{i-1}, s_i)} \|\widehat{\theta}\|_1 \\
&\quad + \frac{s_i - s_{i-1}}{n} \eta_L \|\widehat{L}\|_* \\
&= \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 + O_p\left(n\gamma_n (d_n^{\star\,2} + r^{\star 2}) \right).
\end{aligned} \tag{28}$$

Taking the union of all $m_0 + 1$ estimated intervals leads to the result (22).

Applying Lemma 3 and noting that under the conditions specified in assumption H4, we obtain

$$\begin{aligned}
&\mathrm{IC}(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}) = L_n(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}; \eta_n) + \widetilde{m}\omega_n \\
&> \sum_{t=1}^{n} \|\epsilon_t\|_2^2 + c_1\Delta_n - c_2\widetilde{m} n\gamma_n (d_n^{\star\,2} + r^{\star 2}) + \widetilde{m}\omega_n \\
&\ge L_n(\widehat{t}_1, \ldots, \widehat{t}_{m_0}; \eta_n) + m_0\omega_n + c_1\Delta_n
\end{aligned}$$

$$- \ c_2 m_0 n\gamma_n({d_n^\star}^2 + r^{\star 2}) - (m_0 - \widetilde{m})\omega_n$$
$$\geq L_n(\widehat{t}_1, \ldots, \widehat{t}_{m_0}; \eta_n) + m_0\omega_n, \tag{29}$$

which leads to the proof of case (a).

For case (b), by using a similar procedure as above, we get

$$L_n(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}; \eta_n) \geq \sum_{t=1}^{n} \|\epsilon_t\|_2^2 - c_2\widetilde{m}n\gamma_n({d_n^\star}^2 + r^{\star 2}). \tag{30}$$

Then, we compare $\mathrm{IC}(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}})$ and $\mathrm{IC}(\widehat{t}_1, \ldots, \widehat{t}_{m_0})$

$$\sum_{t=1}^{n} \|\epsilon_t\|_2^2 - c_2\widetilde{m}n\gamma_n({d_n^\star}^2 + r^{\star 2}) + \widetilde{m}\omega_n$$
$$\leq \mathrm{IC}(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}) \leq \mathrm{IC}(\widehat{t}_1, \ldots, \widehat{t}_{m_0})$$
$$\leq \sum_{t=1}^{n} \|\epsilon_t\|_2^2 + Km_0 n\gamma_n({d_n^\star}^2 + r^{\star 2}) + m_0\omega_n, \tag{31}$$

which implies that

$$(\widetilde{m} - m_0)\omega_n \leq (Km_0 + c_2\widetilde{m})n\gamma_n({d_n^\star}^2 + r^{\star 2}),$$

which contradicts assumption H4. Now we proved the first part of Theorem 3.

For the second part, we let $B = 2\mathrm{K/c}$, if there exists a point t_j such that $\min_{1\leq j\leq m_0} |\widetilde{t}_j - t_j| \geq Bm_0 n\gamma_n({d_n^\star}^2 + r^{\star 2})$, then by similar arguments as in Lemma 3, we have

$$\sum_{t=1}^{n} \|\epsilon_t\|_2^2 + cBm_0 n\gamma_n({d_n^\star}^2 + r^{\star 2})$$
$$< L_n(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}) \leq L_n(\widehat{t}_1, \ldots, \widehat{t}_{m_0})$$
$$\leq \sum_{t=1}^{n} \|\epsilon_t\|_2^2 + Km_0 n\gamma_n({d_n^\star}^2 + r^{\star 2}), \tag{32}$$

which contradicts to $B = 2\mathrm{K/c}$. Therefore, we complete the proof. ■

Proof of Theorem 4: It follows along the lines of the proof of Proposition 4 in [8]. We need to firstly verify two important conditions. (1) the restricted eigenvalue (RE) condition for $\widehat{\Gamma}_j = \mathcal{X}_j'\mathcal{X}_j/N_j$; (2) the deviation bound condition for $\|\mathcal{X}_j'\mathcal{E}_j/N_j\|_\infty$. These two conditions can be verified by Lemma 6 directly. Therefore, we can derive the following result

$$\frac{1}{N_j}\|\mathcal{Y}_j - \mathcal{X}_j(\widehat{L} + \widehat{S}_j)\|_F^2 + \rho_j\|\widehat{S}_j\|_1 + \rho_L\|\widehat{L}\|_*$$
$$\leq \frac{1}{N_j}\|\mathcal{Y}_j - \mathcal{X}_j(L^\star + S_j^\star)\|_F^2 + \rho_j\|S_j^\star\|_1 + \rho_L\|L^\star\|_*,$$

we define the same weighted regularizer as in Lemma 3 and the same norm decomposition as in the previous proof. Define $\widehat{\Delta}_L = \widehat{L} - L^\star$ and $\widehat{\Delta}_{S_j} = \widehat{S}_j - S_j^\star$ to obtain

$$\frac{1}{N_j}\|\mathcal{X}_j(\widehat{\Delta}_L + \widehat{\Delta}_{S_j})\|_F^2$$
$$\leq \frac{3}{2}\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}|_{\mathcal{I}_j}, \widehat{\Delta}_L^A) - \frac{1}{2}\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}|_{\mathcal{I}_j^c}, \widehat{\Delta}_L^B). \tag{33}$$

By the RE condition and Lemma 6 and substituting interval $[t_j, s_i]$ with I_{j+1}, there exists a positive constant $\tau > 0$ such that

$$\frac{1}{N_j}\|\mathcal{X}_j(\widehat{\Delta}_L + \widehat{\Delta}_{S_j})\|_F^2$$
$$\geq \frac{\tau}{2}(\|\widehat{\Delta}_L\|_F^2 + \|\widehat{\Delta}_{S_j}\|_F^2) - \frac{1}{2}\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}, \widehat{\Delta}_L);$$

substituting the inequality above in (33) and according to Lemma 4, we have

$$\frac{\tau}{2}(\|\widehat{\Delta}_L\|_F^2 + \|\widehat{\Delta}_{S_j}\|_F^2) \leq 2\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}, \widehat{\Delta}_L)$$
$$\leq 2(\rho_L\|\widehat{\Delta}_L\|_* + \rho_j\|\widehat{\Delta}_{S_j}\|_1)$$
$$\leq 2\sqrt{2r^\star\rho_L^2 + d_j^\star\rho_j^2}\sqrt{\|\widehat{\Delta}_L\|_F^2 + \|\widehat{\Delta}_{S_j}\|_F^2}.$$

Therefore, we get

$$\|\widehat{\Delta}_L\|_F^2 + \|\widehat{\Delta}_{S_j}\|_F^2 \leq \frac{16}{\tau^2}(2r^\star\rho_L^2 + d_j^\star\rho_j^2). \tag{34}$$

Combining the choices for the tuning parameters specified in Theorem 4 and (34), we can obtain the posited result. ■

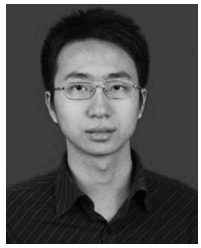
ACKNOWLEDGMENT

The authors would like to thank Associate Editor Prof. Elias Aboutanios and three anonymous referees for many constructive comments and suggestions.

REFERENCES

[1] A. Agarwal *et al.*, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *Ann. Statist.*, vol. 40, no. 2, pp. 1171–1197, 2012.
[2] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 277–286, Mar. 1998.
[3] A. Aue and L. Horváth, "Structural breaks in time series," *J. Time Ser. Anal.*, vol. 34, no. 1, pp. 1–16, 2013.
[4] J. Bai, X. Han, and Y. Shi, "Estimation and inference of change points in high-dimensional factor models,"*J. Econometrics*, to be published.
[5] M. Barigozzi, H. Cho, and P. Fryzlewicz, "Simultaneous multiple change-point and factor analysis for high-dimensional time series," *J. Econometrics*, vol. 206, no. 1, pp. 187–225, 2018.
[6] M. Basseville, "Detecting changes in signals and systems—A survey," *Automatica*, vol. 24, no. 3, pp. 309–326, 1988.
[7] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, vol. 104. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[8] S. Basu, X. Li, and G. Michailidis, "Low rank and structured modeling of high-dimensional vector autoregressions," *IEEE Trans. Signal Process.*, vol. 67, no. 5, pp. 1207–1222, Mar. 2019.
[9] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, vol. 43, no. 4, pp. 1535–1567, 2015.
[10] K. Bleakley and J.-P. Vert, "The group fused lasso for multiple change-point detection," 2011, *arXiv:1106.4199*.
[11] L. Boysen *et al.*, "Consistencies and rates of convergence of jump-penalized least squares estimators," *Ann. Statist.*, vol. 37, no. 1, pp. 157–183, 2009.
[12] N. H. Chan, C.-K. Ing, Y. Li, and C. Y. Yau, "Threshold estimation via group orthogonal greedy algorithm," *J. Bus. Econ. Statist.*, vol. 35, no. 2, pp. 334–345, 2017.
[13] N. H. Chan, C. Y. Yau, and R.-M. Zhang, "Group lasso for structural break time series," *J. Amer. Statistical Assoc.*, vol. 109, no. 506, pp. 590–599, 2014.

[14] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.
[15] H. Cho, "Change-point detection in panel data via double cusum statistic," *Electron. J. Statist.*, vol. 10, no. 2, pp. 2000–2038, 2016.
[16] H. Cho and P. Fryzlewicz, "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *J. Roy. Statistical Soc.: Ser. B (Statistical Methodology)*, vol. 77, no. 2, pp. 475–507, 2015.
[17] M. Csörgö and L. Horváth, *Limit Theorems in Change-Point Analysis*, vol. 18. Hoboken, NJ, USA: Wiley, 1997.
[18] E. F. Fama and K. R. French, "A five-factor asset pricing model," *J. Financial Econ.*, vol. 116, no. 1, pp. 1–22, 2015.
[19] E. F. Fama and K. R. French, "The cross-section of expected stock returns," *J. Finance*, vol. 47, no. 2, pp. 427–465, 1992.
[20] M. Frisén, *Financial Surveillance*, vol. 71. Hoboken, NJ, USA: Wiley, 2008.
[21] C.-D. Fuh and Y. Mei, "Quickest change detection and Kullback–Leibler divergence for two-state hidden markov models," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4866–4878, Sep. 2015.
[22] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, pp. 424–438, 1969.
[23] Z. Harchaoui and C. Lévy-Leduc, "Multiple change-point estimation with a total variation penalty," *J. Amer. Statistical Assoc.*, vol. 105, no. 492, pp. 1480–1493, 2010.
[24] Z. Harchaoui and C. Lévy-Leduc, "Multiple change-point estimation with a total variation penalty," *J. Amer. Statistical Assoc.*, vol. 105, no. 492, pp. 1480–1493, 2010.
[25] S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models," *Econometrica: J. Econometric Soc.*, vol. 59, pp. 1551–1580, 1991.
[26] B.-H. Juang and L. Rabiner, "Mixture autoregressive hidden markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1404–1413, Dec. 1985.
[27] H. Kato, M. Taniguchi, and M. Honda, "Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3414–3425, Sep. 2006.
[28] C. Kirch, B. Muhsal, and H. Ombao, "Detection of changes in multivariate time series with application to EEG data," *J. Amer. Statistical Assoc.*, vol. 110, no. 511, pp. 1197–1216, 2015.
[29] L. Koepcke, G. Ashida, and J. Kretzberg, "Single and multiple change point detection in spike trains: Comparison of different CUSUM methods," *Frontiers Syst. Neurosci.*, vol. 10, pp. 51–69, 2016.
[30] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating time-varying networks," *Ann. Appl. Statist.*, vol. 4, no. 1, pp. 94–123, 2010.
[31] F. Leonardi and P. Bühlmann, "Computationally efficient change point detection for high-dimensional regression," 2016, *arXiv:1601.03704*.
[32] J. Lin and G. Michailidis, "Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4188–4236, 2017.
[33] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos, "Snare: A link analytic system for graph labeling and risk detection," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1265–1274.
[34] A. Meucci, *Risk and Asset Allocation*. Berlin, Germany: Springer, 2009.
[35] G. Michailidis and F. d'Alché Buc, "Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues," *Math. Biosci.*, vol. 246, no. 2, pp. 326–334, 2013.
[36] B. Moghimi, A. Safikhani, C. Kamga, W. Hao, and J. Ma, "Short-term prediction of signal cycle on an arterial with actuated-uncoordinated control using sparse time series models," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2976–2985, Aug. 2019.
[37] S. N. Negahban *et al.*, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Sci.*, vol. 27, no. 4, pp. 538–557, 2012.
[38] F. F. Nobre and D. F. Stroup, "A monitoring system to detect changes in public health surveillance data," *Int. J. Epidemiology*, vol. 23, no. 2, pp. 408–418, 1994.
[39] H. Ombao, M. Fiecas, C.-M. Ting, and Y. F. Low, "Statistical models for brain signals with properties that evolve across trials," *NeuroImage*, vol. 180, pp. 609–618, 2018.
[40] R. Opgen-Rhein and K. Strimmer, "Learning causal networks from systems biology time course data: An effective model selection procedure for the vector autoregressive process," *BMC Bioinformatics*, vol. 8, no. 2, p. S3, 2007.
[41] P. Qiu, *Introduction to Statistical Process Control*. London, U.K.: Chapman & Hall, 2013.
[42] A. Rinaldo *et al.*, "Properties and refinements of the fused lasso," *Ann. Statist.*, vol. 37, no. 5B, pp. 2922–2952, 2009.
[43] S. Roy, Y. Atchadé, and G. Michailidis, "Change point estimation in high dimensional Markov random-field models," *J. Roy. Statistical Soc.: Ser. B (Statistical Methodology)*, vol. 79, no. 4, pp. 1187–1206, 2017.
[44] A. Safikhani and A. Shojaie, "Joint structural break detection and parameter estimation in high-dimensional non-stationary var models," 2017, *arXiv:1711.07357*.
[45] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 587–597, Mar. 2013.
[46] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6926–6938, Dec. 2015.
[47] C.-H. Zhang *et al.*, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Ann. Statist.*, vol. 36, no. 4, pp. 1567–1594, 2008.



Peiliang Bai received the B.Sc. degree in mathematics and statistics from Peking University, Beijing, China in July 2015, and the M.Sc. degree in statistics from University of Florida, Gainesville, FL, USA, in May 2017. He is currently a Ph.D. candidate in statistics with the University of Florida. His research interests include high-dimensional time series, change point detection, machine learning and their applications to neuroscience and finance.

Abolfazl Safikhani received his B.Sc. and M.Sc. in mathematics from the Sharif University of Technology, Iran, in 2008 and 2010, respectively. He holds a Ph.D. degree in statistics from Michigan State University (2015). He joined Columbia University as a term Assistant Professor afterwards. In 2019, he joined the Statistics Department at the University of Florida as an Assistant Professor and is also affiliated with the UF Informatics Institute. His research interests include time series models, high-dimensional statistics, change point detection, spatio-temporal models, and applied probability.

George Michailidis (Member, IEEE) received the B.S. degree in economics from the University of Athens, Greece, in 1987. He holds M.A. degrees in both economics and mathematics from UCLA and the Ph.D. degree in mathematics from UCLA. After a postdoc in operations research at Stanford University, he joined the Department of Statistics at the University of Michigan, in 1998, where he became Full Professor in 2008. In 2015, he joined the University of Florida as the Founding Director of the Informatics Institute. He is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics and the International Statistical Institute. His research interests include network analysis, queueing theory, stochastic control and optimization, applied probability and machine learning.